Integrated Research Infrastructure for Social Science Project Report

ARDC HASS RDC and Indigenous Research Capability Program

4.1 WP4 User requirements_IRISS_meeting_notes_workshops

2022-03-30

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)

[Work package 4: Demonstrators - brainstorm board]

Project Proposal

Overall: The Demonstrator work package involves three small scale demonstration projects to establish and illustrate the integrative capabilities of the combined services in IRISS.

D1 Spatial: This demonstrator will leverage the census data outputs of ACDC, and the integration services of the GeoSocial service, to establish a new data product, and apply person-based and place-based methods to the product.

D2 Sensitive: This demonstrator will assess the new outputs of IRISS services (such as new surveys generated through SPIRE, and curation tools from CARDSS) to assess their suitability for use in sensitive data environments, which are subject to higher levels of data and information security, ethics assessment and confidentiality risks. The project will leverage the Melbourne Institute Data Laboratory to conduct these analyses, providing a pilot study for future sensitive data support within IRISS.

D3 Census (ACDC): This demonstrator will include three pilot activities (a) collection migration and machine actionable formats (b) preservation program each year of census (c) metadata creation program (for data integration and harmonisation).

Success Factors

Research Requirement	*Project Activity*
Use geospatial information in data integration where there are gaps in datasets i.e., for inferencing	*Identify an example dataset (non-sensitive) with known gaps to demonstrate this data integration technique and that are not at risk of disclosure*
Develop data integration methodologies aligned to cybersecurity sensitivity controls and safe settings	*Create a method for integrating geospatial information that can be brought into the Melbourne Institute Data Lab and used with a sensitive dataset*
Socialise best practices on working with and sharing sensitive data	*Align with CADRE project (Five Safes training) and data access requirements*
Definition of sensitive data that covers proprietary and personal info in relation to disclosure risks	*Identify example datasets with known issues to reflect this definition*

Research Requirement	*Project Activity*
Alignment between data products coming out of IRISS and affordance of CADRE platform	*Use IRISS data products as examples for an extension model to the CADRE researcher training (access request – where that involves data integration)*
Develop new curatorial models (governance, technical and research process) for external validation	*Create a new model/template to capture curatorial techniques and ethical concerns to socialise (add into CADRE training)*
Develop a continuum for sensitive safe data and safe settings aligned with and/or complementary to suitable research environments	*Create a diagram to capture the intersection of safe data-safe settings (add into CADRE training)* *Use the CADRE platform technical architecture as an example (add into CADRE general communications)*
Develop a longer vision (10 years) for multiple research environments to be interoperable and complementary	*Propose a CADRE phase 2 that reflects a range of options for researchers using sensitive data in different environments*

Summary

[TBC]

Key Points

- Researchers analyse multiple datasets to drivers e.g., for poverty and interactions with other datasets. In field experiments the spatial info aids with structuring the underlying research design. Analysis of pattern changes over time where context stays the same e.g., family level. Units of analysis e.g., school catchments and aggregation levels. What are the common points where spatial info is being integrated? What are common data curation processes and tools used to integrate the data?
- Integration of survey data with spatial info – research interest lies at individual level (HILDA & LSAY) and the generation of derived variables based on spatial info. Reverse process can occur integration of spatial info with survey data. Where data is missing e.g., in HILDA adding spatial layers can add information value.
- Data scientist needs to understand the datasets before integration with spatial info – both attributes and parameter space above at individual and aggregate levels. Data aggregation changes in space and time – key question is what are the imputation measures e.g., averaging. Spatial structure dictates the analysis method used e.g., multi-level or data cube. HILDA is structured in multiple levels and has mesh blocks.
- Three different data users: (1) individual unit records (2) spatial info (3) intersection – who is the audience?

Follow Ups

D1. Spatial

- Showcase/socialise AURIN's data integration work with PHRN to integrate spatial info to project team.
- Identify the target audience/s for these data products with spatial info added.
- Identify the unit of analysis to use in the example dataset. Develop a framework can be developed to enable multi-level views.
- Set out the scenarios for different data integration techniques and the tools that fit with the different data types.

D2. Census

- Showcase/socialise AURIN's data integration work with ADA for the 1981 Census to integrate spatial info to project team.
- Identify 2-3 Census datasets to integrate spatial info with. Identify when Census data is at aggregate vs micro level and when a risk line is crossed. Capture method for dealing with boundary changes in Census data.
- Identify areas of IRISS project work to enter into dialogue with federal agencies e.g., ABS (use of Census data, dictionaries and classifications).

D3. Sensitive

- Describe examples where the addition of spatial info to individual unit records (SA2) increases the potential for identifiability. Identify the attributes of individual unit records where identifiability may be geographic or non-geographic.
- Identify key geographic info types e.g., households, post codes, school catchment areas, electoral areas to work with.
- Develop two use cases for data treatment: 1. Data integration of non-sensitive data with spatial info (products) 2. Capture of methodologies for integrating spatial info that demonstrate identification risks (process). Link into CADRE project work with sensitive data – safe data and safe settings.
- Explore the role of synthetic data as a means to define spatial info to introduce into a secure environment.
- Explore where code/pseudo code in R, Stata, SASS (and methodology) can be stored for transparency and reuse.

WP4 Demonstrators

Different research audiences (their interactions with different vocabularies and services*) and structure/terminology changes

*Vocabularies - AU

ABS, Census of population and housing dictionary 2016
ABS, Research codes (ANZSRC)
ABS, Geographical classification (ASGC)
ABS, Statistical geography (ASGS)
ABS, Industries (ANZSIC)
ABS, Occupations (ANZSCO)
ABS, Socio-Economic indicators (SEIFA)

*Vocabularies – INT

UNESCO ICD 10 & 11, International Classification of Diseases
Hasset (UKDA), Humanities and social sciences
ELSST (CESSDA), European languages for social science
Sage (UKSG), Social science
SNOMED, Clinical Terms

*Services

ABS Classifications
Research Vocabularies Australia
Loci Index
Education Data Portal
IPUMS