



PLIDA DATA SUPPLY GUIDELINES

Data Integration Services Branch

June 2026



Contents

Introduction.....	Error! Bookmark not defined.
General Information about PLIDA Data Provision	3
Governance information	3
Legislative framework	3
Privacy Impact Assessments	3
Key governance requirements	4
Data requirements	4
Scope and target population	4
Separation principle	4
Technical requirements	5
Linkage data requirements	6
Analytical data requirements	7
Metadata: Data Item List	8
Files	8
Fields	8
Code Tables	9
Data validation	10
Validate against metadata	10
Validate the Person IDs	10
Validate record integrity	10
Check for unsolicited personal information	11
Reviewing of frequency counts	11
Data transfer requirements	11
Informatica	11
Other data transfer facilities	12

Introduction

The Person Level Integrated Data Asset (PLIDA formerly known as MADIP) combines information on health, education, government payments, income and taxation, employment, and population demographics (including the Census) over time.

This document provides Data Custodians with information about how to prepare and validate data so that it is in a suitable format for linkage to PLIDA. It consists of three sections:

1. General Information about PLIDA Data Provision, which outlines the underlying principles about how data should be provided to the ABS for linkage to PLIDA.
2. Example PLIDA Data Specifications, which can be used by Data Custodians to expedite the data provision process (where appropriate).
3. The PLIDA Data Provision Checklist provides step-by-step instructions for the provision of PLIDA data, to help guide Data Custodians through the data provision process.

When you supply data to ABS, you will need to fill out a separate form called a Data Integration Request (DIR) form to provide us with information about your dataset, and a Data Item List (DIL) to provide us with the metadata for your supply.

You will need to provide these documents with new datasets as well as updates to old datasets, to indicate if they have changed since previous supplies.

General Information about PLIDA Data Provision

Governance information

Legislative framework

The ABS is authorised to collect, compile, analyse, and publish statistics under the *Australian Bureau of Statistics Act 1975* (the *ABS Act*) and the *Census and Statistics Act 1905* (the *C&S Act*). This legislation:

- Establishes the ABS as an independent statutory authority;
- Requires that information collected under the *C&S Act* must not be published or disseminated in a manner that is likely to enable the identification of a particular person or organisation;
- Prescribes penalties if an ABS officer or former ABS officer divulges or communicates any information collected under the *C&S Act* (except as required by the *C&S Act*); and
- Requires that ABS staff cannot disclose data collected under the *C&S Act* to an agency, court or tribunal.

Privacy Impact Assessments

The ABS and the PLIDA Board are committed to upholding the privacy, confidentiality and security of information in the PLIDA. The ABS continually assesses PLIDA operations for potential privacy impacts and will continue to take a privacy by design approach for PLIDA.

Privacy Impact Assessments (PIAs) are an important part of the governance and accountability framework for PLIDA. PIAs help identify and manage the privacy impacts of PLIDA.

The ABS may request additional information about the data collection practices and legislative arrangements to ensure that PIA obligations are satisfied. Information about the PLIDA PIA is available on the [ABS website](#).

Key governance requirements

Governance requirements will vary depending on the nature of the dataset being acquired, and the ABS will provide specific guidance to Data Custodians during the data supply process.

The key governance requirements for PLIDA data acquisitions are:

- The ABS and the Data Custodian must agree on the contents of a data sharing agreement. All PLIDA data **must** be collected under the authority of the *C&S Act*.
- The data sharing agreement **must** be signed by authorised delegates of the ABS and the Data Custodian. The data sharing agreement must include a complete data item list (see [metadata requirements](#)).
- The ABS **must** have assessed the privacy risks associated with the data acquisition, confirming that this is within the scope of the PLIDA Privacy Impact Assessment.

In some instances, the ABS may need to undertake additional work to manage the privacy risks associated with a data acquisition. The ABS will provide specific guidance on issues as they arise.

Data requirements

Scope and target population

PLIDA is a person-centred data asset. It is structured around the ABS Person Spine, which aims to cover all people who were resident in Australia at any point from 2006 onwards (the “ever resident” population). This means that datasets provided for linkage to PLIDA:

- **Must** be person-centred data. PLIDA datasets may include transaction, business, or environmental data, provided this is structured around person-level data.
- **Should** align with the scope of the Spine. If data for the period prior to 2006 is provided for linkage, scoping data items should be provided to remove records that don’t align with the scope of the Spine.¹ This information is needed to maintain a high standard of data quality.

Separation principle

PLIDA data **must** be provided in accordance with the Separation Principle.² The Separation Principle distinguishes between two types of data:

- **Linkage Files:** these include any of the personal identifiers required for either a probabilistic link or a direct merge to the PLIDA Spine. The ABS Linkage Centre links these to the spine and then provides a de-identified concordance file to the Assembly section
- **Analytical Files:** these include all non-personally identifying variables that are required for analysis, such as occupation, income, health indicators, etc. Some variables such as sex/gender can be used for both linkage and analysis, in which case they should be present on both files. The analytical files are used to create the products that are available to researchers in DataLab.

¹ For example, data items that indicate when a person has died, or when they migrated.

² This is a requirement of the PLIDA Privacy Impact Assessment.

Linkage Files and Analytical Files are provided separately (though some data items may be used for both linkage and analytical purposes). Linkage Files and Analytical Files must be provided with a common primary key (that is unique for each individual on the data) that allows us to join the files back together.

Technical requirements

Both linkage and analytical files must satisfy the following requirements:

Requirement	Details
Column names have alphanumeric and underscores only	No special characters such as "?", "/", "%", etc.
Column names cannot start with a number	-
Column names are max length 32 characters long	-
Max file size 40GB	Applies to each individual file Partition on something logical (e.g. state, vaccine type)
No fully blank columns	All variables must contain at least some data
No duplicate rows	Remove exact duplicates before supply
Standard missing value formats	Represent missing values as blank. There are conceptual differences between a zero value and a missing value. This difference should be formatted correctly and consistently across all variables and datasets supplied. For both numeric and text variables, please denote missing variables with a blank cell. Do not enter alternate text such as NA, NULL, '?' etc. No variables should be missing for all records.
Zip files not password protected	If legally required to password protect, send the password with file/supply details to plida.data@abs.gov.au
Use Latin-1 encoding if supplying SAS files	If using another encoding, specify it clearly.
Leading-zero numeric categories stored as text	Numeric categorical variables with leading zeros (e.g. postcodes) are stored as text. If the data is supplied in CSV or other file type that doesn't have embedded metadata, indicate in the DIL they have leading zeros.
Identify primary keys in DIL	Required when multiple files are supplied. Primary keys for joining multiple files must be specified in the DIL.
All dates must be accompanied with a format	E.g. yyyy-mm-dd
Only variables listed in the DIL may appear	Any changes to data types or categorical values require a DIL resupply
Primary key must be anonymous and stable over time	No personal identifiers like name letters or DOB digits
Files must be provided in one of the following formats: - Parquet (preferred) - CSV	Use Latin-1 encoding if supplying SAS files

Requirement	Details
• .sas7bdat • .txt (if using .txt, you will need to indicate the delimiter)	

Linkage data requirements

The ABS has two main pathways to link a person-level dataset to the ABS Person Spine for integration with the PLIDA asset:

1. A direct merge using **one** of the following variables
 - a. Scrambled Medicare PIN
(note that this is not the same as Medicare Number)
 - b. Centrelink CRN
 - c. ATO Scrambled Identifier
2. Probabilistic linkage using **all** the following variables
 - a. First name
 - b. Family name
 - c. Full residential address
 - d. Sex/gender
 - e. Date of birth
 - f. Country of birth
 - g. Indigenous status³

³ We collect Indigenous status so we can report on our Closing the Gap initiatives. It is not used to perform the link to the PLIDA Spine.

If using a probabilistic linkage, the variables listed above will normally provide a sufficiently high linkage rate. However, if you have any of the following optional variables, they may be used to further improve linkage rates:

- Middle names
- Aliases or preferred names
- Previous names (with start and end dates)
- Previous sex/gender (with start and end dates)
- Previous addresses (with start and end dates)

Regardless of which pathway you choose, each row of the linkage file must be unique, referring to a different real-world person without duplicates. If there are duplicates, we will need to use the primary key (which must be unique for each real-world person) to remove them.

Analytical data requirements

Requirements for analytical data will vary depending on the dataset. Analytical data should be specified in line with the following principles:

Requirement	Details
No free text allowed	All text variables should be either a primary key, a categorical variable where the values have been outlined in the Data Item List, or a publicly available standard such as Australian Standard Classification of Languages (ASCL)
Use official standards	To make the data interoperable with other PLIDA datasets, official statistical and industry standards should be used when providing categorical data
Consequences of non-compliance	If text variables do not meet the above requirements, it can delay our ability to deliver the data. We will either require a re-supply, or we will need to remove the variables altogether. In rare cases we may be able to perform a custom treatment to clean the free text, but this can be complex and costly and is not guaranteed to succeed.
No personally identifying information	No personal names, addresses, email addresses, Medicare numbers, Tax file numbers, etc.
No business names	Replace names with an anonymised identifier or ABN. If using ABN, they must be placed in a separate file for separate de-identification before being joined to the rest of the analytical data.
No place names	Replace with anonymised identifiers or addresses. Addresses must be supplied in a separate file for separate geocoding before being joined to the rest of the analytical data.
No addresses in analytical file	As in the point above, addresses must be supplied separately for geocoding before joining to analytical data
Data of birth (DOB) handling	If DOB is supplied, ABS will remove the day component and only load month and year to DataLab as part of our privacy treatments
Split compound categorical variables into binary variables	e.g. if a single variable can take on the values "Home care", "In-patient care", and "Home care and in-patient care", split them into two variables, one variable for "Home care" and one for "In-patient care", where each takes on the values of "yes" and "no"

Requirement	Details
Numeric variables stored as numbers and not text	No thousands-separator commas and no dollar signs. E.g. "\$1,000" should be stored as 1000
All variable names should be lowercase	Variables that appear on more than one file should have the same name
CRS required	Latitude/longitude variables must include a Coordinate Reference System (CRS)
Point-level data handling	ABS can receive point-level coordinates, but only meshblock-level or higher geographies can be loaded into DataLab for privacy purposes

Metadata: Data Item List

Data supplied for linkage with PLIDA **must** be supported by a complete Data Item List (DIL), using the PLIDA Data Supply DIL Template, which will be supplied to you when you begin the integration request. This section of the document provides guidance on how to complete this template.

Files

The Files Tab helps us check that we have received the correct files, and that they have followed the Separation Principle for privacy reasons. It **must** contain a complete list of all files provided to the ABS, including the following attributes:

- File Name – the name of the file being provided (e.g. CLIENT-FILE)
- Destination – whether the file will be provided to the Data Linkage Centre (for files containing Linkage Data, addresses to be geocoded, or ABNs to be de-identified), or to the Data Integration Assembly section (for files containing Analytical Data)

An example may look like:

File Name	Destination
Student Analytical File	Assembly
Student Linkage File	Linkage Centre
School Address File	Linkage Centre

Fields

The Fields Tab helps us assess what privacy treatments we will need to apply to keep the data safe, and helps us assess scheduling and costing. It **must** contain a complete list of each field on each file, including the following attributes:

- File Name – the name of the file being provided (e.g. "CLIENT-FILE")
- Field Name – the name of the field (e.g. "GENDER")
- Field Description – a plain English description of the field (e.g. "Gender of client")
- Data Item Type – indicates whether the field is stored as a number, text, date, etc.
- Data Item Maximum Length – indicates the maximum length of text fields. Not required for non-text fields.
- Personal or Sensitive – indicates whether the data field is personal information or sensitive information, as per [OAIC definitions](#). Note that Data Custodians (not the ABS) are responsible for determining whether a field is personal or sensitive information.

If the same field appears on multiple tables, it should be listed against each file in the Fields Tab, and so should appear in multiple rows in the Fields Tab.

An example may look like:

File Name	Field Name	Field Description	Data Item Type	Length (for text fields)	Personal or Sensitive ⁴
Analytical_Person	Gender	Gender of survey respondent	Text	1	No
Analytical_Person	Income	Annual pre-tax income	Number		No
Linkage_Person	Pers_id	Unique identifier for a person	Text	50	No
Linkage_Person	F_name	First Name	Text	50	Yes
Linkage_Person	M_name	Middle Name	Text	50	Yes
Linkage_Person	S_name	Surname	Text	50	Yes
Linkage_Person	DOB	Date of Birth in dd/mm/yyyy format	Date		Yes
Linkage_Person	Gender	Sex of person	Text	1	No
Linkage_Person_Address	Pers_id	Unique identifier for a person	Text	50	No
Linkage_Person_Address	Res_Start_Date	Start Date of Residential Address in dd/mm/yyyy	Date		No
Linkage_Person_Address	Res_End_Date	End Date of Residential Address in dd/mm/yyyy	Date		No
Linkage_Person_Address	Res_Addr_line_1	Residential Address line 1	Text	300	Yes
Linkage_Person_Address	Res_Addr_line_2	Residential Address line 2	Text	300	Yes
Linkage_Person_Address	Res_Suburb	Residential Suburb	Text	300	Yes
Linkage_Person_Address	Res_State	Residential State	Text	30	Yes
Linkage_Person_Address	Res_Postcode	Residential Postcode	Text	4	Yes

Code Tables

The Code Tables Tab helps us verify the categorical and text data we have received, making sure it meets our privacy requirements and has not been truncated. It **must** contain complete code lists for each field containing categorical data, accompanied by a description of the value:

⁴ Personal or Sensitive indicates whether the data field is personal information or sensitive information, as per [OAIC definitions](#). Note that Data Custodians (not the ABS) are responsible for determining whether a field is personal or sensitive information

- Field Name – the name of the field (e.g. “GENDER”)
- Value – the code associated with a particular response category (e.g. “F”)
- Description – the description associated with a particular response category (e.g. “Woman or Female”)

An example may look like:

Field Name	Value	Value Description
Gender	F	Woman or female
Gender	M	Man or male
Gender	OT	Other, neither male nor female
Gender	X	Not stated
Gender	Z	Prefer not to say

Data validation

When the ABS receives the data, we run several validation steps to ensure the data is of sufficient quality for us to process. This section outlines some key checks that custodians can run before supplying the data, to ensure it meets our needs.

Validate against metadata

Compare the files against the metadata to ensure:

- All required files and variables have been provided
- Linkage data and analytical data have been appropriately named and separated
- File names and variable names are consistent with metadata
- Formatting and code lists/values are consistent with metadata
- Date formats are consistent across all records

Validate the Person IDs

Check that Person IDs meet the standard requirements for linkage data by:

- Confirming that each Person ID contains a single date of birth, and investigating any records that contain multiple dates of birth
- Reviewing aggregate counts of records per unique Person ID, and investigating Person IDs with a high count of records

Validate record integrity

Check that there are no glaring quality issues by:

- Removing duplicates (records that are the same on every variable)
- Removing blank columns (columns that are blank on every record)
- Removing records where the date does not match the date in the reference period laid out in the data supply agreement
- Confirming that missing records are blank, not “null” or “?”, etc. There are conceptual differences between a zero value and a missing value. This difference should be formatted correctly and consistently across all variables and datasets supplied. For both numeric and text variables, please denote missing variables with a blank cell. Do not enter alternate text such as NA, NULL, ‘?’ etc. No variables should be missing for all records.

Check for unsolicited personal information

Reduce the risk of unsolicited personal information being supplied by:

- Confirming that all variables supplied appear on the final data supply agreement
- Confirming that all text variables should be either a primary key, a categorical variable where the values have been outlined in the Data Item List, or a publicly available standard such as Australian Standard Classification of Languages (ASCL); there must be **no** free text variables in the dataset
- Confirming that following types of information have been removed from the dataset:
 - Other personal IDs that are not included in the Data Item List
 - Phone numbers
 - Email addresses
 - Bank account numbers
 - Trust numbers
 - Australian Business Numbers (ABNs) (if not included in the Data Item List)
 - Tax File Numbers (TFNs)

Reviewing of frequency counts

We run frequency counts for the following data items to assess overall population counts, missing data, as well as any coding errors:

- Sex/gender
- Year/Month of birth⁵;
- State/Territory
- Indigenous Status (if available)
- Other variables of interest

Frequency counts are run at the record and Person ID levels. This is to ensure that the data is complete and that it is coherent with relevant published outputs produced from the dataset (i.e. no missing records or additional records).

Data transfer requirements

Informatica

The ABS preference is to receive PLIDA data using the Informatica data transfer facility.

Informatica is a contemporary off-the-shelf system used to perform extract-transform-and-load operations on administrative data. Informatica uses the latest security protocols and has been configured to conform to the additional security requirements of the ABS. Informatica also offers a large file upload limit, as well as streamlined account maintenance.

Informatica is highly secure and has been built and configured in accordance with the ACSC's Information Security Manual. It has undergone an IRAP assessment, which involves an independent ASD accredited assessor reviewing the compliance of the environment against controls outline in the Information Security Manual (ISM). The environment has a strong emphasis on access controls and hardening of baseline platforms and operating systems in line with security best practices and ISM compliance requirements.

⁵ The ABS recommends using the following cohorts for age analysis: Less than 1 year old; 1 – 14 years old; 15 – 24 years old; 25 – 34 years old; 35 – 44 years old; 45 – 54 years old; 55 – 64 years old; 65 – 74 years old; 75 – 84 years old; 85 years old and over.

More detailed information about Informatica is available from the ABS on request.

Other data transfer facilities

In some circumstances, the ABS may be able to receive data from other data transfer facilities. The ABS will undertake a suitability assessment of these portals on a case-by-case basis.



www.abs.gov.au