Australian Bureau of Statistics

PLIDA DATA SUPPLY GUIDELINES

Data Integration and Services Branch

CONTENTS

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 PLIDA Data Provision Requirements, 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.

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- Requires that information collected under the *C&S Act* must not be published or disseminated in a manner that is likely to enable the identification of a particular person or organisation;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 these issues as they arise.

Data requirements

Scope and target population

PLIDA is a person-centred data asset. It is structured around the Spine, which aims to cover all people who were resident in Australia at any point from 2006 onwards. 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.

General guidance on data provision

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

- Linkage Files, which includes personal identifiers such as name, address, sex/gender, and date of birth. This information is used to enable person-centred datasets to be linked together.
- Analytical Files, which includes variables of interest for analysis, such as occupation, income or health services use. This information is used to create products that are available to analysts in the ABS DataLab.

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.

Standard requirements for linkage data

Standard requirements for linkage data are set out below. The Data Integration and Services Branch will assess linkage feasibility on a case-by-case basis where these requirements cannot be met.

[1] For example data items that indicate when a person has died, or when they migrated.
[2] This is a requirement of the MADIP Privacy Impact Assessment.

Data item	Required	Optional	Notes on provision	Data item type
IDs	- Unique person ID	- ABN - Other IDs	- Person IDs **must** be provided on both linkage and analytical files. Person IDs **must** be stable over time and be same length/case as previous supplies. - ABNs **must** be provided on linkage file only. - The ABS will provide specific guidance on the provision of other IDs as required. - Ensure that IDs are not truncated.	- Character
Name information	- Personal name - Family name	- Middle name - Aliases, preferred names, etc. - Previous names (with start and end dates)	- Names **must** be provided on linkage file only. Names **must** be provided in discrete fields. - If previous names are provided, start and end dates **must** be provided as discrete variables. - Ensure that names are not truncated.	- Character
Date of Birth	- Full date of birth	- n/a	- Format of date **must** be provided, eg dd-mmm-yyyy - Date of Birth **must** be provided on linkage file and **should** be provided on analytical file. - Ensure that unknown date of birth values are shown as blanks (not a placeholder date or text value).	- Date (DD/MM/YYYY)
Sex or gender	- Sex or Gender, coded to ABS Standard.	- Previous sex/gender (with start and end dates)	- Sex or gender **must** be provided on linkage file and **should** be provided on analytical file. - If previous sexes/genders are provided, start and end dates **must** be provided as discrete variables.	- Character
Address	- Full residential address	- Previous addresses (with start and end dates) - Other address type - Geocoded addresses	- Address **must** be provided on linkage file only. Address **should** be broken into component fields: address line 1, address line 2, suburb/locality, state, postcode, country. - If previous addresses are provided, start and end dates **must** be provided as discrete variables. - Geocoded address information **must not** be used as a substitute for full residential address. Geocoded address information should only be used for analytical purposes only.	- Character

Standard requirements for analytical data

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

- Personal Identifiers **must not** be provided on analytical data.
- The content of the data **must** be consistent with the governance requirements and metadata requirements discussed elsewhere in this document.
- Free-text fields **should not** be provided in analytical data. These fields should be dropped, or recoded to defined categories (see Code Tables below).
- Where possible, official ABS standards **should** be used when providing categorical data.
- The ABS **must** be consulted to ensure that analytical data is structured appropriately.

File formats

Files **must** be provided in one of the following formats (preferably with UTF-8 character encoding):

- .csv;
- .parquet
- .sas7bdat
- .txt

File values **must** be consistent with the chosen file format. For example, for .csv files the data in each field must not contain additional commas in data values.

Metadata requirements

Data supplied for linkage with PLIDA **must** be supported by complete data item list, using the PLIDA Data Supply DIL Template, which is available for download on the ABS website. This section of the document provides more guidance about how to complete this template.

Files

The Files Tab **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), the Data Integration Assembly section (for files containing Analytical Data), or both.

Fields

The Fields Tab **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 filed (e.g "Gender of client")
- Purpose – whether the field will be used for linkage purposes, analytical purposes, or both.
- Data Item Type – indicates whether the field is stored as a number, character, date, etc.
- Data Item Maximum Length – indicates the maximum length of character fields. Not required for other types of field.
- 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 included in also appear in multiple rows in the Fields Tab.

Code Tables

The Code Tables Tab **must** contain complete code lists for each field containing categorical data, including the following attributes:

- Field Name 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")

Changes

Changes to data specifications between supplies **should** be marked in the changes tab.

Data validation requirements

Data **must** be validated before being provided to the ABS. This section outlines some key checks to undertake to ensure that the data is of sufficient quality.

Validation against metadata

Data **must** be compared against metadata to ensure:

- All required files and variables have been provided;
- Linkage data and analytical data has been appropriately named and separated;
- File names and variable names are consistent with metadata; and
- Formatting and code lists/values are consistent with metadata.

Validating file formats

Data **must** be checked to ensure that delimiters are consistent with the chosen file format:

- File delimiters do not appear in data values (for example, commas in the values of a .csv file);
- Unpaired quotes do not appear in data values; and
- Date formats are consistent across all records.

Validating the integrity of Person IDs

Data **must** be checked to ensure 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; and
- Reviewing aggregates counts of records per unique Person ID, and investigating Person IDs with a high count of records.

Validating record integrity

Data **must** be checked for duplicates (records that match on every variable). Rows that are complete duplicates of another row **must not** be provided.

Checking for Unsolicited Personal Information

Data **must** be checked to reduce the risk of Unsolicited Personal Information being supplied by:

- Confirming that all variables supplied appear on the final data supply agreement;
- Free text fields have been kept to an absolute minimum; and
- Confirming that following types of information have cleaned from free text fields.
    - Other personal IDs that are not included in the Data Item List
    - Phone numbers and fax numbers
    - Email addresses
    - Bank account numbers
    - Trust numbers
    - Australian Business Numbers (ABNs) (if not included in the Data Item List)
    - Tax File Numbers (TFNs)
    - Indicators that a person is subject to a court order
    - Indicators that a person in a prisoner

Reviewing of frequency counts

Frequency counts for the following data items **should** be reviewed to assess overall population counts, missing data, as well as any coding errors:

- Sex/gender;
- Age[3];
- State/Territory;
- Indigenous Status (if available); and
- Other variables of interest.

Frequency counts **should** be reviewed 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).

[3] 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.

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.

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.

EXAMPLE PLIDA DATA SPECIFICATIONS

The tables below set out some examples for the provision of PLIDA data. Note that these examples are not exhaustive or complete. They have been provided as a point of reference for Data Custodians to better understand data provision requirements. Analytical data has not been included. Data specifications will vary, depending on the particularities of individual datasets.

Table 1 - Fields

File name	Field Names	Field Description	Purpose	Data item Type	Data item maximum length	Personal or Sensitive
Linkage_Person	Pers_id	Unique identifier for a person	Primary Key	Character	Data item lengths will vary according to source system design. Data item lengths must be specified in Data Item List.	Data Custodians are responsible for determining whether the information they disclose is personal or sensitive information. Refer to the OAIC guidelines for more information.
Linkage_Person	F_name	First Name	Linkage	Character		
Linkage_Person	M_name	Middle Name	Linkage	Character		
Linkage Person	S_name	Surname	Linkage	Character		
Linkage_Person	DOB	Date of Birth in dd/mm/yyyy format	Linkage and Analytical	Character		
Linkage_Person	Sex	Sex of person	Linkage and Analytical	Character		
Linkage_Person_Address	Pers_id	Unique identifier for a person	Primary Key	Character		
Linkage_Person_Address	Res_Start_Date	Start Date of Residential Address in dd/mm/yyyy	Linkage	Character		
Linkage_Person_Address	Res_End_Date	End Date of Residential Address in dd/mm/yyyy	Linkage	Character		
Linkage_Person_Address	Res_Addr_line_1	Residential Address line 1	Linkage	Character		
Linkage_Person_Address	Res_Addr_line_2	Residential Address line 2	Linkage	Character		
Linkage_Person_Address	Res_Suburb	Residential Suburb	Linkage	Character		
Linkage_Person_Address	Res_State	Residential State	Linkage	Character		
Linkage_Person_Address	Res_Postcode	Residential Postcode	Linkage	Character		
Linkage_Person_Address	Res_Country	Residential Country	Linkage	Character		

PLIDA DATA PROVISION CHECKLIST

Step	Complete?
1. Data requirements met	□
2. Metadata finalised	□
3. Data Sharing Agreement signed by ABS delegate	□
4. Data Sharing Agreement signed by Data Custodian delegate	□
5. Data prepared in an appropriate file format	□
6. Validation (can be completed in any order)	□
Validation against metadata complete	□
File format validation complete	□
Person ID validation complete	□
Unsolicited Personal Information checks complete	□
Frequency counts reviewed	□
7. Data transfer facility set up	□
8. Data transfer complete	□