Integrated Research Infrastructure for Social Science Project Report

ARDC HASS RDC and Indigenous Research Capability Program

3.1 User requirements_ WP3 _IIRISS User requirements v3.

2022/03/30

Tomasz Zajac, Michael Rigby, Angela Ryan, Matthias Kubler, Denise Clague, Jonathan Corcoran, Wojtek Tomaszewski

LEAD ORGANISATION: Australian National University (ANU), Australian Data Archive (ADA), ANU Centre for Social Research and Methods (CSRM), Research School of Social Sciences (RSSS)

PARTNER ORGANISATIONS:

- University of Queensland, Institute for Social Science Research (ISSR)
- University of Melbourne, Melbourne Institute
- Australian Urban Research Infrastructure Network (AURIN)
- Australian Consortium for Social and Political Research Inc (ACSPRI)

PROJECT LEAD: Assoc. Prof. Steven McEachern (Project Lead)

PROJECT COORDENATOR: Keila Silva

ARDC IRISS User requirements note

Purpose

The purpose of this document is to discuss the audience profile relevant for 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 too, but would also 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.

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). • Confidentiality and privacy protection measures. • Secure environment to process data if data require protection. • Search & discovery – particularly important as the list of datasets grows.	• 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.
Medium skills level - A user advanced in applying statistical	In addition to the previous: • Linked ready-to-use datasets (e.g. individual-level survey data	• Wider access to the data.

sophisticated statistical methods but less comfortable with data manipulation. IRISS personas: Martin – researcher (data collector) Danielle – researcher (data analyst) Yosef – data analyst	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.	• 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: Serena – policymaker Yosef – data analyst	• 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). • 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.	• 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.

Thematic requirements

The work package aims to create a data product that integrates people, place, time, and space. There are numerous ways in which data on people can be linked with data on places. There are researchers working in a number of disciplines across the social sciences with interest in such data, including sociologists, education researchers, political scientists, economists, social planners, human geographers and others. Attempting to cater to too many types of users and to integrate too many data sources at once is flawed with risks related to feasibility and the robustness of the data infrastructure that is being developed. Instead, an initial stage of the project should focus on maximising the impact/ usefulness of the tool while keeping feasibility at the forefront. This can be achieved by focusing on a small number of datasets with relatively broad user base and topic coverage. Future expansions of the system should involve a bigger number of, and more diverse data sources.

Recommendations for WP3 and Demonstrator 1

The precise data selection criteria are still being developed (for more details see the 'Data requirements' document), however, the following core considerations taken from the project proposal will be applied to select the data[1]:

- The Service developed as part of WP3 and Demonstrator 1 will integrate data on people and spaces, while also capturing the longitudinal dimension.
- For the purpose of the Demonstrator, data on people will come from a 'high-profile' longitudinal survey available through ADA, with demonstrated wide user base (see 'Data requirements' document for more detail).
- Spatial data, which will be at least in part based on the Census and supplied by AURIN, will need to use geographies that are identical to those used in the selected survey(s), or that can be easily converted to those used in the survey(s).
- The datasets will be selected to produce useful and methodologically sound analysis.

It is recommended that the core pilot service developed under WP3, and illustrated with Demonstrator 1, is developed with a relatively high-skilled user in mind, to ensure the feasibility of delivering these within the project timeframe. To this end the following parameters are proposed for WP3 and Demonstrator 1:

- The Demonstrator will consist of:
    - an integrated dataset, combining a longitudinal survey from the ADA with spatially-structured data derived from the Census and other relevant sources;
    - a set of scripts performing data integration, employing relevant data vocabularies, including those developed as part of WP2;
    - a technical report documenting the steps that need to be undertaken to perform data integration and outlining issues that will need to be resolved as part of WP3 service (including legal and technical issues);
    - statistical analyses to demonstrate the utility of the integrated dataset – these can be shared with the research community in the form of an academic paper,

[1] Other future expansion could cover creating 'community profiles' out of other sources of administrative data, Land use data, Satellite data, Private sector data (e.g. CoreLogic, Uber) and more.

and shared with the relevant government stakeholders to demonstrate potential policy applications and impact.

Although the ultimate goal is to offer an online system that researchers can use to access enhanced data, it might not be feasible within the project timeframe. Establishing governance frameworks and addressing related legal issues is a time-consuming process. Therefore, for the purpose of the Demonstrator, the data integration process will be carried out locally at the researcher/user end (UQ) and the integrated dataset will also be deposited locally.

- The GeoSocial service prototype will consist of:
    - A set of scripts – such as Python scripts packaged as a Jupyter notebook – to carry out data integration, employing relevant data vocabularies, including those developed as part of WP2;
    - a technical report documenting the steps that need to be undertaken to perform data integration and outlining issues that will need to be resolved as part of further scaling up/industrialising the service (including legal and technical issues);

In WP3, consideration will be given first to the data integration process being carried out and the integrated dataset being deposited outside of the local user environment – for instance to make it available to the broader research community as part of the ADA Dataverse. However, each step of the development process will include exploring the options for transferring the system to an online environment and mapping out potential challenges.

While the core service and the Demonstrator will be initially designed with a relatively skilled academic/researcher user in mind, consideration will also be given to provide solutions for relatively lower-skilled policy user, which might include tools to facilitate spatial visualisations of data.