Australian Bureau of Statistics

CAPACITY TO CONTRIBUTE: LINKAGE IMPROVEMENTS

Direct measure of income refinement working group paper

November 2020

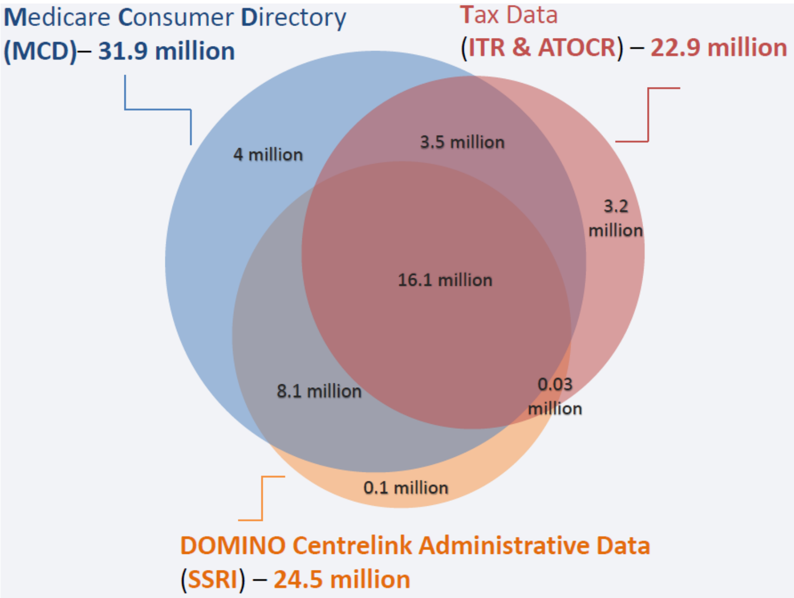
DATA LINKAGE FOR CAPACITY TO CONTRIBUTE

Data that is linked for CTC

1. The direct measure of income used for Capacity to Contribute (CtC) relies on an annual linkage of the CtC Address Collection to MADIP (Multi-Agency Data Integration Project). MADIP is an integrated data asset combining information on health, education, government payments, income and taxation, employment and population demographics over time. It provides person-centred data to support policy analysis and research.

2. Data is linked to MADIP via its person linkage spine, which is comprised of administrative data from Taxation, Medicare and Social Security to cover the majority of people in the Australian population (Figure 1). The ABS is trusted as the accredited integrating authority for MADIP, and updates the person linkage spine on an annual basis to maintain and improve its coverage of the Australian population and ensure key linkage information is kept up to date.

Figure 1: MADIP Spine data sources



3 For CtC, anonymised administrative data on income from the Australian Taxation Office and the Department of Social Services is then integrated with the CtC Address Collection via the Spine in order to derive a person-level direct measure of income for parents and guardians.

Data linkage method

4. The ABS links the CtC Address Collection to MADIP using a deterministic linking method. The variables used for linking CtC are anonymised name and address. Age (or date of birth) is also a common variable used in other MADIP linkage projects, and generally improve linkage rates and quality, however is not included in the CtC Address Collection.

5. This linkage process is imbedded within a productionised sequence for linkage, whereby address strings are coded to a location based on the ABS Address Register, and names are

standardised to account for known issues with discrepancies across administrative data (e.g. ‘Chris’ may appear in one data source, yet represent ‘Christopher’ or ‘Christian’ across other administrative data sources) then anonymised so it cannot be recognised by the linker.

6. Deterministic linkage involves matching records on each dataset that have the same and unique combination of linking variables. The search criteria are gradually broadened to identify more matches and the final parameters are chosen to maximise both linkage rate and quality. For CtC, link quality is defined as:

- quality 1 links predominantly match on anonymised parent name and address location or meshblock;
- quality 2 links match on anonymised parent name and a higher level of geography (i.e. SA1);
- quality 3 links are made at a broader level of geography. As this introduces uncertainty in the accuracy of the link, quality 3 links are not used in the direct measure of income.

7. Throughout the linkage process, the quality of the work has progressed through a series of Quality Gates. Quality gates are check points placed throughout the statistical production process to support the identification and treatment of statistical quality risks. These Gates are outlined in more detail in section 4 of the *Data Quality Framework for the Australian Government’s Direct Measure of Income for Capacity to Contribute*.

Data linkage results

8. Linkage rates underpinning the direct measure of income are high, and results for the 2020 CTC linkage have significantly improved compared to 2018 and 2019 (Table 1). This largely reflects improvements that have been made to the coverage and quality of the spine over recent years, and closer alignment of the spine time period with the CtC Address Collection period.

Table 1: CTC Linkage rates for 2018, 2019 and 2020

Address Collection data year	Linkage rate to MADIP Spine
2018 Address Collection	Quality 1 = 80.7% Quality 1 & 2 = 85.7%
2019 Address Collection	Quality 1 = 77.4% Quality 1 & 2 = 83.2%
2020 Address Collection	Quality 1 = 85.7% Quality 1 & 2 = 90.8%

Quality 1 links predominantly match on anonymised parent name and Address location or Mesh Block.

Quality 2 links match on anonymised parent name and a higher level of geography (e.g. SA1).

9. Overall, the majority of schools had very high linkage rates in 2020, with 68.5% of schools achieving a linkage rate above 90% and only 1.1% of schools with a linkage rate of 70% or below (Table 2). The linkage rate is comparable across all states and territories, with just the Northern Territory with a slightly lower linkage rate (Table 3).

Table 2: School linkage rates 2020

Linkage %	Number of Schools	% of Schools
<= 70%	28	1.1
71 - 90%	805	30.4
> 90%	1,815	68.5
Total	**2,648**	**100**

Table 3: Linkage rates by State 2020

	Linkage rate (%)
NSW	91.0
Vic.	90.6
Qld.	91.0
SA	91.3
WA	91.4
Tas.	92.2
NT	84.6
ACT	89.5
Total	**90.8**

10. Linking rates between the CtC Address Collection and MADIP are not expected to be 100%, as a match may not be possible for the following reasons:

- a small number of people may not be represented in the MADIP person linkage spine;
- there may be differences in how a name is recorded on two different datasets which are not resolved by standardisation;
- a person may have moved and may have a different address on each dataset;
- linkage information may be missing or invalid for a small number of people;
- in the case of non-unique matches, where two people with the same name live in the same geographic area, ABS attempts to find the true match using information available such as age. However in some cases it may not be possible to identify the true link.

PLANNED LINKAGE IMPROVEMENTS FOR CAPACITY TO CONTRIBUTE

11. While overall linkage rates are generally high, and the majority of schools have quality linkage results, some schools do have lower linkage rates. Investment in improving linkage methods for some sub-populations may result in further quality improvement for some school communities.

12. As part of the suite of DMI refinement work program, the ABS is undertaking the following work with an aim to further understand the reasons why unlinked records did not match to the MADIP spine and improve linkage quality in future CtC cycles:

1. investigate linkage outcomes and the characteristics of Address Collection records which did not link in order to inform potential solutions to further improve linkage rates;
2. implement an automated non-standard geocoder for addresses in Aboriginal and Torres Strait Islander Community localities;

3. review the names index used to standardise and match given and surnames, to account for recent new names and cultural diversity changes in Australia.

13. The proposed work has been scheduled for the 2020-21 financial year and it is anticipated that improvements will be implemented for the annual linkage process for the 2021 CtC cycle, where possible.

1. Investigate linkage outcomes to inform potential solutions

14. The linkage of the 2020 Address Collection to the MADIP Person Spine achieved a linkage rate of 90.8% (for links of an acceptable quality). Given the data available for linking (name, address) this is a very good linkage rate. Further improvements to the linkage rate require reviewing the linkages achieved with reference to both the Address Collection and the MADIP Person Spine, as well as the unlinked records. This will support better understanding of the reasons why unlinked records did not match to the MADIP spine, the characteristics of schools with lower linkage rates and inform potential solutions to further improve linkage outcomes.

15. Key milestones for the work include:

- analysis of linkage: December 2020;
- identification of potential solutions, data sources and linking variables: February 2021;
- testing and reporting: From March 2021 (depending on data access and supply).

2. Non-standard Geocoder

16. For CtC 2020, the ABS employed a manual process whereby residential addresses that failed to match to a location on the ABS Address Register were manually mapped to a separate list of Aboriginal and Torres Strait Islander Community localities. This process increased the number of records with valid addresses for linking purposes and supported improvements in the linkage rate.

17. ABS plans to optimise and productionise this activity so that it can be incorporated into the annual linkage work for CtC. The implementation of a non-standard geocoder proposes to supplement the current automated coding of addresses and replace the need for manual mapping.

18. Key milestones for the work include:

- framework for quality metrics: December 2020;
- testing: February 2021;
- implementation: March 2021.

3. Review names index

19. An individual's name is a key variable for data linkage in CtC, especially where name and address are the only available linkage variables. It is therefore very important to have both high-quality name data from the CtC Address Collection, as well as robust processes to find valid matches between records across different data sources. As the cultural diversity of Australian society changes and evolves over time, it is important to be able to incorporate associated changes in naming conventions and patterns in the linkage process.

20. The ABS will investigate Address Collection records that do not successfully link to MADIP to investigate whether there is a relationship between linkage and the existing name indexes used,

review the available literature regarding cultural or linguistic diversity and may engage cultural consultants to investigate gaps in the name Index currently used for linkage.

21. Key milestones for the work include:

- literature review and initiate consultation: December 2020;
- analysis and review: February 2021;
- implementation: March 2021.