
# Introducing the IRISS Geosocial Data Integration Service: Integrating information on people, places, time, and space.

Australian National University

ADA AUSTRALIAN DATA ARCHIVE

ACSPRI Australian Consortium for Social & Political Research Inc.

AURIN aurin.org.au

THE UNIVERSITY OF MELBOURNE

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Australian Research Data Commons

Integrated Research Infrastructure for Social Science

# Acknowledgement of Country

We acknowledge the Traditional Owners of the land on which this event is taking place and pay respect to their Elders (past and present) and families.

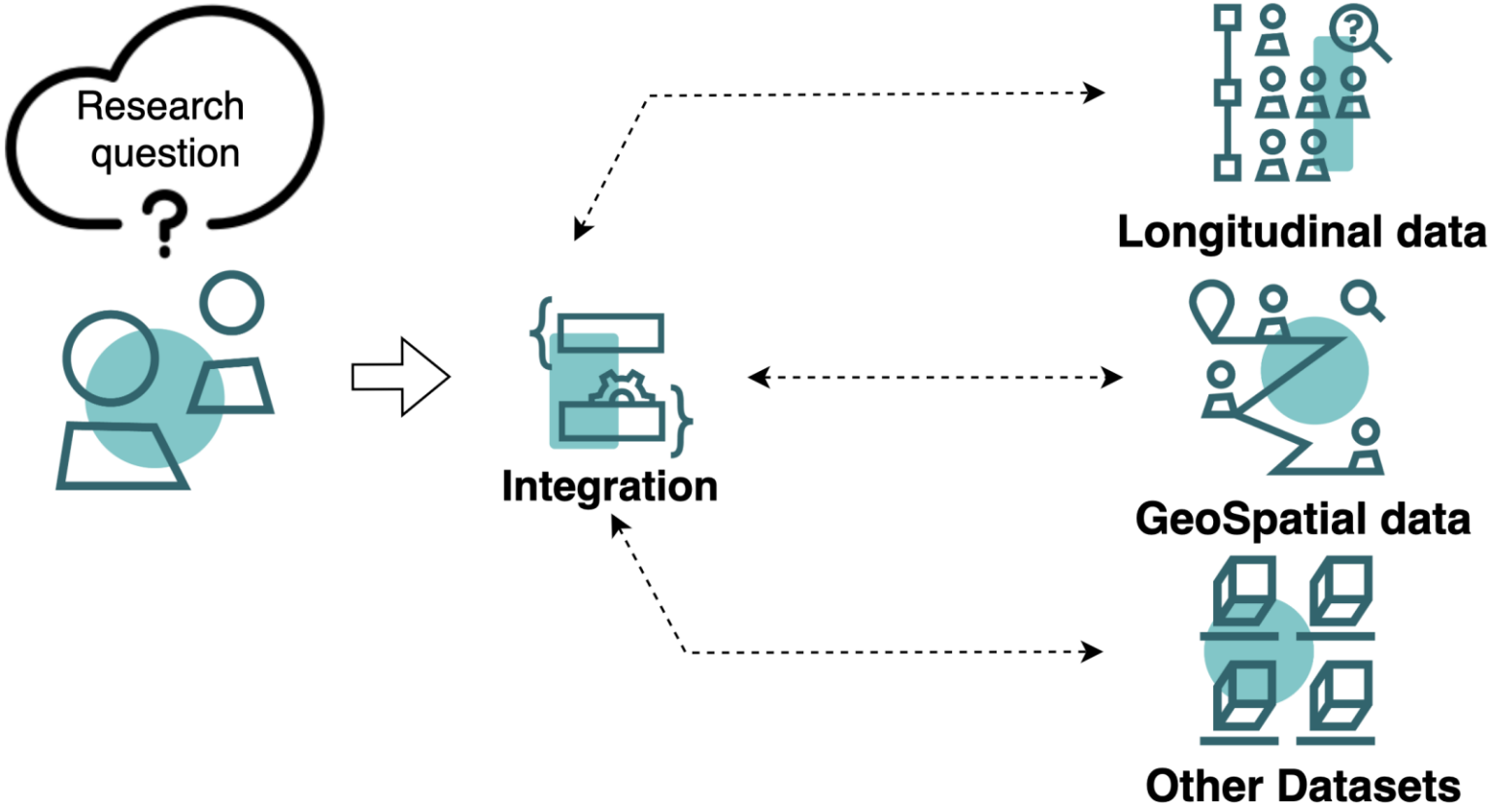
# The Integrated Research Infrastructure for the Social Sciences (IRISS)

**Objective**: Address the fragmentation of the Australian social science research infrastructure, establishing a new foundation for integrating data, analysis and platforms for social science research in Australia.

- **WP1:** Project Management
- **WP2:** VASSSAL (Vocabulary Access Service for Social Science in Australia)
- **WP3:** GeoSocial
- **WP4:** Demonstrator Projects
- **WP5:** SPIRE (Survey Project Integrated Research Environment)
- **WP6:** CARDSS (Curation of Australian Research Data in the Social Sciences)

# GeoSocial: Problem

The researchers want to bring data on people and places together, but don’t know how to do it and what the issues might be.



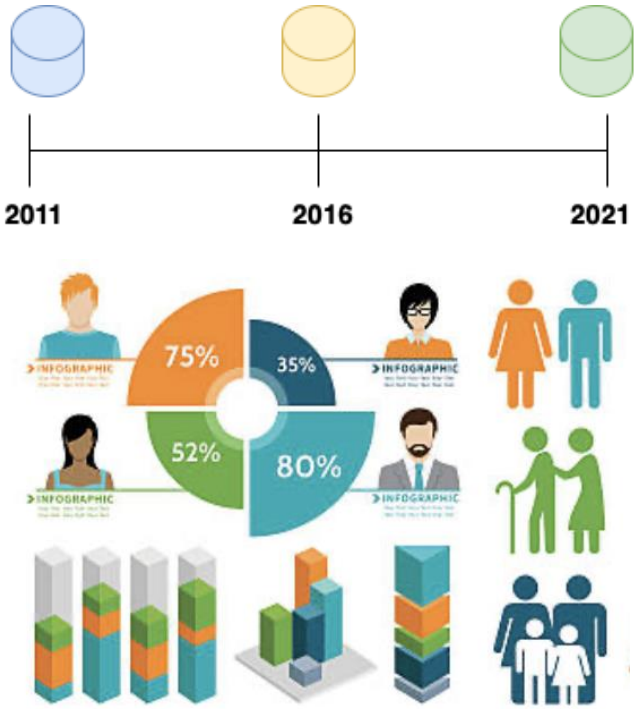
# Motivation

The GeoSocial solution allows researchers to link 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.

**Longitudinal data**

**Geographical data**

**Data custodians**




# Personas/users

## Low skills level

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

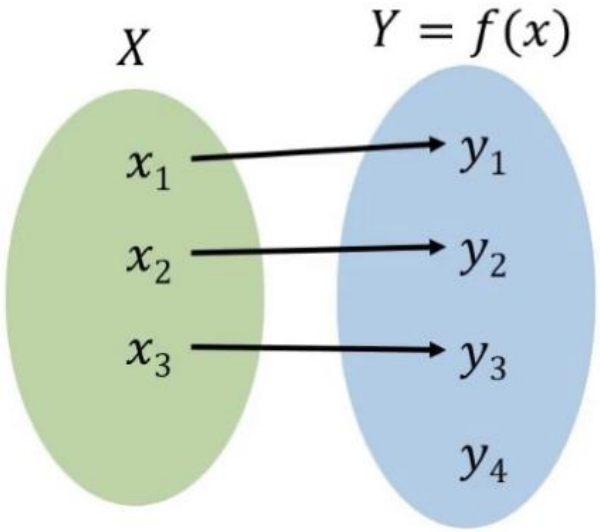
## Mid-level user

- 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 researchers to achieve goals

## 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

# Methodological Considerations



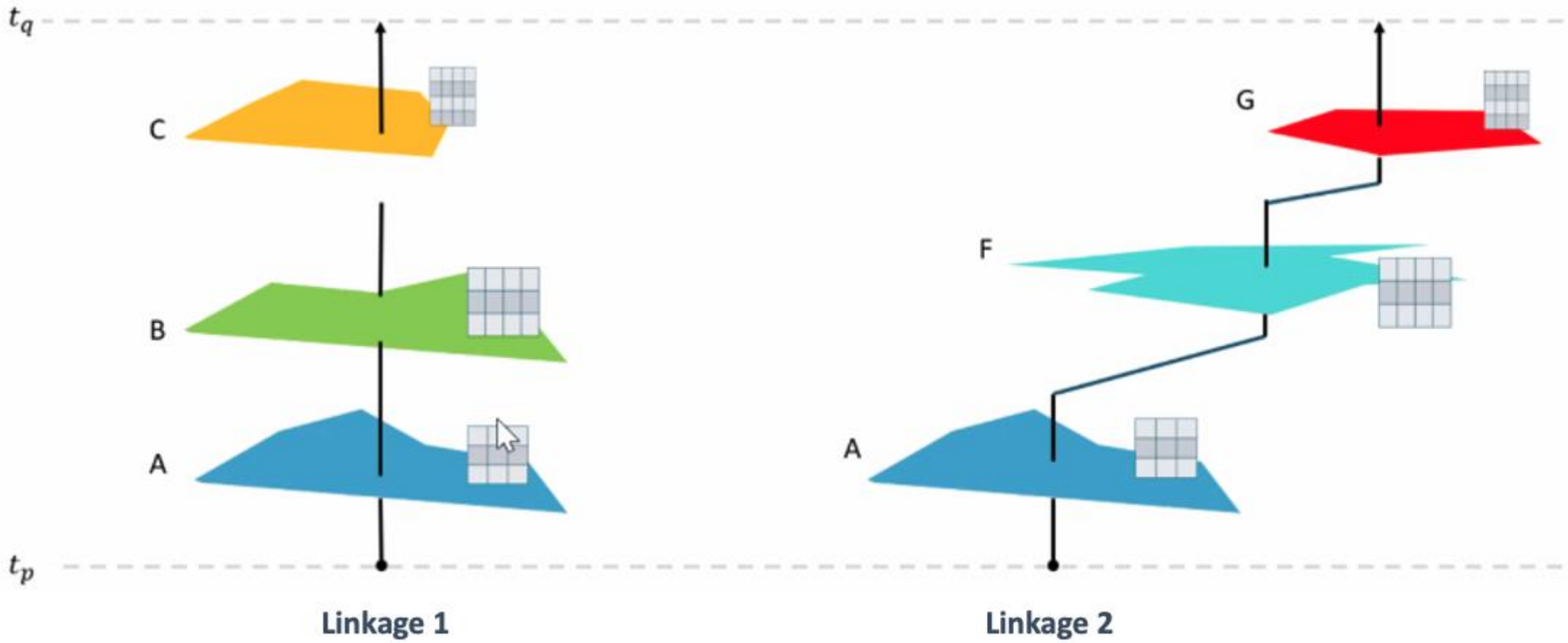

**Concordance/Correspondence**

Comparison between two spatial variables collected in two different moments of the time.

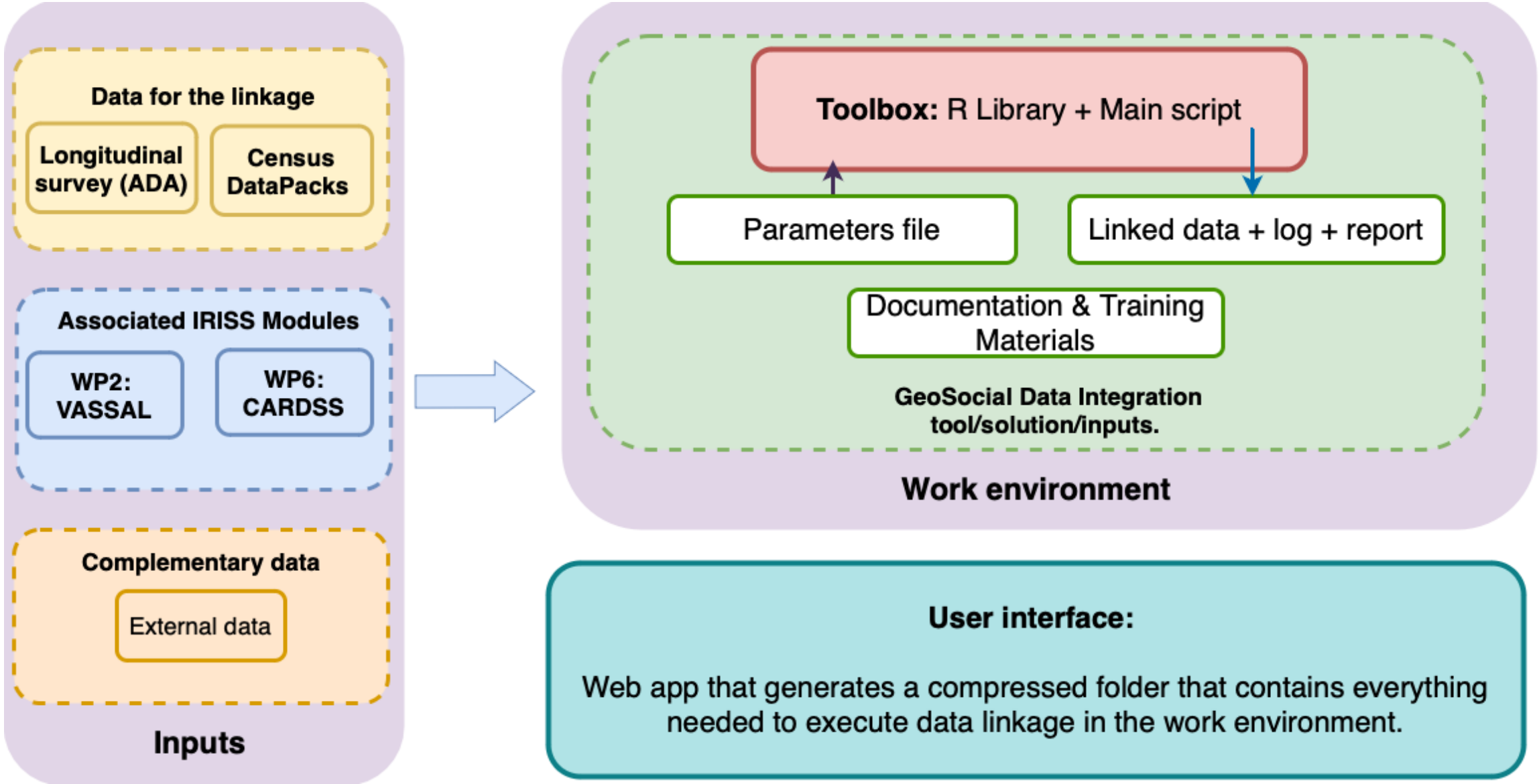
**Semantics**

Classifications and vocabularies change with time and are not in a machine-readable format.

# Methodological Considerations



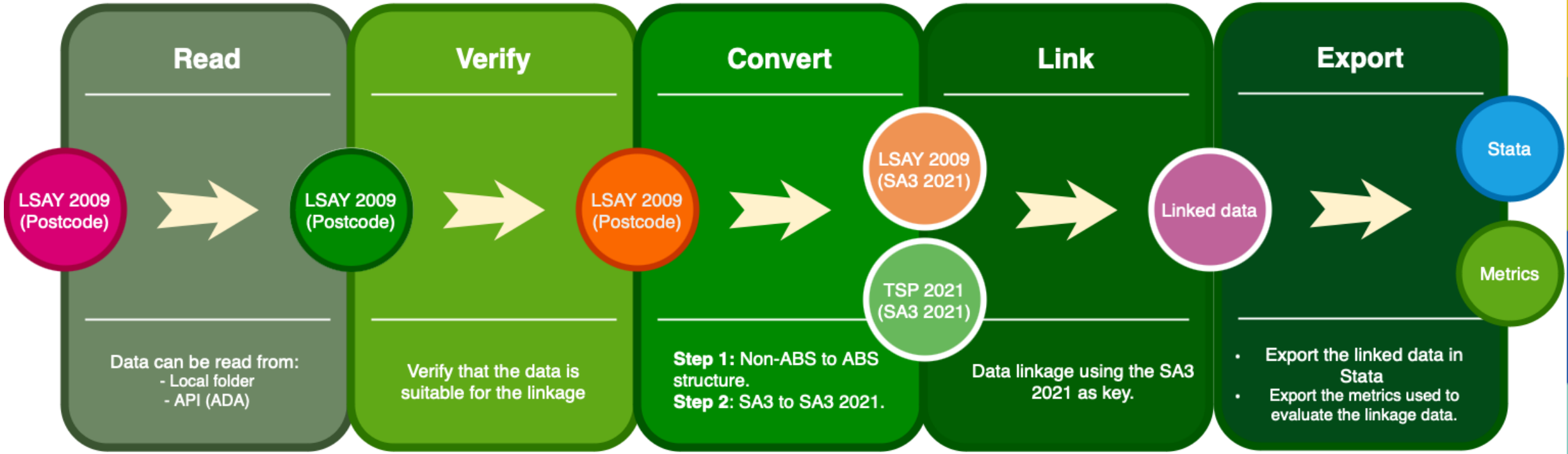
# GeoSocial service design



# Demonstrator:

**Longitudinal Survey:** LSAY 2009

**Geospatial dataset:** Time series profile 2021



## IRISS project

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

## GeoSocial

Longitudinal survey

Geospatial data

The GeoSocial solution allows researchers to link 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.

GeoSocial

START

# Package 'geosocial'

June 30, 2023

**Type** Package

**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

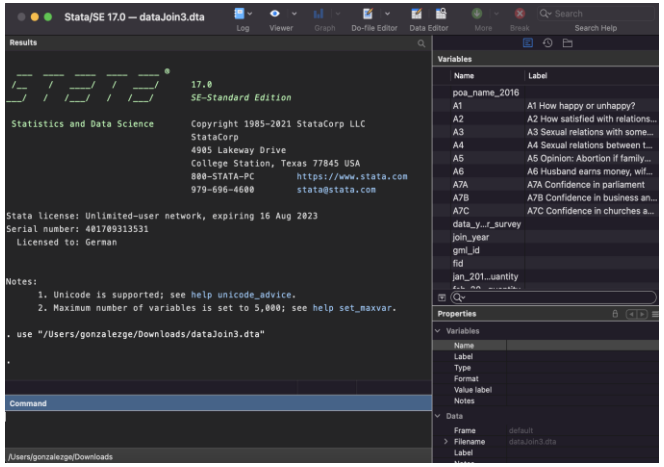
**Depends** haven,
dplyr,
sjlabelled,
openxlsx,rjson,dataverse,tidyverse,
R (>= 2.10)

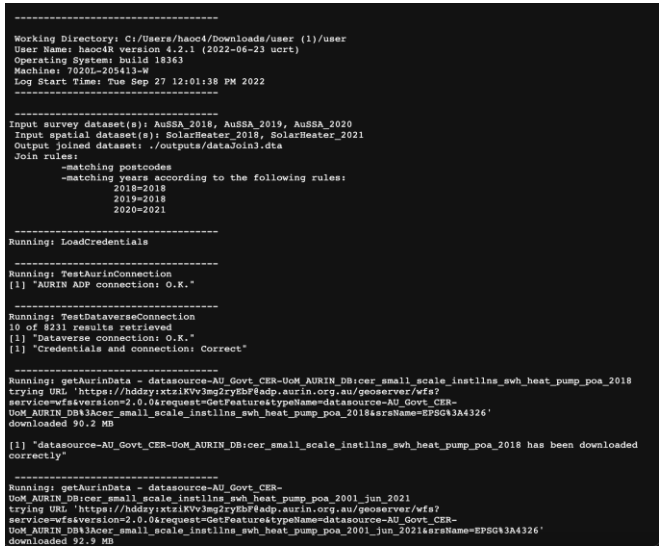
**RoxygenNote** 7.2.3

## R topics documented:

checkLSAY . . . . . 2
checkNamesDuplicates . . . . . 2
checkPostcodeStructure . . . . . 3
checkVariableNames . . . . . 3
concordances . . . . . 4
CreateFolders . . . . . 4
downloadDataverseData . . . . . 5
FilterConcordance . . . . . 5
GenerateLog . . . . . 6
GetTerm . . . . . 6
LoadLSAY . . . . . 7
LoadParameters . . . . . 7
LoadTSP2021 . . . . . 8
LSAY_metadata . . . . . 8
PotentialCensus . . . . . 9



# Next steps

- Consultations/road testing with a broad range of stakeholders
- Re-engaging with the policy and service agencies that provide and control access to various relevant datasets.
- Collect user training needs, and develop a forward plan for user training and community engagement
- Create a web app that executes the flow to low-skilled users


# Thank you
# Follow us:@ausiriss

## IRISS project

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

## GeoSocial

Longitudinal survey

Geospatial data

The GeoSocial solution allows researchers to link 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.

GeoSocial

START

## 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.

SELECT

## 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.

SELECT

BACK

# 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

BACK

CONTINUE

# Where would you like your integrated data stored?

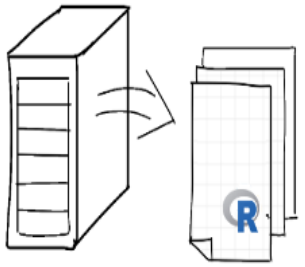
The toolbox allows the user to load the survey data from one of the following sources:

## Australian Data Archive


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

CLOUD

## Local environment




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

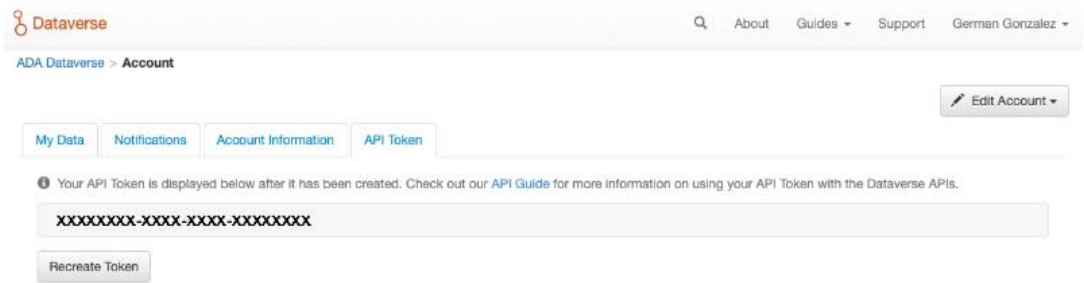
LOCAL

BACK

# 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.




**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.**

BACK

CONTINUE

# 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:**

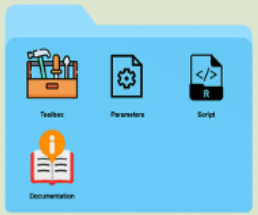
```
1  Please introduce your absolute path. For example: C:\Users\example\Documents\LSA09\
```

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

BACK

CONTINUE

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




**Step 1: Download GeoSocial**

DOWNLOAD

**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.

BACK