Loc-I - Location Integration Capability

User research report

Paul Box, Ashlin Lee CSIRO
14 May 2019

Land and Water
www.csiro.au

This work is licensed under a Creative Commons Attribution 4.0 International License
Introduction: Location Integration Capability – Loc-I

Everything happens somewhere

The Data Integration Partnership for Australia (DIPA) program aims to maximise the use of government data assets to improve policy advice through integrating data assets held by the Commonwealth.

As part of DIPA, the Loc-I project (Loc-I) was conceived to enable and improve location-based integration of data. Originally, Loc-I was designed as a ‘spine’, following from the development of a people spine (a mechanism to associate multiple identifiers for people across different government systems) and a business spine (which sought to associate multiple identifiers used for businesses). However, it was recognised at the outset of the project that Loc-I was significantly different from the people and business spines, as location identifiers: are highly variable, with many representations and identifiers used to represent the same place; and are governed at a data product level with a lack of any overarching identifier governance mechanisms.

Therefore, in order to succeed, Loc-I must:

- establish a governance framework and authorising environment for identifiers for spatial features
- operate across an open, federated and distributed web-based environment with a need to ensure that data managed at points of truth can be delivered using reliable and consistent identifiers; and
- (at present) primarily focused on creating links between different types of spatial features organised in separate reporting hierarchies and geographies, to assist in the data integration process.

Given the differing contexts and challenges of Loc-I compared to the other spines, Loc-I was conceived as a socio-technical infrastructure, comprising three inter-related architectures:

1. Information architecture - identifier information products and ontologies (or information models) that render them into Linked Data forms and formats;
2. Technical architecture – describing the technology to store, publish, access and use identifiers and the links between them;
3. Social architecture - a set of stakeholder decision-making bodies and processes, policies, standards and other agreements that enable the reliable delivery, sustainability and use of authoritative government identifiers for spatial features.

This report focuses on one component of Loc-I’s social architecture, describing the results of user interviews based on the current environment of end users (data analysts) working with location-based data. The following pages describe the current and future states of data analysis around geography, the user research approach and methodology, and the results of the user research. These results include a set of user archetypes (a typology of different users) and associated experience maps for each archetype.
Current state
Data analyst workflow

Geospatial data production and distribution is fragmented, creating a challenging environment for policy analysts. Individual geospatial data products are fit for purpose, but issues occur when attempting to use them to integrate statistical data.

Human and environmental data are typically recorded and aggregated using different administrative or reporting geographies (different ways of sub-dividing space) e.g. ASGS for population data and the physical hydrological catchments for environmental data. Each user needs to calculate the intersection between these different geographies to be able to re-apportion, integrate and use the two data sources in a calculation. This involves manual effort and cost, and may be done differently by different analysts with likelihood of human error, inconsistency between approaches and reduction in ability to replicate.

Consequently analysts, spend a large proportion of their time accessing, preparing and integrating data before they can start to use it. By working in parallel each analyst spends time solving data problems that other analysts are also solving, with solutions to common problems unshared, shared because of siloing. This represents significant duplication of work, and entrenches silos of data practice that prevents others from value adding to data.

WHY THIS PATHWAY IS DIFFICULT NOW

Silos approach limits the ability to find and use data outside specific domains of expertise.

Multiple ways of identifying the same place.

Re-projection of data between different reporting geographies is challenging.

Manual processes are prone to error and limit opportunities for repeating analysis methods or applying analysis to new contexts.

WHAT THE CURRENT PROCESS LOOKS LIKE

Data takes a long time to access, prepare and integrate.

Differing methods developed to integrate data.

Analysis has taken a long time and consumed considerable resources.

Trying to compare and analyse data at different scales.

• Each analyst solves same challenge independently
• Cleaned, linked, and value-added data is not shared
Loc-I will provide a Whole of Government approach to geospatial data, addressing many of the issues that policy analysts face. It will offer a means of identifying and connecting locations using unique, persistent identifiers, allowing data to be linked and made available across domains in a timely manner. Integrated data will be reliable and reputable through the use of authoritative identifiers and links between them, allowing analysts to answer policy questions consistently across time, with like-for-like comparisons possible across reporting geographies. Loc-I will also improve information governance by standardising policies and approaches to identifiers across Commonwealth agencies.
To understand the data ecosystem into which Loc-I will be incorporated, user research was conducted. This research focused on exploring end users (data analysts) and their contexts, practices and issues around the integration and analysis of data using location.

The approach adopted here uses the DTA’s Digital Service Standard (DTA 2019) to structure the research process. The steps of the Digital Service Standard are presented below, and the tasks undertaken as a part of the user research are described on the right. By adhering to the Digital Service Standard, Loc-I adopts a co-design approach to user research by designing the product with users.

**Discovery**
- Start mapping the broader service landscape, researching the real needs and problems faced by your users, and understanding the policy intent and technology constraints.
  - Learn about the Discovery stage
  - Read the Discovery guide

**Alpha**
- Test out your hypotheses by building prototypes in code to explore different ways you might be able to meet your users' needs. Explore multiple ideas. Do user research to learn which approach works best and iterate your solution as you learn more.
  - Read the Alpha guide

**Beta**
- Define a minimum viable product from the successful prototype in Alpha. Build this as an accessible and secure service. Allow the public to trial the beta alongside the existing service. Use their feedback to improve the service.

**Live**
- Put the team and processes in place to continue operating and improving the service. Phasing out the old services, and consolidating existing non-digital channels.

- Define stakeholder cohorts
- Define assumptions
- Develop research protocols
- Identify users
- Conduct interviews
- Synthesise
- Review insights around user needs
- Develop minimum viable product (MVP) (June ’19)
- Iterate with users and selected use cases
In accordance with the Discovery phase of the Digital Service Standard, a map of the service landscape was developed. This map identifies the key stakeholder groups that Loc-I must engage with in order to succeed, and highlights the pertinent research questions for each group. Loc-I project stakeholders were partitioned into four main cohorts for the purposes of research and engagement.

The remainder of the report describes the results of user research with data analysts and technical stakeholder cohort, stakeholders identified as crucial to improving outcomes for end user data analysts.

### Institutional stakeholders
- Data policy actors – Finance, AGLDWG, DTA
- Spatial data governance actors - ANZLIC, ICSM, FSDF

#### Research questions
- How can we establish an authorising environment for identifiers?
- Who decides on ‘minting’ of identifiers?
- How do we delegate to existing authority structures with mandates?

### Spatial data providers
- ABS
- GA
- BoM
- PSMA
- States and Territories

#### Research questions
- How do we manage change in spatial data?
- How is this data provided to users?
- Can we stabilise and improve efficiency of spatial data supply chains to deliver stable spatial features for Loc-I?

### Loc-I users (data analysts)

#### Data practice research questions
- What tools do you use?
- What data do you use?
- How do you use spatial data?
- What are your pain points?
- How might Loc-I address these pain points?

### Technical stakeholders

#### Research questions
- How might we integrate / interact with the Loc-I capability in our organisation?
User Research Methods

Key assumptions to be tested

Key assumptions underpinning Loc-I were identified and tested through user research with end user and technical stakeholders.

1. The current ways in which we deliver and use spatial data as the basis for integrating data is inefficient/sub-optimal.
   a) Delivery of spatial data: (data provider and end user cohort assumptions)
      i. A lack of stable, authoritative, unambiguous spatial identifiers creates data integration problems for users.
      ii. Relying on geometry to explore topological relationships is inefficient.
      iii. The way we manage changes in spatial data and identifiers for spatial objects creates challenges for users.
      iv. Non-GIS users are not well served due to the way we currently deliver geospatial data (for use in GIS).

2) Currently spatial data governance arrangements are fragmented and need to be improved (institutional cohort assumption).

3) Loc-I could supplement/support/fit into the existing ecosystem to improve the situation (institutional/technical cohort assumption).

4) This capability will be sustained institutionally by GA/Spatial authority (ANZLIC and ICSM)? (institutional/technical cohort assumption).

5) Linked Data (HTTP URIs web resource ID) is the solution to stabilising identity and linking data both on and off the web (technical cohort assumption).

Research methodology

Recruitment: Participants were recruited via purposeful, snowball sampling. Participants who worked as data analysts on DIPA projects were sought through a call for participants made by members of DIPA to their respective agencies, and through recommendations by colleagues who operate in the space.

Sample: Commonwealth Government data analysts working both in their agency environments and in the ABS DataLab on Data Integration Partnership in Australia (DIPA) Policy Delivery Plans and other analytic tasks. Twenty-three data analysts were volunteered for interview, and a further seven technical stakeholders were also interviewed due to their role and proximity to data analysis tasks.

Method: Semi-structured interviews were used to collect data, and to test project assumptions and explore the realities of data practice. Interviews lasted between 60 and 90 minutes, and were audio recorded and transcribed. Demographic survey information was also collected and included in the analysis. Interviews were undertaken in teams, with at least one member of both the social architecture and the technical teams being present to capture both social and technical insights. An interview guide was developed and used by interviewers to ensure a general consistency to interviews, while also allowing opportunities to probe and explore issues as they develop. Full ethics approval was sought and granted by the CSIRO’s ethics review board.

Using a User Centred Design approach, the results of the interviews were synthesised to produce:

- five ‘user archetypes’ – categories of users in the data ecosystem - each has their own practices, contexts, and environments to negotiate
- user experience maps for each archetype.

These are presented in the next section of this report.

The research methodology was designed and supported by consultants at ThinkPlace in Canberra and Melbourne.
Archetypes and experience maps
User Archetypes, Interactions and Environments for Loc-I Users and Technical Stakeholders

The DataLab is the ABS's Big Data Analytics capability. It includes the MADIP and BLADE data products. Access is available on request to the ABS and is subject to extensive security clearance. It exists in parallel to the operations of departments.

The analytic environments for Government Departments reflect a general move towards a centralised Enterprise Data Warehouse, with analysis access this environment. Although individual departments operations vary, the generally reflect the configuration described below:

Government Department

DataLab

Safe

People

Environment

Data

Outputs

Sam Sterling
DataLab analyst

Gordan Abara
GIS analyst

Alex Ward
Data analyst

Joan Kosta
Analyst manager/broker

Helga Santoso
Enterprise data warehouse administrator

Commission enquiry
Develop enquiry
Sourcing data
Prepare data
Perform the analysis
Develop output
Share knowledge

ANALYTIC WORKFLOWS

POLICY

DATA PREP

ANALYSIS

SHARE

INFORMATION PRODUCTS

OTHER DATA SOURCES

OTHER INTERNAL DATA SOURCES

DATA ACCESS

DATA ACCESS

DATA SHARING

Enterprise Data Warehouse (EDW)
“When people want something cool done with their data, they come to me.”

My role is to translate policy questions into tractable analytical tasks, and to manage the process of delivering the analytical product to clients.

Data is a lot messier than my clients think; integrating data can be a challenge, and determining what might be done with data (if data is available and fit for purpose) is challenging. Communicating data realities to clients is difficult, particularly as it isn’t always clear what they are asking for. I have to manage the expectations of my client against issues with the data. This includes: Do we have the right data? What can this data and analysis really tell us? How can we best present this data for the client?

While I have analytical abilities and experience, and do some analysis work, I mainly oversee my colleagues as they work with the data, and the associated process of discovering, integrating and analysing data.

My work involves a lot of tricky negotiation between both internal and external stakeholders, and sometime with a team of analysts. I am therefore more a manager with analytical skills, than an analyst. With many competing needs at play my role is challenging and fast paced!

I sometime work in the ABS DataLab environment, so I can use MADIP or BLADE, but that’s another story (see Analyst 3).

**USER TOOLS**

- Python
- SAS
- Qlik
- Excel

**DATA**

- MADIP
- BLADE

“Different states and territories have different permissions that kind of allow us to access the data or not. It requires just a lot of leg work to negotiate in a kind of bilateral way with both, each of the states and territories, and then within that, each of the kind of departments that are the custodians, that takes up a lot of time. That’s the major barrier.”

“So what we want are farm boundaries I think, which would allow us to match more spatial data sets and things like satellites.”

“One of the things I think that is missing with a lot of geospatial data is often having some sort of metadata associated word. I mean I could say good meta, you know just having great metadata but in reality sometimes it's just having some metadata.”

**USER INSIGHT**

- Collaboration is essential; I need a variety of different data sources and analyses performed, and I don’t always have the ability to do this. This includes GIS! Must rely on my team.
- Managing client needs is crucial; clients often want something ‘cool’ with data that aligns with their world view. What we can give them may not match their needs.
Joan Kosta – Analyst Manager/Broker

**STAGES**

<table>
<thead>
<tr>
<th>MANAGEMENT TASKS</th>
<th>ACTIVITIES AND INTERACTIONS</th>
<th>ANALYSIS TASKS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STEPS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commission enquiry</td>
<td>Develop enquiry</td>
<td></td>
</tr>
<tr>
<td>Source data</td>
<td>Prepare data</td>
<td></td>
</tr>
<tr>
<td>Perform the analysis</td>
<td>Develop output</td>
<td></td>
</tr>
<tr>
<td>Share knowledge</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**OPPORTUNITIES FOR LOC-I**

- Analytic task brief
  - Currently limited, but in future may contribute to improved understanding of data availability and potential use

**SATISFACTION POINTS**

- Acting as a broker for analytic functions and task brief shaping the questions
- Working backwards from the agreed brief to find the data needed for analysis

**PAIN POINTS**

- Lack of client understanding of what could be done with data
- Negotiating policy questions can be difficult to manage
- Geography inherent in many questions, but geospatial analytical resources not necessary factored into analytic task resourcing

**BLOCKERS & SUPPORTS**

- I assemble and present the product to the stakeholder, including context and narrative for the data analysis
- I create information products to present analytic outputs and insights

**MANAGEMENT TASKS**

- Assign scoped task to team
- Support and monitor team’s progress on acquiring data, negotiating with stakeholders for appropriate external data, and with EDW for internal data for team
- Coordinate and collaborate with my team, other analysts and data scientists to prepare analytical products (tables, reports, graphics etc)
- Conduct analysis

**ANALYSIS TASKS**

- I create information products to present analytic outputs and insights
- I assemble and present the product to the stakeholder, including context and narrative for the data analysis
- I create information products to present analytic outputs and insights

**DATA ACQUISITION & PREPARATION**

- Request datasets needed to contextualise the scope of the task and the narrative the team will present

**ANALYSIS**

- Coordinate and collaborate with my team, other analysts and data scientists to prepare analytical products (tables, reports, graphics etc)
- Data acquisition & preparation
  - Support and monitor team’s progress on acquiring data, negotiating with stakeholders for appropriate external data, and with EDW for internal data for team
  - Conduct analysis

**INFORMATION PRODUCTION AND EXCHANGE**

- Request from stakeholder (internal or external client), co-development of analytic task brief defining needs, scope and potential output
- Assign scoped task to team
- Request datasets needed to contextualise the scope of the task and the narrative the team will present

**SUPPORT & RESOURCES**

- Supports and monitors team’s progress on acquiring data, negotiating with stakeholders for appropriate external data, and with EDW for internal data for team
- Supports and monitors team’s progress on acquiring data, negotiating with stakeholders for appropriate external data, and with EDW for internal data for team
- Supports and monitors team’s progress on acquiring data, negotiating with stakeholders for appropriate external data, and with EDW for internal data for team

**SATISFACTION POINTS**

- Lacking trust in reliability of data and creates efficiencies in data use by providing access to documentation, methods, metadata in human and machine readable ways
- Longer term
  - Find observational data that is coded using Loc-I spatial data products
  - Address data fragmentation through federated delivery

**PAIN POINTS**

- Being told to ‘do something cool with my data’ sometimes signifies lack of understanding/desire for engangement in/respect for the analytic process
- Sometimes analysis is narrowly scoped to support a client’s agenda/narrative and this constrains ability to generate real insights

**KEY**

- Connection or handover to another person
- System dependency
- Data product interaction

**Loc-I Experience Maps | ANALYST**
“Sometimes I need location data to answer complex regional questions, and improve the quality of my analysis”

I am an economist at an undisclosed government agency. I conduct econometric and economic analysis in response to department/project/team policy questions. Depending on the task, I work alone or with a team, collaborating with internal and external stakeholders as needed.

My analysis involves sourcing the right kinds of data, checking/cleaning the data, then analysing the data in the environments provided, before creating an output from this. A fair amount of my work involves national level statistics so location is not an issue (I get a table of nationally aggregated statistics). In some cases (e.g. regional economic or innovation systems analysis) I am looking for spatially enabled data (firm level data with addresses). However, there are problems with getting appropriate location data for regional analysis, including multi location enterprises, and the geocoding of business data using postcodes rather than addresses.

I want to create outputs (products, insights, reports etc) from my analysis that satisfy the needs of the stakeholder(s), and ensure that my policy question is answered in the most efficient/effective way possible. The quality and accessibility of data is therefore critical to my work.

I sometime work in the ABS DataLab environment, so I can use MADIP or BLADE, but that’s another story (see Analyst 3).

**USER TOOLS**
- ArcGIS
- Python
- SAS
- Qlik
- Artificial intelligence/Machine learning
- Excel
- R
- STATA

**DATA**
- ABS economic statistics
- ABS Socio-Economic Indexes for Areas (SEIFA)

**USER INSIGHT**
- Analysts see cleansing data as part of their jobs and not necessarily a frustration as they want to build trust that the data is fit for purpose.
- They need address and finer grained location data to support regional analysis, and postcodes don’t provide high enough spatial resolution.

“Good quality location data will also help us capture any spill over effects that are happening within a region.”

“…unfortunately, because our data set’s [are] not temporal, we can’t see how things have changed in specific locations over time.”

“We’d like the data to be cleaner than it usually is. We’d like for it to be in a form that we can put in to the model directly.”
Alex Ward – The Economist (Data Analyst 1)

**Analytic task brief**
- I receive an analysis brief from a policy officer or manager (See Data Broker Archetype).
- I research the scope of enquiry to identify what data is needed to explore the policy questions/issues.

**Data acquisition & preparation**
- I search for and access data from my organisation’s EDW and/or from external sources (e.g. data.gov.au).
- I clean and filter data so it is fit for purpose.
- I realise I need additional data as there are gaps or the data I have is deficient. I go back to our data services teams and ask for help.

**Analysis**
- I explore potential approaches to performing the analysis based on the data I have.
- I perform the analysis and go back to the data iteratively.

**Information production and exchange**
- I sense-make, interpret and visualise results so that multiple audiences can engage with the findings.
- I validate the findings with project manager/client.
- I hand over the final data product to project manager/client to add context and present to end user.

**SATISFACTION POINTS**
- The EDW is the light at the end of the tunnel because they direct, cleanse and enable analyst’s work
- Able to access data via EDW or via open data repositories
- Being witness to a change in culture and attitudes to sharing data and data products
- Available datasets are increasingly spatially enabled
- I believe my analytic products inform policy process
- Producing information products that effectively visualise data for policy and publication is valuable

**PAIN POINTS**
- May not know the context and scope of enquiry developed between the project manager/policy officer and client – which limits comprehension of the task at hand
- Managing client expectations around what can be done with available data and tools
- Barriers to collaboration and innovation impact on work culture and new products
- Inefficiency in data supply chains (enterprise) causes gaps and fragmentation in datasets making it difficult to find relevant data for policy questions
- Corporate environments are not set up for more advanced analytic techniques (e.g. AI/ML)

**Blockers & Supports**
- Data stewardship is unclear so I can’t provide feedback on issues like data quality
- Integrating and using curated ‘tame’ corporate data with ‘wild caught’ data can be difficult
- Internal IT bureaucracy limits access to appropriate data sets and tools that impact analysts’ workflows and products
- When hardware fails, data is lost
- It is more difficult to find relevant data for policy questions because of data fragmentation
- Problems in data identified by analysts are not addressed upstream by EDW/stewards

**OPPORTUNITIES FOR LOC-i**
- Analytic task brief
  - Currently limited, but in future may assist in discovery and assessment of available data to inform analysis brief
- Data acquisition & preparation
  - Availability and access to data linked to locations using Loc-I identifiers
  - Access to Loc-I identifiers for self-serve geocoding
  - Transformation/reprojection of data between geographies using pre-defined spatial relationships, without using a GIS
  - Providing machine accessible data products to improve efficiencies
  - Access to documentation, methods, metadata in human and machine readable ways builds trust in reliability of data and creates efficiencies in data use
- Longer term
  - Find observational data that is coded using Loc-I spatial data products
  - Addressing data fragmentation through federated delivery
GIS analyst (Data Analyst 2)

“I want to create consistent products that are meaningful to users.”

I’m a GIS analyst and I take (statistical) data and geocode it by linking it to spatial locations to create spatial data products like colour coded maps. I also conduct (geo)statistical analyses on data and then present the results in graphical form i.e. maps and graphs.

Map-based visualisations are a powerful communication tool that appeals to decision makers, and I get a lot of satisfaction from producing these products that communicate complex issues in a compelling way.

To deliver these outputs, I want to have consistent and reliable data that builds confidence in the products I create. I also need access to the appropriate data sources, and have the right tools to work with this data – including open source and cloud based products.

Gordan Abara

USER TOOLS

- ArcGIS (or QGIS)
- R
- Python
- Maplato
- FME
- PowerBI
- Tableau
- Excel

DATA

- ABS Data
- Open data sources
- Satellite imagery data

“…there’s no GIS. software. So I have had to learn R. which is fine… [It’s] free and open source. I downloaded it but couldn’t install it as I don’t have admin rights. To get permission from IT was just a nightmare…”

“We are really, really struggling actually with [limited processing power].”

“…all the data that we receive is in a tangle and needs to be spatialised - given a geocode of some sort like an address, so that it can be mapped.”

USER INSIGHT

- Coordination is required to address lack of standardisation in data needed for analysis.
- GIS remains a specialist activity and a lack of geospatial skills constrains analytic capacity.
- Working with data geocoded using different administrative geographies presents so many challenges.
Gordan Abara – GIS Analyst (Data Analyst 2)

**STAGES**

**ACTIVITIES AND INTERACTIONS**

**STEPS**

- **Commission enquiry**
- **Develop enquiry**
- **Sourcing data**
- **Prepare data**
- **Perform the analysis**
- **Develop output**
- **Share knowledge**

**SATISFACTION POINTS**

- Collaborative knowledge exchange internally within organization is good

**PAIN POINTS**

- There is a lack of geospatial capacity within the broader analyst community and consequently GIS analysts are asked to do a lot of GIS grunt work – producing basic map visualisations

**BLOCKERS & SUPPORTS**

**OPPORTUNITIES FOR LOC-I**

**OPPORTUNITIES**

- Provide efficient access to ‘the right’ datasets

**SATISFACTION POINTS**

- The practice of exploring and understanding data are useful steps in enabling better analyses

**PAIN POINTS**

- Lack of consistency across data sources requires significant cleansing and ‘munging’ efforts are required
- Data extract, transform and load is a painful process
- Collaboration to obtain data is problematic
- Data quality is often uncertain and can limit analytical capability and quality of outputs
- Inconsistent use of location identifiers

**KEY**

- Connection or handover
- System dependency
- Data product interaction

**Information production and exchange**

**Data acquisition & preparation**

- I access the data from my organisation’s EDW. Occasionally I obtain data from external sources, and data I have saved internally.

**Analysis**

- I link demographic and other data to spatial data to create a spatially enabled products. This is time consuming.

**Information production and exchange**

- I hand over the results of my analysis (often a map) to the end user and sometimes provide a narrative interpretation for the product.

**Data preparation**

- Enabling faster more reliable spatial data integration (no need to perform time consuming overlay analysis)
- Reliable identification and reuse of spatial features
- A means to find and connect to the observation data related to specific geographies

**SATISFACTION POINTS**

- Use of satellite imagery in analysis
- Geography is a key structuring concept for our data so the geographic and other data fit together well

**PAIN POINTS**

- Desktop workstations are not powerful enough to perform some geo-analysis
- Sharing of data does not always happen
- Metadata to make sense of data provides insufficient detail
- Timely analysis hampered by lack of spatially enabled data
- Lack of coherent standardised data
- Difficult to connect spatial data (i.e. geometry) and observation data to enable analysis due to inconsistent IDs

- Map based visualisations are recognised as being valuable to support policy processes

- The final information product (produced by the end users) is sometimes not shared with the GIS analysts so we don’t see how our work is used
- Many people don’t understand geography and want data represented in both map and table formats
"BLADE is a big revolution, we can do a lot more things now than we could five years ago"

I work in an undisclosed Federal Government agency as a policy analyst.

For much of my work I utilise the Australian Bureau of Statistics DataLab environment to access MADIP and BLADE data.

These datasets are incredibly detailed, and enable me to access invaluable data to explore social and economic issues. I primarily use statistical and econometric techniques to explore policy issues using the data. I take great satisfaction in leveraging the power of these data assets to provide new insights to inform policy.

However, working in the DataLab presents some unique challenges. Because MADIP and BLADE combine many different types of data, there are some inconsistencies that take a lot of time to fix. Geographical information is not provided or inconsistently applied, placing a vital analytical tool just out of reach. The Virtual Machine access has limited software tools, and runs slowly during business hours.

I like my work, and the DataLab is a great resource, but would love to see it streamlined. It is a highly secure environment, and while this is very appropriate, there are challenges to this environment. This includes long administrative times, a lack of appropriate software in the DataLab environment, and limitations in what can be input and output. It’s a revolution, but not without its complications!

**USER TOOLS**
- R
- Stata
- SAS
- Python
- GIS

**DATA**
- MADIP
- BLADE
- Excel

**USER INSIGHT**
- Users need more streamlined and immediate support in the DataLab Environment
- GIS analysis is not available in DataLab and location data is limited, but if it were available consistently and easily, analysts would be enthusiastic users of GIS functions
- Limits in computing power and bandwidth sometime limit analysts capacities

“I think you could be onto a real winner, just in uniting all that information by location. That would be really good.”

“I think it’s really good because if you want to do evidence based policy... Access to the data, such large datasets is great because it gives you more certainty within your regressions.”

“[Getting Access to MADIP/BLADE] is time consuming. And it seems not to be a very fixed process. And some things have changed along the way. So it’s taken some time to get everything squared away...”
# Sam Sterling – DataLab Analyst (Data Analyst 3)

<table>
<thead>
<tr>
<th>STAGES</th>
<th>Analytic task brief</th>
<th>Data acquisition &amp; preparation</th>
<th>Analysis</th>
<th>Information production and exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIVE SAFES</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Safe project / people</td>
<td>Safe data</td>
<td>Safe environment</td>
<td>Safe output</td>
</tr>
<tr>
<td>ANALYSIS TASKS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTIVITIES AND INTERACTIONS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS INTEGRATION AND DATA LAB TEAM TASKS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STEPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Commission enquiry</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Develop enquiry</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## SATISFACTION POINTS
- Enabling institutional and government support for use of DataLab and ABS capacities, shining light on work that can be done with data

## PAIN POINTS
- Policy questions being posed may not match capacities of BLADE/MADIP e.g. time-scale, spatial resolutions
- Processes for gaining access to DataLab takes time

## BLOCKERS & SUPPORTS

## OPPORTUNITIES FOR LOC-I
Loc-I identifiers (that enable reliable linking of data to locations) and pre-defined relationships between spatial features (enabling transformation of data between reporting geographies) can be used by the ABS data integration team to support preparation of data for delivery into the DataLab.

## SATISFACTION POINTS
- MADIP and BLADE are revolutionary and offer enormous analytic possibility
- Important datasets, such as transactional administrative data, are now integrated, linked, and available in a usable format through MADIP and BLADE

## PAIN POINTS
- MADIP and BLADE cannot be explored in totality - slices must be requested and used, with additional slices requested if these are not sufficient
- Turn-around times for new slices vary; can impact workflow
- Data fitness depends on nature of analysis and cleaning filtering may be required before use.
- Deficiencies in the granularity of spatial location data e.g. multi-enterprise businesses cannot be analysed

## SATISFACTION POINTS
- Large amounts of data available, which can be sliced for specific analysis tasks

## PAIN POINTS
- Limited analytic software available in the DataLab
- Computing capacity and bandwidth to run complex analysis can be limited, some delays possible
- Collaboration is difficult in Virtual Machine environment
- Cannot share or collaborate with colleagues who don’t have permission for DataLab
- Cannot perform geospatial analysis in the DataLab

## SATISFACTION POINTS
- Insights from granular data can be tremendous for policy output and design
- DataLab safe output vetting is really efficient

## PAIN POINTS
- Analysis performed outside the DataLab is disconnected from the analysis performed inside

---

With my policy team we develop a ‘safe’ project proposal and form the analysis team (safe people).

I request data from the ABS Integration team, who prepares slices of data from MADIP and BLADE based on availability and my analytical needs.

As additional data is needed for analysis, I ask ABS to prepare new slice or bring new data into DataLab.

I perform analysis using tools provided in the DataLab.

I create outputs from the analysis that I want to bring out of the DataLab.

I write a narrative (report, brief) about the analysis results and share with the policy area.

The DataLab reviews outputs and grants permission to bring the analysis output out of the DataLab.

Loc-I identifiers (that enable reliable linking of data to locations) and pre-defined relationships between spatial features (enabling transformation of data between reporting geographies) can be used by the ABS data integration team to support preparation of data for delivery into the DataLab.
Enterprise data warehouse administrator

“I work behind the scenes to make data work in our organisation.”

I work as a systems/database administrator in a large government organisation. My role involves the on-boarding, curation and supply of data to users in my organisation. I sometimes on-board data from external sources, and prepare this and internal data for integration in our enterprise data warehouse. For example, making sure health records stored in disparate systems are cleaned, structured, and integrated, into our environment for use by analysts.

I want to be able to effectively and efficiently provide data to my organisation’s users. I also want the information I supply to support operations in a timely manner, while maintaining all necessary compliance measures (legal, regulatory, policy, etc).

There is sometimes a (data) culture gap between my area and the broader organisation. Because I work across numerous systems and sources I am exposed to a range of issues and agency wide concerns that analysts don’t experience or understand. This sometimes creates a gap between myself and our analysts which isn’t helpful. I sometimes feel my work isn’t valued.

Helga Santoso

USER TOOLS
- FME
- Infomatica
- Cognos
- ESRI - ArcGIS / QGIS

DATA
- Python
- R
- Organisation-specific datasets of varying quality

USER INSIGHT
- Much data wrangling cost and effort is borne by back-of-house staff (e.g. maintenance of enterprise datasets) to enable analysts to have analysis-ready data.
- Approaches and architectures vary between organisations, creating confusion through differing data products, versions, and approaches between administrators.

“...geocoding, it's kind of a bit all over the place at the moment, there's not an enterprise geocoding strategy... we have four different geocoding software out there and they all give you slightly different longitude and latitudes.”

“So there was kind of a challenge to get access to all these different data sets lying around because people won't disclose, or won't want to disclose.”

“[The analysts] are supposed to come and talk to me. [But] in reality they just go off and do it. And hence we wind up with 14 copies [of the same data].”
Helga Santoso – Enterprise Data Warehouse (EDW) administrator

**Commission enquiry**

It takes a lot of time to get data into an EDW. Internal workplace practices vary around sensitive data.

Much of the cost and effort is in ensuring governance of the data and managing and use of data is consistent. These issues are often addressed by EDWs.

Growing recognition of how EDW contributes to informed decision making and performance improvement.

Cleaned data does not flow back into the EDW (it may be archived, or discarded).

There is a constant need to reference social and economic data with data derived from multiple sources.

Tools to rapidly and reliably reapportion data to multiple organisations are required.

**Analytic task brief**

I receive datasets (or request for data) to be added into the Enterprise Data Warehouse (EDW), by authorised people in my organisation.

I explore the requirements looking at the data to determine if/how it can be on-boarded and brought into our EDW.

I may (typically) clean, transform and QC data in order to onboard it into the EDW.

I liaise with IT and business units to ensure the on-boarded data is fit for purpose (and meets enterprise standards).

I produce regular reports of my organisation’s data holdings and their usage.

**Data acquisition & preparation**

I source data from our organisation’s EDW to perform an analysis task.

I clean and join geospatial and demographic data to perform analysis.

I perform required analysis.

Handover analytic information product (e.g. report with graphics and maps)

Cleaned data does not flow back into the EDW (it may be archived, or discarded)

**Analysis**

Satisfaction points:

- Use of business intelligence platforms in my organisation to visualise data, has highlighted the key role of EDW and well-governed data

- PAIN POINTS
  - Governing and shaping peoples’ data practices is time consuming
  - On-boarding numerous copies of the same data outside of the EDW, creates unnecessary duplicates
  - There is no desire to change entrenched data practices and behaviours leading to a lack of compliance with internal policies
  - Transforming data between different referencing geographies (social and environmental) is complex and time consuming
  - There is a constant need to reference social and economic data with data derived from multiple sources, including GNAP, ASGS, GeoFabric etc.

**Information production and exchange**

I produce regular reports of my organisation’s data holdings and their usage.

**Sourcing data**

Stages:

- Commission enquiry
- Develop enquiry
- Commission feedback
- Develop feedback
- Produce feedback
- Review feedback

Activities and interactions:

- Connection or handover
- System dependency
- Data product interaction

**Satisfaction points**

- Growing awareness at executive level of the importance of strong and unified enterprise data management and governance
- When leadership and executive staff engage with our work, assisting in positioning direction; this motivates others in the organisation to engage
- The common information model (like the FSDF common model) are useful to help inform the design of my (spatial data) warehouse and data product developed from it.

**Pain points**

- Lack of technology capacity amongst users limits engagement with our work and the EDW
- Variable culture and institutional arrangements across the organisation leads to variable data practice and lots of duplication of effort

**Opportunities for Loc-i**

- An easy to use set of user oriented tools to provide access points to Loc-i data
- Data exposed by Loc-i should have already dealt with the institutional arrangements and be ‘ready to use’
- Improved ability to manage spatial and other data separately; easily integrated using location IDs
- Well managed spatial feature identifier delivered in consistent ways can be more easily used and on-boarded by EDWs
- Reliable and consistent machine-oriented access to spatial data (based on published ontologies ) and consistent metadata
- Tools to rapidly and reliably reapportion data to multiple organisations

**Longer term**

- Publishing data via Loc-i provides a new level of visibility of datasets. Loc-i could produce usage metrics for data owners to report against

**Blockers & Supports**

**Steps**

- Commission enquiry
- Develop enquiry
- Commission feedback
- Develop feedback
- Produce feedback
- Review feedback

**Satisfaction points**

- Commission enquiry
- Develop enquiry
- Commission feedback
- Develop feedback
- Produce feedback
- Review feedback

**Pain points**

- It takes a lot of time to get data into an EDW
- Standardising the data and approaches to its management and use takes time and resources
- Managing multiple formats, structures and ‘shapes’ of data is tricky and takes time

**Opportunities for Loc-i**

- An easy to use set of user oriented tools to provide access points to Loc-i data
- Data exposed by Loc-i should have already dealt with the institutional arrangements and be ‘ready to use’
- Improved ability to manage spatial and other data separately; easily integrated using location IDs
- Well managed spatial feature identifier delivered in consistent ways can be more easily used and on-boarded by EDWs
- Reliable and consistent machine-oriented access to spatial data (based on published ontologies ) and consistent metadata
- Tools to rapidly and reliably reapportion data to multiple organisations

**Longer term**

- Publishing data via Loc-i provides a new level of visibility of datasets. Loc-i could produce usage metrics for data owners to report against
Next Steps – MVP and User Testing

- Development of three Minimum Viable Product (MVP) applications is underway. Initial engagement with potential user testers (identified through the user research process) has been commenced to identify use cases for Loc-I that application development will be tailored to address.
- The Loc-I team will seek to visit user testers in the field, observing and identifying workflows and use cases with them, using an MVP as the basis for co-designing solutions using Loc-I.
- If you’d like to be involved in this process please get in touch - contact details on the next page.
References:


Acknowledgements:

With great thanks to the participants who volunteered their time for this user research, to the DIPA board and members of the Loc-I project and their associated Commonwealth agencies, and to Sally Tetreault-Campbell for editorial support.
Land and Water
Paul Box
Loc-I Project Lead
t  +61 2 9325 3122
e  paul.j.box@csiro.au

Land and Water
Nicholas Car
Loc-I Technical Lead
t  +61 7 3833 5632
e  nicholas.car@csiro.au

Geoscience Australia
Shane Crossman
Loc-I Project Manager
t  +61 2 62499499
e  Shane.Crossman@ga.gov.au

https://research.csiro.au/ei/

Loc-I Project Team
CSIRO - Paul Box, Nicholas Car, Ashley Sommers, Shane Seaton, Ashlin Lee, David Lemon, Edmond Chuc
GA - Shane Crossman, Joseph Bell, Mathew Purrs, Irina Bastrakova, Will Francis, Alex Ip