Document 1

Data Integration Oversight Forum Minutes – Meeting #23

Minutes: 1 December 2025

Date & Time	**Monday 1 December 2025, 130pm to 230pm AET**
Chairperson(s)	Emily Walter and Luke Hendrickson
Location	Canberra 4N 431 Mudda Meeting Room
Dial in	MS teams
Invitees	**Co-chairs:** Emily Walter **Members:** Helen Robson, Adam Harris, Jodie Stevenson, Andrew McMahon, Deborah Miliszewski^ **Observers:** [redacted] **Apologies:** Andrew McMahon, Jodie Stevenson, Helen Robson, Siobhan Campbell^, Deborah Miliszewski ^, [redacted] **Proxies:** [redacted] for Andrew McMahon, [redacted] for Jodie Stevenson, [redacted] for Helen Robson **Presenters:** Emily Walter **Secretariat:** [redacted] [redacted]
Attachments	[redacted]
Links	[redacted]

#	Agenda Item	Discussion and outcomes
1	Welcome, Introductions, Acknowledgement of Country (Chair)	Emily commenced the meeting with an Acknowledgement of Country and welcomed all attendees.
[redacted]		

S22

S22 [redacted]

Methodology S22 [redacted]

- A paper focusing on the recent spine build using Splink was presented to the Methodology Advisory Committee last week. Some suggestions received which will be considered and prioritised in conjunction with other possible improvements for the next spine build identified by Methodology and the DLC.

S22 [redacted]

Document 2

Data Integration Oversight Forum Minutes – 3 June 2024

Date & Time	Monday 3 June 2024, 14:00 – 15:00
Chairperson(s)	Luke Hendrickson
Location	Canberra 4N 431 Mudda Meeting Room
Dial in	MS Teams
Invitees	**Chair(s):** Luke Hendrickson* (this meeting) **Members:** Michael Meagher, Helen Robson, S22 [redacted], Jenny Telford*, S22 [redacted], Adam Harris* **Proxies:** **Guests:** **Observers:** S22 [redacted] **Presenters:** S22 [redacted] **Secretariat:** S22 [redacted] S22 [redacted] ** indicates participants present in the first group*
Papers	S22 [redacted]
Attachments	PowerPoint presentation “Introducing Splink” (Agenda Item 4) S22 [redacted]

#	Agenda Item	Discussion and outcomes
1	Welcome	The Chair opened the meeting and made an Acknowledgement of Country in Ngunnawal language. *Concurrently, Helen Robson provided an Acknowledgment of Country in the second group meeting.*
S22 [redacted]		

[redacted]

4	Splink presentation [redacted]	A presentation (attached) regarding Splink was provided by [redacted] with an introduction by [redacted]
		
		The presentation provided:
		• An overview of probabilistic vs deterministic linking
		• Current ABS linkage tools (DMAC, FEBRL)
		• An overview of Splink - a Python package for probabilistic linking
		• Results from investigations to date, including outputs from NLS spine build using Splink, and Medicare and Domino linkages using Splink and DMAC
		• Timeline for ongoing work
		• Recommendations
		
		Questions from the Forum:
		• Whether confirmation can be given that cloud scalability and speed benefits are not being conflated when comparing Splink and DMAC? (Luke H)
		  o Confirmation not possible, there is currently no evidence to support. DLC are working on rebuilding DMAC so it can function in the Cloud, once complete, comparisons will be clearer. [redacted]
		• Luke has indicated it would be good to test a five-way linkage method to include Births and migration data into a spine, if this data can be supplied in a timely manner.
		• Regarding social licence – noting Adam anticipating doing the national linkage map with Splink – have there been public discussions regarding the probabilistic method? And noting upcoming work such work re the PLIDA PIA, do we need to raise this? (Luke H)
		  o At this stage work has been focused on assessing the utility of the Splink tool for NLS. Working Group established around the NDDA to look at the quality of work to build the spine which has led to conversations about the tools, but this has not led to public conversations about social licence. Potential for this to occur re NDDA and ANDII with update of the Privacy Impact Assessment next year – this will be one mechanism available for these conversations (Adam H)
		  o Question is how we need to engage with the public with these analytical processes. In consideration of privacy work to establish SEAD and update of privacy assessments to cover this, informing rather than consulting on this change would be the recommendation [redacted]

- Don't think we need permission to use a different method, but consider what warranty we give when we release the spine – same, or better? Consider low risk, that we could publicise a change in method without raising public alarm (Michael M)

- What is the success point to put Splink into production? (Michael M)
  - There are discussions to roll this out now. Noting 98% identical to outputs from current tools - need to consider what is the 2% and is this heavily biased in a certain area (i.e., if probabilistic matching doesn't handle certain criteria well). Need to understand the bias, and if a bias correction required. But no red flags to not utilise Splink, if we can de-couple migrating to Cloud from Splink (Luke H)
- Noting the use of integrated data and in consideration of the full spine rebuild each year and comparability over time - do we know enough about Splink to know if we would get the same linkage result year-on-year, or do things like the ordering of data change how those links get defined? (S22 [redacted]
  - Expect results largely similar as probabilistic linking is not a random process – imagine a somewhat similar risk of change as with deterministic linking. After a year, some data has changed, some records have been updated, some records are present which were not there before, and in the presence of those records, some links would be made differently. But don't envisage this would cause us to make a lot of different decisions about links. S22 [redacted]
  - Currently the ABS Person Linkage Spine outputs are approximately 98.4%– the same between Splink and DMAC. Support for the ABS Linkage Spine and NLS Linkage Spine are on our FWP for next year. Noting stakeholder feedback, even compared with DMAC in the Cloud, there is a remarkable improvement in speed using Splink. Splink replaces several aspects in the process and provides a marked improvement in time, not just over DMAC, but overall. S22 [redacted]
- In discussion is whether a full spine rebuild every year is necessary if 98% of the data is not changing. This does not directly relate to Splink but does impact on the questions whether a full spine rebuild would lead to a bias or drift in population year-on-year. (Luke H)

5 S22 [redacted]

Document 3

Introducing Splink

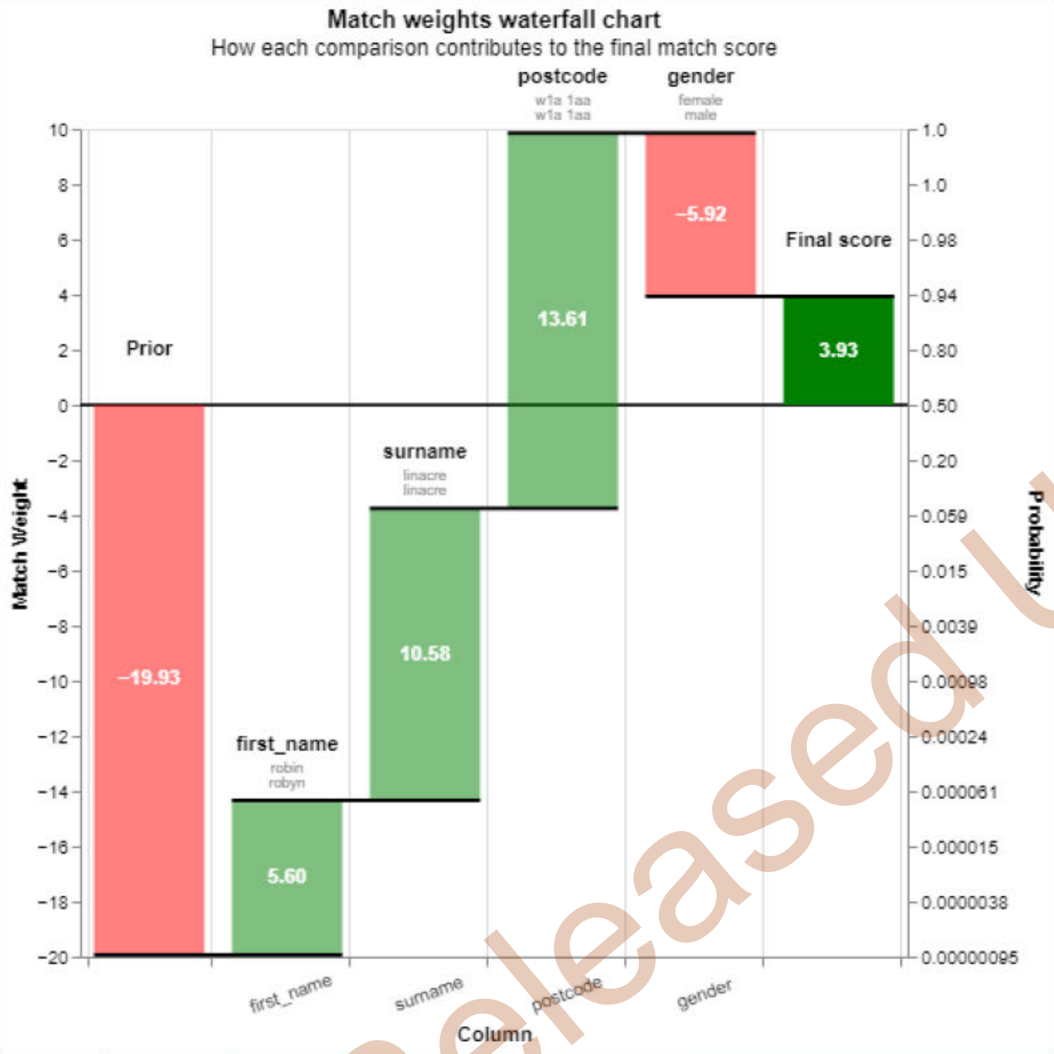
Data Integration Program Board
June 2024

Australian Bureau of Statistics
Informing Australia's important decisions

Deterministic Linking

- Deterministic linking relies on rules
  - Eg if a record pair is a unique match on first name, surname, DOB, sex then it is a link
- In order to account for a range of circumstances we typically need a lot of rules and many passes through the data
  - Eg link on first name, surname, DOB, sex
  - Then first name, surname, DOB only
- Selecting and ordering rules can be subjective and deriving quality measures is tricky

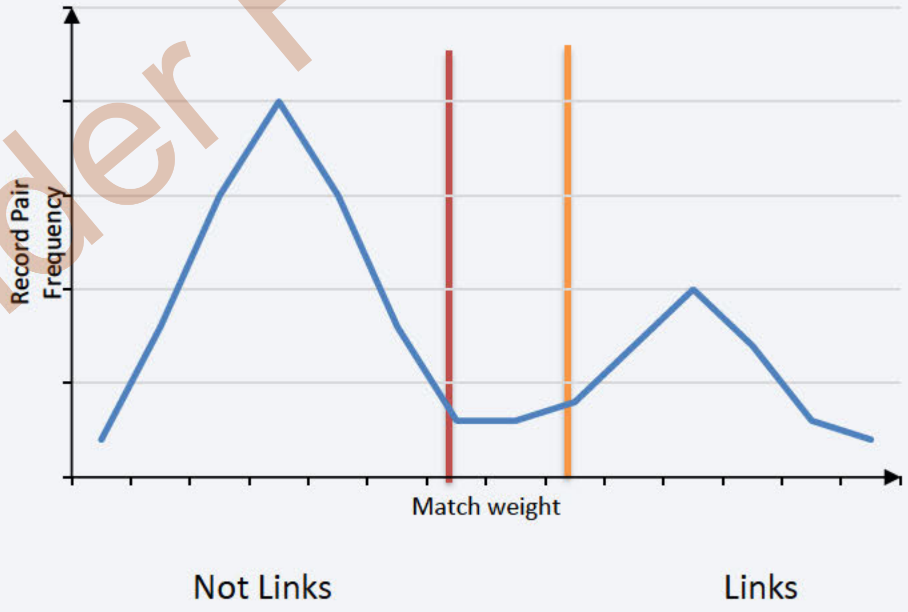
Probabilistic Linking



- Probabilistic Linking calculates a ‘comparison weight’ for each linking variable for each record pair
  - This is derived from the closeness of the comparison and the discriminatory power of the variable
- These weights are summed to create a match weight for each record pair

Probabilistic Linking

- We choose a threshold for match weights balancing the need to find more links with the risk of false links
- Anything above the threshold we assign to be a link



Probabilistic vs Deterministic Linking

Australian Bureau of Statistics

- Probabilistic linking is considered best practice by most NSOs
- Probabilistic linking makes use of more information when determining a link than deterministic linking can
- Probabilistic linking can form links when linking variables are of poor quality
- Probabilistic linking provides better, more robust quality measures for a linkage process
- Probabilistic linking is more computationally intensive

Current ABS linkage tools

- DMAC
  - Deterministic linking SAS macro
  - Used for most large-scale linkages including ABS Spine creation
  - Currently being rewritten in Python/PySpark for cloud

- FEBRL
  - Probabilistic linkage Python (v2) package
  - Currently used for PES-Census and Census-Mortality linkages
  - Very slow and inflexible
  - No longer supported (tool or language)

Splink

- A Python package for probabilistic linking
  - Fast and highly scalable via parallel computing in the cloud

- Flexible and easier to use than FEBRL

- Capable of multi-way linkages in a single run

Results so far

- On a standard spine deduplication and linkage task Splink in SEAD is ~10x faster than DMAC in NGI

- For the NLS spine build Splink made 98.7% links the same as the current NLS process

- Medicare and Domino linkage (first two steps of ABS spine) was 98.3% the same between Splink and DMAC

Timeline for ongoing work

- [Currently]: Investigate Splink as a replacement for FEBRL in PES-Census linkage
- [Currently]: Investigate further into the links Splink and DMAC make differently for NLS spine
- [Soon]: Understand Splink's capacity to do 4+ linkages simultaneously
- [June-Dec 2024]: Implement the 2024 NLS spine build in Splink
- [June-Dec 2024]: Develop quality measures from Splink outputs to inform future linkage quality assurance
- [2024-25]: Understand Splink's capacity to improve linkage
  - Especially for poor quality linking variables

Recommendations

- That the ABS consider adopting Splink to replace FEBRL
- That the ABS consider adopting probabilistic linking and Splink more widely
  - Splink be considered for use in the 2025 ABS Spine build

Questions for the Board

- Does the board have any questions about Splink?

- What further information would be helpful to support a broader adoption of Splink?

Document 4

Data Integration Oversight Forum Minutes – Meeting #22

Minutes: 14 November 2025

Date & Time	**Friday 14 November 2025, 2pm to 3pm AET** *# Rescheduled from Monday 3 November, 130pm-230pm AET*
Chairperson(s)	Luke Hendrickson and Emily Walter
Location	Canberra 4N 431 Mudda Meeting Room
Dial in	MS teams
Invitees	**Co-chairs:** Luke Hendrickson and Emily Walter **Members:** Helen Robson, Adam Harris, Siobhan Campbell, Jodie Stevenson, Andrew McMahon **Observers:** S22 [redacted] **Apologies:** S22 [redacted] **Proxies:** S22 [redacted] **Presenters:** S22 [redacted] **Secretariat:** S22 [redacted]
Attachments	
Links	

#	Agenda Item	Discussion and outcomes
1	Welcome, Introductions, Acknowledgement of Country (Chair)	Emily commenced the meeting with an Acknowledgement of Country and welcomed all attendees.

[redacted]

S22

S22

[redacted]

		Methodology (Andrew) • First spine build with Splink is now complete. A paper will be presented at a methodology advisory committee in a fortnight seeking feedback from international peers and academic experts.

S22

[redacted]

Document 5

Data Integration Oversight Forum Minutes – Meeting #21

Minutes: 13 October 2025

Date & Time	Monday 13 October 2025, 1:30pm to 2:30pm AET
Chairperson(s)	Emily Walter
Location	Canberra 4N 431 Mudda Meeting Room
Dial in	MS teams
Invitees	**Chair:** Emily Walter **Members:** Luke Hendrickson (co-chair)^, Helen Robson, Adam Harris^, Siobhan Campbell^, Jodie Stevenson^, Andrew McMahon^ **Observers:** [redacted] **Proxies:** [redacted] for Luke Hendrickson, [redacted] for Adam Harris, [redacted] for Siobhan Campbell, [redacted] for Jodie Stevenson, [redacted] for Andrew McMahon **^ Apologies:** Luke Hendrickson, Adam Harris, Siobhan Campbell, Jodie Stevenson, Andrew McMahon, [redacted] **Presenters:** [redacted] **Secretariat:** [redacted]
Attachments	[redacted]
Links	[redacted]

#	Agenda Item	Discussion and outcomes
1	Welcome, Introductions, Acknowledgement of Country (Chair)	The Chair commenced the meeting with an Acknowledgement of Country. The Chair noted apologies for several members and their nominated proxies.

S22 [redacted]

		TSD and Methodology update (S22 [redacted]) • Currently working through QA process for the latest spine build, the first using Splink. Progressing well – complete linkage of all three datasets finalised within two days (previous timeframe 6 weeks approx.)

Document 6

Methodology Design Committee Meeting - December 2024 [SEC=OFFICIAL]
MDMD Admin WDB [redacted] 23/09/2024 12:09 PM

OFFICIAL

Basics

Protective Mark	OFFICIAL		
Information management markers	☐ Personal privacy ☐ Legal privilege ☐ Legislative secrecy	Caveat	☐ N/

Methodology Design Committee (MDC)
Agenda

Date | Time	4th December 2024 |3:00pm - 4:30pm (AEDT)
Location	Teams meeting with link in calendar invitation
Convenor | Chair	Anders Holmberg
Minutes	[redacted]
Invitees	**MEMBERS:** Anders Holmberg (GM), [redacted], Paul Schubert, Julian Doak, Jodie Stevenson, Tom Joseph, David Zago, [redacted] (for Robert Ewing), [redacted] **INVITED CONTRIBUTORS/GUESTS:** [redacted] (for Luke Hendrickson), [redacted] (for Emily Walter), [redacted] Denise Carlton, [redacted] **APOLOGIES:** Robert Ewing, [redacted]

#	Time (Minutes)	Agenda Item	Presenter	
1	[redacted]			
2				
3				
4	20	SPLINK and its evaluation	[redacted]	Splink and its evaluation.d
5	[redacted]			

Document 7

Minutes MDC - 4th December 2024 (Response to: Methodology Design Committee Meeting - December 2024 [SEC=OFFICIAL])
MDMD Admin WDB [redacted] 05/12/2024 11:48 AM

OFFICIAL

Basics

Protective Mark	OFFICIAL
Information management markers	☐ Personal privacy ☐ Legal privilege ☐ Legislative secrecy
Caveat	☐ NATIONAL-CABINET

Methodology Design Committee
Minutes: 4th December 2024

Agenda:

Attendees e [redacted] Paul Schubert, David [redacted] (for Robert Ewing), [redacted] (for Luke Hendrickson), [redacted] (for Emily Walter), [redacted]

Apologies: Robert Ewing, [redacted] Julian Doak

# Agenda Item	Paper/Presentation	Notes and Actions
[redacted]		

S22 [redacted]

4	SPLINK and its evaluation	Splink and its evaluation.docx Splink slides for MDC.pptx	The MDC endorsed that probabilistic linkage using the SPLINK tool be the recommended approach for ABS linkage projects. The following comments were also given: • The MDC was very supportive of probabilistic linkage using SPLINK and keen to see it move forward. • This is a really well-written paper • It is great to see MDSD appropriating existing tools and methods rather than inventing new ones ourselves. • S22 [redacted] • It was asked if there are any applications of SPINK for business surveys/data but it was not obvious that there were very many (given the availability of deterministic identifiers) • It was clarified that MDC 'endorsement' is of the methodology but that support from other areas will be needed (and should be easily obtained) to keep moving forward the adoption of SPLINK and probabilistic linking.

S22 [redacted]

The case for updating linking methods in the ABS

Purpose of this paper

Data Linkage Methodology (DLM) are seeking MDC endorsement on the adoption of the probabilistic linking tool Splink for future ABS linkages. Splink is a Python library that uses the Fellegi-Sunter algorithm to probabilistically link datasets together. Testing has shown that Splink is a much more efficient tool than the existing linkage tools used in the ABS, and produces results of a comparable quality to our existing linkages.

Probabilistic and Deterministic Linking

The two main linkage methods used in data linking are **probabilistic** linkage and **deterministic** linkage.

Probabilistic linkage considers a range of identifiers (variables like first name, surname, date of birth, and address) and assigns weights to each identifier based on its ability to correctly identify a match. These partial weights are combined into what is called a match weight. Records are linked if their match weight is greater than a set threshold. Notably, this allows records to be linked even if they disagree on certain identifiers. It also allows the degree of disagreement to be incorporated into the match weight. For example, a name that differs by 5 letters from the original could get a lower partial match weight than a name that differs by 3 letters. In calculating match weights probabilistic linkage also includes the 'm' and 'u' probabilities. That is, the probability that a unit is a match or not a match given the identifier in question is the same. These probabilities are usually calculated from the data to be linked using the Expectation-Maximisation (E-M) algorithm. This approach is often referred to as the Fellegi-Sunter method for linking as it was first proposed by Fellegi and Sunter in 1969.

Deterministic linkage differs from probabilistic linkage in that it requires agreement on **all** identifiers to determine that two records are a link. To account for errors in the data (such as spelling mistakes in names), multiple passes are run with different combinations of identifiers. For example, pass 1 may link all pairs of observations that match on first name and surname, pass 2 may link all pairs of observations that match on first name, date of birth, and address, and so on. The ordering of passes to achieve the most optimal set of links can often be down to trial and error. At the ABS it is done by comparing the 'marginal uniqueness rate' – the proportion of records that will be uniquely matched in a pass given the other passes that have already happened. Generally, we expect that if more unique links can be made in a particular pass, then it is a better, higher quality linkage. A lot of work has been done at the ABS to establish best practice for this approach to deterministic linking, given the linking variables we typically use for linking records.

While the ABS has traditionally performed deterministic linkage, there are several reasons why we need to maintain a probabilistic linkage method and tool alternative to our existing deterministic linking tools:

- **Alignment with external best practices:** Most NSOs use probabilistic linkage as their primary linkage method. For example, Statistics Canada has recommended using probabilistic linking to match records without unique IDs for over a decade, and the UK

Office of National Statistics recommends probabilistic linkage approaches as, compared to deterministic approaches, they are more robust against errors and more effective with large datasets.

- **The ABS already performs probabilistic linkage:** PES-Census linkage currently uses probabilistic linkage through the Febrl tool (Freely extensible bio-medical record linkage) and will need to use this methodology into the future (this is very important to ABS business as it supports the legislated output of population estimates).
- **Better quality linkages:** Probabilistic linking incorporates disagreements as well as agreements into deciding whether a record pair should be linked, hence leading to better quality results.
- **Linking low-quality data:** Deterministic linking approaches struggle to link low-quality data and using a probabilistic method may allow us to maximise the number of links found in those challenging cases while avoiding too many false links.
- **More sound quality metrics:** Having a probabilistic model allows us to straightforwardly get estimates of precision (the proportion of all links that are true links).
- **Relationships with researchers:** Most researchers' approaches to post-linkage analysis usually assumes that the linked dataset has been probabilistically linked and that linkage weights are available.

The Deterministic Linkage Macro (DMAC)

In the recent past, deterministic linkage through DMAC has been the main method and tool used for linking in the ABS. This was largely driven by logistic concerns as the ABS did not have ready access to a linkage tool capable of running probabilistic linkage at the scale typically required for creating or linking to the ABS Person Spine. However, along with the considerations listed above, there are reasons why the ABS should be looking to move away from DMAC:

- **DMAC cannot be used to link PES records to the Census:** Analysis in 2023 showed that, due to the different methodologies used by Febrl and DMAC, DMAC missed too many key links and would have required significant retooling of the PES automated linking components to be a feasible replacement for Febrl.
- **DMAC doesn't produce robust quality estimates:** When performing a linkage, DMAC requires the user to manually determine at which point we stop accepting links, and our current methods for setting these cut-offs are very ad-hoc.
- **DMAC is built in SAS:** As part of the Australian National Data Integration Infrastructure project, the ABS is hoping to run future linkages in virtual environments. DMAC is written in SAS, which is not expected to be available in these environments. Consequently, DMAC will need to be rewritten in a supported language (such as Python) for use in future linkages.

Freely extensible bio-medical record linkage (Febrl)

The ABS has historically used Febrl to do probabilistic record linkage. Febrl is a python 2 library developed by Peter Christen at ANU under contract from the NSW Health Department in the early

2000s. There are several reasons why Febrl is currently not routinely used for linkages at the ABS and why we should not rely on it into the future:

- **Run time of linkages:** Febrl is very slow, meaning that Febrl projects may have to use an excessively strict blocking strategy to run in an acceptable time, which potentially compromises the quality of the linkage. It also means that we cannot realistically use Febrl for large linkages like Census-Spine and the Spine build.
- **Configuration is prone to errors:** Probabilistic linking requires estimation of parameters to compute a prediction of whether two records are a match. Unfortunately, Febrl does not include an implementation of the Expectation Maximisation algorithm for parameter estimation. ABS linkage projects using Febrl require the user to calculate the *m* and *u* probabilities in SAS and then manually enter them into Febrl. This process is prone to introducing errors and undermines the reliable implementation of probabilistic linking on a production scale for the Data Linkage Centre.
- **Inflexibility of linkage strategy:** Lack of flexibility in how comparisons between linking variables are specified means that we must run multiple passes. This increases the complexity of projects, and leads to very long runtimes for linkage processes, as a single pair of records may be compared in several different passes. We also have no theoretically defensible way of combining results from multiple passes.
- **Febrl is out of date:** Febrl is implemented in python 2, which has been beyond its end of life since January 2020. Python 2 will not be present in virtual environments where we hope to run future production code. Upgrading Febrl to python 3 would be a substantial project. Additionally, Febrl is not used outside the ABS, and Peter Christen has not updated it in 10+ years.
- **Difficulty of use:** Even simple projects require the user to interact with a large amount of code. This creates key-person risks when people with experience in using Febrl move out of sections that perform linkage projects using Febrl.

Splink

Splink is a fast and scalable Python library for probabilistic linkage. It was developed by a team in the UK Ministry of Justice, funded by the UK Government Data Science Partnership. Our testing has shown that Splink has many benefits compared to DMAC or Febrl. The key benefits we have identified are as follows:

- **Run time of linkages:** Splink can perform a linkage and deduplication of 300 to 500 million record pairs in 60-90 minutes, while linking the same datasets with DMAC would take 12 hours.
- **Quality metrics**: The presence of pairwise match probabilities opens the potential of several micro-quality metrics that can be derived which are not possible with DMAC.
- **Flexibility of linkage strategy**: Unlike Febrl, comparisons in Splink are not limited to yes/no agreements of the linking variables, and instead you can create multi-level comparisons.

- **Ease of use**: Running Splink requires writing a short Python script where the user defines how the linking variables will be compared, what blocking rules will be used, and above what threshold links will be accepted. This is significantly simpler than Febrl, which requires a configuration file and a separate script for each pass in the linkage project. The slow speed of Febrl also means that testing different linkage strategies is very onerous.
- **Simplifying the linkage process**: Splink can link more than two datasets at the same time, and link and deduplicate multiple datasets at the same time.
- **Expectation-Maximisation (EM) algorithm**: Splink includes an implementation of the EM algorithm for estimating required probabilistic linkage parameters, which is a marked advantage over Febrl.

Results

While it is vital to find a replacement for Febrl for those linkages where the ABS currently applies probabilistic linking, Splink opens the opportunity to gain the benefits of probabilistic linking for a wider range of linkages. This section describes the outcomes of testing Splink on several ABS linkages.

ABS Spine

The ABS Person Spine is a three-way linkage between DOMINO data from DSS, the Medicare Consumer Directory, and Personal Income Tax (PIT) data from the ATO, all from 2022. DMAC can only link two datasets together at one time, meaning this linkage is done in several steps involving first deduplicating each dataset and then linking them together one at a time. Deduplication is usually performed as a separate linkage process in which a dataset is linked to itself to find any duplicate records.

Splink can link any number of datasets together in one go and can also deduplicate records while the linking is being performed, meaning that the entire ABS Spine can be built with one linkage step instead of five. DLM recreated the ABS Person Spine in Splink, linking all 3 datasets together in a single linking step. It took Splink 1.5 days to perform the full linkage, which compares favourably with the current approach where the linking process of the spine build took 73 days.

Splink could reproduce 98.44% of the original Spine elements unchanged from the current approach. A manual review of a sample of the records that were linked differently across the two approaches found that Splink produced a better outcome in a majority of cases (56% vs. 18% for the current approach), with the remaining 26% of cases being impossible to determine which method was better due to the inherent limitations of the data used for linking records and being unable to discern common and derivative names due to de-identification.

Some changes were made to deal with the increased memory and computation requirements of linking the third dataset. A blocking rule developed for the NLS linkage concatenated the full name in a way that included null values. This rule created very large comparisons that crashed the linkage due to memory timeouts. Removal of the rule meant 13,604 records, or 8,156 distinct IDs would not be involved in blocking. Inspection of these records showed they had first and last name information only, with missing values elsewhere, so there wasn't value in keeping the rule. Three more blocking rules were dropped as they did not provide new comparisons. Comparisons on date-of-death and sex were added to ensure the u values trained fully after dropping the four blocking rules.

The Splink build was significantly better when it merged clusters that were initially split in the original spine build, and slightly worse when it split clusters that were merged in the original spine build. When looking at the results at a finer grain, examining the individual cluster structures, we see relatively similar results to the overall sample totals with the majority of outcomes being positive changes by the Splink build.

The National Linkage Spine

The National Linkage Spine, which is part of the National Disability Data Asset, was initially built using DMAC, followed by the Australian Institute of Health and Welfare's probabilistic linkage tool, and then their machine learning model to predict the match status of the remaining ambiguous links. Splink has the capacity to replace all these steps. Earlier this year Splink was approved for use in building the 2024 NLS. This build commenced in October 2024 and has been progressing well, and is estimated to have saved 1 month of processing time, with more savings forecast in the future as the process is bedded down.

Another key benefit of Splink is that it is able to create the National Linking Spine without using SAS. This is important as SAS is currently not supported in the virtual environments that will be used for future linkage projects.

Results from the National Linkage spine testing showed that Splink could reproduce 98.81% of the original Spine elements unchanged from the current approach. A manual review of a sample of the records that were linked differently across the two approaches found that the Splink build and the current method produced a better outcome in a similar share of cases. The Splink build was mostly better when it merged clusters that were split in the original spine build, and worse when it split clusters that were merged in the original spine build. These differences were mostly due to Splink more harshly penalising disagreements in geography levels and being better able to merge duplicate records that had a slight disagreement in the date-of-birth field.

Overall, the fact that Splink replaces three different methods, yields significant time savings, simplifies the linkage process, incorporates a more methodologically defensible data linking method, does not rely on SAS, and produces results that are of comparable quality to the existing method were all factors in the decision to build the National Linkage Spine in Splink.

PES-Census

Splink has been endorsed to replace Febrl to link PES records to Census records in 2026. Testing on PES and Census data from 2021 saw both methods produce a similar number of high-quality links and drop a similar number of 'true' person links (and for both methods these missed links would still be found via clerical linking). In addition, Splink also produced 10k fewer junk links that would have been flagged for clerical review, which represents a saving of approximately 333 staff hours. The Splink run took 12 hours, which is comparable to a Febrl run that uses 4 processing cores. Given that Splink was not run using Databricks and a move to Databricks would result in a much-reduced runtime (for comparison's sake, the National Linkage Spine – a much larger linkage than PES-Census – takes 5-6 hours using Databricks), the results from the testing combined with the expected decrease in runtime and the issues with Febrl highlighted above mean we are confident that Splink is the best method to link PES to the Census in 2026.

Survey of Income and Housing

Staff in the Data Linkage Centre (DLC) investigated using Splink to link data from the Survey of Income and Housing (SIH) with the 2022 PLIDA Person Linkage Spine. The key findings were:

- Splink was significantly faster than DMAC: The total runtime for the SIH linkage using Splink was approximately 1.5 hours, compared to over 28 hours using D-MAC.
- Splink produced a higher linkage rate than DMAC: Splink produced a linkage rate of 93.64% compared to 91.45% using DMAC.
- Splink improved linkage rates for populations with historically low linkage rates, including increasing the linkage rate for the 25-34 age group, the Northern Territory, and Remote Australia, by 5.03%, 4.29%, and 5.47% respectively.

Ongoing Work

Improvements to Splink

DLM has developed a set of tools built around Splink which are intended to be made available for use in future linkages. For example, staff in DLM created a tool that allows the user to sweep through a range of different parameters to detect the optimal settings for a Splinkage.

Splink is open source, which allows users to make changes to it. Several staff in DLM have submitted improvements to Splink that improve its functionality from an ABS context, and these have been incorporated in subsequent Splink releases.

There will be some further development required to set out appropriate quality measures and build the functions to produce these. We are expecting to be able to produce some quantitative measures from Splink which are not available in the DMAC/deterministic paradigm.

LFS linking

DLM are using Splink to relink the 12-month Labour Force file that was originally linked using DMAC. Preliminary results show that the Splinkage produced a similar number of total links, including approximately 10,000 more links assessed as high quality. The processing time for Splink also appears to be faster than DMAC.

Twins Adjustment

The Fellegi-Sunter model divides all record pairs into two classes: matches and non-matches. Then there are parameters which specify the probability of a given agreement pattern of the record pair conditional on whether the pair is a true match or not.

This can be problematic given the presence of twins (and to a lesser extent family members) in the data, since we have pairs of records that agree on several fields (like surname, date of birth, and address), and the prevalence of these agreements is much more than one would expect from a random sample of non-matching records. We can add an extra class to the model, meaning that pairs of records can be classified as matches, twins, or non-matches. This was tested in the NLS build and it was successful in removing most of the twin links that we had found. It was decided to not use this approach since there was additional complexity in implementing it, and the adjustment had some unintended consequences such as ranking links which disagreed on surname higher than links which agreed on everything.

Quality Measures

Splink produces estimated match probabilities based on the model implemented under the Fellegi-Sunter framework. The estimated match probabilities are particularly useful in the absence of ground truth data (this is the case for ABS linkages) because they enable the calculation of several key metrics (such as precision and recall) used to evaluate and improve the quality of the linkage.

The match probabilities are also useful for:

- Forming clusters in more complex linkage scenarios such as one-to-many and many-to-many links e.g. spine build and PES-Census linking.
- Enabling the provision of more useful quality measures to users so that if desired they can account for linkage uncertainty in their analyses. There is academic literature on this topic, but it requires analysts to be provisioned with estimated match probabilities.

It should be noted that the estimated match probabilities rely heavily on the accuracy of the linkage model set up by the user of Splink. Hence, model specification and potentially some form of calibration of the estimated match probabilities is essential to avoid biased estimates.

Linking Aboriginal and Torres Strait Islander data

DLM have commenced a project that involves improving linkage quality for the Aboriginal and Torres Strait Islander population. When key linkage variables are of poor quality (such as in the Indigenous population) then probabilistic linkage is known to be more flexible and effective than deterministic, and therefore we will be using Splink in these investigations.

Discussion and recommendations

Splink has proven that it can more efficiently produce very similar results to linkages carried out using deterministic linking. There are significant benefits to the use of Splink in terms of improved speed and functionality, and probabilistic linkage also offers numerous benefits compared to the deterministic linkage methods currently used in the ABS data integration programme. Splink offers a way past the barriers that have prevented the ABS from adopting probabilistic linking in the past.

Two major ABS linkage projects have approved the use of Splink for their upcoming linkages, and for the above reasons we recommend that the ABS continues down this path and, unless there is a specific reason to use deterministic linking, adopts Splink as the primary linking tool for all future ABS linkages.

Australian Bureau of Statistics

Splink and Its Evaluation

Data Linking Methodology

AUSTRALIAN BUREAU OF STATISTICS
Informing Australia's important decisions

Two decisions requiring endorsement

01

Shifting from **deterministic** linking to **probabilistic** linking

02

Endorsing Splink as the **default** probabilistic linking tool

Why probabilistic linking?

Alignment with **external best practices**

Incorporation of **disagreements** as well as **agreements**

Better for **low quality data** (e.g. Aboriginal and Torres Strait Islander Linking)

Gives **straightforward** estimates of precision

Why Splink?

ABS has tested 3 linkages with Splink:

- ABS Spine
- National Linkage Spine
- PES-Census

Results have shown that...

Splink is fast

NLS linkage went from 5 days to 7 hours

ABS Spine linkage went from 73 days to 1.5 days

Splink is accurate

Splink reproduced **98.44%** of ABS spine elements.

Splink reproduced **98.81%** of NLS elements.

Splink reproduced a similar number of good links to the existing PES-Census process, and also produced **10,000 fewer** junk links.

Splink is comprehensive

Splink contains an implementation of the EM algorithm.

Splink can link multiple datasets together and deduplicate in a single linking step.

Splink is customisable

Splink allows for **multi-level comparisons**

Users can **add classes** to the model to allow for things like twins

DLM staff have contributed **directly** to the Splink code base

Why we are seeking endorsement

Splink has been endorsed for 2 ABS linkage projects, and we expect it to be adopted for next year's ABS Spine build.

The methodological endorsement of Splink is somewhat ad-hoc, and we would like MDSD's **formal** position to be that Splink is the default linking tool in the ABS.

MDC endorsement would be highly valued in this space.

Person Linkage Spine June 2025

Methodology and quality assessment

October 2025

Contents

Executive Summary

The Person-Level Integrated Data Asset (PLIDA) is a secure data asset combining information on healthcare, education, government payments, income and taxation, and population demographics (including the Census) over time.

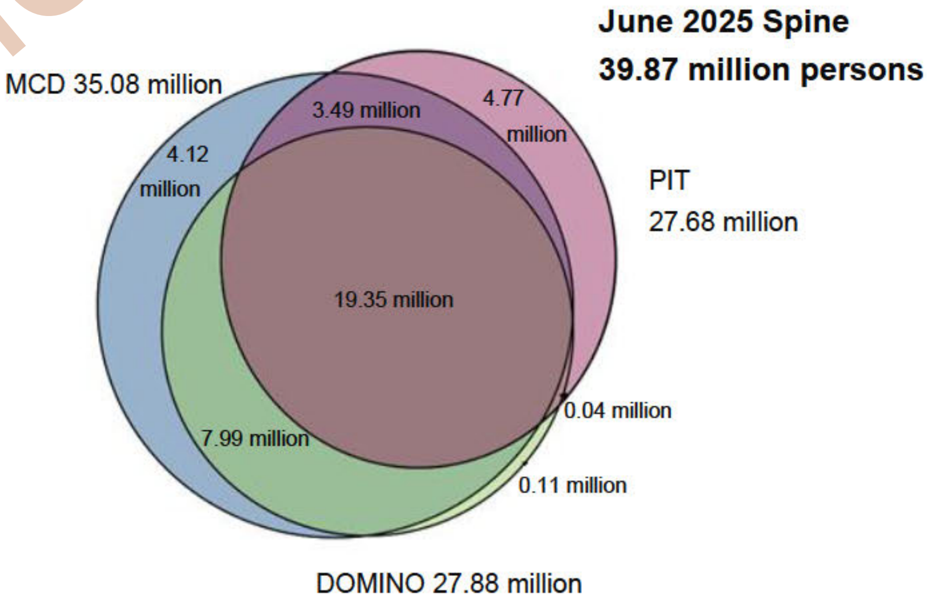
The Person Linkage Spine (the 'Spine') is the core infrastructure enabling integration of these datasets and is a combination of data sourced from the Medicare Consumer Directory (MCD), Data Over Multiple Individual Occurrences Centrelink Administrative Dataset (DOMINO) and Personal Income Tax (PIT). The Spine is updated annually to include an additional year of data from these three datasets. This paper describes the methods and quality assessment of the June 2025 Spine.

For the June 2025 Spine build, the ABS used a probabilistic linkage tool called "Splink" for the first time. All previous versions of the Spine were created using an ABS developed deterministic linkage tool. The June 2025 Spine update includes data from January 2006 through June 2025; it comprises of 39.87 million unique persons (elements). Compared to the 2024 Spine, the additional year of data incorporates a further 0.86 million unique persons on the Spine. This is consistent with population growth estimates.

The overall structure of the Spine is consistent with previous versions, with the contribution from each source dataset remaining within one per cent of last year. Linkage rates across the three datasets were consistent with previous years. The change to probabilistic methodology resulted in a marginally higher linkage rate across all three datasets, and a small increase in duplicate[1] record identification, though duplicates remained very minor components of the overall datasets.

Figure 1 below illustrates the overlap of the three datasets and their contribution to the 2025 Spine, with an overview of the underlying data shown in Table 1. Most persons on the Spine have information from all three data sources, a significant proportion contain information from two different sources with a minority containing records from just MCD or PIT. Table 1 details complete Spine composition and comparison to previous years.

Figure 1: Spine element composition for the June 2025 Spine



[1] Duplicate records occur when multiple unique IDs within the same source dataset represent the same person.

Table 1: Spine element composition across versions

	2023		2024		2025	
	Count	**%**	**Count**	**%**	**Count**	**%**
Elements containing all three datasets	18,262,060	47.90	18,686,556	47.90	19,352,400	48.54
Elements containing MCD and DOMINO only	8,004,491	21.00	8,000,536	20.51	7,987,855	20.03
Elements containing MCD and PIT only	3,496,127	9.17	3,616,320	9.27	3,486,499	8.74
Elements containing DOMINO and PIT only	41,299	0.11	40,490	0.10	41,428	0.10
Elements containing MCD only	4,043,480	10.61	4,102,643	10.52	4,123,094	10.34
Elements containing DOMINO only	108,677	0.29	108,363	0.28	110,224	0.28
Elements containing PIT only	4,167,376	10.93	4,455,770	11.42	4,769,687	11.96
Total Spine elements (Unique Persons)	**38,123,510**	**100**	**39,010,678**	**100**	**39,871,187**	**100**
MCD entities	33,870,075	39.17	34,472,080	39.03	35,078,039	38.70
DOMINO entities	26,622,831	30.79	27,038,808	30.61	27,884,400	30.76
PIT entities	25,985,647	30.05	26,818,899	30.36	27,683,755	30.54
Total Spine entities	**86,478,553**	**100**	**88,329,787**	**100**	**90,646,194**	**100**

Background

The Spine is the central linkage infrastructure that underpins PLIDA. All datasets in PLIDA are linked to the Spine independently. These linkages are then reused to produce various integrated data products. This 'supply once use many times' approach enables efficient, high-quality linkages and maintains a high standard of data security.

The Spine aims to represent all persons who were resident in Australia at any point during the given reference period. To achieve comprehensive population coverage, the Spine is built through a three-way linkage of the following administrative datasets:

- The Medicare Consumer Directory (MCD) – Services Australia
- Data Over Multiple Individual Occurrences Centrelink Administrative Dataset Centrelink Administrative Data (DOMINO) – Department of Social Services
- The Personal Income Tax Client Register (PIT) – Australian Tax Office (ATO).

Use of the Spine is tightly controlled. As outlined in "Keeping integrated data safe", the Australian Bureau of Statistics (ABS) enforces a robust framework of protections. The Spine is stored within a secure, access-controlled environment that has strict data ingress and egress controls.

The Spine datasets are collected under the authority of the *Australian Bureau of Statistics Act 1975* and the *Census and Statistics Act 1905*, with the use of the datasets being supported by *Australian Privacy Principles (APP)*. Currently, the Spine is managed under PLIDA governance frameworks. Within the ABS, the Data Integration Services Branch is responsible for the ongoing maintenance and improvement of the Spine and has authority over its management and governance.

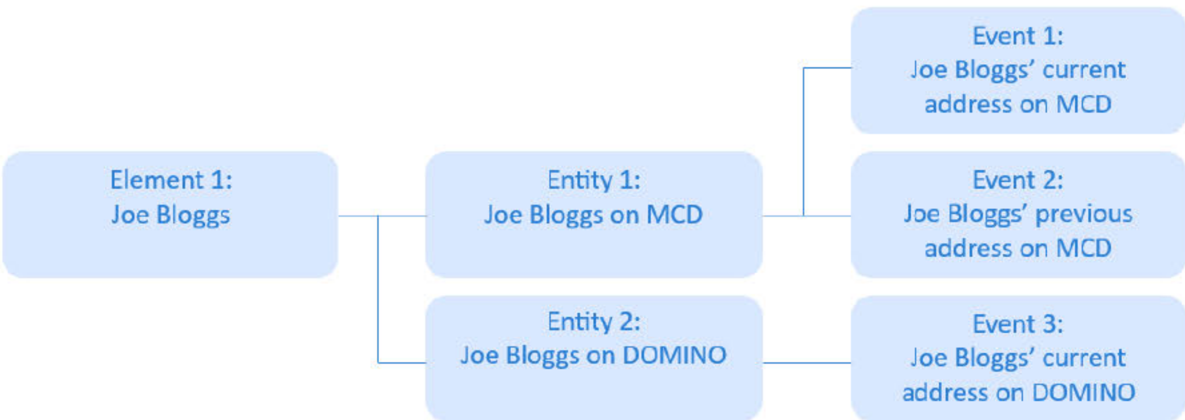
Data Model of the Person Linkage Spine

The Spine data model has been developed to establish connections between the MCD, DOMINO and PIT datasets. This structure supports data linkage where people have changed names or addresses over time.

As illustrated in Figure 2, the Spine contains three levels of data:

- Elements – each Spine element represents the statistical unit of a unique person.
- Entities – each Spine entity represents a unique person appearing in one of the Spine datasets.
- Events – each Spine event represents a transaction of a unique person appearing in one of the Spine input datasets.

Figure 2: Conceptual model of the Person Linkage Spine



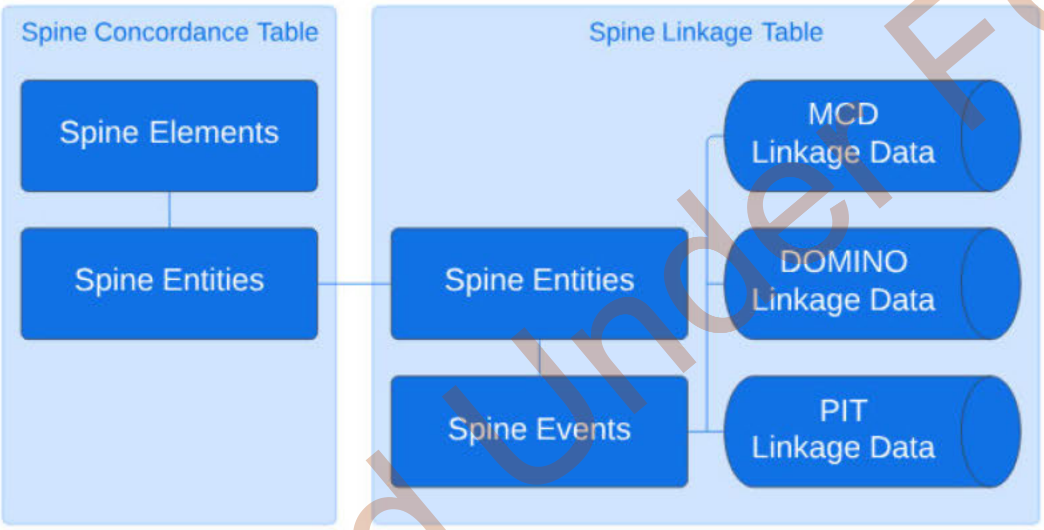
As shown in Figure 3, the Spine itself is structured across two tables:

- The Spine Concordance Table, which maps entities to elements.

- The Spine Linkage Table, which contains Spine entities, alongside the identifiers that are used for person-centred data linkage. These include name information, address information, sex, and date of birth.

The two tables are combined as needed to facilitate linkage to the Spine, for an example of the format of this combined table, see Appendix A.

Figure 3: Table structure of the Person Linkage Spine

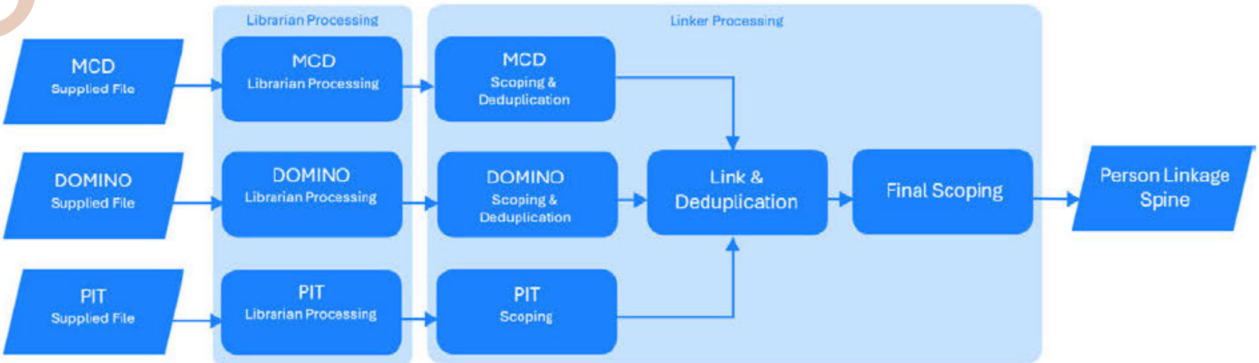


Spine build process

Building the Spine involves combining data from the MCD, DOMINO and PIT datasets into a structure that supports data linkage. The Spine build process applies appropriate scoping before resolving duplicates within, and establishing links between, the Spine datasets.

The key steps of the Spine build process are illustrated in Figure 4. Each step is discussed in more detail below.

Figure 4: Spine build process



Librarian processing

During librarian processing, the personal identifiers used for data linkage are cleaned and standardised. Librarian processing is detailed in Appendix B.

Scoping and initial Deduplication of Spine datasets

Once librarian processing is complete, the Spine datasets are descoped to remove records that were outside the target population of the Spine (the ever-resident population of Australia for the given reference period). Scoping rules are detailed in Appendix C.

Next, the MCD and DOMINO Spine datasets undergo an initial deduplication process, which involves identifying and resolving incidents of unique persons with more than one associated ID. For more details see Appendix D.

Link and Deduplication of Spine datasets

Once scoping and initial deduplication is complete, the MCD, DOMINO and PIT datasets are linked together in a three-way linkage. During the linkage process, additional duplicate records are also identified. More details are provided in Appendix D.

For the 2025 Spine build process, the probabilistic method of linking was applied for the first time via the Python package, Splink. Previously, the Spine was built via the deterministic method utilising an internal tool. Deterministic linkages follow assigned rules to form links, whereas probabilistic linkages use an evidence-based logic to link records.

Because each of the Spine datasets covers different parts of the population, it's not expected that there will be a 100 per cent linkage rate between the Spine datasets. Where a link is found, Spine entities from different datasets are grouped together under the same element (person). Where a link cannot be found, a new Spine element is formed.[2] Further details on the linkage process are detailed in Appendix E.

Final Scoping of the Spine

Following linkage of the datasets, based on the scoping rules a subset of DOMINO records is removed if they have failed to find a link with MCD or PIT records. For full details on the scope of the Spine refer again to Appendix C.

[2] See Data Model of the Person Linkage Spine for the difference between Spine elements and Spine entities.

The 2025 Person Linkage Spine

The June 2025 Spine update includes data from January 2006 through June 2025; it contains 39.87 million unique persons. Compared to the 2024 Spine, the additional year of data incorporates an additional 0.86 million unique persons on the Spine.

Both demographic (Appendix F) and structural (Appendix G) characteristics of the Spine are consistent with previous versions.

Data quality

Missingness

Missingness refers to the level of null data within a specified field. If a dataset has high levels of missingness, the affected variables will not be able to be utilised when forming links, and therefore linkage precision will be reduced.

For the 2025 Spine, missingness rates were very low to moderate across the three datasets. Of note was the slightly elevated missingness rate for date of birth and ABS's Address Register Identifier (ARID – the finest level of geo-coding) within the DOMINO dataset when compared to MCD and PIT. The missingness rates were otherwise consistent with historical data supplies.

Any records missing more than three key linkage variables were considered out of scope and removed from the input dataset. See Appendix H for results of scoping applied to each dataset.

Intertwined records

An intertwined record refers to an occurrence of a singular source ID containing information from more than one unique person. If intertwined records are left untreated, they will adversely affect the quality of analytical outputs as each of the records will not correspond uniquely and completely to a single individual. Any IDs within a single source dataset containing multiple dates of birth and significantly different names are identified as intertwined and removed from the linkage data.

Rates of intertwined records was minimal (0.02%) for the MCD collection and there were no intertwined records for PIT, which is consistent to the previous supply. There were no intertwined records found for DOMINO due to a change from the previous supply. See Appendix H for figures.

Duplicate rates

Duplicates occur when the same unique person has been allocated multiple IDs within a source dataset. Duplicates are identified through two different processes outlined in Appendix D.

The level of duplicates continues to be low for all three datasets. However higher levels of duplicates IDs were identified during the linkage stage due the change in methodology to Splink. Deduplication is implicitly included in Splink process as the links can be formed within and between datasets during a single process.

The most significant change in the number of duplicates discovered was for DOMINO. 1.39 per cent of DOMINO IDs were identified as duplicates, compared to 0.70 per cent in the last supply. This is still a relatively minor increase, and the overall duplication rate is similar to previous supplies. See Appendix H for figures.

Initial Scoping

Various scoping rules are applied for each of the input datasets to ensure that the final Person Linkage Spine represents the target population. A summary of these scoping rules is detailed in Appendix C. In applying these rules, the percentage of DSS benefit recipients and related persons for DOMINO remain consistent with last supply. PIT has a similar distribution of migrants, and persons interacting with Single Touch Payroll and Income Tax Return compared to previous years. See Appendix H for figures.

Linkage and final scoping

In the three-way linkage process, links were found for 99.60 per cent of all DOMINO records, 88.23 per cent of all MCD records, and 82.77 per cent of all PIT records. A difference in linkage rates between datasets is expected, as most persons eligible for social services are also able to receive Medicare benefits, whereas there are many persons required to pay tax but cannot receive any government benefits – for example recent migrants.

Table 2 outlines the proportion of links formed for each dataset at each clustering stage. Clustering is the process of grouping records that likely refer to the same entity, based on pairwise match probabilities. After match weights are assigned, the clustering process uses a threshold to determine which links are strong enough to form a group, helping to balance precision and recall in the final linkage output. A high proportion of populations linked with the highest quality in stage one, having a strong agreement on first name, surname, sex, and full date of birth. See Appendix E for details of the linkage process.

Following linkage of the datasets, 443,520 DOMINO records were removed where they failed to find a link with MCD or PIT records and sat outside the scoping criteria outline in Appendix C. For the remaining records (or clusters of records) from a single entity that were unable to link to another entity, and are subsequently unlinked records, these were "birthed" to the Spine. More specifically, birthed refers to a Spine Element that only contains records from one entity.

Table 2: Number and percentage of links by dataset

	Spine elements with MCD entities		Spine elements with DOMINO entities		Spine elements with PIT entities	
Clustering Stage	**Count**	**%**	**Count**	**%**	**Count**	**%**
Stage 1	30,621,857	87.62	27,245,361	99.10	22,717,213	82.18
Stage 2	204,897	0.59	136,322	0.50	163,114	0.59
Total Links	30,826,754	88.20	27,381,683	99.60	22,880,327	82.77
Unlinked records birthed to Spine	4,123,094	11.80	110,224	0.40	4,769,687	17.23
Total links + new records birthed to the Spine	34,949,848	100	27,491,907	100	27,650,014	100

Spine linkage rates by demographics

The table below shows the linkage rates for all records on the 2025 Spine, calculated after the final stage of the build. A record is considered linked when the Spine element it is a part of contains more than one dataset. An element that contains only one dataset is not considered a link, even if multiple entities are present.

Table 3: Linkage rates for all records on the June 2025 Spine, by Age and Sex

Linkage Variable	MCD linkage rate (%)	DOMINO linkage rate (%)	PIT linkage rate (%)
Sex			
Male	88.44	99.55	81.71
Female	87.96	99.65	83.89
Age			

Under 15	92.92	99.98	97.44
15-24	97.59	99.84	81.30
25-34	97.64	99.87	65.39
35-44	96.84	99.83	72.51
45-54	95.38	99.81	90.15
55-64	92.47	99.92	96.18
65-74	91.90	99.73	97.80
75-84	89.86	98.69	98.48
Over 84	50.94	98.11	98.64
Missing	0.00	0.31	0.21
State			
NSW	89.39	99.89	81.62
Vic.	88.68	99.87	83.99
Qld	88.80	99.85	83.01
SA	85.11	99.85	89.65
WA	86.36	99.82	81.34
Tas.	82.07	99.87	92.09
NT	67.79	97.95	71.66
ACT	71.21	99.67	75.50
OT (not used)	50.57	96.02	80.48
Missing	0.00	0.14	0.22
Total	88.20	99.60	82.75

As the table illustrates, linkage rates were generally stable for all populations across the three datasets. The declining rate for the MCD linkage from the age of 75 years is expected (for example self-funded retirees may no longer interact with either the ATO or the Department of Social Services). The decrease in rates between the ages of 25 to 44 for the PIT linkage is also expected, as many individuals in this age cohort are temporary migrants (for example International Students) who are generally not eligible for Medicare or social security benefits. Linkage rates across states were variable for both MCD and PIT. This is consistent with previous years and may be explained by differences in distribution of migrant populations across the states. For linkage rates across demographic information for historical Spine versions, see Appendix I .

Linkage rates were lower for new records when compared to existing entities. This is unsurprising as individuals are unlikely to engage with all government services simultaneously and therefore would not appear on all three datasets at once. This effect is significantly less pronounced for DOMINO where benefit recipients generally qualify for Medicare first, and more pronounced for PIT as there is a higher representation of migrants. See Appendix J for full breakdowns of linkage rates for old and new records by demographic information.

Spine composition

Figure 1, presented in the Executive Summary, shows the contributions of the input datasets to the June 2025 Spine. As can be seen from the overlap between these three datasets, most DOMINO records are linked to an MCD record. A further majority of these records then formed a link to PIT. Spine elements containing entities from all three datasets are mostly Australian residents who are of working age.

The remaining records are either MCD-only or PIT-only. Many of the PIT-only Spine elements are recent migrants generally not eligible for Medicare or Social Services payments, so no link is expected to be found for these records. The MCD-only Spine elements are mostly older persons who are retired and no longer interacting with the ATO and perhaps are not eligible for any Social Services benefits.

When compared with previous versions of the Spine, the composition of the various Spine elements is quite stable (within one percent). See Executive Summary for a full breakdown of element composition across Spine versions.

As outlined in the Data Model, Spine elements may include one or more identifiers from the MCD, DOMINO, or PIT datasets. These identifiers enable data to be merged directly onto PLIDA without requiring the supply of personal identifying information such as name or address.

Spine coverage

The target population of the Person Linkage Spine, June 2025 includes all persons who were resident in Australia at any point during 1 January 2006 to 30 June 2025. The June 2025 Spine comprised of 39.87 million persons over that reference period. The number of persons represented on the Spine increased by 2.21 per cent (0.86 million) from the June 2024 Spine. This increase was primarily due to babies born during the 2024-25 financial year as well as migrants interacting with the taxation system for the first time. While population statistics have not yet been released for this period, this is consistent with the demographic inflow pattern of approximately 600,000 migrants[3], and just under 300,000 births[4] annually.

There is some under coverage of the Australian population in the Spine. For example, migrants (particularly those on temporary visas) are known to be underrepresented on the Spine as they may not be eligible for inclusion in any of the core datasets. The implications of this for analysis will depend on the population of interest. Weighting methods may assist in improving the representativeness of the data in these circumstances.

Future work

The ABS is currently exploring opportunities to enhance the Spine for the 2026 Spine build. The most critical areas for enhancement concern the Spine's coverage and reference period.

As mentioned above in the Spine coverage section, one of the key populations of interest that experience under coverage on the Spine are migrants. This under coverage will be addressed directly by assessing the feasibility of incorporating datasets sourced by the Department of Home Affairs. Additional coverage may also be found in the Births or Deaths registries.

Another key population of possible under coverage are newborns that are yet to be enrolled in Medicare. There is believed to be a lag between birth and being present on MCD, which generally coincides with first immunisations. Although this potential issue pertains to under coverage, it only is so for the short lag period between birth and MCD enrolment. The effectiveness of adding births data to the Spine to address under coverage will be explored in the coming months.

Expanding the reference period of the Spine to earlier than 2006 (likely to 2000) is also a focus for Spine enhancement. The scalability of cloud infrastructure means the considerably larger datasets shouldn't hinder processing times. Ensuring that the pre-2006 data is comparable quality to post-2006 is the key consideration for incorporating these extended files into the Spine build.

The 2026 Spine is expected to be delivered in the Australian National Data Integration Infrastructure (ANDII) environment. ANDII is a purpose-built data integration infrastructure that allows for secure data processing at scale.

---

[3] Overseas Migration, 2023-24 financial year | Australian Bureau of Statistics
[4] Births, Australia, 2024 | Australian Bureau of Statistics

Appendices

Appendix A: The Spine as used for linking

To facilitate linkage, the Spine concordance and linkage tables are combined. The combined table includes status (event) information for all records (entities) from the three Spine input datasets, alongside mapping to the Spine ID (element) they belong to. This is used for all PLIDA linkages to the Spine.

SPINE_ID	AEU_ID_SPINE	status_spine	id_group	over_st	over_end	arid_hash_trunc	fn_repaired_anon	sn_repaired_anon	DOB	sex
1	1357	1	ScMedPIN	18/06/04	20/11/08	70818BD09	Joe	Bloggs	1/10/78	1
1	1357	2	ScMedPIN	21/11/08	.	896TS0810	Joe	Bloggs	1/10/78	1
1	2468	3	ScSocSec	04/02/03	.	70818BD09	Joe	Bloggs	1/10/78	1
1	1470	4	EncTFN	21/11/08	.	896TS0810	Joe	Bloggs	1/10/78	1
2	9753	5	EncTFN	27/09/05	20/11/08	40921LE93	Jane	Best	25/03/79	2
2	9753	6	EncTFN	21/11/08	08/10/09	896TS0810	Jane	Best	25/03/79	2
2	9753	7	EncTFN	09/10/09	.	896TS0810	Jane	Bloggs	25/03/79	2

71

Appendix B: Librarian processing

During librarian processing, the personal identifiers used for data linkage are cleaned and standardised:

- Name processing. Special characters and titles are removed (such as '&' or 'Dr') from name fields, and standardised names are added (for example, 'Beth', 'Eliza' and Libby' are all standardised to 'Elizabeth').
- Demographic information (sex or gender, date of birth, country of birth, etc) is standardised to fit within existing ABS classifications.
- Address information (street, suburb, postcode, etc) is geocoded to the ABS Address Register May 2025 Index[5], which provides a comprehensive list of all physical addresses in Australia. Address information is assigned an anonymised Address Register ID (ARID), as well as a corresponding Mesh block, SA1, SA2, and SA4 on the Australian Statistical Geography Standard (ASGS).[6]

[5] ABS 2025 May Address Register - Gif Boundaries 2024 - Street Range 51
[6] Australian Statistical Geography Standard (ASGS) | Australian Bureau of Statistics

Appendix C: Scoping of datasets

The Spine aims to represent all persons who were resident in Australia at any point during the given reference period, built using the MCD, DOMINO and PIT administrative datasets. Table 4 details the scope of these datasets.

Table 4 – Scope of Spine Input Datasets

Dataset	Reference period	Scope
MCD	01/01/2006 – 30/06/2025	Persons who were enrolled in the Medicare system.
DOMINO	01/01/2006 – 30/06/2025	Persons who: • Were the primary beneficiary of at least one benefit or entitlement; • Had an active concession card; or • Were related to anyone who was the primary beneficiary of at least one benefit or entitlement, or where related to anyone who had an active concession card AND obtained a link during the maintenance process[7].
PIT	01/07/2006 – 30/06/2025	Persons who: • Had any Income Tax Return, Payment Summary or Single Touch Payroll tax activity; or • Held a migration visa that had been granted from 2006 onwards.

[7] Related persons have higher rates of missingness and duplication and therefore unlinked records aren't considered to have enough quality information to confirm they are a unique person.

Appendix D: Deduplication processes

Two deduplication processes were performed for Spine data. The initial deduplication process is applied prior to linkage.

MCD data is deduplicated using a 'Record Integrity Details' dataset, which is provided to the ABS with the MCD linkage data. This file contains a list of records that have been identified by the data custodian as duplicates.

DOMINO data is deduplicated using relationship information. Records are identified as duplicates if they:

- Have the same relationship type to the same entity (i.e. daughter of John Smith); and
- Match exactly on date of birth; and
- Match strongly on first name and surname.

During the linkage process, duplicate records are clustered together should IDs from the same dataset match strongly using the following data items:

- First name;
- Surname;
- ARID or Mesh Block (MB)[8];
- Date of Birth (DOB)
- Sex;
- Birth Country;
- Country of Citizenship; and
- Indigenous Status.

Deduplication is illustrated in Table 5 below. In this example, the entities D4 and D5 could be resolved because they both match the same MCD entity on first name, surname, mesh block, date of birth, and sex.

Table 5: Example deduplication

Dataset	Deduplicated Entity ID	Original Entity ID	First Name	Surname	ARID	MB	DOB	SEX
MCD	M9	M8	JANE	BROWN	ARID 2	MB 2	19/01/1990	FEMALE
MCD	M9	M9	JANE	SMITH	ARID 3	MB 2	19/01/1990	FEMALE
DOMINO	D4	D4	JANE	BROWN	ARID 2	MB 2	19/01/1990	FEMALE
DOMINO	D4	D5	JANE	SMITH	ARID 3	MB 2	19/01/1990	FEMALE

[8] The ABS Address Register provides a comprehensive list of all physical addresses in Australia. The term ARID (Address Register ID) is the identifier assigned by the Address Register to each physical address. Mesh Blocks are the smallest geographical area defined in the Australian Statistical Geography Standard (ASGS). For more information see: Main Structure and Greater Capital City Statistical Areas | Australian Bureau of Statistics.

Appendix E: Linkage methodology

The Spine was constructed using a probabilistic linkage approach, named Splink, which estimates the likelihood that records from different sources refer to the same individual - even when identifiers are missing, inconsistent, or contain minor errors. This method applies a statistical model to assess similarities across key fields such as names, address details, dates of birth, and other demographic information.

To account for differences in data quality and structure across datasets, the linkage process is carried out in multiple stages, with match probability thresholds and comparison criteria progressively refined at each step. This allows for flexible and accurate identification of candidate record pairs.

Once comparisons have been made and match weights assigned, clustering is used to group records that are likely to refer to the same entity. Clustering is the step that transforms pairwise match probabilities into linked groups of records, based on a specified match weight threshold. This threshold determines which links are strong enough to be included in a cluster, helping to balance precision and recall.

Minimum matching requirements, based on common variables such as names, dates of birth, and addresses, are outlined in the table below. These serve as baseline conditions for establishing potential links in cluster groups. Depending on the characteristics of the datasets involved, additional comparison rules or thresholds may be introduced to further improve linkage quality.

Table 6: Minimum criteria at each clustering stage

	Name	Demographics	Sex
Stage 1	Strong agreement on First name and Surname[9]	Partial Agreement on DOB[10]	Agreement on Sex
	Strong agreement on First name and Surname[8]	Strong agreement on DOB	No agreement on Sex
Stage 2	The sum of any combination of comparisons between data items reached a predetermined threshold.		

A link formed during Stage 1 clustering (match probability > 0.99) reflects very high confidence in the match. Stage 2 indicates lower confidence however these links meet the required similarity threshold (match weight ≥ 1.5) and represent the best available match for that record. The chosen match weight threshold was validated by reviewing links formed at that level.

[9] Includes standardised name as well as swapping first name/surname.
[10] Partial DOB agreement refers to either year of birth (YYYY) or day of birth (1-366).

Appendix F: Historical demographic distribution of Spine elements

Table 7: Historical demographic distribution of Spine elements

	2023		2024		2025	
	Count	**%**	**Count**	**%**	**Count**	**%**
Sex						
Male	19,394,927	50.87	19,853,884	50.89	20,319,487	50.96
Female	18,721,189	49.11	19,147,585	49.08	19,545,821	49.02
Missing	7,394	0.02	9,209	0.02	5,879	0.01
Age group						
Under 15	4,852,405	12.73	4,850,974	12.43	4,830,334	12.11
15-24	3,668,041	9.62	3,769,464	9.66	3,879,940	9.73
25-34	5,778,383	15.16	5,810,273	14.89	5,844,131	14.66
35-44	5,583,855	14.65	5,845,146	14.98	6,085,753	15.26
45-54	4,154,463	10.90	4,225,304	10.83	4,316,536	10.83
55-64	3,762,462	9.87	3,814,720	9.78	3,854,519	9.67
65-74	3,193,863	8.38	3,266,361	8.37	3,334,490	8.36
75-84	2,373,927	6.23	2,476,527	6.35	2,577,483	6.46
Over 84	4,746,191	12.45	4,941,863	12.67	5,134,588	12.88
Missing	9,920	0.03	10,046	0.03	13,413	0.03
State						
NSW	15,166,515	39.79	15,524,419	39.80	16,077,054	40.32
Vic.	9,599,161	25.18	9,839,331	25.22	10,096,983	25.32
Qld	6,603,476	17.32	6,747,687	17.30	6,795,700	17.04
SA	2,226,972	5.84	2,256,952	5.79	2,268,009	5.69
WA	3,238,196	8.49	3,337,855	8.56	3,372,598	8.46
Tas.	567,999	1.49	571,884	1.47	559,181	1.40
NT	215,680	0.57	219,129	0.56	204,944	0.51
ACT	384,673	1.01	389,207	1.00	378,074	0.95
OT (not used)	1,550	0.00	1,555	0.00	1,435	0
Missing	119,288	0.31	122,659	0.31	117,209	0.29
Total	**38,123,510**	**100**	**39,010,678**	**100**	**39,871,187**	**100**

Appendix G: Missingness of input datasets

Table 8: Missingness rates for MCD key linking variables (%)

Data item	2023	2024	2025
First name	0.18	0.20	0.20
Surname	0.00	0.02	0.02
Sex or Gender	0.00	0.00	0.00
Geography			
ARID	4.32	4.07	3.93
State	0.01	0.00	0.00
Date of birth	0.00	0.00	0.00

Table 9: Missingness rates for DOMINO key linking variables (%)

Data item	2023	2024	2025
First name	0.08	0.08	0.07
Surname	0.02	0.03	0.02
Sex or Gender	0.01	0.01	0.00
Geography			
ARID	4.49	4.82	5.03
State	0.67	0.66	0.67
Date of birth	1.69	1.63	1.62

Table 10: Missingness rates for PIT key linking variables (%)

Data item	2023	2024	2025
First name	0.13	0.28	0.32
Surname	0.03	0.05	0.05
Sex or Gender	0.03	0.03	0.03
Geography			
ARID	1.92	1.95	1.94
State	0.22	0.24	0.22
Date of birth	0.03	0.03	0.03

Table 11: Missingness benchmarks, key linkage variables (%)

Data item	Very Low	Low	Moderate	High	Very High
First name	0.0 – 0.69	0.70 – 1.99	2.00 – 4.99	5.00 – 9.99	10.00 – 100.00
Surname	0.0 – 0.69	0.70 – 1.99	2.00 – 4.99	5.00 – 9.99	10.00 – 100.00
Sex or Gender	0.0 – 0.69	0.70 – 1.99	2.00 – 4.99	5.00 – 9.99	10.00 – 100.00
Geography					
ARID	0.0 – 2.99	3.0 – 4.99	5.00 – 9.99	10.0 – 14.99	15.00 – 100.00
State	0.0 – 0.29	0.30 – 1.99	2.00 – 4.99	5.00 – 9.99	10.00 – 100.00
Date of birth	0.0 – 0.49	0.50 – 0.99	1.00 – 1.99	2.00 – 4.99	5.00 – 100.00

Appendix H: Application of scoping and resolution of duplicates

Table 12: Scoping and deduplication of MCD dataset

	2023	2024	2025
Records on source file	34,624,079	34,478,913	35,085,113
Less records removed due to missingness	33,875,908	34,478,913	35,085,095
Less intertwined records	33,870,075	34,472,080	35,078,039
Less initial duplicates	33,815,594	34,415,398	35,028,471
Less linkage duplicates	33,806,158	34,406,055	34,949,848

Table 13: Scoping and deduplication of DOMINO dataset

	2023	2024	2025
Records on source file	32,006,419	32,482,855	32,963,055
Less records removed due to missingness	32,004,101	32,480,683	32,962,760
Less intertwined records	31,997,830	32,474,892	32,962,760
Records after scoping	27,118,391	27,520,954	28,328,430
Benefit recipient	18,184,121	18,453,346	19,417,938
Related person	8,891,700	9,067,608	8,910,492
Less initial duplicates	26,958,391	27,369,405	28,294,351
Less linkage duplicates	26,918,126	27,329,178	27,935,427
Less lone related records	26,428,900	26,847,274	27,491,907

Table 14: Scoping and deduplication of PIT dataset

			2023	2024	2025
Records on source file			34,188,464	35,155,507	36,025,712
Less records removed due to missingness			34,039,588	34,998,762	35,876,998
Records after scoping			25,985,647	26,818,899	27,683,762
ITR & PS	**STP**	**Migrants**			
Yes	Yes	Yes	2,917,256	3,369,263	3,742,153
Yes	Yes		11,960,025	12,540,614	13,007,914
Yes		Yes	2,648,598	2,603,414	2,675,936
	Yes	Yes	379,897	314,215	279,253
Yes			6,201,830	6,116,239	6,056,819
	Yes		278,417	317,442	238,064
		Yes	1,599,624	1,557,712	1,683,620
Less initial duplicates			25,974,548	26,807,507	27,683,762
Less linkage duplicates			25,966,974	26,799,243	27,650,014

ITR – Income Tax Return, PS – Payment Summary, STP – Single Touch Payroll

Appendix I: Linkage rates by demographic characteristics across Spine versions

The following three tables show the linkage rates for all records on the Person Linkage Spine. A record is considered linked when the Spine element it is a part of, contains more than one dataset. An element that contains only one dataset is not considered a link for the table.

Table 15: Linkage rates by demographic information for MCD entities (%)

	2023	2024	2025
Sex			
Male	88.37	88.37	88.44
Female	87.73	87.77	87.96
Age group			
Under 15 years	93.81	93.46	92.92
15-24	97.64	97.57	97.59
25-34	97.49	97.53	97.64
35-44	96.46	96.58	96.84
45-54	94.56	94.79	95.38
55-64	91.86	91.87	92.47
65-74	91.78	91.68	91.90
75-84	89.13	89.39	89.86
Over 84	48.87	50.11	50.94
State			
NSW	89.04	89.10	89.39
Vic.	88.49	88.50	88.68
Qld	88.90	88.90	88.80
SA	85.36	85.29	85.11
WA	86.49	86.48	86.36
Tas.	82.85	82.72	82.07
NT	67.80	67.30	67.79
ACT	73.45	73.16	71.21
Total	**88.05**	**88.08**	**88.20**

Table 16: Linkage rates by demographic information for DOMINO entities (%)

	2023	2024	2025
Sex			
Male	99.53	99.54	99.55
Female	99.65	99.65	99.65
Age group			
Under 15 years	98.98	99.98	99.98
15-24	99.73	99.73	99.84
25-34	99.83	99.81	99.87
35-44	99.91	99.92	99.83
45-54	99.90	99.91	99.81
55-64	99.88	99.90	99.92
65-74	99.52	99.61	99.73
75-84	98.38	98.51	98.69
Over 84	98.12	98.09	98.11
State			
NSW	99.90	99.91	99.89
Vic.	99.87	99.87	99.87
Qld	99.85	99.85	99.85
SA	99.86	99.86	99.85
WA	99.84	99.85	99.82
Tas.	99.89	99.89	99.87
NT	98.55	98.47	97.95
ACT	99.79	99.76	99.67
Total	**99.59**	**99.60**	**99.60**

Table 17: Linkage rates by demographic information for PIT entities (%)

	2023	2024	2025
Sex			
Male	82.91	82.36	81.71
Female	85.12	84.50	83.89
Age group			
Under 15 years	97.52	97.67	97.44
15-24	83.80	82.64	81.30
25-34	65.98	65.76	65.39
35-44	75.47	73.84	72.51
45-54	92.45	91.42	90.15
55-64	96.81	96.54	96.18
65-74	98.06	97.98	97.80
75-84	98.62	98.58	98.48
Over 84	98.73	98.71	98.64
State			
NSW	82.54	82.16	81.62
Vic.	85.30	84.65	83.99
Qld	84.34	83.72	83.01
SA	90.73	90.20	89.65
WA	83.38	82.26	81.34
Tas.	92.55	92.22	92.09
NT	73.80	72.48	71.66
ACT	77.92	76.82	75.50
Total	**83.96**	**83.37**	**82.75**

Appendix J: Linkage rates for existing and new IDs on the Spine

Table 18: Linkage rates for existing and new IDs on the Spine (%)

	MCD linkage rate			DOMINO linkage rate			PIT linkage rate		
	Existing	**New**	**Total**	**Existing**	**New**	**Total**	**Existing**	**New**	**Total**
Sex									
Male	88.54	82.62	88.44	99.59	98.47	99.55	82.77	48.06	81.71
Female	88.08	81.25	87.96	99.70	98.34	99.65	84.96	51.25	83.89
Age group									
Under 15	94.15	76.21	92.92	99.98	99.99	99.98	97.45	97.43	97.44
15-24	97.77	83.08	97.59	99.85	99.07	99.84	85.60	58.27	81.30
25-34	97.74	94.72	97.64	99.89	99.31	99.87	67.25	22.84	65.39
35-44	96.89	93.02	96.84	99.95	97.25	99.83	72.79	48.16	72.51
45-54	95.42	87.27	95.38	99.95	96.79	99.81	90.37	56.17	90.15
55-64	92.54	68.83	92.47	99.95	98.94	99.92	96.31	66.02	96.18
65-74	92.02	45.56	91.90	99.74	99.37	99.73	97.90	73.83	97.80
75-84	89.91	36.89	89.86	98.69	99.10	98.69	98.52	83.35	98.48
Over 84	50.94	40.24	50.94	98.11	98.57	98.11	98.64	96.55	98.64
State									
NSW	89.47	83.99	89.39	99.93	98.70	99.89	82.45	51.02	81.62
Vic.	88.80	82.77	88.68	99.91	98.73	99.87	85.21	48.00	83.99
Qld	88.90	82.87	88.80	99.90	98.29	99.85	84.17	50.03	83.01
SA	85.21	78.64	85.11	99.90	98.14	99.85	90.64	58.45	89.65
WA	86.56	76.95	86.36	99.88	98.12	99.82	82.86	46.00	81.34
Tas.	82.18	74.38	82.07	99.92	98.02	99.87	92.80	66.55	92.09
NT	67.65	72.57	67.79	98.32	90.26	97.95	73.31	45.20	71.66
ACT	71.47	61.79	71.21	99.79	96.64	99.67	77.38	40.34	75.50
Total	**88.31**	**81.94**	**88.20**	**99.64**	**98.41**	**99.60**	**83.82**	**49.58**	**82.75**

Australian Bureau of Statistics

www.abs.gov.au