Integrated Research Infrastructure for the Social Sciences

Work Package 3

Technical Report 2: Demonstrator #1

Germán González, Pascal Perez, Masoud Rahimi, Tomasz Zajac, Matthias Kubler, Denise Clague, Jonathan Corcoran, Wojtek Tomaszewski

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
AuSSA	The Australian Survey of Social Attitudes
CARDSS	Curation of Australian Research Data in the Social Sciences
DSDM	Dynamic Systems Development Method
DLOD	Dwelling Location
DWTP	Dwelling Type
EETP	Engagement in Employment, Education and Training
EMPP	Number of Employees
FAIR	Findability, accessibility, interoperability, and reusability
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
MoSCoW	Stands for “must-have,” “should-have,” “could-have,” and “won't-have”
MVP	Minimum Viable Product
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
SA3s	Statistical Area Level 3
SEIFA	Socio-Economic Indexes for Areas
SIEMP	Status in Employment
STRD	Structure of Dwelling
STUP	Full-Time/Past-Time Student Status
VASSAL	Vocabulary Access Service for Social Science in Australia
WP3	Work package 3

1 Introduction

The Integrated Research Infrastructure for the Social Sciences (IRISS) Project aims to address the fragmentation of the Australian social science research infrastructure. There is a dearth of data integration of information from various sources on people, places, time, and space. The Work Package 3 (WP3) of the IRISS project aims to close this gap by developing a data integration service called GeoSocial. The service is designed to allow researchers to augment Australia's largest longitudinal surveys with geospatial statistical data derived from the Australian Census of Population and Housing (Census). By doing so, GeoSocial will empower Australia's large cross-disciplinary social research community to identify patterns, make predictions, and inform social policy using rich integrated GeoSocial data.

The purpose of this report is to showcase GeoSocial Demonstrator #1. As part of the demonstrator, we built a Minimum Viable Product (MVP) which implements the preliminary GeoSocial service design and allows the integration of selected data sources. Both the GeoSocial service and demonstrator involved an iterative approach to drawing on Agile methods. Such an approach allows outputs from different stages, comprising several activities, to feedback on the design and development of the solution in a way that caters for flexibility in functionality under time and cost constraints. Previous outputs on which this report builds include:

- *Technical Report 1* published on 31 August 2022 presented the correspondent framework and principles of the design, followed by the user profiles, data flows, and users' needs translated into prioritized requirements and key features for the GeoSocial service.
- *The preliminary GeoSocial service design* published on 31 March 2023 presented a theoretical framework for the service. The report explained the workflow within the service, the technical aspects, and interactions between the service and other IRISS project work packages.
- *Technical Report 2: Data Linkage report*, published on 31 May 2023 and included as Appendix A, described in detail the data selection process, the survey data to be augmented with information about places drawn from the Census, and conceptual, methodological, and practical issues related to linking area-based data to person-level data.

This report focuses on the final stage of prototype development and the creation of the demonstrator dataset. It documents the development process of the MVP and explains the design of the tool and the reasoning behind selected solutions. It has been organised into six chapters, including the introduction. The second chapter summarizes the solution development and the framework used to approach the problem. It highlights the minimum requirements that the GeoSocial service should include. The third chapter recaps the preliminary GeoSocial service design, explaining each component of the solution. It also proposes the demonstrator, which implements the solution description and contributes to the design of the GeoSocial service solution architecture. Chapter four offers an overview of the demonstrator, from the start to the end of the user experience, providing the pipeline and user experience. Chapter Five covers the technical requirements and explains how to deploy and maintain the demonstrator. The last chapter will provide a conclusion to the stage one of the project.

2 Development process

This chapter summarises the GeoSocial service development process and highlights the key findings that informed the service design and demonstrator. First, it describes our iterative approach to service development. Second, it discusses aspects of the service design that needed to be considered before the development of the demonstrator including user profiles, user needs and their translation into prioritised requirements, the selection of datasets, types of data linkage, geospatial units, and the FAIR principles for research software.

2.1 The iterative approach to software development

We adopted an iterative approach to software development and the implementation of the service. It allowed us to be flexible with the functionalities of the service, collect feedback through peer review by the members of the WP3 and wider IRISS teams at each stage, and implement improvements informed by this feedback. Figure 1 presents the stages of the iterative process.

- Planning
- Iterative development
- Understand researchers, their contexts, and their needs.
- Identify, categorise, and prioritise user requirements.
- Design, develop and increment solutions.
- Evaluate solutions against requirements.
- Completion

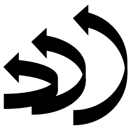


Figure 1: Iterative approach to software development used in WP3.

The first stage of the project focused on understanding the future service users, that is social researchers, as well as their contexts and needs. We gathered information through interviews, reviews of past projects, and data audits. We analysed this data to create preliminary user profiles, contextual information, and their needs. This work took place between March and August 2022 and resulted in the initial user profiles for the GeoSocial service.

The next stage of GeoSocial service development focused on identifying user requirements. During the Workshop held on August 16, 2022, AURIN and ISSR teams worked collaboratively to define functional requirements and prioritize them using the Agile method MoSCoW. We categorized the requirements into 'Must have', 'Should have', 'Could have', and 'Won't have' categories. The resulting categorisation of the requirements was included in Technical Report 1.

Besides the user requirements, Technical Report 1 presented the theoretical framework of solution development, and conceptualisation of the initial solution, highlighting the key elements towards the design. All these guided the development of the preliminary service design and the demonstration implementation. We discuss the most essential elements of Technical Report 1 in the next section.

Furthermore**,** Technical Report 1 laid out the key stages of the software development process and assigned responsibilities. We chose to follow the Agile Dynamic Systems Development Method[1] (DSDM) while developing software for the GeoSocial service. Key team members involved in this process included the Social Data Scientist and Software Developer at AURIN, with oversight from the Partner Lead (AURIN) and Work Package Manager (ISSR). Specific outputs from development are described in Appendix B.

Throughout the project, we held regular meetings to continuously track project progress, gather feedback on implemented solutions, and revise the service design. These included weekly technical meetings of the Australian Urban Research Infrastructure Network (AURIN) developer team, fortnightly WP3 meetings between the Institute for Social Sciences Research (ISSR) and AURIN teams, and project-level meetings organised by the Australian Data Archives (ADA) at the Australian National University (see Table 1). Those meetings allowed

[1] The DSDM Agile Project Framework, https://www.agilebusiness.org/page/TheDSDMAgileProjectFramework

us to identify several further issues affecting the service design that were not addressed in Technical Report 1. We discuss the most important of them in the next section.

Type	Frequency	Coordinator
Technical Development	Weekly	AURIN
Work Package	Fortnightly	ISSR
Project	Monthly	ADA

Table 1: WP3 meeting schedule

2.2 Key design considerations

While working on the service, we discovered several factors that influence the design and require careful consideration. These factors include:

- Service users and user requirements.
- A longitudinal survey to be augmented with GeoSocial data.
- Geographical dataset to be linked to the survey data.
- Type of linkage used to merge survey data.
- Geographical unit.

The following sub-sections present the main findings for each of these factors.

2.3 Service users and user requirements.

One of the first steps in the service design process was identifying the potential users. The initial user requirements are available in Appendix C. It was decided that the service should first cater to the needs of social science researchers, who have intermediate to advanced levels of data manipulation skills. Table 2 presents the profiles of those users.

Social science researcher – Advanced user	- Confident with using Python and/or R for data wrangling, integration, and analysis. - Good understanding of geospatial data - Needs to integrate longitudinal and geospatial data for analysis. - Supports other social science researchers

Social science researcher – Mid-level user	- Confident with understanding and tweaking R scripts - Experienced in the use of Stata software. - Limited understanding of geospatial data - Needs to integrate longitudinal and geospatial data for analysis. - May consult with data science researchers to achieve goals.

Table 2: GeoSocial user profiles

The discussions about the users and their needs helped us identify and categorise 49 functional requirements. The full list of requirements is available in Appendix D. The functional requirements can be separated into different categories that determine each aspect of the solution. For instance, some user requirements define which format would read the data, the type of linkage, how would the data be exported, the user interface, etc. We would like to highlight four key requirements that were crucial for shaping the service design:

- Accessible coding style - widely used language (R), scripts easy to understand and modify.
- Login-free access to the service allows users to discover what the service has to offer without investing too much time.
- Example Script
- Design and develop the following FAIR principles for research software.

These requirements determined that the preliminary GeoSocial service designs needed to be delivered through a code language and executed using a script.

2.4 Longitudinal survey data

Understanding the needs of social researchers in Australia includes identifying the most relevant longitudinal surveys that should be included in the GeoSocial service. The selection of datasets for the demonstrator was preceded by data audits reviewing metadata and assessing the current usage of various datasets. The process has been documented in Technical Report 1 and the Data Linkage report. Below we summarise the main conclusions from our data reviews.

HILDA: In the first report, we found that, after a data audit focused on the ADA data collection, the Household, Income and Labour Dynamics in Australia (HILDA) was by far the most often downloaded survey available in ADA with 33,924 downloads2 (See Appendix 3). The structure and contents of the HILDA dataset made it a good example dataset for the service prototype. The consistency across the multiple waves, the size of the sample, and the relevance of the variables to the social research community were the main reasons that the survey was considered the main candidate for the prototype.

After identifying the HILDA as suitable data for the demonstrator, the team applied to get access to the restricted data containing the geographical identifiers. Unfortunately, the Australian Government Department of Social Services, which is the custodian of HILDA data, declined our request for access to the restricted version.

We were unable to use HILDA data, so we looked for other surveys that could help us showcase the capabilities of the GeoSocial service. This was difficult because it required developing a new application and making substantial changes to the project's timeline, causing delays. We outlined these changes in the Preliminary GeoSocial service design, which was published on March 31, 2023.

LSAY: Following a process like the initial data audit, we selected the Longitudinal Surveys of Australian Youth (LSAY) as a replacement for HILDA. Although it is not as popular as HILDA (see Appendix E for the most downloaded ADA surveys), it is a large-scale longitudinal study, which allows following the same individuals over an extended period and covers a wide range of topics, making it interesting to a good section of the research community. Appendix A describes LSAY data in more detail and explains why it is a suitable dataset for the demonstrator.

AuSSA: Gaining access to survey data proved time-consuming. To continue developing the service while waiting first for HILDA and then for LSAY data, we used the Australian Survey of Social Attitudes, 2020 (AuSSA). This allowed us to work with real survey data that includes

[2] Data as of 20 April 2022.

geographical identifiers, explore potential issues, and start experimenting with different solutions and frameworks that would later inform the preliminary GeoSocial service design.

Accessing survey data:

While applying for the above datasets, we learned several invaluable lessons. We found that the data custodians are always concerned with data security and will ask not only about how the data will be used but also about how the data will be stored and processed. That is especially true in the case of longitudinal surveys comprising information about the location of respondents, as such data increases the risk of reidentification of survey participants. It was also clear that the GeoSocial service could not host any survey data. That determined one of the key features of the service, namely that the data needs to be processed in a local computer environment that satisfies the data custodians' requirements. Another issue was that we needed to be transparent about the process so that the users would be aware of their responsibilities regarding data access and management.

We have determined that the GeoSocial solution should be constructed in a manner that does not store any data and does not disrupt the procedures in place for the data custodians to access the data. To achieve this, we have specified that the solution should be delivered through code snippets that can be easily integrated into the existing workflow.

2.5 Geographical datasets:

To identify which geographical datasets GeoSocial should include in enhancing the data, we evaluated geospatial data sets available on AURIN and determined which ones to include based on usage and the number of downloads. We identified several options, such as the Australian Bureau of Statistics (ABS) Socio-economic Indexes for Areas (SEIFA), to enhance the surveys with geospatial data.

However, we discovered that a significant part of the datasets hosted by AURIN, and available in Australia specialise in providing information about specific characteristics of the population and don’t provide a comprehensive picture of an individual’s socio-demographic characteristics. For instance, SEIFA adds only socio-economic indexes for the areas in which the participants of the longitudinal survey reside. One possible solution to create a comprehensive picture of the socio-demographic characteristics of the individuals of the longitudinal survey is to do multiple linkages across different datasets.

To simplify the previous approach, we decided that the initial version of the GeoSocial should include the variables available in the last three censuses to enrich the longitudinal survey. However, it is important to note that not all census variables can be compared across different periods and locations. This is due to various reasons:

- Changes in the questions: the latest census has made changes to the questionnaires by either updating or removing some questions from the previous census.
- Changes in possible responses: For instance, the latest census collected responses on non-binary sex as part of the question on sex.
- Changes in the geographic definitions: Some geographical boundaries of certain places change between 2011 to 2021.

For more information on this matter, please refer to Appendix N, which provides examples of inconsistencies and changes in categories found in the ABS census.

Based on the above findings, we determined that the Time Series Profile 2021 dataset from ABS would be the best option to enhance the longitudinal survey during its initial stage. The description of the dataset indicates that:

> "*The Time series profile contains the Census characteristics of persons, families, and dwellings over time. The data is based on the place of usual residence.*
>
> *The 2021 Time series profile contains data from 2011, 2016, and 2021 Censuses. Where classifications have been revised, the output is based on the classification used for the 2021 Census.*
>
> *When interpreting the results from different periods, take care as censuses are based on a point in time. Changes to the Census form design, collection procedures and processing may impact the comparability of data."* (ABS, 2021)

We chose the Time Series Profile 2021 because it provides sociodemographic characteristics that can be compared across censuses, based on the usual place of residence for individuals.

2.6 Type of linkage:

After analysing potential use cases, we determined that the solution should be offering multiple options for joining surveys and GeoSocial data. The available linkage methods can be divided into two categories:

- **Cross-sectional spatial data linkage**: This refers to integrating spatial information from a particular cohort and wave of the longitudinal survey with a particular geospatial data captured at a particular point in time**.** For example, merging 2011 Census data to any wave of the 2009 LSAY cohort. This type of data linkage is simpler to interpret, and the data requirements are lower, given that the linkage does not include a temporal component -methodological changes across waves or rounds of the data collection do not have to be considered.
- **Longitudinal spatial data linkage**: This refers to integrating spatial information from multiple rounds of the Census into different waves of the longitudinal survey. This linkage type is more complex and requires data using consistent geographies and variable definitions. In other words, the data must be collected for the same areas over time. Furthermore, they need to be collected in the same way. For example, merging the variables of the Time Series Profile to 2011, 2016 and 2021 to different waves of the 2009 LSAY cohort. This type of data linkage opens new research opportunities, such as identifying temporal trends captured in spatial data and those observed in survey data.

2.7 Geographical unit:

Another factor to consider in data linkage is the geographical unit used. In Australia, there are two categories of geographical units available:

- **Non-ABS structures:** administrative regions that are not defined or maintained by the ABS.
- **ABS structures:** administrative regions that are defined and maintained by the ABS.

The following Figure 2 shows the ABS structures and the main non-ABS structures.

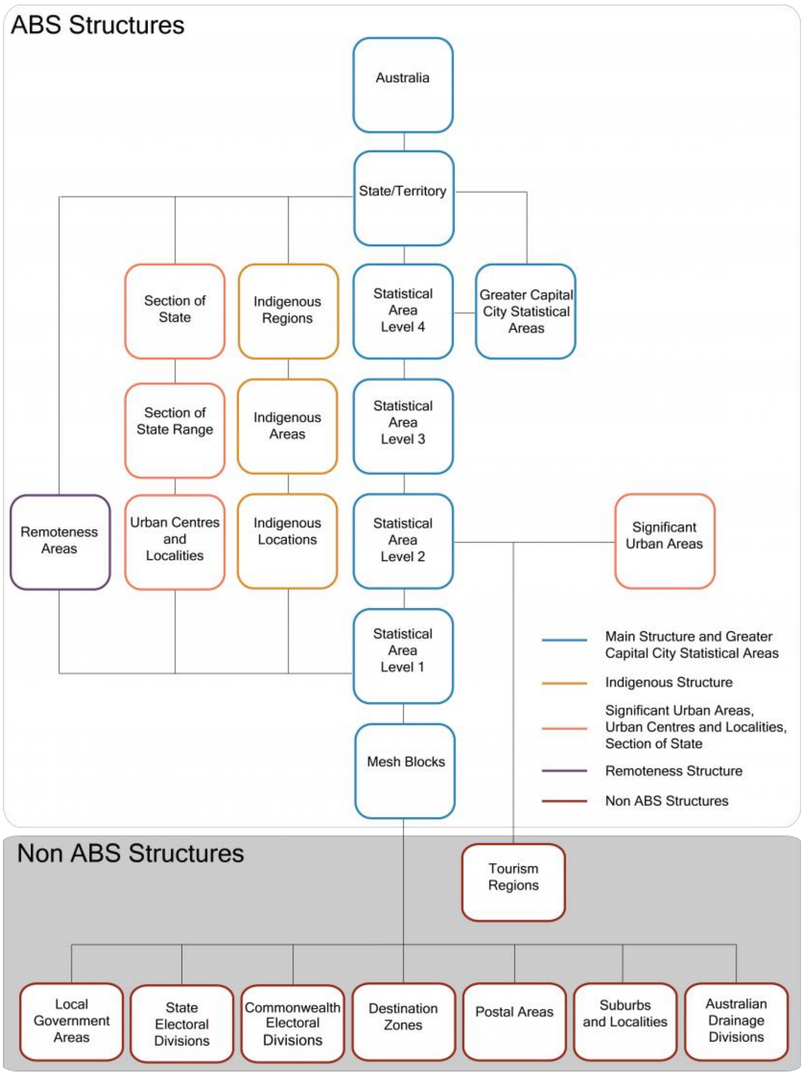


Figure 2: ABS structures vs non-ABS structures. From: ABS.

We identified that a significant number of the longitudinal surveys use geographical identifiers for non-ABS structures such as postcodes, given the wide use of postcodes across Australia. To perform any type of data linkage with ABS structures, the non-ABS structures within the longitudinal surveys need to be transformed into ABS structures. The transformation is possible using postal areas as a proxy variable of the postcodes.

> *"Postal Areas are an ABS Mesh Block approximation of a general definition of postcodes. They enable the comparison of ABS data with other data collected using postcodes as the geographic reference. ABS approximations of administrative boundaries do not match official legal boundaries and should only be used for statistical purposes."* (ABS, 2021)

Data linkage requires the use of the same geographical unit between the two datasets. Through correspondence files, it is possible to aggregate the postal areas to statistical areas levels 2, 3 and 4. To simplify the development of the GeoSocial service, the initial version of the demonstrator will support statistical areas level 3 (SA3s). SA3s provide enough information to find statistical patterns while protecting participants' privacy in the longitudinal survey.

2.8 FAIR Principles for Research Software

To meet user needs, address requirements and make the GeoSocial service demonstrator valuable to the research community, the service design will seek to adopt the FAIR Principles for Research Software (FAIR4RS) defined under findable, accessible, interoperable, and reusable categories in Table 3.

Findable: Software, and its associated metadata, are easy for both humans and machines to find

F1. Software is assigned a globally unique and persistent identifier.

F1.1. Components of the software representing levels of granularity are assigned distinct identifiers.

F1.2. Different versions of the software are assigned distinct identifiers.

F2. Software is described with rich metadata.

F3. Metadata clearly and explicitly includes the identifier of the software they describe.

F4. Metadata are FAIR, searchable, and indexable

Accessible: Software, and its metadata, is retrievable via standardized protocols

A1. Software is retrievable by its identifier using a standardized communications protocol.

A1.1. The protocol is open, free, and universally implementable.

A1.2. The protocol allows for an authentication and authorization procedure, where necessary.

A2. Metadata are accessible, even when the software is no longer available

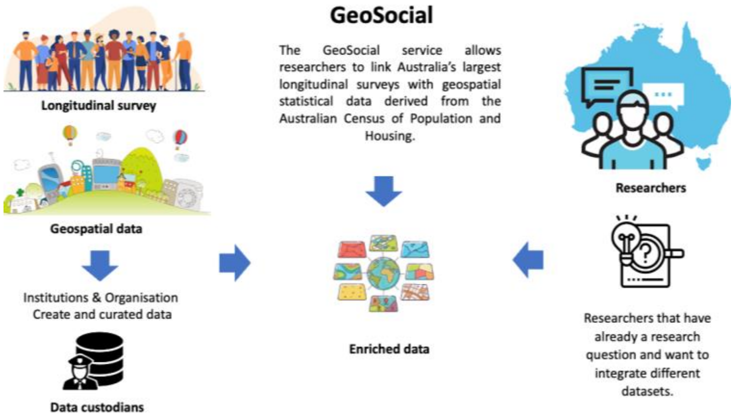
Interoperable: Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards. I1. Software reads, writes, and exchanges data in a way that meets domain-relevant community standards. I2. Software includes qualified references to other objects.
Reusable: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software). R1. Software is described with a plurality of accurate and relevant attributes. R1.1. Software is given a clear and accessible license. R1.2. Software is associated with detailed provenance. R2. Software includes qualified references to other software. R3. Software meets domain-relevant community standards.

Table 3: FAIR Principles for Research Software (Chue Hong et al., 2022)

3 Preliminary service design

In this chapter, we will provide a summary of the GeoSocial service design, including its different components. We will start by giving an overview of the service design, then we will explain each element of the service. Finally, we will conclude with a detailed explanation of the data linkage workflow.

The preliminary GeoSocial service design outlines the technical structure of the service by standardising the main task of the data linkage into code functions, that are delivered through a software toolkit and online service. The solution considers the user requirements, as well as the findings mentioned in the previous section. The following graph shows the relationship between the different actors involved, and how the GeoSocial service interacts with them.



GeoSocial
The GeoSocial service allows researchers to link Australia's largest longitudinal surveys with geospatial statistical data derived from the Australian Census of Population and Housing.
Longitudinal survey
Geospatial data
Institutions & Organisation
Create and curated data
Data custodians
Enriched data
Researchers
Researchers that have already a research question and want to integrate different datasets.

Figure 3: GeoSocial

The GeoSocial service is designed for researchers who need enhanced data to support their research questions or want to explore the data but may not have the ability or desire to integrate the data on their own. This could include individuals lacking the technical skills necessary for performing data integration, not having sufficient knowledge of the data (e.g., not knowing what geographical identifiers should be used in data linkage), or simply preferring to save time by using readily available integration tools.

To perform the data linkage and produce enriched data using the GeoSocial service, users must have access to survey and geospatial data collected by other institutions and organizations. Those data custodians have established certain conditions and procedures for accessing their resources. The solution is designed to be integrated into the workflow of the researchers without interfering with the standards and procedures set by the data custodians.

The elements that compound the GeoSocial service design, and the interaction across the different work packages of the IRISS project, were defined in the *Preliminary GeoSocial service Design* published on 31 March 2023. Figure 4 shows an overview of the proposed architecture of the GeoSocial service and how it interacts with other services, including those developed by other IRISS work packages.

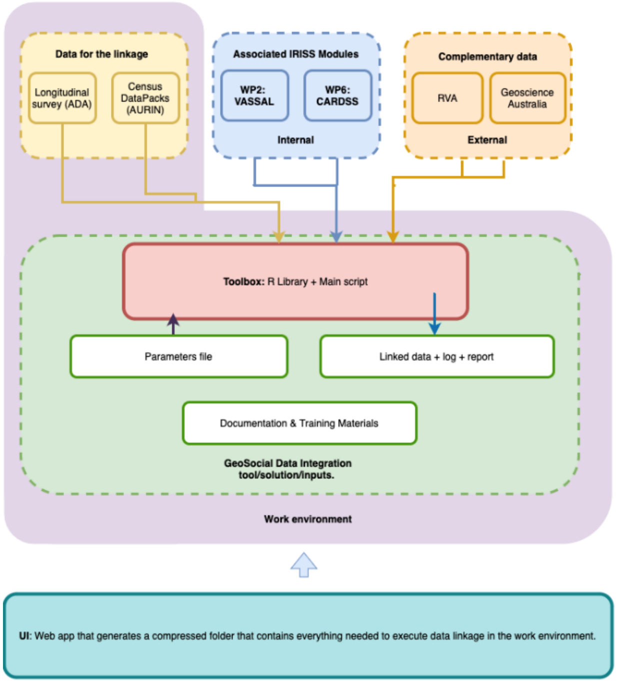

Data for the linkage
Longitudinal survey (ADA)
Census DataPacks (AURIN)
Associated IRISS Modules
WP2: VASSAL
WP6: CARDSS
Internal
Complementary data
RVA
Geoscience Australia
External
Toolbox: R Library + Main script
Parameters file
Linked data + log + report
Documentation & Training Materials
GeoSocial Data Integration tool/solution/inputs.
Work environment
UI: Web app that generates a compressed folder that contains everything needed to execute data linkage in the work environment.

Figure 4: GeoSocial overview diagram

This graphic demonstrates how the service reads and links the data using an R library, main script, and parameters file in a work environment. The linked data can be enhanced by interacting with other work packages associated with the IRISS project. Users can access the solution through a user interface that provides a compressed folder containing everything necessary to execute the linkage. The following sections provide more detailed explanations of each element.

GeoSocial data integration:

The GeoSocial data integration is carried out by an R script, which loads the standardised data linkage tasks through an R library, and a set of parameters, and executes the data linkage and enhancement task. The solution's components are shown in Figure 5.

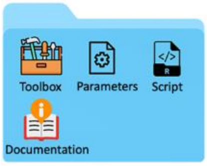


Figure 5: GeoSocial data integration

Commented [GU1]: Make sure you mention this Figure 5 in the text.

Users download a data integration tool from the service website to perform data linkage in their local machines. The package includes the following elements.

- **Toolbox: R** library which standardizes data linkage tasks through code functions.
- **Parameters file**: The parameters file standardises all the definitions that are relevant to the data linkage process. Some of these definitions are the location of the longitudinal data, API credentials, wave, cohort, type of linkage, and variables of interest.
- **Script**: The main script executes the data linkage.
- **Documentation**: Clear documentation of the R library in case the user desire to personalise their pipelines to their necessities.

Work environment:

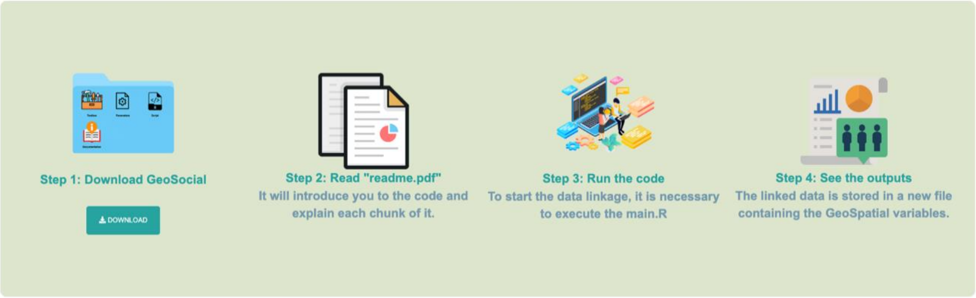
The solution must be executed in a work environment that includes data, scripts, and parameters. This work environment can be hosted on a local or virtual machine that meets the requirements of data custodians. Researchers must set up, use, and maintain this environment through various processes and procedures, such as data-sharing agreements.

User interface:

The solution is introduced to the users through a user interface that allows them to customise the data linkage. The solution offers two options:

The first option guides the user through the data linkage process using a standard pipeline that meets the needs of most users, increasing data usability and utility to untrained users, and minimizing the risk of analytical errors. Users can choose the specific datasets, years, variables, and other relevant information they want to link via drop-down lists.
The second option allows users to do advanced linkages through a self-guided option that enables the customisation of the pipeline and personalises the data linkage.

The user interface generates a compressed file that contains everything needed to execute the data linkage in the work environment. Figure 6, illustrates this process.



Step 1: Download GeoSocial

Step 2: Read "readme.pdf"
It will introduce you to the code and explain each chunk of it.

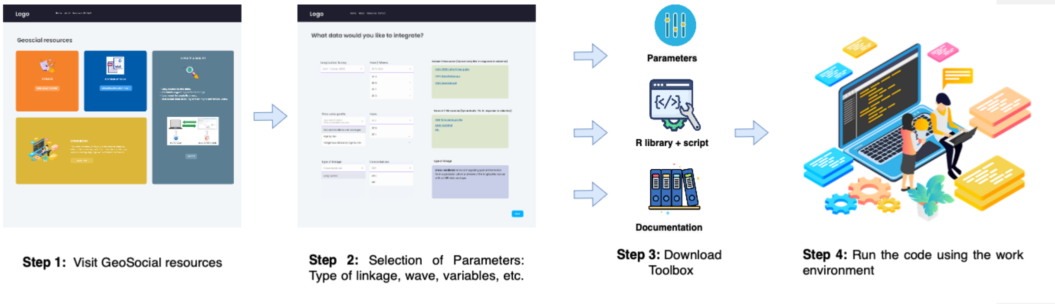
Step 3: Run the code
To start the data linkage, it is necessary to execute the main.R

Step 4: See the outputs
The linked data is stored in a new file containing the GeoSpatial variables.

Figure 6: Output - GeoSocial

Data linkage workflow:

The data linkage workflow can be summarised by the following graphics:



Step 1: Visit GeoSocial resources

Step 2: Selection of Parameters: Type of linkage, wave, variables, etc.

Step 3: Download Toolbox

Step 4: Run the code using the work environment

Figure 7: Data linkage workflow

Once the linkage is executed, the linked data is exported in the format of the longitudinal survey, including the Geospatial variables and corresponding metadata. Moreover, a log file is created throughout the process to record each linkage step. It flags any error or warning messages to assist with debugging and ensure transparency and reproducibility of the results.

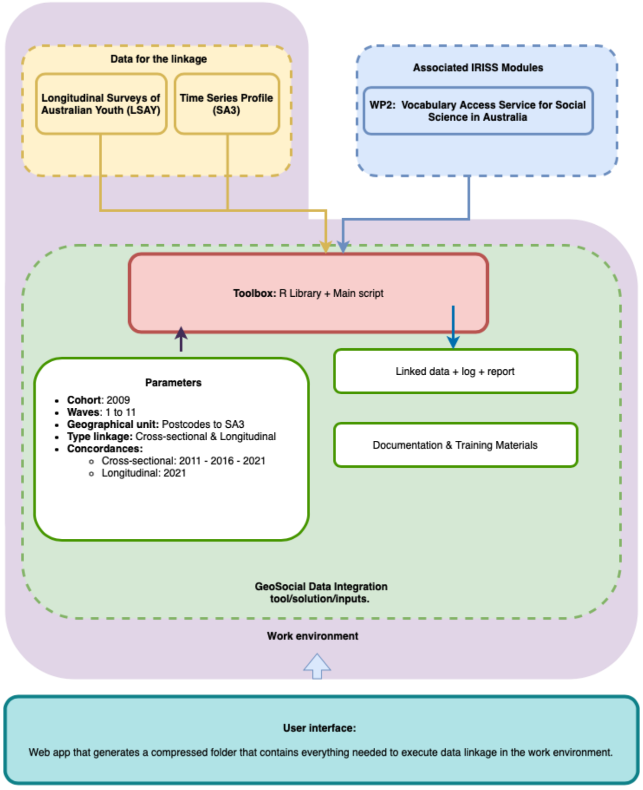
For more details about the preliminary GeoSocial service design and the elements that compose it, please refer to the Preliminary GeoSocial Service Design published on 31 March 2023.

4 Demonstrator

This chapter discusses the creation of a Minimum Viable Product (MVP) that implements the preliminary GeoSocial service design. It discusses the decisions we made about various parameters involved in the data linkage, such as which cohort and waves we selected for the longitudinal survey data as well as the type of linkage and concordances for Census DataPacks. The chapter is divided into two sections. The first section explains the guiding principles and assumptions used in building the demonstrator. The second section focuses on the development of the demonstrator.

4.1 Guiding principles:

Figure 7 presents a simplified illustration of the GeoSocial demonstrator diagram.



Data for the linkage
Longitudinal Surveys of Australian Youth (LSAY)
Time Series Profile (SA3)
Associated IRISS Modules
WP2: Vocabulary Access Service for Social Science in Australia
Toolbox: R Library + Main script
Parameters
Cohort: 2009
Waves: 1 to 11
Geographical unit: Postcodes to SA3
Type linkage: Cross-sectional & Longitudinal
Concordances:
Cross-sectional: 2011 - 2016 - 2021
Longitudinal: 2021
Linked data + log + report
Documentation & Training Materials
GeoSocial Data Integration tool/solution/inputs.
Work environment
User interface:
Web app that generates a compressed folder that contains everything needed to execute data linkage in the work environment.

Figure 8: Simplified GeoSocial demonstrator diagram

As illustrated, the demonstrator can integrate the 11 waves of the LSAY 2009 cohort with the ABS Time Series Profile (TSP) using SA3 as the geographical unit to join the datasets. The employed Time Series Profile presents data from the 2011, 2016 and 2021 Censuses based on the geographical boundaries from the 2021 Census.

The Longitudinal Survey of Australian Youth (LSAY) has been selected as the primary survey dataset for the pilot given the completeness of the survey and the rigour in the collection of the data with more than 6 cohorts, with around 11 waves for each wave. The first LSAY cohort began in 1995 and these individuals were contacted once a year until they were 25 years old. LSAY includes a wide range of information such as (Detailed in Appendix A):

- Demographics (such as gender, country of birth, indigeneity, and socioeconomic status)
- Education (including school characteristics, attitudes, engagement, and subject choices)
- The transition from school (including post-school plans)
- Post-school study and training (including pathways, and tertiary education)
- Employment (including hours worked, job search activity and history, wages, and benefits)
- Social life, living arrangements, finance, and health.
- General attitudes (including life satisfaction, and aspirations)

In addition, the 2021 Time Series Profile (TSP) DataPack was selected as the main Census DataPack given that it maintains consistent geospatial and semantic definitions over time. For instance, definitions of *gender* or *income* may vary from one census to another. However, the TSP DataPacks standardise such definitions while preserving the geospatial units, enabling interpretation, and facilitating comparability across different censuses.

Data linkage types:

As discussed in Section 1.3, two categories of data linkage are considered in this demonstrator:

- Cross-sectional spatial data linkage.
- Longitudinal spatial data linkage.

Spatial concordances:

To facilitate spatial data linkage, one of our initial steps involved mapping data associated with postcodes (POA) to their respective Statistical Area Level 3 (SA3) 2021 regions (further details in Appendix A).

Interface:

The demonstrator offers a comprehensive solution to users by providing them with both an R library and a user-friendly graphical interface. This dual approach allows users to choose the most convenient and accessible method for utilizing the demonstrator's functionalities. The R library offers a powerful toolset, enabling users to programmatically access and analyse data, leveraging the flexibility and versatility of the R programming language. The graphical user interface provides an intuitive and visually driven interface that simplifies the interaction with the demonstrator's features, catering to users who may prefer a more visual and interactive approach. By offering these two options, the demonstrator supports a wide range of applications, regardless of the programming expertise or preferences of the researchers, to effectively derive meaningful insights from the data.

- Code: R library + script
- GUI: Shiny application

Documentation:

The demonstrator offers comprehensive documentation that guides users in effectively utilizing the R library, scripts, and user interface. The documentation includes clear explanations and instructions for researchers. It covers the functions, parameters, and usage of the R library, provides insights into the logic, and offers instructions on navigating the user interface.

Fully documented code, i.e., library, script, and Shiny application

Readme: file that introduces the code and explains each part of the code.

Technical considerations:

The following additional design considerations have been identified for the GeoSocial service:

- **Security**: The GeoSocial service does not include authentication/authorisation or access management mechanisms. The library and the data integration scripts are publicly available and will operate within a working environment. Any privacy considerations are the responsibility of the user.
- **Programming language:** Scripts used within the GeoSocial service are written in the R programming language, as per user profiles.
- **R Library:** The R library can be installed using a compressed file. At this stage, the project would not be published in R CKAN.

4.2 Development of the demonstrator:

Building the demonstrator started with generating the R library, standardising the data linkage tasks into R functions, enabling users to enhance the library with new functionalities, and documenting data and methods in a standard way.

The R library is designed to be modular and maintainable. This includes creating functions that are self-contained, well-documented and extendable. Moreover, the R library allows the creation of downstream products that are maintainable, scalable, and transferable while staying interoperate across various platforms. The following are some examples of the best practices and approaches employed to make this happen:

- Clean Code Principles
- Good and clear documentation
- Flowcharts and code hierarchy diagrams
- Documenting all the files, classes, methods, and variables effectively
- Git version controlling systems to track and manage the code.

In addition, the implementation of the solution through an R library brought with it additional benefits that improved the user experience, such as:

Context-specific help menu inside R, allowing the user to better understand different functionalities.
data storage, verification, and management in the library, allowing automatic data loading and easy usage.

Figure 9 summarises the capabilities and functions of the R library into six categories:

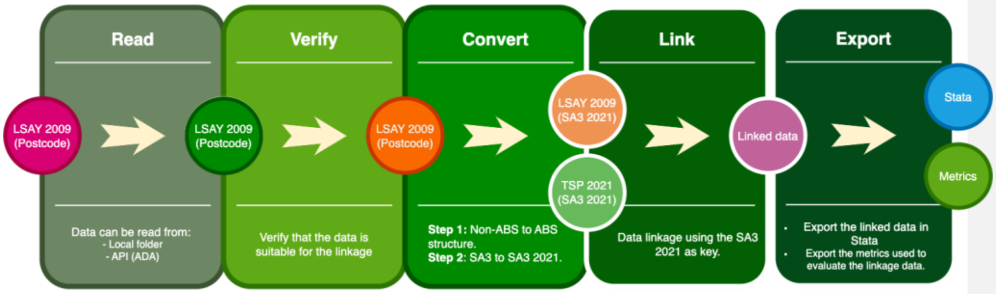


Figure 9: Workflow - GeoSocial

The specifications and capabilities of the main functions were outlined in the *Preliminary GeoSocial service design*, published on 31 March 2023. The main script allows execution of this data linkage pipeline, using the functions stored in the R library. While some of these functions are initially designed to handle the LSAY 2009 dataset, they possess significant potential for generalization and can be applied to other types of datasets from various contexts, applications, and demonstrators. In what follows, we explain the implementation of these categories within the demonstrator. Detailed descriptions, inputs, and outputs of each function are outlined in the Manual of the R library, presented in Appendix F.

4.2.1 Read

LSAY: The demonstrator supports loading data locally or through the Australian Data Archive (ADA) Dataverse API, which provides easy access to research data. As we mentioned before, it is the responsibility of the user to request access to LSAY data and load it in a Stata (.dta)[3] format.

- **Loading local data:** The demonstrator reads the data from a local folder. The user needs to download the dataset from the source and keep it in a local folder. The user should then specify the path where the data is contained to load it into the environment.
- **API:** The demonstrator downloads the data from the Dataverse API and stores it in a local folder in the work environment. The user needs to provide their corresponding database credentials and access to the data.

In both cases, the data loads into the R environment, preserving all the metadata, and values that are readable in Stata. This process is generalisable for other longitudinal surveys and does not require any previous knowledge about the dataset.

TSP 2021: The geospatial data used for the data linkage is sourced from the ABS Time Series Profile 2021. This dataset is presented in different .csv files, containing key Census characteristics of persons, families, and dwellings. In collaboration with the ISSR team, some variables of potential interest to link with the LSAY 2009 cohort were selected for the GeoSocial demonstrator. These variables include:

- Selected Person Characteristics by Sex

[3] We identified in the previous report: "*Work Package 3 Technical Report 1 published on 31 August 2022*" that Stata is widely used by the potential users.

- Selected Medians and Averages
- Age by Sex Indigenous
- Status by Age by Sex
- Proficiency in Spoken English by Age
- Type of Educational Institution Attending (Full-Time/Part-Time Student Status by Age) by Sex
- Dwelling Structure by Household Composition and Family Composition
- Tenure and Landlord Type by Dwelling Structure
- Highest Non-School Qualification: Level of Education by Age by Sex
- Highest Non-School Qualification: Field of Study by Age by Sex
- Labour Force Status by Age by Sex
- Industry of Employment by Sex
- Occupation by Sex

We store the filtered data and corresponding metadata as a list that is accessible by the user as “TSP2021” when the R library is loading. Please see Appendix F to find more details about the structure of the data and other descriptions.

4.2.2 Verify

Data verification is a crucial step to ensure that the data entering the demonstrator remains unaltered and free from any modifications that could potentially impact the quality of the results. In this regard, it is first essential to understand the structure of the data and identify the required elements for data linkage. To this end, we analysed the structure of the LSAY 2009, and the different variables of each of the 11 waves. This stage is detailed in Appendix N, reported by ISSR. The report reviews the related work utilising LSAY to identify research themes and explore related methodologies. This was followed by identifying conceptual, methodological, and practical challenges in linking area-based data (here the Census data) to person-level data and using spatial characteristics as predictors in the analysis of individual outcomes.

The outcome of the explanatory work enabled us to incorporate public metadata of the LSAY into the R library. We developed a main function that takes the longitudinal survey as input and uses the public metadata to verify the following:

- The data is coming from an LSAY survey.
- The dataset does not have modifications that interfere with the linkage process.
- Check for duplicate variable names.
- Check that the variable name is accepted by Stata.
- Verify if the dataset has a valid geographical identifier.
- Check that the year columns are in a consistent format – similarly, as the data will also be joined by year in some cases, it is important to check that both datasets have the year column in the same format.

The function produces a Boolean value that shows whether the incoming data is suitable for data linkage. If the incoming data is inadequate, the function will stop and notify the user with an alert, summarizing which conditions the data does not meet.

The metadata is stored in the R library as “LSAY_metadata” and is always accessible by the user when the R library is loading in the working environment. Please see Appendix F, to see the structure in more detail.

4.2.3 Convert

As mentioned in the second chapter, before the data linkage, the geographical units of the longitudinal survey and the target geographical unit need to be unified. In our case, the LSAY 2009 data is presented at the postcode level for the corresponding wave years, whereas the geographical unit of the TSP dataset is SA3 2021. To enable the conversion of postcode to SA3, a two-step process needed to be accomplished:

Stage 1. From Postcodes to Postal Areas (Non-ABS structure): Postcodes are exclusive data structures created and maintained by Australia Post. Access to the geographic boundaries of Postcodes is restricted by a commercial license, specifically for commercial and business purposes. For more information about this please refers to this URL. This is while to enable the data comparison, there is a crucial need to have an exact match between the Postcodes and an ABS structure.

Given the academic purpose of Geosocial demonstrator, we use it as a proxy of the Postcodes by ABS, so-called Postal Areas As we mentioned before, the Postal Areas are a non-ABS data structure that approximates the geographic boundaries of Postcodes. One limitation of using POAs as a proxy for Postcodes is the lack of data availability. While Postcodes can be updated annually, the Postal Area data may not be as frequently updated compared to actual Postcodes. To address this issue, we employed a workaround where we assumed that the closest Postal Area each year serves as the equivalent of a Postcode. This approach allowed us to approximate the correspondence between Postcodes and Postal Areas when dealing with data limitations or discrepancies. For instance, by utilizing the aforementioned assumption, we assumed that the closest non-ABS structure corresponding to Postcode data from 2010 would be Postal Area 2011. The table below displays the equivalent information for other waves.

	Wave 2	**Wave 3**	**Wave 4**	**Wave 5**	**Wave 6**	**Wave 7**	**Wave 8**	**Wave 9**	**Wave 10**	**Wave 11**
Postcode	Postcode 2010	Postcode 2011	Postcode 2012	Postcode 2013	Postcode 2014	Postcode 2015	Postcode 2016	Postcode 2017	Postcode 2018	Postcode 2019
POA (non-ABS)	Postal areas 2011	Postal areas 2011	Postal areas 2011	Postal areas 2011	Postal areas 2011	Postal areas 2011	Postal areas 2016	Postal areas 2016	Postal areas 2016	Postal areas 2016

Table 4: Postcodes to non-ABS structure

Stage 2. From POAs (Non-ABS Structure) to SA3 (ABS Structure): After data conversion to an equivalent non-ABS structure, we needed to aggregate the data into an ABS structure, here ABS SA3s. The best way to do this using the population-weighted concordance estimated by the ABS. The concordance enables the transformation between the Postal Areas and SA3-based on the ratio that expresses the percentage of the population that is contained within the two geographical units. In other words, the ratio expresses the percentage of the POA's population contained within the SA3 that intersects it. The following Table 5 shows an example of the concordance tables between Postal Areas 2011 and SA3 2011.

POA_CODE	POA_NAME	SA3	SA3_NAME	RATIO	PERCENTAGE
0800	0800	70101	Darwin City	1.0	100.0
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
7470	7470	60403	West Coast	1.0	100.0

Table 5: Concordances: Postal area 2011 to SA3 2011

Using the previous table, we assume that conformity is guided by the highest ratio of correspondence. For example, the postal area, 4053 in 2011 could be represented by five SA3s in 2011. Each SA3 covers a percentage of the postal area of 4053. The following table shows the five SA3s and the correspondence percentages.

Postal area 2011	SA3 2011	SA3 name 2011	Ratio correspondence
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

Table 6: Example: Postal area 2011 to SA3 2011

In this case, the postal area 4053 will be mapped to SA3 30201, Bald Hills – Everton Park, with a higher percentage of correspondence – i.e., 37.19%. The following map illustrates the aggregation process for the LSAY 2009, cohort 2011 from postal areas to SA3 2011:

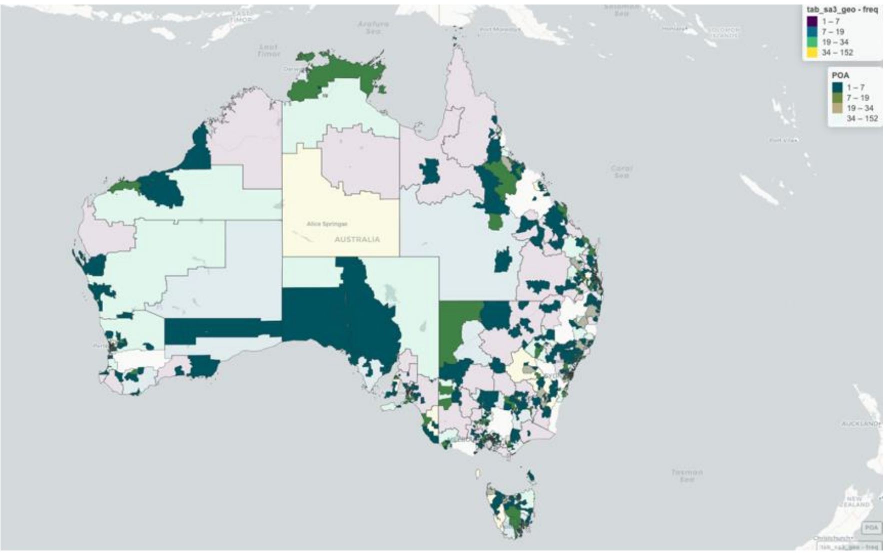


Figure 10: LSAY 2009 – Wave 2011 - Postal area 2011 to SA3 2011

The opaque colours represent the aggregate data in each postcode, and the light colours represent the equivalent SA3 2011. As we can see, some SA3s are represented by a postcode that covers a small area. For example, the SA3 70201, Alice Springs in North territory is only represented by a postcode 0870, which has a different size. However, in this case, the population is concentrated in postcode 0870, which increases the correspondence ratio. Even though the area covered by the SA3 is bigger in comparison to the corresponding postcode since the population representation is similar.

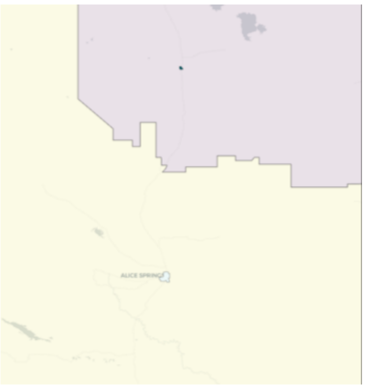


Figure 11: Correspondence between Postcode 0870 to SA3 70201 – (LSAY 2009 – Wave 2011)

The ABS also emphasizes that when using weighted population concordances, one assumption is that the accuracy of the results relies on the spatial distribution of the variable being concorded aligning with the population distribution across the Postal Areas (POAs). In the case of Alice Springs in the Northern Territory, for instance, it is evident that the population is spatially distributed in the same area across both geographical units. This alignment further supports the accuracy of the concordance results. This pattern is likely to happen in remote areas of Australia, particularly where the SA3s represent a larger geographical extent to the corresponding postcodes, while the population remains equally distributed in both spatial areas.

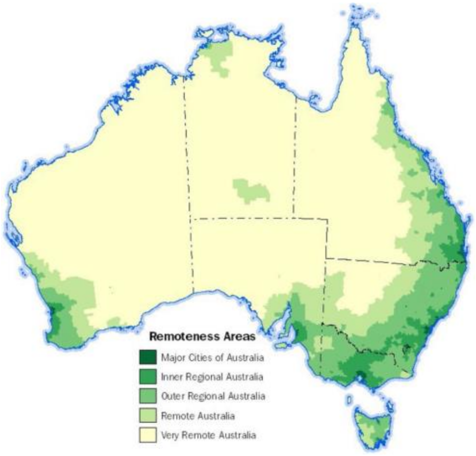


Figure 12: ASGS Edition 3 Remoteness Areas for Australia 2021

In contrast, the SA3s and postal areas cover similar areas in the major cities. The following map shows the aggregation process for the LSAY 2009, cohort 2011 from postal areas to SA3 2011 in greater Melbourne.

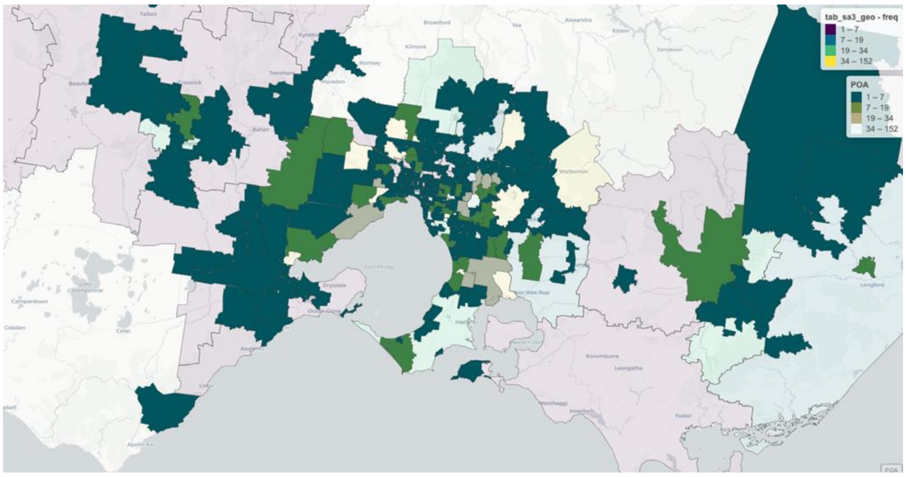
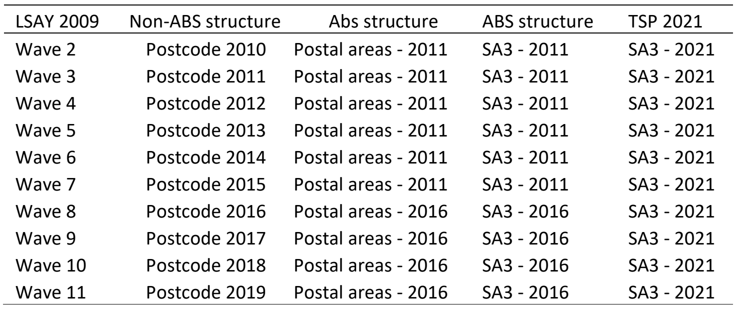


Figure 13: Postcodes to SA3 2011.

To better serve advanced users and enable them to personalise the data linkage, the R function provides an optional parameter for aggregating the data from postal areas to SA3s. This parameter allows the user to define a threshold ratio for the correspondence matching, allowing the user to restrict the matching based on the cut-off value. For example, if the user sets a cut-off ratio of 70% for correspondence, the candidate matches would need to have a ratio higher than that value. Otherwise, the matching would not be included in the analysis. For instance, in the previous example, the ABS matching between postcode 4053 and SA3 30201 would not happen as the 37.19% correspondence is lower than the 70% threshold set by the user.

SA3 to SA3 2021: To align the previous versions of geographical units with the TSP 2021, which reflects the latest changes in ABS statistical boundaries and geographical units, we need to convert them into the SA3 2021 standard. ABS updates the definition of these geographical units based on the changes in demographic distribution. Similar to Postal Areas, the ABS provides concordance tables to enable such conversion between different spatial units. The table below shows the transformation candidates for the LSAY cohort 2009 and its related waves, enabling linkage with the TSP 2021.



LSAY 2009	Non-ABS structure	Abs structure	ABS structure	TSP 2021
Wave 2	Postcode 2010	Postal areas - 2011	SA3 - 2011	SA3 - 2021
Wave 3	Postcode 2011	Postal areas - 2011	SA3 - 2011	SA3 - 2021
Wave 4	Postcode 2012	Postal areas - 2011	SA3 - 2011	SA3 - 2021
Wave 5	Postcode 2013	Postal areas - 2011	SA3 - 2011	SA3 - 2021
Wave 6	Postcode 2014	Postal areas - 2011	SA3 - 2011	SA3 - 2021
Wave 7	Postcode 2015	Postal areas - 2011	SA3 - 2011	SA3 - 2021
Wave 8	Postcode 2016	Postal areas - 2016	SA3 - 2016	SA3 - 2021
Wave 9	Postcode 2017	Postal areas - 2016	SA3 - 2016	SA3 - 2021
Wave 10	Postcode 2018	Postal areas - 2016	SA3 - 2016	SA3 - 2021
Wave 11	Postcode 2019	Postal areas - 2016	SA3 - 2016	SA3 - 2021

Figure 14: Concordances for residential postcodes in LSAY 2009

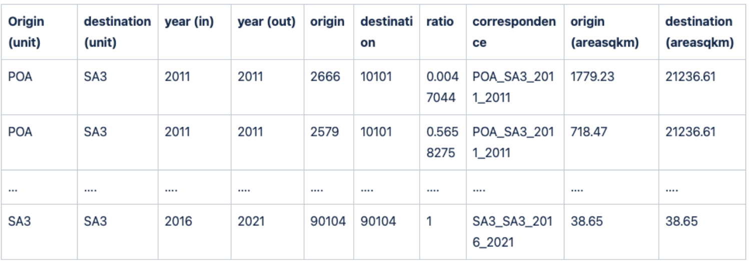
The transformation process from SA3-to-SA3 2021, uses the same methodology described previously in the aggregation from postal areas to SA3. The only difference is using a different concordance table.

Standardisation:

To standardise the previous process, we consolidate all the ABS correspondence tables available between 2011 and 2021 into the R library. In particular, the following concordance tables are used in the demonstrator:

- Postal areas 2011 to SA3 2011
- Postal areas 2016 to SA3 2016
- SA3 2011 to SA3 2016
- SA3 2016 to SA3 2021

Using these correspondence tables, we create a unique file to consolidate the correspondences in the following format:



Origin (unit)	destination (unit)	year (in)	year (out)	origin	destination	ratio	correspondence	origin (areasqkm)	destination (areasqkm)
POA	SA3	2011	2011	2666	10101	0.0047044	POA_SA3_2011_2011	1779.23	21236.61
POA	SA3	2011	2011	2579	10101	0.5658275	POA_SA3_2011_2011	718.47	21236.61
...	....	....	....	....	....	....	....	....	....
SA3	SA3	2016	2021	90104	90104	1	SA3_SA3_2016_2021	38.65	38.65

Figure 15: Concordance structure - R library

The previous table is stored in the R library as “concordances” and is accessible by the user when the R library is loading in the working environment. Please see Appendix F, to see the structure in more detail.

Using this table, we create a unique R function that takes all the previous transformations as input:

- Concordances table
- Geographical unit of the actual data
- Geographical unit to transform.
- Year of the data of the actual data
- Year of the data to transform.
- Geographical identifier of the actual data

For example, if we want to transform Postal areas to SA3, 2011, the inputs of the function would be:

- Geographical unit of the actual data = Postal areas
- Geographical unit to transform = SA3
- Year of the data of the actual data = 2011
- Year of the data to transform = 2011

The details of the implementation of this function are presented in the manual of the R library.

Quality indicator:

In addition, to assist users to interpret the quality of the conversion of their data by using the concordance tables, we included the Overall Quality Indicator created by the ABS. The quality indicator categorises the ratio of each concordance into one of three values:

- **Good** (Ratio greater than 0.9): The correspondence will convert data to a high degree of accuracy and users can expect the converted data will reflect the actual characteristics of the geographic regions involved.
- **Acceptable** (Ratio between 0.75 to 0.9): The correspondence will convert data to a reasonable degree of accuracy, though caution needs to be applied as the quality of the

converted data will vary and may differ from the actual characteristics of the geographic regions involved.

- **Poor** (Ratio lower than 0.75): There is a high likelihood that the correspondence will not convert data accurately and that the converted data should be used cautiously as it may not reflect the actual characteristics of many of the geographic regions involved.

We provide this information to the user as an Excel file, containing all the transformation details, the matching correspondence, and the quality ratio for each participant in the longitudinal survey.

4.2.4 Link

After unifying the geographical units, the R library can conduct cross-sectional and longitudinal data linkage as mentioned in the second chapter. The following figure shows a spatial-temporal representation of the two types of linkages.

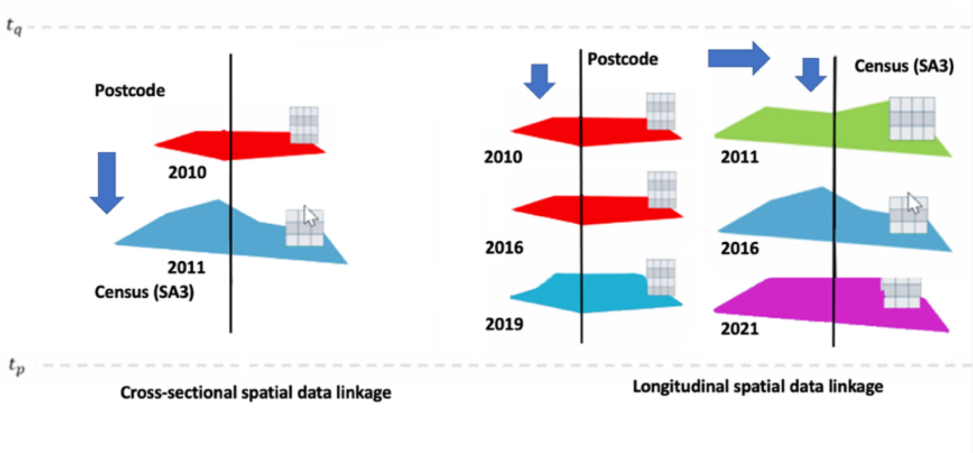


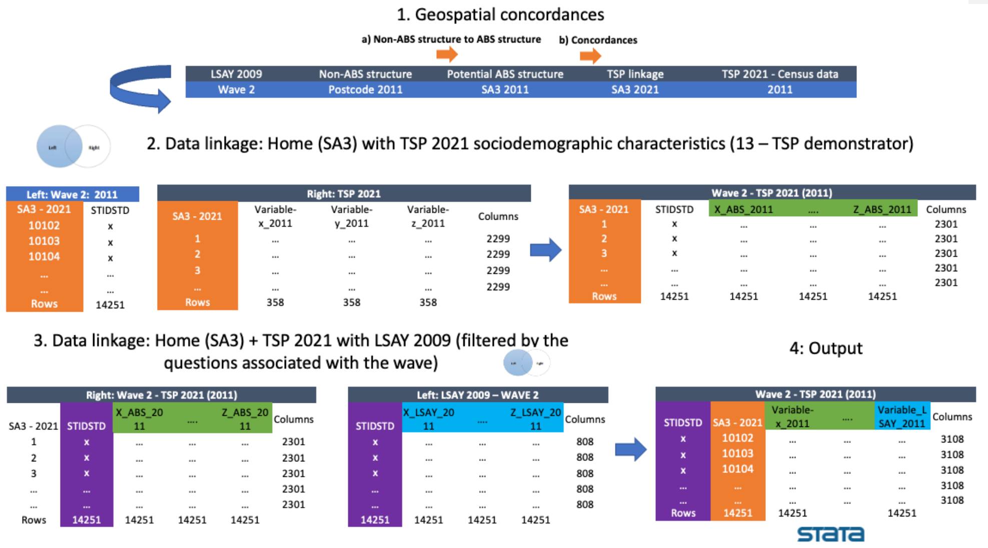
Figure 16: Spatial-Temporal representation of the types of linkage

The following sub-sections provide a detailed explanation of each type of linkage.

Cross-sectional data linkage: Refers to integrating spatial information from a particular cohort and wave of the longitudinal survey with an ABS data package. For example, merging any wave of the LSAY 2009 with the data census 2011 using the TSP 2021. In this case, we need to do the following steps:

1. Convert postcode into SA3 2021 (LSAY 09) using the concordance tables.
2. Filter TSP 2021 data with the variables from 2011.
3. Use a left join to combine both datasets, using SA3 2021 as the linkage key.
4. Incorporate metadata and export the data in Stata.

The process is shown step by step in the Figure below.



1. Geospatial concordances

a) Non-ABS structure to ABS structure

b) Concordances

LSAY 2009	Non-ABS structure	Potential ABS structure	TSP linkage	TSP 2021 - Census data
Wave 2	Postcode 2011	SA3 2011	SA3 2021	2011

2. Data linkage: Home (SA3) with TSP 2021 sociodemographic characteristics (13 – TSP demonstrator)

Left: Wave 2: 2011

SA3 - 2021	STIDSTD
10102	x
10103	x
10104	x
...	...
...	...
Rows	14251

Right: TSP 2021

SA3 - 2021	Variable-x_2011	Variable-y_2011	Variable-z_2011	Columns
1	...	...	...	2299
2	...	...	...	2299
3	...	...	...	2299
...	...	...	...	2299
Rows	358	358	358	

Wave 2 - TSP 2021 (2011)

SA3 - 2021	STIDSTD	X_ABS_2011	....	Z_ABS_2011	Columns
1	x	...	...	...	2301
2	x	...	...	...	2301
3	x	...	...	...	2301
...	...	...	...	...	2301
...	...	...	...	...	2301
Rows	14251	14251	14251	14251	

3. Data linkage: Home (SA3) + TSP 2021 with LSAY 2009 (filtered by the questions associated with the wave)

Right: Wave 2 - TSP 2021 (2011)

SA3 - 2021	STIDSTD	X_ABS_2011	....	Z_ABS_2011	Columns
1	x	...	...	...	2301
2	x	...	...	...	2301
3	x	...	...	...	2301
...	...	...	...	...	2301
...	...	...	...	...	2301
Rows	14251	14251	14251	14251	

Left: LSAY 2009 – WAVE 2

STIDSTD	X_LSAY_2011	....	Z_LSAY_2011	Columns
x	...	...	...	808
x	...	...	...	808
x	...	...	...	808
...	...	...	...	808
...	...	...	...	808
14251	14251	14251	14251	

4: Output

Wave 2 - TSP 2021 (2011)

STIDSTD	SA3 - 2021	Variable-x_2011	....	Variable_LSAY_2011	Columns
x	10102	...	...	...	3108
x	10103	...	...	...	3108
x	10104	...	...	...	3108
...	...	...	...	...	3108
...	...	...	...	...	3108
Rows	14251	14251		14251	

Figure 17: Cross-sectional linkage – workflow

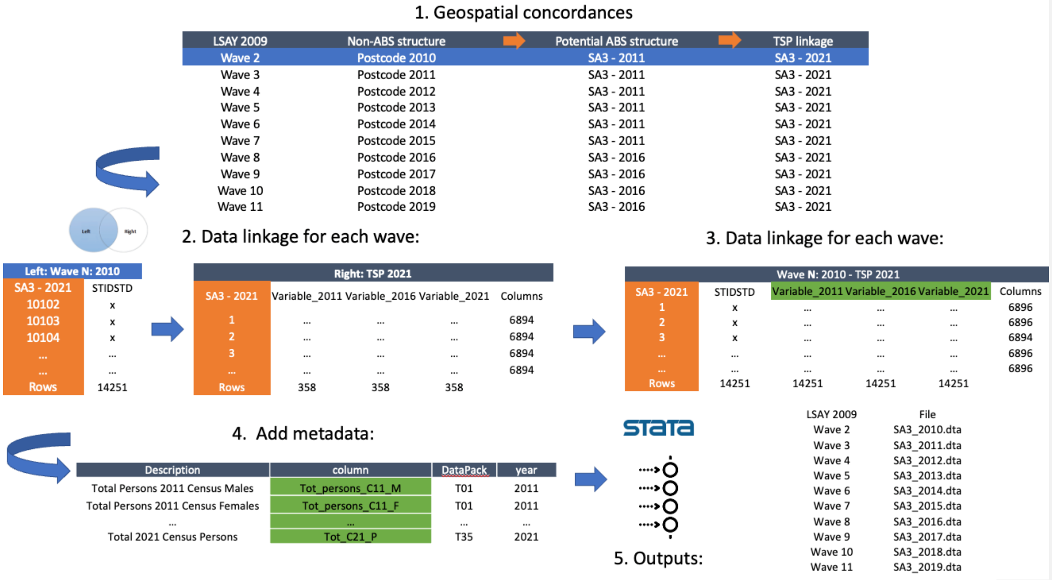
The output of this type of linkage is a Stata file where each row is a student, and the columns contain information about the TSP census 2011 (linked using SA3 2021) and all the variables of LSAY 2009 wave 2011.

Longitudinal: Refers to integrating spatial information from multiple rounds of the Census to different waves of the same LSAY cohort based on consistent geography and consistent Census data definitions. (e.g., when merging longitudinally consistent 2011, 2016 and 2021 Census data from the TSPs to different waves of the 2009 LSAY cohort).

In this case, we need to do the following steps:

1. Convert the postcode of each wave into SA3 2021 (LSAY 09) using the concordance tables.
2. Use a left join to combine the wave and TSP for each wave, using SA3 2021 as the linkage key.
3. Incorporate metadata and export the data in Stata.

The process is shown step by step in the Figure below.



1. Geospatial concordances

LSAY 2009	Non-ABS structure	Potential ABS structure	TSP linkage
Wave 2	Postcode 2010	SA3 - 2011	SA3 - 2021
Wave 3	Postcode 2011	SA3 - 2011	SA3 - 2021
Wave 4	Postcode 2012	SA3 - 2011	SA3 - 2021
Wave 5	Postcode 2013	SA3 - 2011	SA3 - 2021
Wave 6	Postcode 2014	SA3 - 2011	SA3 - 2021
Wave 7	Postcode 2015	SA3 - 2011	SA3 - 2021
Wave 8	Postcode 2016	SA3 - 2016	SA3 - 2021
Wave 9	Postcode 2017	SA3 - 2016	SA3 - 2021
Wave 10	Postcode 2018	SA3 - 2016	SA3 - 2021
Wave 11	Postcode 2019	SA3 - 2016	SA3 - 2021

2. Data linkage for each wave:

Left: Wave N: 2010	
SA3 - 2021	STIDSTD
10102	x
10103	x
10104	x
...	...
...	...
Rows	14251

Right: TSP 2021				
SA3 - 2021	Variable_2011	Variable_2016	Variable_2021	Columns
1	...	...	...	6894
2	...	...	...	6894
3	...	...	...	6894
...	...	...	...	6894
Rows	358	358	358	

3. Data linkage for each wave:

Wave N: 2010 - TSP 2021					
SA3 - 2021	STIDSTD	Variable_2011	Variable_2016	Variable_2021	Columns
1	x	...	...	...	6896
2	x	...	...	...	6896
3	x	...	...	...	6894
...	...	...	...	...	6896
...	...	...	...	...	6896
Rows	14251	14251	14251	14251	

4. Add metadata:

Description	column	DataPack	year
Total Persons 2011 Census Males	Tot_persons_C11_M	T01	2011
Total Persons 2011 Census Females	Tot_persons_C11_F	T01	2011
...	...	...	...
Total 2021 Census Persons	Tot_C21_P	T35	2021

5. Outputs:

LSAY 2009	File
Wave 2	SA3_2010.dta
Wave 3	SA3_2011.dta
Wave 4	SA3_2012.dta
Wave 5	SA3_2013.dta
Wave 6	SA3_2014.dta
Wave 7	SA3_2015.dta
Wave 8	SA3_2016.dta
Wave 9	SA3_2017.dta
Wave 10	SA3_2018.dta
Wave 11	SA3_2019.dta

Figure 18: Longitudinal linkage – workflow

Given the high dimension of the data, we split the data into the following structure:

Questionnaire information: lsay_y09.dta (the file containing the survey with all the waves)

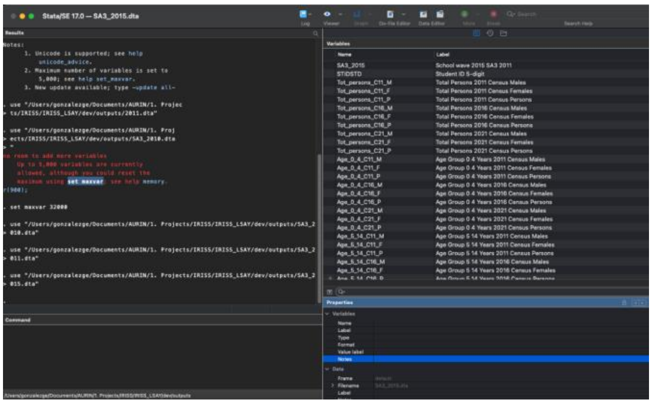
Geospatial data linkage: SA3_2010.dta, SA3_2011.dta, SA3_2012.dta, SA3_2013.dta, SA3_2014.dta, SA3_2015.dta, SA3_2016.dta SA3_2017.dta, SA3_2018.dta and SA3_2019.dta.

Each Stata file contains the following variables:

- SA3 2021: after changing the postcode from a non-ABS structure to an ABS structure and using the concordances to transform SA3 2011 to 2021.
- STD: Student id
- All the TSP variables were selected in the demonstrator.

Export:

Finally, the outputs of the data linkage are saved in Stata format, with the structure previously described. The Stata file preserves the longitudinal survey's metadata and includes the metadata of the new variables linked to the TSP. Figure 19 shows an example of how the output data would look after applying the longitudinal data linkage.



```
Notes:
      1. Unicode is supported; see help
         unicode_advice.
      2. Maximum number of variables is set to
         5,000; see help set_maxvar.
      3. New update available; type -update all-

. use "/Users/gonzalezge/Documents/AURIN/1. Projec
> ts/IRISS/IRISS_LSAY/dev/outputs/2011.dta"

. use "/Users/gonzalezge/Documents/AURIN/1. Proj
> ects/IRISS/IRISS_LSAY/dev/outputs/SA3_2010.dta
> "
no room to add more variables
    Up to 5,000 variables are currently
    allowed, although you could reset the
    maximum using set maxvar; see help memory.
r(900);

. set maxvar 32000

. use "/Users/gonzalezge/Documents/AURIN/1. Projects/IRISS/IRISS_LSAY/dev/outputs/SA3_2
> 010.dta"

. use "/Users/gonzalezge/Documents/AURIN/1. Projects/IRISS/IRISS_LSAY/dev/outputs/SA3_2
> 011.dta"

. use "/Users/gonzalezge/Documents/AURIN/1. Projects/IRISS/IRISS_LSAY/dev/outputs/SA3_2
> 015.dta"

.
```

Name	Label
SA3_2015	School wave 2015 SA3 2011
STIDSTD	Student ID 5-digit
Tot_persons_C11_M	Total Persons 2011 Census Males
Tot_persons_C11_F	Total Persons 2011 Census Females
Tot_persons_C11_P	Total Persons 2011 Census Persons
Tot_persons_C16_M	Total Persons 2016 Census Males
Tot_persons_C16_F	Total Persons 2016 Census Females
Tot_persons_C16_P	Total Persons 2016 Census Persons
Tot_persons_C21_M	Total Persons 2021 Census Males
Tot_persons_C21_F	Total Persons 2021 Census Females
Tot_persons_C21_P	Total Persons 2021 Census Persons
Age_0_4_C11_M	Age Group 0 4 Years 2011 Census Males
Age_0_4_C11_F	Age Group 0 4 Years 2011 Census Females
Age_0_4_C11_P	Age Group 0 4 Years 2011 Census Persons
Age_0_4_C16_M	Age Group 0 4 Years 2016 Census Males
Age_0_4_C16_F	Age Group 0 4 Years 2016 Census Females
Age_0_4_C16_P	Age Group 0 4 Years 2016 Census Persons
Age_0_4_C21_M	Age Group 0 4 Years 2021 Census Males
Age_0_4_C21_F	Age Group 0 4 Years 2021 Census Females
Age_0_4_C21_P	Age Group 0 4 Years 2021 Census Persons
Age_5_14_C11_M	Age Group 5 14 Years 2011 Census Males
Age_5_14_C11_F	Age Group 5 14 Years 2011 Census Females
Age_5_14_C11_P	Age Group 5 14 Years 2011 Census Persons
Age_5_14_C16_M	Age Group 5 14 Years 2016 Census Males
Age_5_14_C16_F	Age Group 5 14 Years 2016 Census Females
Age_5_14_C16_P	Age Group 5 14 Years 2016 Census Persons

Variables
Name
Label
Type
Format
Value label
Notes
Data
Frame
Filename
Label

/Users/gonzalezge/Documents/AURIN/1. Projects/IRISS/IRISS_LSAY/dev/outputs

Figure 19: Example output Stata

In addition, to the output, the demonstrator would store an Excel file called "Metrics.xlsx" that contains the following sheets:

- **Conversion**: Offers an overview of the conversion process applied for each wave. For example, the following table shows the metrics calculated for the conversions for three waves.

Cohort	N	Linkage	Sample attrition	Not Linked Areas	Not Linked Individuals	Good	Acceptable	Poor
2009	14,251	POA to SA3	0	4	192	12,265	464	1330
2010	14,251	POA to SA3	5,492	6	45	7,371	406	937
2011	14,251	POA to SA3	6,625	3	10	6,423	343	850
2019	14,251	POA to SA3	11,318	12	39	2,407	164	323
2009	14,251	SA3 to SA3	192	0	0	13,855	0	204
2010	14,251	SA3 to SA3	5,537	0	0	8,591	2	121
2011	14,251	SA3 to SA3	6,635	0	0	7,523	2	91
2019	14,251	SA3 to SA3	11,357	0	0	2,894	0	0

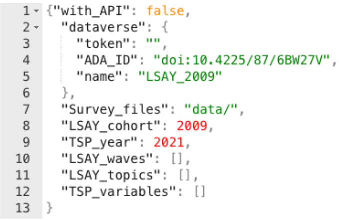
Table 7: Example Conversion metrics

- **Concordances:** This sheet provides a concise overview of the concordances applied in the conversion process. It provides specific details about the conversion parameters, including the geographical units, years, and file usage.
- **TSP variables:** Indicates the TSP 2021 variables used in the data linkage.

- **Ratio postcodes to SA3:** Indicates the correspondence used for each student.
- **Ratio postcodes SA3 to SA3 2021:** Indicates the correspondence used for each student.

4.2.5 Parameters file

The parameters file standardises all the definitions that are relevant to the data linkage process. Some of these definitions are the location of the longitudinal data, API credentials, wave, cohort, type of linkage, and variables of interest. The parameters file also helps users to modify the relevant information used in the data linkage and adapt these parameters to their necessities and study cases. The parameters file has a JSON structure that contains the following structure:



```
{"with_API": false,
  "dataverse": {
    "token": "",
    "ADA_ID": "doi:10.4225/87/6BW27V",
    "name": "LSAY_2009"
  },
  "Survey_files": "data/",
  "LSAY_cohort": 2009,
  "TSP_year": 2021,
  "LSAY_waves": [],
  "LSAY_topics": [],
  "TSP_variables": []
}
```

Figure 20: Parameter File – Structure

- *with_API*: is a Boolean variable:
- **True** if the user chooses to download the data using the ADA API.
    - token: ADA API token
    - ADA_ID: a unique identifier associated with the LSAY 2009.
    - name: name of the longitudinal survey.
- **False** if the user chooses to use the local environment.
    - Survey_files: path where the data is located. For example: “/Users/test/Downloads/data”
- *LSAY_cohort*: Year of the LSAY.
- *TSP_year*: Year of the TSP
- *LSAY_waves*: a vector with the years of the waves that would be included in the linkage. For example: [2011,2012,2013].
- *LSAY_topics*: a vector with the sub-major topic area that would be included in the linkage. For example: ["School","Student","Current"]
- *TSP_variables*: a vector with the TSP variables that would be included in the linkage. For example: [“Highest Non-School Qualification: Field Of Study By Age By Sex”]

4.2.6 Main Script

To read the parameters file, and execute the data linkage, we created a main script that loads all methods into the R working environment and executes the steps respectively. Appendix G explains the various functionalities of the main script in more detail.

4.2.7 Interface:

In this chapter, we present the interface designed for the demonstrator. The user interface provides an intuitive and visually driven space that simplifies the interaction with the demonstrator's features, catering to users who may prefer a more visual and interactive approach. The interface was designed to have a continuous flow where the user makes decisions at each step to build elements that best suit their needs. The interface serves two user flows, distinguished by the user's level of expertise, accommodating both mid-level and advanced users. The following figure summarises the two user flows.

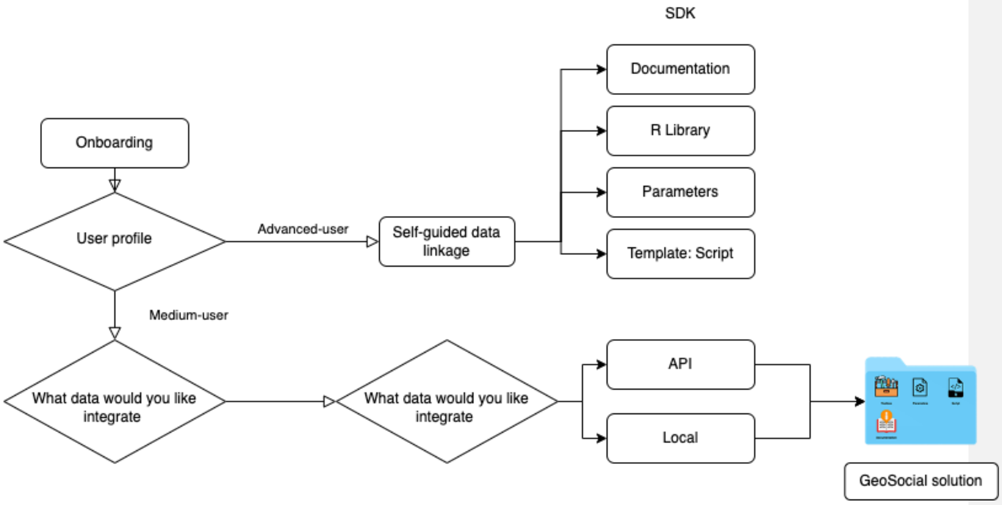
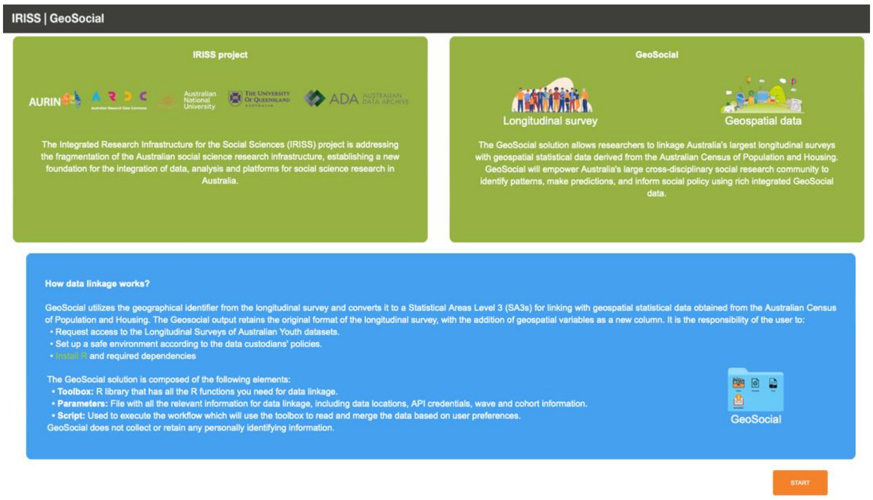


Figure 21: User flow - Interface

The flow starts with an onboarding page that offers an overview of the IRISS project and GeoSocial, followed by a detailed explanation of the data linkage process and the steps that medium and advanced users must take before using GeoSocial.



IRISS project

AURIN
ARDC
Australian National University
The University Of Queensland
ADA Australian Data Archive

The Integrated Research Infrastructure for the Social Sciences (IRISS) project is addressing the fragmentation of the Australian social science research infrastructure, establishing a new foundation for the integration of data, analysis and platforms for social science research in Australia.

GeoSocial

Longitudinal survey
Geospatial data

The GeoSocial solution allows researchers to linkage Australia's largest longitudinal surveys with geospatial statistical data derived from the Australian Census of Population and Housing. GeoSocial will empower Australia's large cross-disciplinary social research community to identify patterns, make predictions, and inform social policy using rich integrated GeoSocial data.

How data linkage works?

GeoSocial utilizes the geographical identifier from the longitudinal survey and converts it to a Statistical Areas Level 3 (SA3s) for linking with geospatial statistical data obtained from the Australian Census of Population and Housing. The Geosocial output retains the original format of the longitudinal survey, with the addition of geospatial variables as a new column. It is the responsibility of the user to:

- Request access to the Longitudinal Surveys of Australian Youth datasets.
- Set up a safe environment according to the data custodians' policies.
- Install R and required dependencies

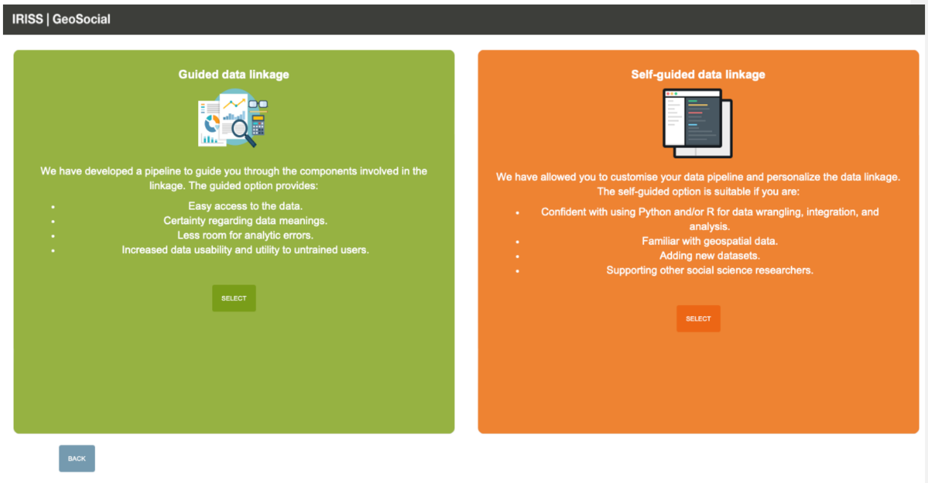
The GeoSocial solution is composed of the following elements:

- **Toolbox:** R library that has all the R functions you need for data linkage.
- **Parameters:** File with all the relevant information for data linkage, including data locations, API credentials, wave and cohort information.
- **Script:** Used to execute the workflow which will use the toolbox to read and merge the data based on user preferences.

GeoSocial does not collect or retain any personally identifying information.

Figure 22: Onboarding

After the user selects ‘Start’, the type of linkage can be selected.



Guided data linkage

We have developed a pipeline to guide you through the components involved in the linkage. The guided option provides:

- Easy access to the data.
- Certainty regarding data meanings.
- Less room for analytic errors.
- Increased data usability and utility to untrained users.

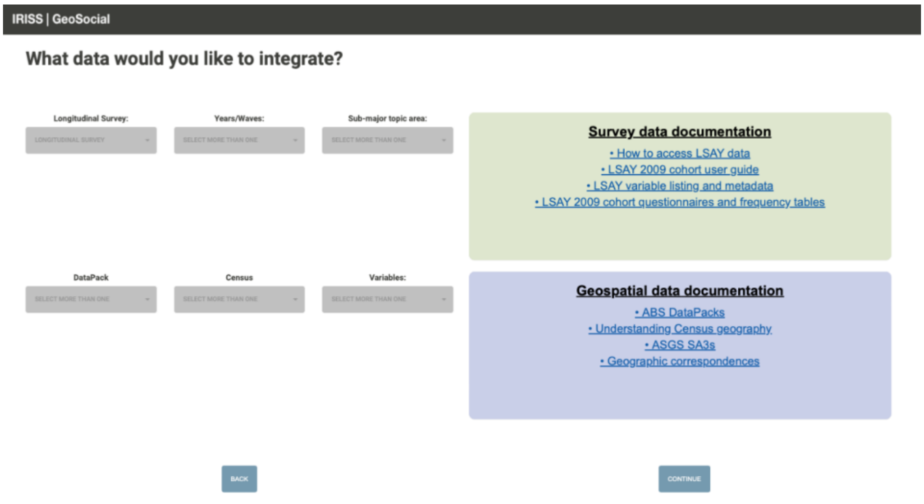
Self-guided data linkage

We have allowed you to customise your data pipeline and personalize the data linkage. The self-guided option is suitable if you are:

- Confident with using Python and/or R for data wrangling, integration, and analysis.
- Familiar with geospatial data.
- Adding new datasets.
- Supporting other social science researchers.

Figure 23: Type of linkage

The guided data linkage process uses a standard pipeline that meets the needs of most users, increasing data usability and utility to untrained users, and minimizing the risk of analytical errors. Users can choose the specific datasets, years, variables, and other relevant information they want to link via drop-down lists. The following figure shows an example of the different elements that the user can personalise in the guided data linkage.



IRISS | GeoSocial

What data would you like to integrate?

Longitudinal Survey: LONGITUDINAL SURVEY

Years/Waves: SELECT MORE THAN ONE

Sub-major topic area: SELECT MORE THAN ONE

DataPack: SELECT MORE THAN ONE

Census: SELECT MORE THAN ONE

Variables: SELECT MORE THAN ONE

Survey data documentation

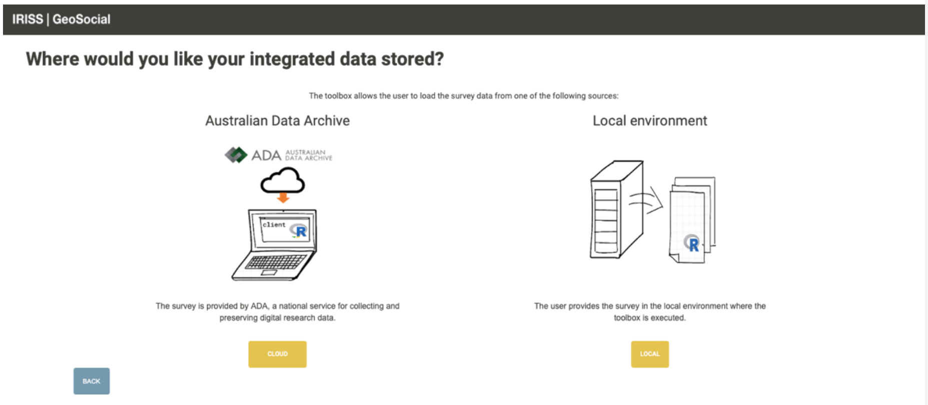
- How to access LSAY data
- LSAY 2009 cohort user guide
- LSAY variable listing and metadata
- LSAY 2009 cohort questionnaires and frequency tables

Geospatial data documentation

- ABS DataPacks
- Understanding Census geography
- ASGS SA3s
- Geographic correspondences

Figure 24: Guided data linkage

After personalising the linkage, the user can select where to locate the data in the work environment.



IRISS | GeoSocial

Where would you like your integrated data stored?

The toolbox allows the user to load the survey data from one of the following sources:

Australian Data Archive

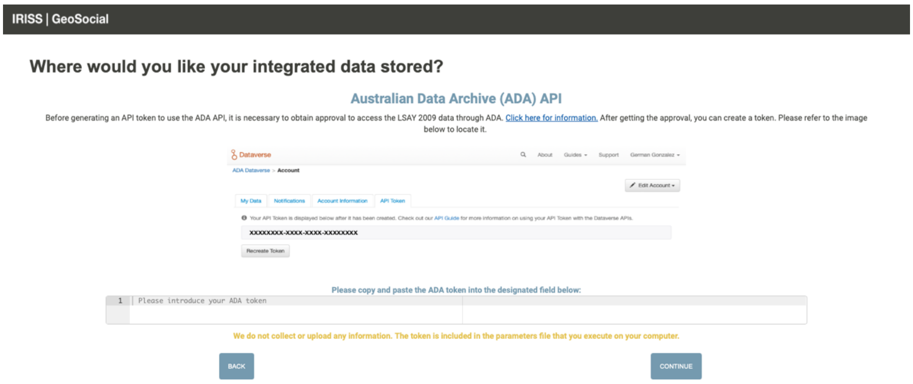
The survey is provided by ADA, a national service for collecting and preserving digital research data.

Local environment

The user provides the survey in the local environment where the toolbox is executed.

Figure 25: Location

The user has the option to choose between utilizing the ADA API or loading the data in the local environment. The following two figures illustrate these two choices.


IRISS | GeoSocial

Where would you like your integrated data stored?

Australian Data Archive (ADA) API

Before generating an API token to use the ADA API, it is necessary to obtain approval to access the LSAY 2009 data through ADA. Click here for information. After getting the approval, you can create a token. Please refer to the image below to locate it.

Your API Token is displayed below after it has been created. Check out our API Guide for more information on using your API Token with the Dataverse APIs.

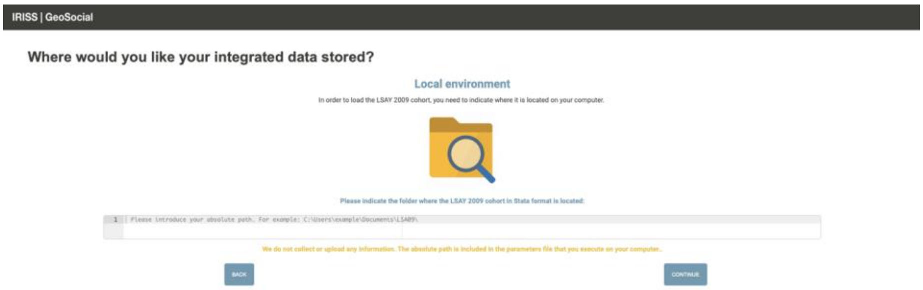
XXXXXXXX-XXXX-XXXX-XXXXXXXX

Please copy and paste the ADA token into the designated field below:

```
1 | Please introduce your ADA token
```

We do not collect or upload any information. The token is included in the parameters file that you execute on your computer.

Figure 26: Load the data from the Australian Data Archive – API


IRISS | GeoSocial

Where would you like your integrated data stored?

Local environment

In order to load the LSAY 2009 cohort, you need to indicate where it is located on your computer.

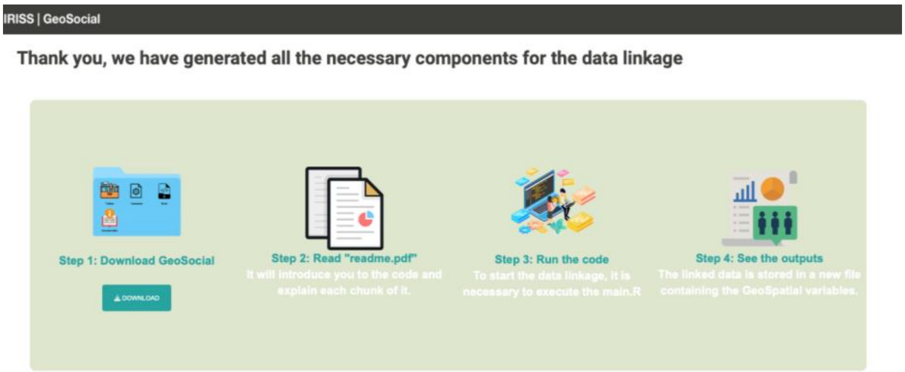
Please indicate the folder where the LSAY 2009 cohort in Stata format is located:

```
1 | Please introduce your absolute path. For example: C:\Users\example\Documents\LSA09\
```

We do not collect or upload any information. The absolute path is included in the parameters file that you execute on your computer.

Figure 27: Load the data from a local environment.

The preferences selected by the user are stored in the parameters file. In the final step, the user can download a folder that contains all the necessary elements to execute the data linkage.


IRISS | GeoSocial

Thank you, we have generated all the necessary components for the data linkage

Step 1: Download GeoSocial

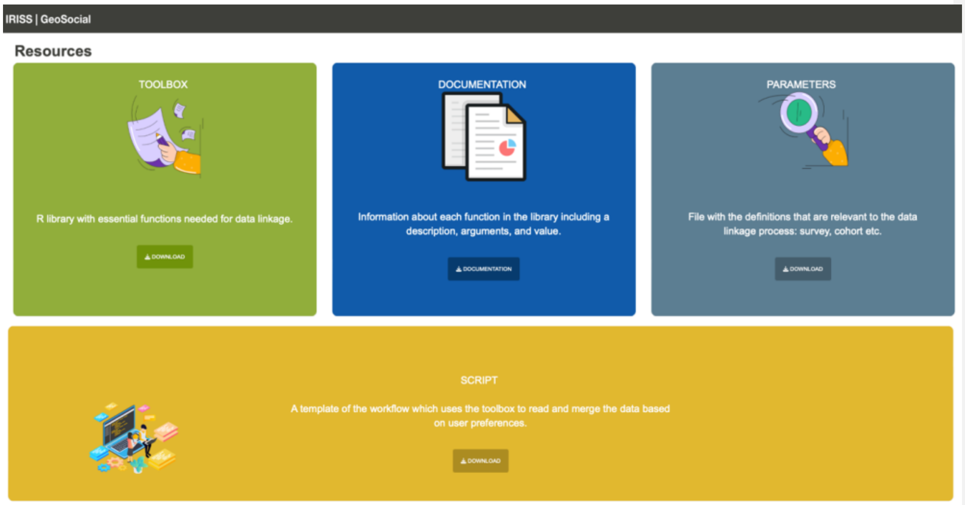
Step 2: Read "readme.pdf"
It will introduce you to the code and explain each chunk of it.

Step 3: Run the code
To start the data linkage, it is necessary to execute the main.R

Step 4: See the outputs
The linked data is stored in a new file containing the GeoSpatial variables.

Figure 28: Download the Geosocial solution – Guided data linkage

The second user flow is self-guided, allowing users to customize the pipeline and personalize data linkage. Users can download each element that makes up the GeoSocial solution, as well as a template of the script that implements the workflow.



R library with essential functions needed for data linkage.

Information about each function in the library including a description, arguments, and value.

File with the definitions that are relevant to the data linkage process: survey, cohort etc.

A template of the workflow which uses the toolbox to read and merge the data based on user preferences.

Figure 29: Self-guided data linkage

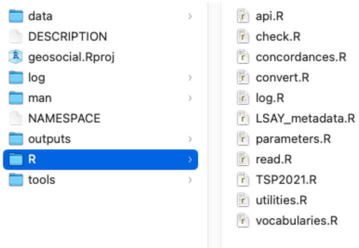
5 Transferability and Deployment:

All the project services were created using a block system that enables new researchers to re-utilise each piece of work or even combine services to create a new structure. This chapter provides guidelines about how to modify the R library's source code and how to deploy the interface.

5.1 Development - R library

The R library that the users use is a tar.gz which compiles all the elements of the library such as functions, documentation, and datasets. This library can be updated/maintained in the future using the source code. The source code is available on GitHub at the following link: https://github.com/AURIN-OFFICE/geosocial/

The R library contains the following files:



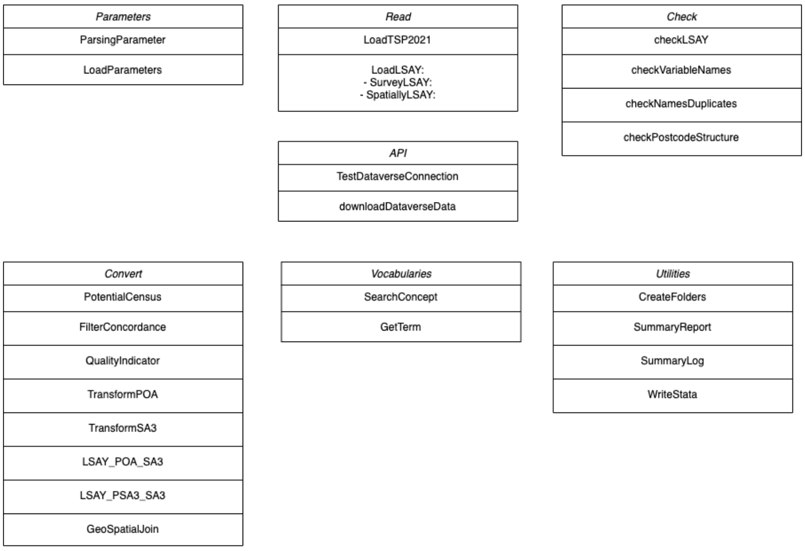
data
DESCRIPTION
geosocial.Rproj
log
man
NAMESPACE
outputs
R
tools

api.R
check.R
concordances.R
convert.R
log.R
LSAY_metadata.R
parameters.R
read.R
TSP2021.R
utilities.R
vocabularies.R

Figure 30: R library Structure

The important elements in the R library are:

- DESCRIPTION file contains all the necessary information about the R library, including its description, authors, dependencies, and other relevant details.
- The Geosocial.Rproject file includes the project environment and compiles the R library.
- The folder named "R" includes all the functions that are grouped together in an R script. Each individual .R file has unique functions that are accessible within the package. A chart is provided below to give an overview of the functions included in each file.



Parameters
ParsingParameter
LoadParameters

Read
LoadTSP2021
LoadLSAY:
- SurveyLSAY:
- SpatiallyLSAY:

API
TestDataverseConnection
downloadDataverseData

Check
checkLSAY
checkVariableNames
checkNamesDuplicates
checkPostcodeStructure

Convert
PotentialCensus
FilterConcordance
QualityIndicator
TransformPOA
TransformSA3
LSAY_POA_SA3
LSAY_PSA3_SA3
GeoSpatialJoin

Vocabularies
SearchConcept
GetTerm

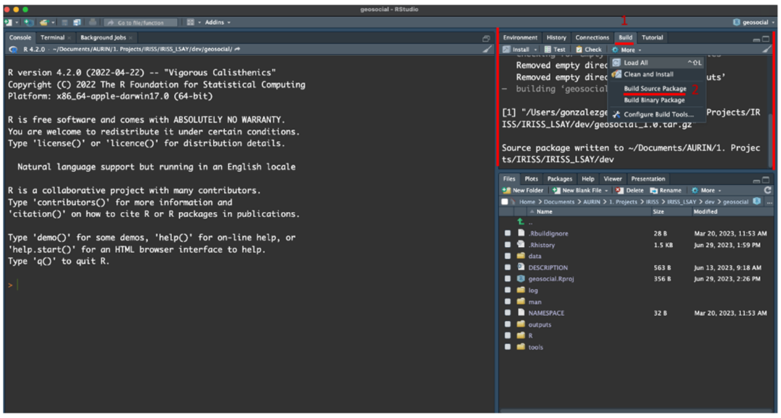
Utilities
CreateFolders
SummaryReport
SummaryLog
WriteStata

Figure 31: “geosocial” – Structure

The description of each function, input and output is described in Appendix F.

Compile the R library:

If you want to make changes to a function in the R library, you can modify the files into .R format and compile the R library into a new tar.gz file. To do this, simply open the project geosocial.Rproj in R studio, go to the right menu, click on the "Build" tab, select "more" and then click on "Build Source package". See the following figure for a visual representation of this process.



```
R version 4.2.0 (2022-04-22) -- "Vigorous Calisthenics"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

```
   Removed empty direc
   Removed empty direc                uts'
─  building 'geosocial

[1] "/Users/gonzalezge                Projects/IR
ISS/IRISS_LSAY/dev/geosocial_1.0.tar.gz"

Source package written to ~/Documents/AURIN/1. Projec
ts/IRISS/IRISS_LSAY/dev
```

```
Home > Documents > AURIN > 1. Projects > IRISS > IRISS_LSAY > dev > geosocial
  Name              Size     Modified
  ..
  .Rbuildignore     28 B     Mar 20, 2023, 11:53 AM
  .Rhistory         1.5 KB   Jun 29, 2023, 1:59 PM
  data
  DESCRIPTION       563 B    Jun 13, 2023, 9:18 AM
  geosocial.Rproj   356 B    Jun 29, 2023, 2:26 PM
  log
  man
  NAMESPACE         32 B     Mar 20, 2023, 11:53 AM
  outputs
  R
  tools
```

Figure 32: R Library - Compilation

5.2 Deployment Interface:

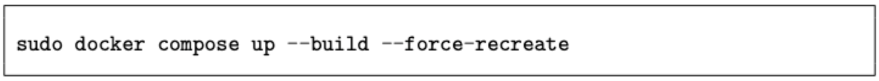
We created the user interface using R and Shiny, along with several dependencies and compiling engines like Latex. To make it easier to deploy and avoid any issues with different environments, we opted to use a containerisation process that provides useful features like:

- **Requirements**: Control of all the dependencies and requirements that the application needs, such as specific versions of programming language run times and other software libraries.
- **Isolation**: Containers give developers the ability to create predictable environments that are isolated from other applications.
- **Agility**: Containers accelerated development, improved consistency across environments, empowered autonomous teams improving productivity and quality.

We have containerized the user interface in a Docker image that utilizes a lightweight version of R in an Ubuntu environment. To learn more, please refer to the Docker repository utilized in the deployment. Additionally, we have included all the necessary functionalities for the interface in the "Dockerfile".

To execute the container, you need to follow the next steps:

- **Install and execute Docker:** Follows the suitable tutorial: Windows, Mac or Ubuntu.
- **Clone the repository:** git clone https://github.com/AURIN-OFFICE/geosocial_interface.git
- **Execute the container**: To execute the container, you need to open a cmd/terminal and access the folder where is the container. Now you can build and execute the container using the command:



```
sudo docker compose up --build --force-recreate
```

Figure 33: Docker compose.

Finally, the container will be executed in localhost (0.0.0.0) using port 80. You can change the port and the IP, by changing the Dockerfile. The Figure shows an example of the execution process.

We recommend hosting the interface in a Nectar machine, with the following characteristics:

- Flavour: m3.medium
- VCPUs: 4
- RAM: 8GB
- Disc size: 30GB
- Image/OS: Ubuntu 20.04.5 LTS (Focal Fossa)

6 Conclusions:

This report summarises important groundwork completed towards the preliminary Geosocial service design and provides an implementation of it through the development of a demonstrator.

The demonstrator was presented to the HASS RDC Technical Advisory Group Meeting, held on Friday 30 June. In this space were presented the user requirements, the preliminary service design, and a live demo of the demonstrator.

Following the demonstrator stage, the formal design and development stage is planned to commence with a pilot program to assess user feedback and identify opportunities for improvement. In the next stage of the project, we want to integrate with the other work package outputs, such as VASSSAL (vocabulary service) and SPIRE (survey package).

The workshops and consultations led to the identification of several issues that are out of the scope of the project's prototyping phase but are important for the long-term future of the service:

- First, a new administrator needs to be chosen to run the service after June 2023. The future administrator will be responsible for maintaining and updating the service. For example, the scripts will need to be reviewed and possibly updated every time a new wave of a panel survey is published. In addition, changes in existing software, e.g., R libraries, must also be monitored.
- Second, the service will need to expand in terms of available data. This means including more ADA survey data and more spatial data in the service, as well as offering new types of linkages, i.e., linking data that use different spatial identifiers or different versions of the same identifiers. Such linkages would require geographical dictionaries with concordances between various classifications, which are currently unavailable.

Appendix

Appendix A – Technical Report 2: Data Linkage Report

This document was created by Matthias Kubler, Denise Clague, Tomasz Zajac, Wojtek Tomaszewski, and Jonathan Corcoran of ISSR on 31/05/2023

7 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.[4]

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

[4] 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).

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, 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.

8 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.

8.1 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 8. 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.

8.2 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.

8.3 2.3 Insights from the scoping review

8.3.1 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.

8.3.2 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.

8.3.3 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.

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

8.3.4.1 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).

8.3.4.2 '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).

8.3.4.3 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).

8.3.4.4 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.

8.4 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.

8.5 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[5]. 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.

[5] 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.

8.6 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.

9 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.

9.1 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.

9.1.1 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.

9.1.2 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.

9.1.3 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).

9.1.4 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

9.1.5 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.

9.1.6 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

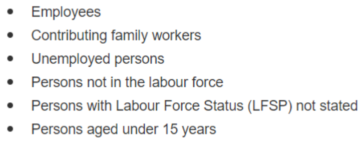
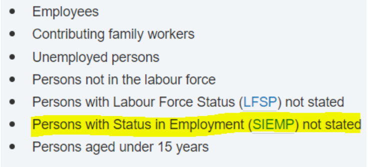
The reported categories include the three categories above, which are directly taken from the responses to the question. However, as the question is asked of a sub-population it is likely that the ABS checks responses to the preceding questions to derive those that should be coded as ‘not applicable’ 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.

9.1.7 Change in categories’ (dollar) ranges

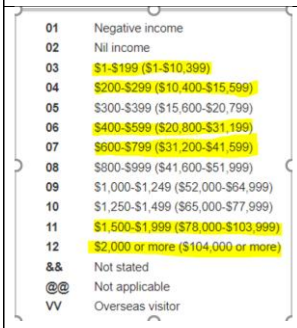
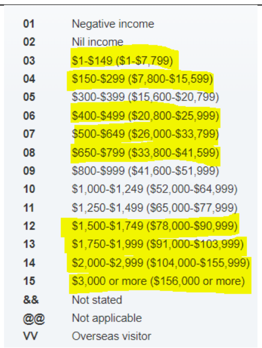
It is not uncommon that dollar ranges are updated for relevant variables (e.g., affecting personal and/or household income or rent/mortgage payment variables) in the Census. The example shown here is for Total Personal Income (weekly). The highlighted categories in the Census 2011 column do not exist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

9.1.8 Change in category order

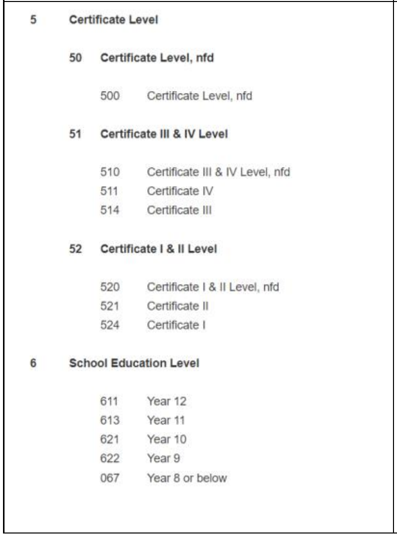
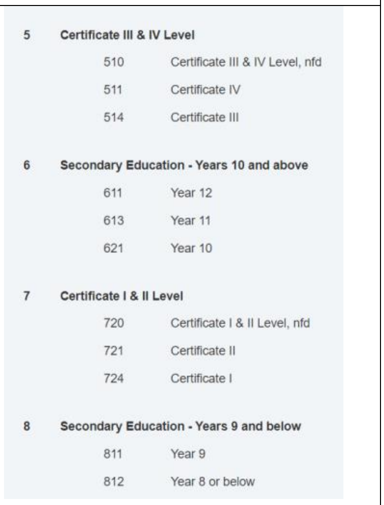
The example here relates to the variable Level of Highest Educational Attainment. The variable is derived from questions on non-school and post-school education and the derivation rules define the order of the categories. In 2016 the order of the categories was changed to align with ASCED. The change 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 reordered 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

9.1.9 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.

9.2 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.

9.2.1 ABS Census documentation (2011-2016[6])

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.

[6] 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.

9.2.2 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.

10 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.

10.1 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:

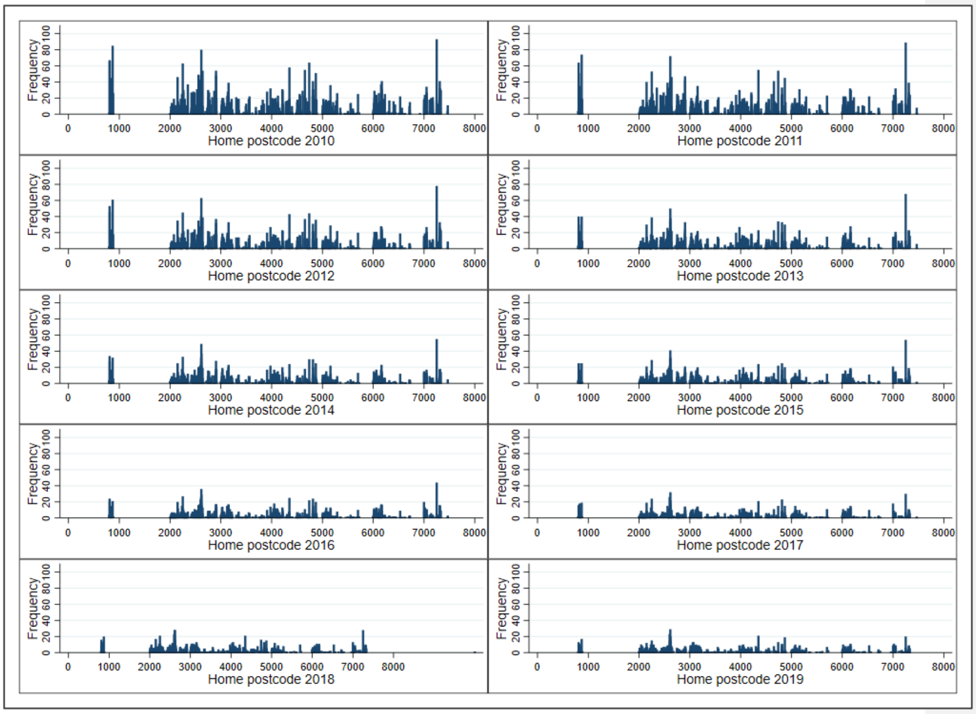
- The number of school postcodes (n=290) in the data (first wave) is much smaller than the number of residential postcodes in the following waves (between 974 and 1,216).
- There are some individual cases with missing information on residential postcodes in 2010, 2011, 2012, 2013, 2014, 2018 and 2019 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 9. 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 34. Residential postcodes, frequencies for waves 2010 to 2019



10.2 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 10. 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 11. 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**

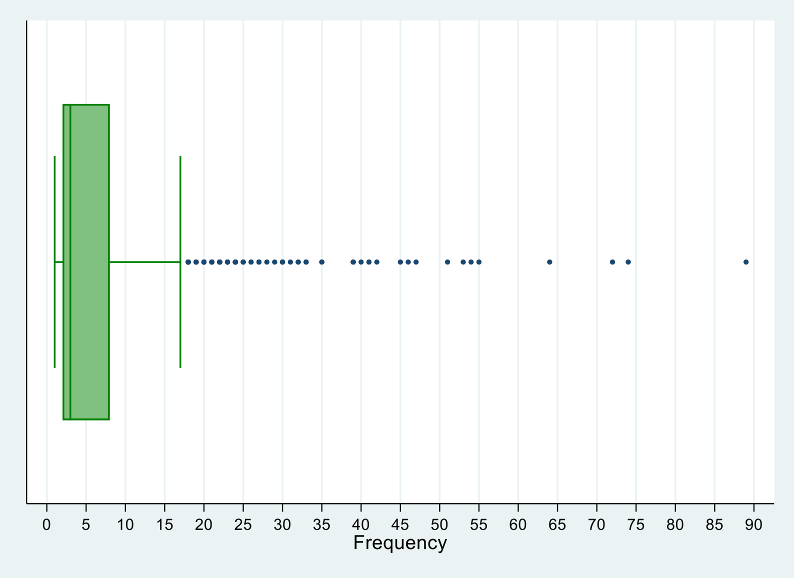
The population aged 17 years in the 2011 ABS Census was 282,055 (excluding people who were migratory, offshore or had no usual address). The 7,620 respondents in the third (2011) wave of LSAY constitute about 2.7% of that population.

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 35. Residential postcodes, distribution in LSAY 2011 wave



10.3 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*)[7] 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 12. 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).

10.3.1 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

[7] 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[8] – 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

[8] 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 13. 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

10.4 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.

11 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.

11.1 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?).

11.1.1 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 [9]. 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

[9] 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[10]. 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:

[10] 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 14. 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.

11.2 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.

11.2.1 Relevant aspects of the data linkage

ABS data in scope

Census rounds

The ABS offers TSPs containing statistics based on consistent data definitions[11] 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

[11] 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 15. 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.

11.3 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.

12 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.

13 Appendix 1: The scope and design of LSAY

13.1 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.

13.2 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 16. 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

13.3 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.

13.4 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.

14 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

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

The below table only considers postcode and POA geographies[12] 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)^

[12] 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

16 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].

Appendix B – Software development outputs and their assignments

Output	Description	Responsible	Accountable	Consulted	Informed
Prioritised requirements list (PRL)	Prioritised list detailing all project requirements. Contents initially drawn from the ISSR-AURIN Workshop 16 August 2022. Items on this list are initially functional, drawn from data, user, environment levels, and will expand to non-functional through iterative development. Administration of this is critical for software development to evolve effectively.	Social Data Scientist (AURIN)	Partner Lead (AURIN)	Work Package Lead (ISSR)	Work Package Members (ISSR + AURIN)
Timeboxes	Structured time periods within which requirements are actioned by the Software Developer. These are expected to be fortnightly and planned, tracked and reviewed during work package meetings.	Software Developer (AURIN)	Partner Lead (AURIN)		Work Package Lead (ISSR) Work Package Members (ISSR + AURIN)
Examples and demonstrators	Developed demonstrating a requirement's functionality. Tracking of these is a responsibility of the Partner Manager with oversight from the Work Package Manager.	Social Data Scientist (AURIN)	Partner Lead (AURIN)	Work Package Lead (ISSR)	Work Package Members (ISSR + AURIN)
Testing	Testing of incremental solutions are key stages in development. Outputs are formal testing procedures for use by the developer to ensure suitable coverage and	Social Data Scientist (AURIN)	Partner Lead (AURIN)		Work Package Members (ISSR + AURIN)

	acceptance criteria are met. To be report on in regular fortnightly meetings	Software Developer (AURIN)			
Versioning	Development will follow version control methods to handle revisions and their tracking.	Social Data Scientist (AURIN) Software Developer (AURIN)	Partner Lead (AURIN)		
Supporting materials	Supporting materials are critical to ensure project quality remains at a certain level. Outputs are well-documented codebases, documentation and user guides.	Social Data Scientist (AURIN) Software Developer (AURIN)	Partner Lead (AURIN)		Work Package Lead (ISSR + AURIN)
Evolving solution	The current solution under development, together with work in progress within timeboxes. Outputs are combined in increments that contribute to the evolving IRISS solution.	Social Data Scientist (AURIN) Software Developer (AURIN)	Partner Lead (AURIN)	Work Package Lead (ISSR)	Work Package Members (ISSR + AURIN)
Project technical reports	Feedback, learnings and benefits from solutions are to be recorded and this continues until the project's successful delivery and closure.	Social Data Scientist (AURIN) Software Developer (AURIN)	Partner Lead (AURIN)	Work Package Lead (ISSR)	Work Package Members (ISSR + AURIN)

Appendix C – Initial user data

The following material is sourced from WP3 Technical Report 1, 31 August 2022

Purpose

The purpose of this document is to discuss the audience profile relevant to the WP4 Demonstrator 1, and the WP3 GeoSocial Service more broadly. It covers the technical skills that different users have, as well as different thematic/substantive interests. Recommendations for the target user profile for the purpose of WP4 Demonstrator 1 and WP3 are also outlined in the document.

Technical requirements

User requirements change with the user's skills levels. Most advanced users have the technical skills and tools to merge and transform data on their own. They are also more likely to be interested in using sophisticated statistical methods that require unit-level data. For such users, ease of access to the data as well as high-quality documentation are essential. Moreover, for such users, flexibility is important.

Less skilled users will require more advanced tools to be able to utilise available data. For example, users less comfortable with data processing need easy access to the data and documentation, which would require more support in data preparation. Ideally, they could access ready-to-use or easy-to-link datasets, with curated content such as derived variables. Such data enhancements (e.g. linked dataset, derived variables) might not be crucial for more advanced users but could be still appreciated as time-saving measures (if flexibility is sufficiently retained).

The least advanced users, who neither can process data themselves nor are interested in complex statistical modelling, would require significantly more features provided as part of the service. They would need not only ready-to-use data but also additional functionalities allowing data analysis and visualisation. Again, such tools might not be necessary for more advanced users but could prove useful, e.g. by allowing quick descriptive analysis.

The user base is likely inversely proportional to the level of technical skill. The tools with more features and aimed at low-skilled users might potentially have a wider user base, possibly stretching beyond the academic and research community – such as policy analysts working in government agencies. However, the service aimed at this type of user would need to be more developed and would require more curated data assets to operate. Developing this kind of service and data requires significant time and resources. Such more curated service and data are also much less flexible. Therefore, more time is required for researching the needs and potential use cases before commencing work on developing them.

Below we present user requirements for three discrete types of users. The actual users will quite often fall in between these categories, but the classification helps to illustrate what types of functionalities are required at various levels of skills.

Table A1.1: User requirements and benefits.

Type of user	User requirements	Benefits
High skills level - Advanced user, capable of performing advanced data transformations, merging datasets, deriving variables, and interested in sophisticated statistical methods. IRISS personas: Evan – data scientist/analyst Martin – researcher (data collector) Danielle – researcher (data analyst)	Easy access to the data frees the user from negotiating data access with data custodians and securing permissions for data linkage. Flexibility in data formats. Information on linkage keys for geosocial data (e.g., SA2 codes) or concordance tables between different geographies allowing individual-to-spatial data linkage (e.g., postcodes to geographical classifications). Data documentation that includes data vocabularies and classifications, data comparability (e.g., akin to IPUMS), description of data limitations (e.g., limitations to survey data aggregation by geography arising from sampling, notes on data quality/ reliability – especially in the case of administrative data, etc.). Instructions for data derivations (e.g. on how to derive income from ATO data).	Faster access to the data. Removes the need for data conversions and some processing. Faster and easier data merging. Certainty regarding data meanings. Reduced time needed for developing data transformations code. Less room for errors leading to data breaches. Improved discoverability of data. Easier access to the right data.

	Confidentiality and privacy protection measures. Secure environment to process data if data require protection. Search & discovery – particularly important as the list of datasets grows.	
Medium skills level - A user advanced in applying statistical sophisticated statistical methods but less comfortable with data manipulation. IRISS personas: Martin – researcher (data collector) Danielle – researcher (data analyst) Yosef – data analyst	In addition to the previous: Linked ready-to-use datasets (e.g. individual-level survey data appended to geographical/spatial data). A user could potentially select from a list of datasets and variables to be downloaded/analysed. Derived variables or a library of code that could be used for data derivations.	Wider access to the data. Even easier and faster access to the analysis-ready data. Standardised and comparable measures. Reduced data processing
Low skills level – A user who cannot manipulate data and who might be more interested in descriptive analysis than advanced statistical modelling. IRISS personas:	Easy access to the data frees the user from negotiating data access with data custodians and securing permissions for data linkage (like for previous types of users). Data documentation explaining the variables and data limitations (focused more on the meaning of the data than the technical process of variable derivation than in the case of more advanced users).	Easy access to the data. Certainty regarding data meanings. Less room for analytic errors. Increased data usability and utility to untrained users.

Serena – policymaker Yosef – data analyst	Safeguards preventing incorrect use of data (e.g. making sure that a dataset can be used to produce representative community profiles) Interface for data analysis and visualisation (e.g. Gapminder data animations, Shiny, Tableau). Built-in confidentiality and privacy protection measures ensuring that only safe outputs are available.	Reduction of the risk of data breaches.

Appendix D – Original prioritised requirements list

The following material is sourced from WP3 Technical Report 1, 31 August 2022

Must have	
UR1	No login required to browse/explore scripts so that users can explore what the service has to offer without investing too much time.
UR2	Example Script
UR3	Script selection (selection of datasets to be integrated)
UR4	Accessible coding style - widely used language (R), scripts easy to understand and modify
UR5	Provide a correspondence flag (warning) when a classification might have changed, e.g., when survey data include an older version of SA2 than geographical data that will be added.
UR6	Ability to read/input survey data (HILDA)
UR7	Ability to browse and select survey data (e.g., respondent/HH files)

UR8	Ability to browse and select spatial data (e.g., year, SA1/2 files)
UR9	Join survey data first to perform left-join
UR10	Integrate multiple waves of survey data
UR11	Add multiple versions (years) of geographical data
UR12	Read spatial data
UR13	Check input data (ids, variable names, e.g., duplicated variable names or whether the variables to be added are already present in the dataset)
UR14	Meaningful error statements if encountered
UR15	Ability to link one survey file with another geospatially (using geographical identifier such as SA2)
UR16	Ask user to define a file input/output location
UR17	Read Stata input files
UR18	Provide a link to correspondence information (ABS)
UR19	Provide a link to signup/access AURIN data
UR20	Provide high level information on what survey datasets can be linked and to what
UR21	Provide links to some survey metadata (HILDA, ABS)
UR22	Retain everything from the input files (e.g., data structure, variable values and labels) for both survey and spatial
UR23	Outputs integrated data (e.g., in Stata and R formats)
UR24	Outputs the script executed
UR25	Outputs a list of input data used and basic information (e.g., id used for integration)
UR26	Design and develop following FAIR principles for research software

UR27	Outputs a brief data linkage report (input data, output data, join ID).
Should have	
UR28	Graphic user interface (in addition to the example script)
UR29	Ability to select loading/save locations etc.
UR30	Join multiple (more than two) input datasets in a single step, i.e., add multiple spatial datasets to the survey file at the same time.
UR31	Outputs detailed data linkage report (input data, output data, join ID, filters, (e.g., temporal, spatial, variables), linkage rate %, lists codes with no match).
Could have	
UR32	Check input data quality (e.g., Numeric/string, value range)
UR33	Vocabulary service integration
UR34	Ability to filter based on labels
UR35	Link to AURIN API (instead of pointing to spatial data in the local environment)
UR36	Create/publish R package and index in CRAN
UR37	Ability for a user to select/map between original vs modified input
Won't have	
UR38	Visualisation
UR39	Loc-I Demonstrator (concordances between areal identifiers)
UR40	Join with flexibility (e.g., right join)
UR41	Full search and discovery of data
UR42	Curated linked ready-to-use datasets

UR43	Service to aggregate individual data to areas
UR44	Maintainability (ongoing service agreement)
UR45	Easy calculation of derived variables
UR46	Data access service and authentication/authorisation (CADRE)
UR47	Secure environment to process sensitive data
UR48	Ability to automatically convert between different input data types/formats
UR49	Easy access to data documentation (vocabularies, classifications, data comparability, limitations etc) ranging from technical info for advanced users to broad info for less advanced users

Appendix E – Most downloaded ADA surveys

Table A3.1: The number of downloads 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

Appendix F – GeoSocial R Library Manual

Package ‘geosocial’

June 30, 2023

Title IRISS WP3 GeoSocial solution - Toolbox

Version 1.0

Author Australian Urban Research Infrastructure Network (AURIN)

Maintainer Pascal Perez <pascal.perez@unimelb.edu.au>,German Gonzalez <german.gonzalez@unimelb.edu.au>, Masoud Rahimi <masoud.rahimi@unimelb.edu.au>

Description License

GLP-3 **Encoding** UTF-8 **LazyData** true

Imports haven, dplyr, sjlabelled, openxlsx, rjson, dataverse, tidyverse

Depends haven, dplyr, sjlabelled, openxlsx,rjson,dataverse,tidyverse, R (>= 2.10)

RoxygenNote 7.2.3

R topics documented:

checkLSAY *checkLSAY*

Description

Check that the dataset is an LSAY dataset and the dataset does not have modifications affecting the linkage process.

Usage

```
checkLSAY(dataset, cohort)
```

Arguments

dataset The dataset that would be analysed.

cohort year of the cohort (For example, LSAY 2009, cohort = 2009).

Value

True if the dataset does not have modifications that affect the linkage process.

checkNamesDuplicates *checkNamesDuplicates*

Description

It’s important to check for duplicate names to avoid errors during data linkage and ensure accurate results. Stata does not allow duplicate variable names, so this process ensures the joined datasets don’t have any variables with the same name.

Usage

```
checkNamesDuplicates(dataset)
```

Arguments

dataset The dataset that would be analysed.

Value

True if none of the variable names are duplicates, otherwise false if overlap exists.

checkPostcodeStructure *checkPostcodeStructure*

Description

Checks that the postcodes are valid values.

Usage

```
checkPostcodeStructure(dataset)
```

Arguments

dataset	The dataset that would be analysed.

Value

True if all the postcodes are valid.

checkVariableNames *checkVariableNames*

Description

Check that the variable name is accepted by Stata: Stata variable names must adhere to the following rules: • Contain 1-32 characters. • Only contain the characters A-Z, 0-9, and underscore (_). • Begin with a letter or an underscore.

Usage

```
checkVariableNames(dataset)
```

Arguments

dataset	The dataset that would be analysed

Value

if the variable names are valid. Prints a message describing problem and specific variable that is the problem if invalid.

concordances *Concordances ABS data The ABS has developed a suite of geographical correspondences, primarily to assist users make comparisons and maintain time series between different editions of the Australian Statis- tical Geography Standard (ASGS). Correspondences are a mathemati- cal method of reassigning data from one geographic region to another geographic region.*

Description

This file combines the concordances

Usage

```
concordances
```

Format

‘concordances‘ A data frame with 8149 rows and 10 columns:

origin_unit Geographical unit - Origin **destination_unit** Geographical unit - Destination **year_in** Year of Geographical unit - Origin **year_out** Year of Geographical unit - Destination **origin** Geographical code - Origin

destination Geographical code - Destination

ratio Year

origin_areasqkm Area in square kilometres - Origin

destination_areasqkm Area in square kilometres - Destination

Source

<https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/correspondences>

CreateFolders *CreateFolders*

Description

This function generates the folders the script needs to run.

Usage

```
CreateFolders(path = getwd())
```

Arguments

name	the path

Value

True if the folders are created correctly.

downloadDataverseData *downloadDataverseData*

Description

Function takes inputs doi (get from dataverse listing of desired data) and the year of the data

Usage

```
downloadDataverseData(id, file)
```

Arguments

id	unique doi (get from dataverse listing of desired data)
file	where the file would be storage

Value

True if the download is correct

FilterConcordance *FilterConcordance*

Description

Narrow down the concordances and select the highest ratio that meets the threshold set by the user.

Usage

```
FilterConcordance(concordances, id, ratio_threshold = NULL)
```

Arguments

concordances	Concorcondaces file.
id	Geospatial identificator in que concordances table.
ratio_threshold	

ratio.

Value

True if the file is storage correctly.

GenerateLog *GenerateLog*

Description

This function generates the log file that would store all the events generated by the code.

Usage

GenerateLog(name)

Arguments

name the name of the log file

Value

True if the log file is generated

GetTerm *GetTerm*

Description

Based on a term, get the metadata associated to it. For example: GetTerm(keywords=’http://example.com/census/IFAGE

Usage

GetTerm(term)

Arguments

term A String with the direction of the term in the ARDC server.

Value

Return label, dcterms_created, dcterms_modified, creator, dc_publisher, dc_source, dcterms_title, has_top_concept, history_note, scope_note, type and identifier.

LoadLSAY *LoadLSAY*

Description

Load or download the LSAY data, given a cohort and wave. (The demonstrator only supports LSAY cohort 2009)

Usage

```
LoadLSAY(cohort, wave, LSAY_topics)
```

Arguments

cohort year of the cohort (For example, LSAY 2009, cohort = 2009)

wave vector of years. (For example, wave = [2012,2013,2014])

LSAY_topics vector with the sub-topics that would be included in the analysis, (For example["School transition" "Current",..]).

Value

a list which contains the survey data and the geospatial data.

LoadParameters *LoadParameters*

Description

Loads the parameters and sets the global environment.

Usage

```
LoadParameters(file)
```

Arguments

file is the path where the JSON file is located.

Value

True if the parameters are valid

LoadTSP2021 *LoadTSP2021*

Description

Load and filter the TSP 2021 by the demonstrator variables and the closest year published census. For example, if the year is [2017,2018,2019], the TSP2021 only will have the closet census to this years that is 2016.

Usage

```
LoadTSP2021(year = NULL, variables = NULL)
```

Arguments

year	Vector of years.

Value

a list which contains the data and metadata

LSAY_metadata *Metadata - Longitudinal Surveys of Australian Youth*

Description

A complete listing of the variables and their associated formats and value labels contained in the LSAY data files for all six LSAY cohorts.

Usage

```
LSAY_metadata
```

Format

‘LSAY_metadata‘ A data frame with 37,895 rows and 15 columns:

Cohort Y95,Y98,Y03,Y06,Y09,Y15

Wave Wave/year

Section

Data element ID Major topic

area Sub-major topic area

Minor topic area Data

element Variable

Type Label
Question Base

Source

<https://www.lsay.edu.au/publications/search-for-lsay-publications/2621>

PotentialCensus *PotentialCensus*

Description

Given a year, find the closest census year.

Usage

```
PotentialCensus(year)
```

Arguments

year year of analysis.

Value

String which contains the ABS_year and the original year.

QualityIndicator *QualityIndicator*

Description

Implementation of the ABS quality indicator which provides a considered measure of the quality of the correspondence. in relation to the weighting unit. (ABS, 2021).

Usage

```
QualityIndicator(ratio)
```

Arguments

ratio This field describes the Ratio of the FROM region that is being donated to the TO region. The Ratio is a figure between 0 and 1. (ABS, 2021).

Value

Returns the equivalent SA3 for each row.

Description

Based on a keyword, query the ARDC vocabs server and return related terms. For example: SearchConcept(keywords=’AGED’).

Usage

```
SearchConcept(keywords)
```

Arguments

keywords A vector of strings that contains keywords.

Value

Terms related to the keywords.

SummaryConcordances *SummaryConcordances*

Description

Create a summary of the concordances involved in the data linkage.

Usage

```
SummaryConcordances(concordances_POA,  concordances_SA3)
```

Arguments

concordances_POA
concordances using in the POA linkage

concordances_SA3
concordances using in the SA3 linkage

Value

a data frame with the concordances involved in the data linkage.

SummaryL *SummaryLog*

Description

Consolidate a log with all the information on the data linkage.

Usage

```
SummaryLog(summaryResults)
```

Arguments

summaryResults List which contains the output after the data linkage.

SummaryMetrics *SummaryMetrics*

Description

Create a summary with important information about the data linkage, such as year, cohort, linkage, missing values, not linked areas, not linked individuals and the quality of the data linkage.

Usage

```
SummaryMetrics(metrics_POA,  metrics_SA3)
```

Arguments

metrics_POA Metrics created in the POA linkage.

metrics_SA3 Metrics created in the SA3 linkage.

Value

a data frame with all the information mentioned before.

SummaryL *SummaryLog*

Description

Create the summary report

Usage

SummaryReport(concordances_POA, concordances_SA3, metrics_POA, metrics_SA3)

Arguments

concordances_POA
Concordances used during the first stage: POA to SA3.

concordances_SA3
Concordances used during the second stage: SA3 to SA3. metrics_POA Metrics used during the first stage: POA to SA3. metrics_SA3 Metrics used during the second stage: SA3 to SA3.

Value

Returns a list with all the information about the data linkage.

TestDataverseConnection
TestDataverseConnection

Description

Tests connection to ADA dataverse. Requires dataverse token to be loaded in the system environ- ment.

Usage

TestDataverseConnection()

Value

True if the connection is correct

Description

Recieve a vector of POAS, This function transformate POA to SA3.

Usage

```
TransformPOA(data, concordances, year, ratio_threshold = NULL)
```

Arguments

year	Year.

Value

Returns the equivalent SA3 for each row.
the metric of missing values, miss matching and POAS that doenst match.

TransformSA3 *TransformSA3*

Description

Recieve a vector of SA3 ABS year, This function transformate POA to SA3.

Usage

```
TransformSA3(data, concordances, year_in, year_out, ratio_threshold)
```

Arguments

data	Year.
concordances	Year.

Value

Returns the equivalent SA3 for each row.
the metric of missing values, miss matching and POAS that doenst match.

TSP2021 *Time series profile SA3*

Description

The Time series profile contains the Census characteristics of persons, families and dwellings over time. The data is based on place of usual residence.

Usage

```
TSP2021
```

Format

'TSP2021' A data frame with 358 Statistical Area Level 3 (SA3) and 35 Sociodemographic characteristics.

Details

The 2021 Time series profile contains data from the 2011, 2016, and 2021 Censuses. Where classi- fications have been revised, output are based on the classification used for the 2021 Census.

When interpreting the results from different time periods, take care as censuses are based on a point in time. Changes to the Census form design, collection procedures and processing may impact the comparability of data.

Source

<https://www.abs.gov.au/census/guide-census-data/about-census-tools/datapacks#:~:text=The

WriteStata *WriteStata*

Description

Write the outcome of the data linkage.

Usage

```
WriteStata(DataJoined, SurveyResponses, waves, path)
```

Arguments

DataJoined	List which contains the output after the data linkage.
waves	vector of years. (For example, wave = [2012,2013,2014])
path	Location where the files are going to be written.

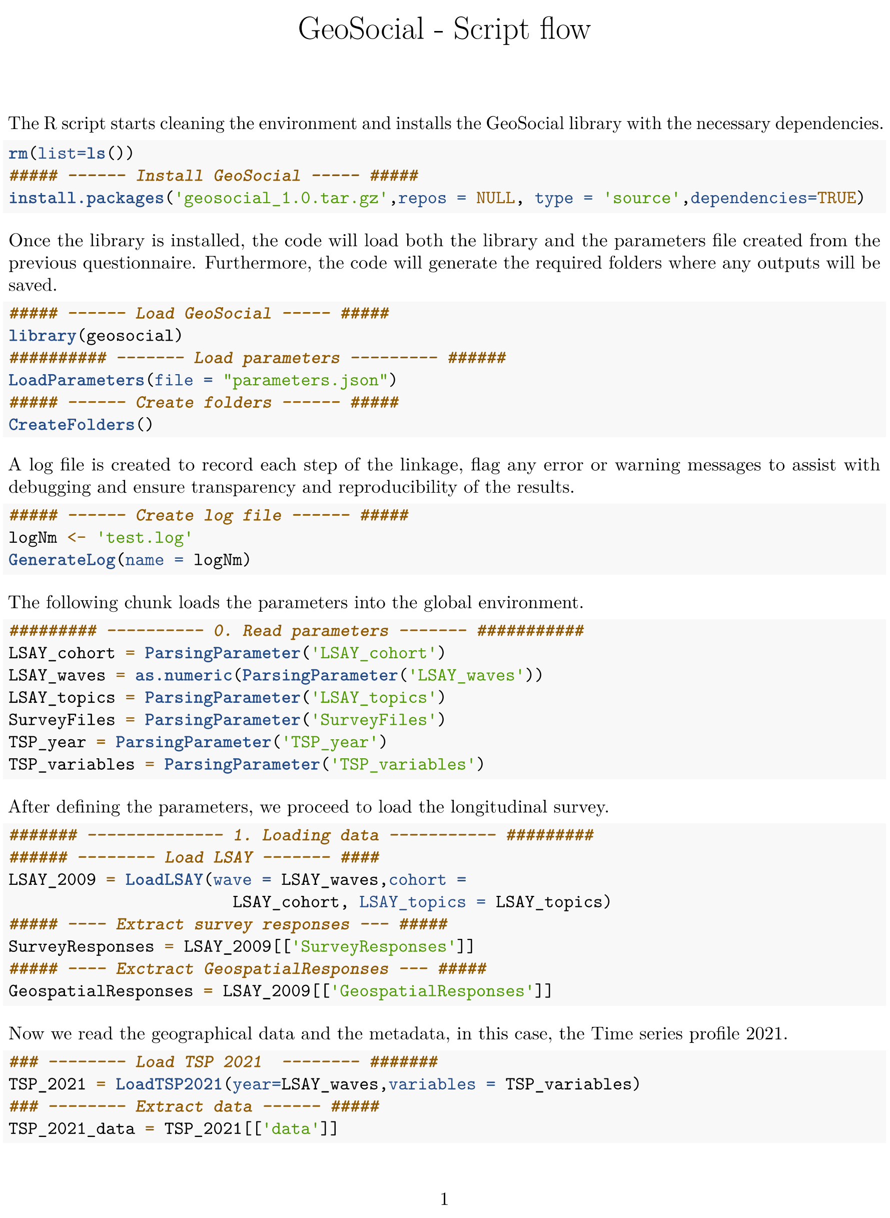
SurveyResponses:
Dataframe which contains the survey responses.

Value

True if the folders are created correctly.

Index

Appendix G – GeoSocial Main script



GeoSocial - Script flow

The R script starts cleaning the environment and installs the GeoSocial library with the necessary dependencies.

```
rm(list=ls())
##### ------ Install GeoSocial ----- #####
install.packages('geosocial_1.0.tar.gz',repos = NULL, type = 'source',dependencies=TRUE)
```

Once the library is installed, the code will load both the library and the parameters file created from the previous questionnaire. Furthermore, the code will generate the required folders where any outputs will be saved.

```
##### ------ Load GeoSocial ----- #####
library(geosocial)
########## ------- Load parameters --------- ######
LoadParameters(file = "parameters.json")
##### ------ Create folders ------ #####
CreateFolders()
```

A log file is created to record each step of the linkage, flag any error or warning messages to assist with debugging and ensure transparency and reproducibility of the results.

```
##### ------ Create log file ------ #####
logNm <- 'test.log'
GenerateLog(name = logNm)
```

The following chunk loads the parameters into the global environment.

```
######### ---------- 0. Read parameters ------- ###########
LSAY_cohort = ParsingParameter('LSAY_cohort')
LSAY_waves = as.numeric(ParsingParameter('LSAY_waves'))
LSAY_topics = ParsingParameter('LSAY_topics')
SurveyFiles = ParsingParameter('SurveyFiles')
TSP_year = ParsingParameter('TSP_year')
TSP_variables = ParsingParameter('TSP_variables')
```

After defining the parameters, we proceed to load the longitudinal survey.

```
####### -------------- 1. Loading data ----------- #########
###### -------- Load LSAY ------- ####
LSAY_2009 = LoadLSAY(wave = LSAY_waves,cohort =
                       LSAY_cohort, LSAY_topics = LSAY_topics)
##### ---- Extract survey responses --- #####
SurveyResponses = LSAY_2009[['SurveyResponses']]
##### ---- Exctract GeospatialResponses --- #####
GeospatialResponses = LSAY_2009[['GeospatialResponses']]
```

Now we read the geographical data and the metadata, in this case, the Time series profile 2021.

```
### -------- Load TSP 2021  -------- #######
TSP_2021 = LoadTSP2021(year=LSAY_waves,variables = TSP_variables)
### -------- Extract data ------ #####
TSP_2021_data = TSP_2021[['data']]
```

```
### -------- Extract metada ------ #####
TSP_2021_metadata = TSP_2021[['metadata']]
```

We start to prepare the longitudinal survey data for the data linkage. The first step is to transform the postcode to SA3. We store all the information about the linkage, to export at the end.

```
####### -------------- 2. Geospatial data linkage ----------- #########
####### ----------- 2.1 POAS TO SA3 ----------- ######
SurveyResponses_SA3 = LSAY_POA_SA3(data=GeospatialResponses,
                                   concordances=concordances)
###### --------- SA3 ------- ########
SurveyResponses_SA3_IN = SurveyResponses_SA3[['sa3']]
###### --------- ABS Metric ------- ########
SurveyResponses_SA3_metric = SurveyResponses_SA3[['metric']]
###### --------- ABS Concordances ------- ########
SurveyResponses_SA3_concordances = as.data.frame(SurveyResponses_SA3[['concordances']])
###### --------- ABS Ratio ------- ########
SurveyResponses_SA3_ratio = as.data.frame(SurveyResponses_SA3[['ratio']])
```

After having all in terms of SA3, we need to transform all to SA3 2021 to link with TSP 2021.

```
####### ----------- STAGE 2.2: SA3 TO SA3 year ----------- ######
SurveyResponses_SA3_TSP = LSAY_PSA3_SA3(data=SurveyResponses_SA3_IN,
                                        concordances=concordances,
                                        year_out = TSP_year)
###### --------- SA3 ------- ########
SurveyResponses_SA3_TSP_OUT =
        SurveyResponses_SA3_TSP[['sa3']]
###### --------- ABS Metric ------- ########
SurveyResponses_SA3_TSP_metric =
         SurveyResponses_SA3_TSP[['metric']]
###### --------- ABS Concordances ------- ########
SurveyResponses_SA3_TSP_concordances =
         as.data.frame(SurveyResponses_SA3_TSP[['concordances']])
###### --------- ABS Ratio ------- ########
SurveyResponses_SA3_TSP_ratio =
         as.data.frame(SurveyResponses_SA3_TSP['ratio'])
```

After transforming the longitudinal survey in the same terms as the Time series profile, we performance the data linkage between the two datasets.

```
####### ------- GeoSpatial join ------ #####
DataJoined = GeoSpatialJoin(year=LSAY_waves,
                            GeospatialResponses=SurveyResponses_SA3_TSP_OUT,
                            SurveyResponses=SurveyResponses,
                            TSP_data = TSP_2021_data,
                            TSP_metadata = TSP_2021_metadata)
```

We calculated consolidated the metrics, and concordances used in the process and storage, and stored it in an xlsx file.

```
############ --------- Geospatial report --------- ############
summaryResults = SummaryReport(concordances_POA= SurveyResponses_SA3_concordances,
                               concordances_SA3 = SurveyResponses_SA3_TSP_concordances,
                               metrics_POA = SurveyResponses_SA3_metric,
                               metrics_SA3 = SurveyResponses_SA3_TSP_metric)
```

```
######## -------- Write the summary report --------- #######
openxlsx::write.xlsx(x=summaryResults,
                     file='outputs/Summary.xlsx')
```

Subsequently, we export the data in Stata, preserving the structure and metadata.

```
####### ------- Write stata ------ #####
WriteStata(path = 'outputs/', waves =LSAY_waves,
           DataJoined = DataJoined,
           SurveyResponses = SurveyResponses)
```

Finally, we write a summary of the data linkage in the log, and storage the metrics.

```
####### -------- Write final log ------ #######
SummaryLog(summaryResults = summaryResults)
```