AUSTRALIA
Australian
Bureau of
Statistics

CAPACITY TO CONTRIBUTE: DATA LINKAGE IMPROVEMENTS

Direct Measure of Income refinement working group paper

April 2021

EXECUTIVE SUMMARY

Objectives

For Capacity to Contribute (CtC), linkage refers to the linking of parent information from the Address Collection to the MADIP Person Spine. While overall linkage rates for the Direct Measure of Income (DMI) used in CtC are generally high, and the majority of schools have quality linkage results, some schools do have lower linkage rates.

As part of the suite of the DMI refinement work program, the ABS undertook the following work to support improvements to linkage in future CtC cycles:

1. Investigated 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. Implemented improved addresses coding for Aboriginal and Torres Strait Islander Community localities;
3. Updated the index used to standardise and match given names and surnames, to account for new names and cultural diversity changes in Australia.

Key Findings

- The linkage of the 2020 Address Collection to the MADIP Person Spine achieved a linkage rate of 90.8%. Given the data available for linking (name, address) this is a very good linkage rate.
- Improvements were seen in the 2020 linkage over linkage rates in 2018 and 2019, which can be largely attributed to the MADIP Spine refresh and the inclusion of more up-to-date address information prior to linkage.
- The majority of Address Collection records that failed to link to the MADIP spine did so because insufficient information was available to distinguish between multiple possible matches in order to create a unique link.
- Apart from considering the inclusion of additional parent information in the CtC Address Collection to assist with distinguishing between multiple possible matches on the MADIP spine, the recommendations presented in this report may only result in marginal improvements to the linkage rate or the quality of successful links.
- Improved address coding for Aboriginal and Torres Strait Islander Community localities enabled an additional 2,700 parent records to be coded to a geographic location, which supported improvements to linkage rates for a subset of schools. ABS has optimised and productionised this process so that it is incorporated into the annual linkage for CtC.
- The new names index developed for CtC results in less erroneous standardisations, has better representation of names with non-European origins and is more efficient to maintain. Overall it can be expected to result in more higher quality links compared with the current index.

Recommendations

Recommendation 1: *Consider the merits of including additional linking information in the CtC Address Collection to increase the potential for achieving more unique matches across datasets. ABS notes that these improvements need to be considered in terms of impact on the existing privacy framework for the Address Collection and CtC.*

Recommendation 2: *Trial the inclusion of spine records over the age of 79 years, previously descoped from linkage, in the 2021 linkage cycle and assess the impact on linkage.*

Recommendation 3: *Use datasets outside the suite used for updating MADIP spine information to provide more up-to-date address information to assist with matching addresses on the Address Collection.*

Recommendation 4: *Implement the new names index in the linkage of the 2021 CtC Address Collection to the MADIP Spine, whilst performing a parallel test linkage using the current index, enable assessment of improvements.*

1. INTRODUCTION

The Direct Measure of Income (DMI) used for Capacity to Contribute (CtC) relies on an annual linkage of the Student Residential Address and Other Information Collection (Address Collection) to the Multi Agency Data Integration Project (MADIP). MADIP is an integrated data asset combining information on health, education, government payments, income and taxation, employment and population demographics over time. It provides data to support policy analysis and research.

A linkage rate of 90.8% was achieved for CtC in 2020. That is, 90.8% of parents in the 2020 Address Collection were able to be linked to the MADIP spine. This represents a high-quality linkage outcome that supports a fit-for-purpose dataset for calculating DMI scores. While overall linkage rates are generally high, and the majority of schools have quality linkage results, some schools do have lower linkage rates. It is important to note that, as part of the CtC policy framework, there is a robust quality assurance process in place for evaluating the fitness-for-purpose of each school's DMI score[1].

As part of the suite of the DMI refinement work program, the ABS undertook the following work to support improvements to linkage in future CtC cycles:

1. Investigated 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. Implemented improved addresses coding for Aboriginal and Torres Strait Islander Community localities;
3. Updated the index used to standardise and match given names and surnames, to account for new names and cultural diversity changes in Australia.

This report reports on findings from this work. It outlines analysis undertaken and presents recommendations to the Department of Education, Skills and Employment in considering further improvements to the CtC linkage process.

Note given the high linkage rates for CtC, work to further increase the number of Address Collection records that link to the MADIP spine may only result in marginal improvements, and these should be considered in balance with the cost and effort of implementation. Adjusting linkage methodology or focusing on discrete populations also has the potential to introduce bias into the linkage process and impact analytical outcomes when calculating the DMI score for school communities. Care can be taken to minimise this risk, however proposed improvements should be considered with potential biases in mind, as well as being assessed for fitness-for-purpose.

[1] Refer Quality Gate 3 outlined in the *Data Quality Framework for the Australian Government's Direct Measure of Income for Capacity to Contribute*.

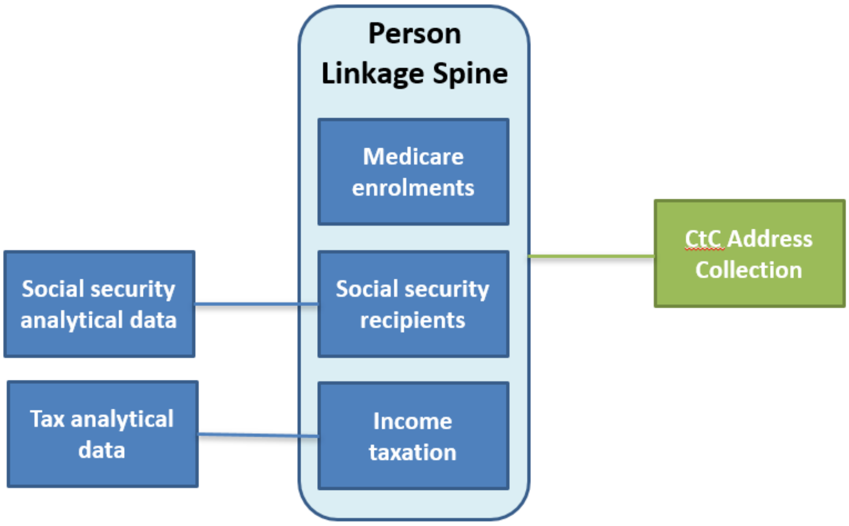
2. DATA LINKAGE FOR CTC

2.1 Data that is linked for CTC

The CtC Address Collection is linked to MADIP via the MADIP Person Linkage Spine (the Spine), which is comprised of administrative data from taxation, Medicare and Social Security datasets and aims to cover all people who were resident in Australia at any point since 2006. The ABS is the accredited integrating authority for MADIP, and updates the spine on an annual basis to maintain coverage of the Australian population and ensure key linkage information is up-to-date.

Data 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 for CtC (Figure 1). Note that while the Medicare enrolments dataset forms part of the Spine, no Medicare or health data contributes to CtC.

Figure 1: Capacity to Contribute data sources



2.2 Data linkage method

The ABS uses a deterministic, or 'exact matching' method to link the CtC Address Collection to MADIP. 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 improves linkage rates and quality, however it is not available in the CtC Address Collection.

This linkage process is part of a productionised process for integrating datasets, which includes:

- coding address information to a location based on the ABS Address Register;
- standardising names to account for possible variations across administrative data (e.g. 'Jon', 'Jonathon' and 'Johnny' are standardised across administrative data sources); and
- anonymising all information so real-world information cannot be recognised by the staff performing data linkage.

Deterministic linkage involves matching records on each dataset that have the same 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 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. Statistical Area Level 1);
- 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.

Quality gates are used throughout the linkage process to identify and manage statistical quality risks, and are outlined in section 4 of the *Data Quality Framework for the Australian Government's Direct Measure of Income for Capacity to Contribute*.

2.3 Linkage Rate vs Coverage Rate

The linkage rates noted in this report refer to the number of unique parent records in the Address Collection that formed an acceptable quality link to the MADIP spine as a proportion of the total number of unique parent records in the Address Collection.

The linkage rate is different from the coverage rate, which is sometimes used in understanding the quality of CtC scores. The coverage rate refers to the proportion of students with parent/guardian income data, as a proportion of the total number of students in the Address Collection.

2.4 Data linkage results

Linkage rates underpinning the DMI are high, and results for the 2020 CtC linkage have significantly improved compared with the 2018 and 2019 iterations (Table 1). This largely reflects improvements that have been made to the coverage and quality of the Spine over recent years, the closer alignment of the Spine time period with the CtC Address Collection period and the inclusion of more up-to-date address information for spine records prior to the 2020 CtC cycle.

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).

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 most states and territories, with the Northern Territory having 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

State / Territory	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**

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 at the time of linkage;
- there may be differences in how a name is recorded on two different datasets which are not resolved by standardisation;
- a person who recently changed address may have a different address on each dataset;
- linkage information may be missing or invalid for a small number of people;
- there may be insufficient information available on the datasets to distinguish a unique match between a number of people with similar characteristics e.g. two people with the same name residing in the same geographic area.

3. LINKAGE IMPROVEMENT INVESTIGATIONS

The objectives of this investigation were to:

- Understand the drivers behind the current linkage rates and the likely reasons for Address Collection records not linking acceptably to the MADIP Spine.
- Recommend potential improvements that are expected to increase the linkage rate and/or quality of existing links.

The investigations into potential data linkage improvements focused on four areas:

1. Matches that did not form accepted links for CtC
2. Review of the linkage strategy
3. Address Collection records that did not form any matches to the Spine
4. Analysis of linkage rates by geography

3.1 Matches that did not form accepted links for CtC

To produce high quality linked analytical datasets a matched pair of records must be unique in order to be accepted as a link. That is, to make a successful link a single Address Collection record needs to match with only one spine record and vice versa.

During the linkage process an ‘agreements file’ is generated, which contains all potential links found between persons on the CtC Address Collection and the Spine, including matches deemed acceptable quality links for final analysis, as well as those deemed to be of poor quality that were not accepted as final links. The spine agreements file can be used to understand the circumstances in which records do not form a unique one-to-one match between data sources.

In the 2020 CtC cycle, 152,449 (9.2%) Address Collection records did not link, or did not link with an acceptable quality, to the Spine. Of these records, 55,298 (3.3%) had unique matches to the Spine, but were deemed low quality links, 76,793 (4.6%) formed non-unique matches across datasets, and 20,358 (1.2%) could not establish a match at all using the available linkage variables (Table 4).

Table 4: Breakdown of 2020 Address Collection population when linked to MADIP spine

	Number of Address Collection records	Percentage of Address Collection records
Records linked to MADIP spine with acceptable quality (quality 1 and 2) – accepted linkage rate	1,503,925	90.8%
Records that had unique links to MADIP spine that were deemed low quality (quality 3) – excluded from accepted linkage rate	55,298	3.3%
Records with potential links that formed non-unique matches between CtC records and MADIP spine records	76,793	4.6%
Records where no link could be established	20,358	1.2%
Total	**1,656,374**	**100%**

Analysis of the agreements file found that of the 76,793 records that formed non-unique matches across datasets, 80% had potential links with more than one MADIP spine record. This indicates that the linkage process was unable to determine which unique MADIP spine record should link to the corresponding Address Collection record based on the available linkage data. The remaining 20% of non-unique matched records identified the reverse, whereby a singular MADIP spine record showed potential matches to more than one Address Collection record, and as a one-to-one match had not been achieved, the records were not accepted as a successful link.

Analysis on the prevalence of non-unique matches at the household level versus higher levels of geography showed that around 50% matched at the household level. This high proportion of possible links to more than one individual within the same household suggests a number of possible scenarios. One possibility is that name data for these records is of low quality and that while the matching of household address was successful, there is not enough information in the name data to distinguish between different household members. Another possible scenario is that the linkage process did not have enough information available to distinguish between two or more individuals.

Observations of these non-unique links showed that in the case of one Address Collection record matching to multiple possible spine records, almost all spine records reported a different age, and approximately half had differing sex information. The addition of more linking variables, for example, sex, age, year or birth or age group, in the Address Collection would likely lead to a reduction in non-unique matches and a higher linkage rate.

Recommendation 1: Consider the merits of including additional linking information in the CtC Address Collection to increase the potential for achieving more unique matches across datasets. ABS notes that these improvements need to be considered in terms of impact on the existing privacy framework for the Address Collection and CtC.

The ABS can evaluate the extent of likely improvement in linkage rates, and the quality of existing links, to inform this recommendation further.

3.2 Review of the linkage strategy

Each ABS data integration project uses a tailored linkage strategy, designed to maximise the quality and number of links for data sources to the Spine. The strategy focuses on the linkage variables available, such as name and address on the Address Collection, and runs linkage attempts in batches called ‘passes’.

The CtC linkage strategy removes persons on the Spine under the age of 15 years, in order to avoid falsely matching children on the Spine to parent records on the Address Collection where child and parent names may be the same or similar. The strategy also removes persons over the age of 79 years on the Spine who are expected to be outside the scope of parent records in CtC. The removal of records by age helps to reduce the number of records involved in the linkage process and facilitates more efficient and timely linkage.

As the scope of CtC includes both parents and guardians, it may be possible that older persons previously removed from the Spine are parents or guardians of children and should be included in the linkage process. On the other hand, increasing the scope of the Spine to include older people may reduce the linkage rate, as it could lead to more non-unique links.

It is recommended the ABS trial the impact of scoping of older persons in the delivery of the 2021 CtC linkage. Specifically, it is suggested that the Spine continue to exclude those aged over 79 years for the initial 2021 Address Collection linkage. Following this, a secondary linkage of unlinked Address Collections records could be attempted to a spine dataset containing persons aged 79 years and over.

Recommendation 2: Trial the inclusion of spine records over the age of 79 years, previously descoped from linkage, in the 2021 linkage cycle and assess the impact on linkage.

3.3 Address Collection records that did not form any matches to the Spine

For records that did not form any potential links to the Spine during the 2020 linkage cycle, there are two variables available for investigation: name and address of the Address Collection records.

Understanding the name information for records that did not link to the Spine is being investigated as part of the name standardisation work, outlined in the introduction of this report, with outcomes expected to be implemented in time for the linkage of the 2021 Address Collection.

For address information, the 2020 linkage cycle of CtC observed that only 1.3% of Address Collection records did not have sufficient information to assign a precise dwelling location for their given address. This indicates high quality address information on the Address Collection and has contributed to the overall high linkage rate. Addresses that do not geocode to the ABS Address Register but concord to a known Aboriginal and Torres Strait Islander Community locality have been incorporated into the non-standard geocoder work outlined in the introduction of this report, and outputs from this work are expected prior to commencement of the 2021 CtC linkage cycle.

The timeliness of the address information on the Spine was also considered as part of this work. Spine updates are now undertaken annually by the ABS, and improvements to the 2020 CtC linkage rate can be largely attributed to the MADIP spine refresh and its inclusion of more up-to-date address information. However, some inherent issues remain with address information in administrative datasets, for example, the time lag between when people change address, notify government service providers, and this information being incorporated into the annual spine update.

The ABS has identified additional datasets that may provide more up-to-date address information for some spine records. These include state and territory driver’s licence data, the Australian Electoral Roll data and a range of state/territory specific data sources.

Recommendation 3: Use additional datasets to update MADIP spine information and provide more up-to-date address information to assist with matching addresses on the Address Collection.

3.4 Analysis of linkage rates by geography

For the 2020 CtC cycle, linkage rates in different geographical areas were analysed to investigate geographic patterns of linkage. Linkage rates across the states/territories and across the remoteness categories were found to be high, in general, with slightly lower rates for very remote areas in NSW, NT and WA (Table 5). In almost all states, the inner regional and outer regional linkage rates are higher than those in the major city areas. The lower linkage rates in the NT compared with other jurisdictions is seen in other data integration projects, and reflects the limitations of administrative data, particularly for address information.

Table 5: Linkage rates (%) by state/territory and remoteness classification 2020 (a)

	Major Cities	Inner Regional	Outer Regional	Remote	Very Remote
NSW	91.3	92.2	92.3	89.0	85.3
Vic.	90.3	93.2	92.9	–	–
Qld	91.3	91.4	92.6	91.2	91.0
SA	91.2	93.2	92.7	94.8	92.4
WA	91.8	92.6	90.7	89.0	85.6
Tas.	–	92.0	93.8	95.0	97.7
NT	–	–	87.4	83.5	85.0
ACT	90.3	83.9	–	–	–

a. excludes records with missing/non-geocoded address information.

When interpreting these findings, it is important to consider that Address Collection records that fall into the very remote geographic category make up approximately 0.3% of the CtC population and it can be seen that small differences reported for a small population base result in larger percentage deviations from those observed in more populous categories. It should also be noted that the production of the DMI score is subject to a robust quality assurance process, which considers the fitness-for-purpose of each school's score and uses a range of quality indicators. This quality assurance process is outlined in section 4 of the *Data Quality Framework for the Australian Government's Direct Measure of Income for Capacity to Contribute*.

The geographical analysis used heatmaps of Greater City areas as a visual tool for identifying patterns of linkage. Findings indicate that, aside from the urban-rural split, no clear, systematic, geographic pattern is apparent for the linkage rates.

4. IMPROVED ADDRESSES CODING FOR ABORIGINAL AND TORRES STRAIT ISLANDER COMMUNITY LOCALITIES

For the linkage of the CtC 2020 Address Collection, the ABS mapped residential addresses that failed to match to a location on the ABS Address Register to a separate list of Aboriginal and Torres Strait Islander Community localities (Community Places Extract). This was driven by small number of schools with very low geocoding rates using the ABS's standard Address Register.

The geographical information in the Community Places Extract includes but is not limited to:

- Community Primary Name;
- Community Secondary Name;
- Address (used to extract state);
- Meshblock;
- Latitude;
- Longitude.

This process increased the number of records with valid addresses for linking purposes and supported improvements in the linkage rate in 2020. For the 2020 Address Collection, there was a single school with very low geocoding rates, for which the separate list was able to code an additional 24 records. Across the whole Address Collection, over 2,700 records were coded with the Community Places Extract.

While this matching only affected a small overall proportion of the Address Collection, it does cover a large part of the population for certain schools. ABS has optimised and productionised this process so that it is incorporated into the annual linkage procedures for CtC.

5. UPDATING THE INDEX USED TO STANDARDISE AND MATCH GIVEN NAMES AND SURNAMES

An individual’s name is a key variable for data linkage in CtC, as name and address are the only available linkage variables. It is therefore 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 the two different data sources. As the cultural diversity of Australian society changes and evolves over time and new names become more common among the Australian population, it is important to be able to incorporate associated changes in naming conventions and patterns in the linkage process.

The linkage process uses a names index to map given and surnames names to a standard form. For example, the names Jon, Johnathan, Johnny and Jonno might all map to the name John. This is done as people may use different variations of their name in interactions with different administrative data custodians, and closely related names (the type of names someone might switch between depending on context) can vary in terms of text similarity metrics. Name variations are reconciled by adding John, for example, as another option for linkage when someone used Jonno in one dataset and Johnathan in another.

The ABS has developed a new names index using data from the *Behind the Name* website. Data from this website contributed heavily to the current index, and the number of names standardisations available on the website has increased significantly since the current index was built. This new index contains more names with standardisations and incorporates recent cultural diversity changes in Australia. the new names index can also quickly accommodate additional sources of names standardisation information if they are identified.

Evaluation of the new index indicates it outperforms the current index in several areas: it results in less erroneous standardisations, has better representation of names with non-European origins and is more efficient to maintain. Overall it can be expected to result in more accurate links compared with the current index. Appendix 1 provides more detail on the outcomes of the evaluation.

It is recommended that for the linkage of the 2021 CtC Address Collection to the MADIP Spine, linkage will use the new index, together with reprocessed name data from the Spine with the new index applied. A parallel “test” linkage will use the current index, with spine names as-is. Once both linkages are complete, a quality assessment will examine linkage rates and link quality to assess the improvement from the new index. In the event of unforeseen quality problems, the test linkage results will be used for final linkage.

Recommendation 4: Implement the new names index in the linkage of the 2021 CtC Address Collection to the MADIP Spine, whilst performing a parallel test linkage using the current index, enable assessment of improvements.

It is important to note that the overall linkage rate is not expected to improve significantly when using one index versus another. This is partly because the standardised name is only used in a subset of the linkage passes, meaning that many links are made without it. Also, newly created links from the improved index may be offset to some extent by a reduction in false-positive links, leading to a higher quality analytical product, even if the linkage rate is unchanged.

APPENDIX 1. EVALUATION OF THE NEW NAMES INDEX

For this work the following metrics are defined as a way to assess the quality of a name standardisation process.

- **Cost** – the number of standardisations, a measure of how lossy the process is. If too many names map to too small a number of standard forms then the risk of false positives in linkage increases.
- **Coverage** – the distribution across standardised names. If one standard name appears significantly more than others, then either too many names map to it or not enough other names are being standardised.
- **Difference** – how different the standardised names are from their inputs (edit distance or some other text similarity metric). We would expect that most standardisations will be close with an edit distance of three or less, but there will be some outliers – examples such as Bess and Elisabeth.
- **Performance** – the time it takes for code to apply standardisations using the index. If run time is excessive then we run the risk of delaying projects for marginal gain.
- **Maintainability** – how easy it is to update and improve the standardisation process as knowledge is gained or requirements change.
- **Representation** – the diversity of the names that we can standardise and the correctness of standardisations for each demographic. If names of a given cultural origin do not standardise correctly then there is a risk that the process adversely affects linkage for those persons.

The current index (currently in production use) is a manually built index that draws from a number of sources, including the Behind the Name website. This index also contains a small number of generated name standardisations that represent small spelling deviations, mostly the types stemming from Optical Character Recognition (OCR) errors. Another unique feature of the current index is the inclusion of standardisations that exist with the assumption that the provided sex variable is wrong: for example, the name Michael will standardise to Michaela if the provided sex is Female.

The new index is entirely made from data downloaded from Behind the Name. The new index has big improvements to scale - there are a large number of names with standardisations, more than the current index. The standard forms of names in the new index are chosen by taking the most popular name in a group.

The new index does not contain the OCR error correction type of standardisation. Those corrections are redundant, because linkers use fuzzy string comparisons on names in some of the linkage passes, and the fuzzy comparisons capture the same type of error. OCR corrections were introduced because of their relevance to Census data, which had a large proportion of names from scanned paper forms. This is less significant for administrative data such as that used for CtC.

For the comparisons below the current index and new index were both used to process ATO Client Register (CR) and CtC. The ATO CR has good coverage of the Australian population with over 67 million rows of data on the input file. The ATO CR data also has a sex variable which is important to the standardisation process. The names index is being updated for use in CtC data linkage, so it is important to test any changes on the CtC Address Collection data.

The table below provides a brief comparison of the two indexes against the above criteria.

Metric	Current Index	New Index	Comments
Cost	High (mean=14)	Low (mean=3)	The new index removes some erroneous standardisations, meaning groups are smaller
Coverage	More total standardisations	Less total standardisations	The new index standardises fewer rows of data overall. Note this does not indicate a poorer outcome and test linkage is required to ascertain the impact.
Difference	Low	Lower	The new index has lower average difference, lower standard deviation of difference and lower maximum differences.
Performance	O(1) runtime complexity for a name	O(1) runtime complexity for a name	The implementation in names processing code is identical. Constant time lookup for a name.
Maintainability	Manually built	Automatically built	The new index is much easier to maintain, as building a new one only requires an updated data in the form of the *Behind the Name* data.
Representation	Lacks standardisation for non-European names	Has better coverage of non-European names	One of the biggest improvements in the new index is the representation of non-European names.

Cost

The cost can be evaluated by looking at how many names map to a given standard name, cost should neither be too high or too low. A high cost means that the process is very lossy[2] and the risk of false links on passes using standard name increases. On the other hand, a low cost means that the process is not very lossy and that there are potentially not enough standardisations to provide any real benefit.

For the current index we observe the following summary statistics for the number of names that map to any given standard name:

Minimum	Maximum	Standard Deviation	Mean	Median	N
2	148	18.676	14.061	8	539

The maximum of 148 is higher than would typically be expected, examining this group in detail indicates that there are a number of erroneous standardisations. On average, names have 14

[2] Lossy encoding groups names together into a desired number of 'bins'. During data linkage, the bin identifiers are used as linking variables instead of names. First names and last names are encoded separately. Bin identifiers are removed from the dataset that is subsequently used by analysts.

standardisations which would be considered high if it weren't for how few standard names there are on the index (539 standard names).

When running the same analysis on the new index very different results are obtained:

Minimum	Maximum	Standard Deviation	Mean	Median	N
2	84	3.379	3.214	2	1529

For the new index the mean is considerably lower and the maximum is a more reasonable 84. The lower mean is not surprising as the new index also captures almost one thousand more standard forms of names. Most of the groups for the new index are of size 2 which is no surprising as often names only have one common variation (Gwyn → Gwynn, McKenzie → Mackenzie, Rodolf → Rudolf, etc.).

Overall the new index has much tighter groupings that the current index while also succeeding in capturing a large number of standardisations. This indicates that the new index contains less erroneous standardisations and can be expected to result in more accurate linkage.

Coverage

ATO CR coverage

The current index does not show any concerning patterns on the ATO CR (no standardisation appears more common than would be expected). The most common standardisation for first names is Catherine to Katherine making up 1.24% of standardisations. For middle names the standardisation of Ann to Anne accounts for 7.69% of standardisations which is a lot higher but is not unexpected as Ann and Anne are both very common middle names.

The new index is largely similar regarding the most common standardisations with the major exception being that Maria to Mary becomes the most common first name standardisation accounting for 2.69% of first name standardisations. For middle names the most common standardisation is the same family of names but instead Anne standardises to Anna. Another standardisation that is among the ten most common for middle name that is exclusive to the new index is that of Frances to Francis - the current index does not capture uncommon names like Frances, unlike the new index.

CtC Address Collection coverage

Assessing coverage on CtC is less meaningful as the sex variable is not available and as such are permuted with male and female versions, this results in a number of incorrect standardisations, especially with the current index where there are mappings to standardise differently if it is suspected that the sex has been recorded incorrectly.

For the current index the most common standardisation (3.44% of the standardisations) is Michael to Michaela, a number of these are likely to be incorrect as every Michael in the data will have a second version of the record where the only difference is that the standardised name is Michaela. The same can be said for the second most common standardisation which is Michelle to Michael.

The standardisations from the next index are more in line with expectations, because the new index does not have standardisations that assume sex to be incorrect – this works far better for this data where the sex is not provided. The most common standardisation is Maria to Mary (1.88%),

Difference

For difference two text similarity metrics are used: cosine similarity and edit distance. The summaries do not include names that standardised to themselves (i.e. identical standardised names).

Cosine similarity measures the cosine of the angle between two vectors, as such the names must first be converted into numerical vectors. To do this a term frequency inverse document frequency (TF-IDF) matrix is created where each row is a sparse vector that represent a name and encodes information about the descriptive capabilities of each letter in the name. Cosine similarity is great for sparse data as it measures the cosine of the angle rather than the distance. This relies on the bag of words assumption (bag of characters in this case). Cosine similarity is always a value in the zero to one range with zero being no similarity and one being identical.

Edit distance is calculated as the minimum number of edits required to make one string equal to another, this is a good measure of string similarity as it is representative of real edits that are required to make the strings match. For edit distance there is strong support in literature for the majority of the data (90 to 95 percent) to have an edit distance of 3 or less, as such we should not expect a high average[3].

While the tables below indicate that both indexes have a minimum cosine similarity of zero it is important to note that the standardisations with zero similarity in the new index are more logical.

The new index groups several names such as Lexi, Lexa and Lexie as standardising to Sandra – this group of standardisations make up all standardisations with a similarity of 0 in the new index. The intuition for this type of standardisation is Lexi → Alexandra → Sandra with Sandra selected as the standard form due to it being the most common name in that grouping.

On the other hand, the current index contains some harder to justify standardisations such as Diogo → James and Ib → James. While there is a logic to these standardisations such as Diogo → Diego → Santiago → James the jump from Diego to Santiago is commonly accepted to be a rebracketing error due to San-Tiago being mistake for San-Diego. As such the new index aims to avoid some of these questionable standardisations.

ATO CR difference

Cosine similarity for the current index:

Mean	Standard Deviation	Median	Minimum	Maximum	N
0.769	0.159	0.793	0	1.000	13216315

[3] Bloothooft, Gerrit, and Marijn Schraagen. "Learning name variants from true person resolution." *Proceedings of the International Workshop on Population Reconstruction. International Institute of Social History* (2014).

Cosine similarity for the new index:

Mean	Standard Deviation	Median	Minimum	Maximum	N
0.799	0.160	0.812	0	1.000	6540978

Comparing the results for the current and new indexes, on average the names are more similar to their standard forms on the new index, with the higher median indicating that a larger proportion of the data has a higher similarity. The larger standard deviation is to be expected, reflecting there are several standardisations that differ greatly from the original names.

The same information is reflected in the summaries of edit distances.

Edit distances with the current index:

Mean	Standard Deviation	Median	Minimum	Maximum	N
2.542	1.506	2	1	12	13316460

Edit distances with the new index:

Mean	Standard Deviation	Median	Minimum	Maximum	N
2.008	1.361	2	1	8	6730428

The minimum and maximum are more interesting when looking at edit distance, a maximum edit distance of 7 in the new index is much more likely to be a valid standardisation than the maximum edit distance of 12 in the current index.

CtC Address Collection difference

For CtC the results are fairly similar. The standardisations unique to the current index that tackle OCR errors and incorrect sex do not throw the number off as these are usually an edit distance no greater than one apart.

Cosine similarity with the current index:

Mean	Standard Deviation	Median	Minimum	Maximum	N
0.784	0.155	0.794	0	1.000	909841

Cosine similarity with the new index:

Mean	Standard Deviation	Median	Minimum	Maximum	N
0.788	0.165	0.811	0	1.000	310707

Edit distances with the current index:

Mean	Standard Deviation	Median	Minimum	Maximum	N
2.680	1.526	2	1	10	913043

Edit distances with the new index:

Mean	Standard Deviation	Median	Minimum	Maximum	N
2.127	1.491	2	1	8	320320

The new index has a closer similarity by both metrics.

Performance

From the perspective of running the name cleaning, both indexes have nearly identical performance despite the new one being larger. For the matching process the indexes are implemented as hash maps in memory which allow for $O(1)$ lookup time for each row. As this lookup time is constant and not affected by the size of the index, each index has the same run time cost.

Maintainability

The new index is built using an automated process. Data is downloaded from Behind the Name and then a program restructures this data into a format where groups of names map to a single standard form. This means that creation of a new index is repeatable and simple, more name standardisations could be added to the data from other sources and the program could also be modified if needed.

Representation

A web crawler is used to scrape information off the Behind the Name website and create a dataset containing name, cultural origin and sex. Note that in this context, cultural origin has no bearing on the cultural background of any given person who has that name and is not meant to represent that – the cultural origin does, however, provide a way to group names that often share specific features, phonetics, spellings or structures.

Using this information, we can identify if there are groups of names where the index is less effective.

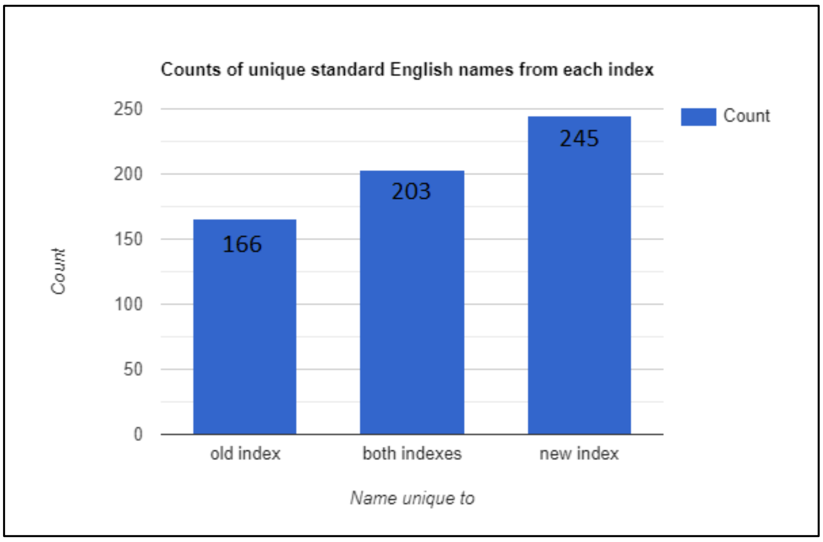
An index that captures many standard forms of names in a given culture is capable of standardising a wider variety of names and is likely less lossy for names in that culture. Counts of unique standard names from each index, for different cultural origin is shown in Figures 1 to 3[4].

[4] Note these only represent a small subset of cultural origins, chosen as they are indicative of the differences between the indexes when comparing European and non-European names

CtC Address Collection representation

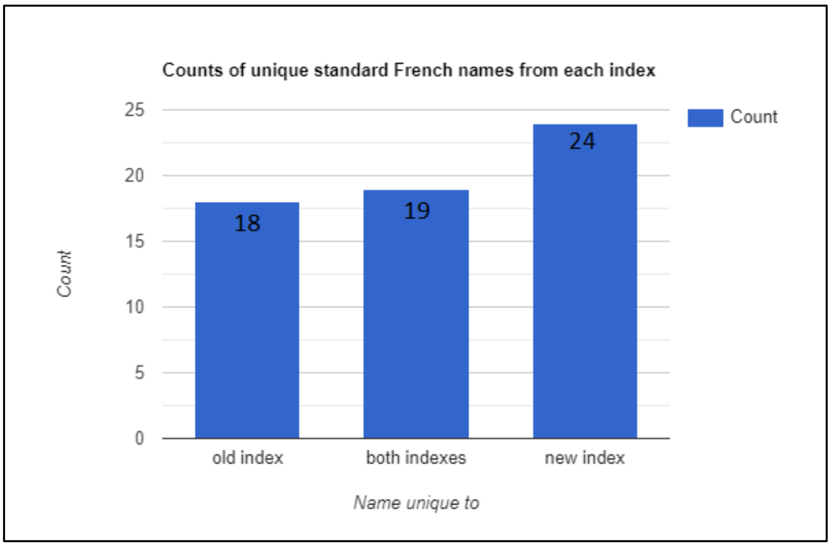
English standardised names are fairly balanced, as is expected given the indexes largely overlap. Note some names that are unique to a given index can belong to the same group of names but differ because the 'standard' version of the name that was chosen for that group is different between the indexes (Figure 1).

Figure 1: Bar plot of distinct English names from the current and new indexes



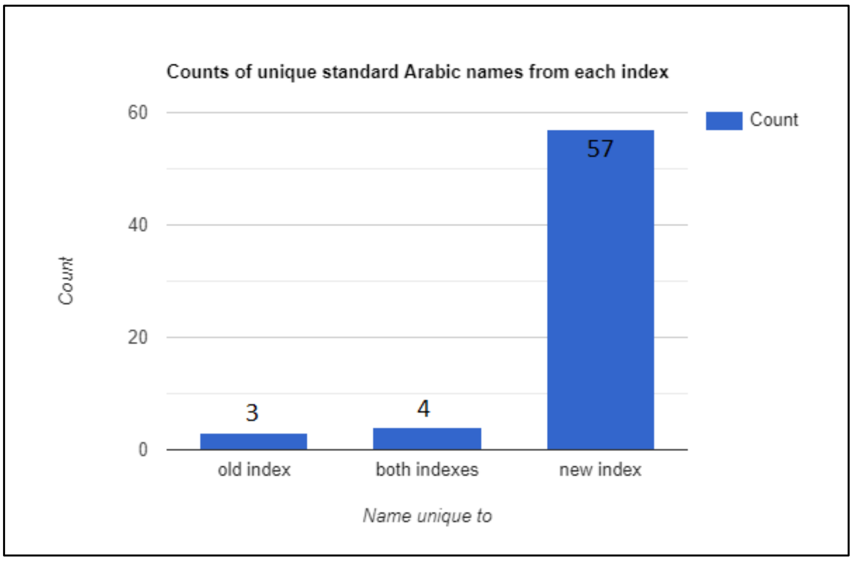
The bar plot for French names is similar in terms of proportion to the one for English names, this is as expected and is a shared amongst European names in general (Figure 2).

Figure 2: Bar plot of distinct French names from the current and new indexes



However, when we move away from European names the distributions change drastically. The new index contains almost nine times as many standardised forms of Arabic names as the current index with more than half the standard forms in the current index also being contained in the new index (Figure 3).

Figure 3: Bar plot of distinct Arabic names from the current and new indexes



Processing of ATO CR data yielded similar results for this metric.