Australian
Bureau of
Statistics

Research Paper

Options for Encoding Name Information for use in Record Linkage

(Methodology Advisory Committee)

Australia

2018

[1351.0.55.162]

ABS Catalogue No. 1351.0.55.162

Produced by the Australian Bureau of Statistics.

INQUIRIES

For further information about these and related statistics, contact the National Information and Referral Service on 1300 135 070.

OPTIONS FOR ENCODING NAME INFORMATION FOR USE IN RECORD LINKAGE

Paul Campbell, Noel Hansen, Charles Au, Jeffrey Wright and Daniel Elazar
Methodology Division

QUESTIONS FOR THE COMMITTEE

1. Does MAC broadly agree that Lossy encoding meets ABS's requirements for a practical encoding method that is secure while still affording a high level of linkage accuracy and flexibility?

2. Given the limited analysis so far undertaken, are there any other major methodological considerations for implementing Lossy encoding (e.g. optimal bin size, degree of uniformity of bin frequencies, linkage accuracy for populations that are difficult to link and allowing fuzzy comparisons such as encoding both repaired and standardised name)? To what extent is encoding multiple versions of names (repaired / standardised) likely to be a security concern?

CONTENTS

OPTIONS FOR ENCODING NAME INFORMATION FOR USE IN RECORD LINKAGE

Paul Campbell, Noel Hansen, Charles Au, Jeffrey Wright and Daniel Elazar
Methodology Transformation Branch

ABSTRACT

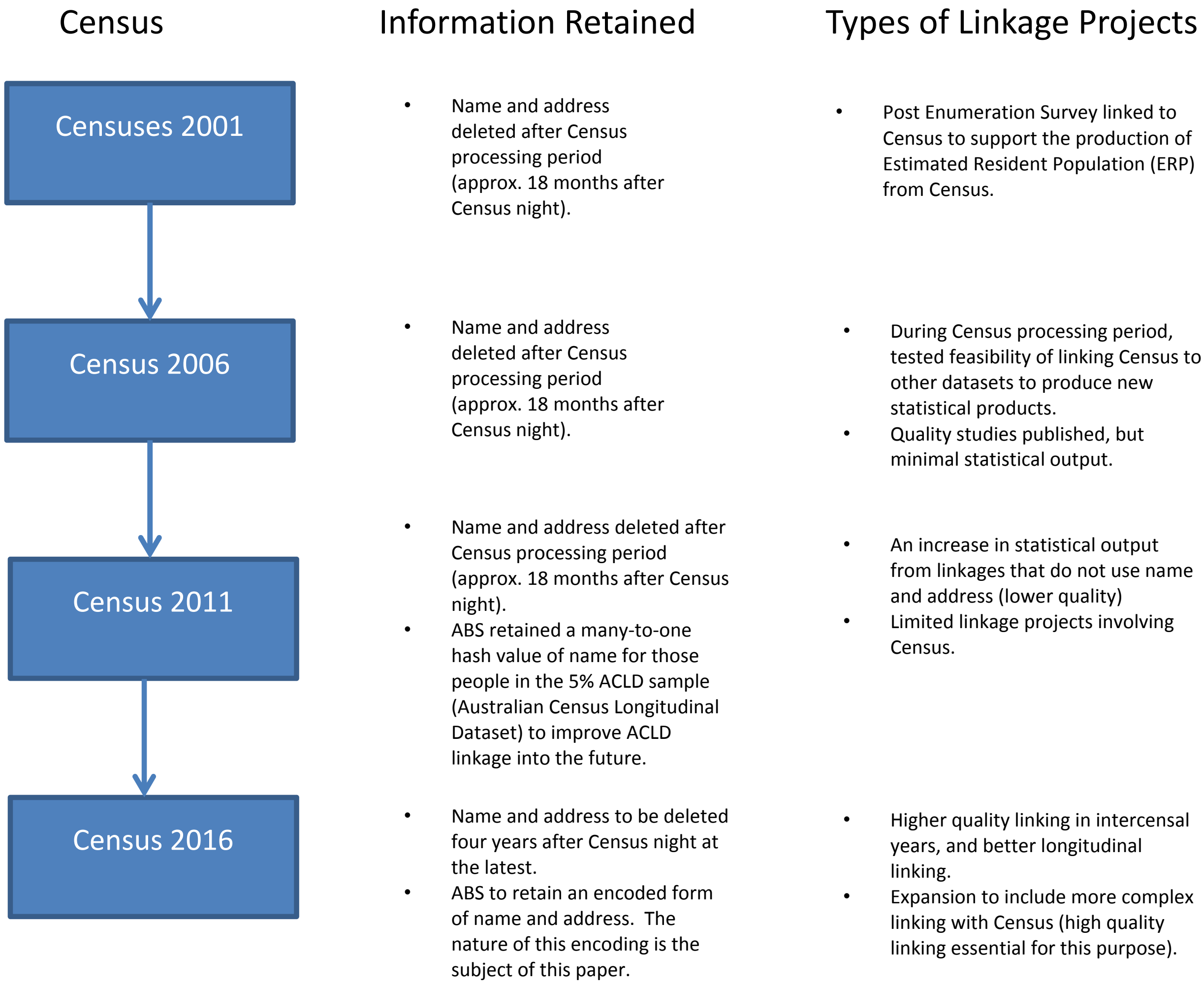
The purpose of this MAC paper is to present and discuss options for encoding Census name for use in record linkage, and seek MAC's views on the preferred option. In particular, we seek advice on whether the preferred method will fulfil the dual aims of providing sufficient security of 2016 Census name information while ensuring that linked datasets are of sufficient quality to support informed policy decision making. We also seek advice on the manner in which the preferred method is implemented.

1. INTRODUCTION

Linking Census data with other survey and administrative data enables the ABS to develop a richer and more dynamic statistical picture of Australia's social and economic landscape. To meet this important data need, ABS has been conducting numerous data linking projects since 2007. The use of name and address information on these datasets is critical for achieving a high level of linkage accuracy for data quality purposes. With this in mind, on 18th December 2015, the ABS announced it would be retaining names and addresses collected in the 2016 Census of Population and Housing for up to four years after collection (August 2020), with the possibility of earlier deletion if it is deemed there is no longer any community benefit (ABS, 2015). The intention is to retain encoded (anonymised) forms of name and address for the foreseeable future, after deletion of the original names and addresses. The focus of this paper is the method to be used to create an encoded form of name for use in record linkage.

The approach for 2016 Census data is an incremental change to the Census Data Enhancement (CDE) program undertaken with the 2006 and 2011 Censuses. The purpose of the change is to enable higher quality linking in the intercensal years, higher quality longitudinal linking, and to generally enhance the value of Census data as a long term data resource. Figure 1 below shows the changes through time in the retention and use of Census name and address information for linking.

Figure 1 - Timeline of Census name and address retention.



Census	Information Retained	Types of Linkage Projects
Censuses 2001	• Name and address deleted after Census processing period (approx. 18 months after Census night).	• Post Enumeration Survey linked to Census to support the production of Estimated Resident Population (ERP) from Census.
Census 2006	• Name and address deleted after Census processing period (approx. 18 months after Census night).	• During Census processing period, tested feasibility of linking Census to other datasets to produce new statistical products. • Quality studies published, but minimal statistical output.
Census 2011	• Name and address deleted after Census processing period (approx. 18 months after Census night). • ABS retained a many-to-one hash value of name for those people in the 5% ACLD sample (Australian Census Longitudinal Dataset) to improve ACLD linkage into the future.	• An increase in statistical output from linkages that do not use name and address (lower quality) • Limited linkage projects involving Census.
Census 2016	• Name and address to be deleted four years after Census night at the latest. • ABS to retain an encoded form of name and address. The nature of this encoding is the subject of this paper.	• Higher quality linking in intercensal years, and better longitudinal linking. • Expansion to include more complex linking with Census (high quality linking essential for this purpose).

To inform the decision on name encoding, ABS sought advice from cryptography experts at University of Melbourne. The University of Melbourne provided advice on a range of options to ensure security of Census name information whilst enabling high quality record linkage (Culnane et al. 2018).

Of the options provided, two were considered suitable for further exploration. The first method considered involves applying a many-to-one function to names, whereby many distinct names map to a single value, and the loss of raw name information protects the data (we term this *lossy encoding*). The second involves combining name with other fields to create a number of unique (or near unique) identifiers, and encrypting these identifiers using a Hashed Message Authentication Code (HMAC) encryption algorithm (we term this approach *HMAC-based linkage identifiers*). These methods were considered with respect to their:

- impact on linkage accuracy
- ability to meet security requirements, and
- ease of implementation.

At this point, lossy encoding appears best place to meet the objectives across these requirements, acknowledging that University of Melbourne had a preference for HMAC-based linkage identifiers from a security point of view. These issues are discussed further in this paper.

In this paper we lay out our rationale for the proposed approach and then discuss in more detail the challenges in implementation. Section 2 describes the three main requirements for any proposed encoding method, these being linkage accuracy, security and ease of implementation. Section 3 describes lossy encoding in general and in terms of these requirements. Section 4 discusses the HMAC encryption option and how it would perform in terms of the broad requirements. Section 5 provides a summary of the implementation issues for lossy encoding. Section 6 gives a summary of the proposed implementation, and Section 7 contains some concluding remarks.

The ABS seeks MAC's views on the content of this paper, particularly whether the proposed method is likely to fulfil the aims of ensuring the security and privacy of personal data, meeting the ABS's public assurance on the retention of 2016 Census name information (ABS (2015) and ABS (2016a)), whilst ensuring that linked datasets are of sufficient quality to support informed policy decision making.

2. ENCODING METHOD REQUIREMENTS

2.1 LINKAGE ACCURACY

In the 2017–18 Budget, the Government funded the Data Integration Partnership of Australia (DIPA) as a coordinated Australian Public Service-wide investment to maximise the use and value of the Government's data assets through data integration. Through DIPA, the Government is enhancing data assets and analytical capability to deliver better policy outcomes and better targeted and more effective services. Building trust and user support in integrated datasets is of strategic importance to ABS and depends upon their level of quality and fitness for purpose.

ABS currently essentially uses two data linkage methods. The first is a multiple pass deterministic linkage method, the Deterministic linking Macro (D-MAC) written in SAS (ABS, 2016b), which is used for the efficient and reliable generation of most linked datasets. However probabilistic linking (Fellegi and Sunter, 1969) is still used as the failsafe method when a similar linkage exercise has not been attempted before or when the quality of previously linked datasets is believed to have deteriorated significantly. Probabilistic linking is the preferred option in the case of poor quality linkage variables and it provides more methodologically defensible measures of linkage quality, with which to quality assure those produced by D-MAC.

These linkage methods use fairly sophisticated linkage strategies involving multiple passes / runs that are tailored to the quality and nature of the linking fields and the objectives of specific linkage exercise. Multiple linkage passes are necessary to help link persons who have moved, in cases where linking fields are subject to higher levels of error as well as produce improved quality measures. In probabilistic linking, match scores are used to determine the optimal cut off between links and non-links. These rely on calculating diagnostic measures for each linking variable.

Methods that involve encrypting or hashing plain text name in combination with other linking fields (e.g. age, sex, data of birth etc.) make it infeasible to calculate match scores for name and hence undermine the value of probabilistic linking. This is due to the fact that when name and other linking variables are encoded together, it is not possible to distinguish agreement from disagreement on name, when there is disagreement on at least one other linking field involved in the combined encoding. (There will be no restriction on observing agreement/disagreement on any of the non-sensitive component fields)

Lastly, it is desirable for the encoding method to allow fuzzy matching to deal with, for example typographical errors in names. However the rigorous processes of cleaning, repair and standardisation applied to plain text names, tend to compensate for a lack of fuzzy matching. Name repair, for example, involves using an approximate string comparator to compare each plain text name against an index of names gleaned from multiple administrative data sources, acquired by ABS.

2.2 SECURITY

In discussing the merits of the proposed name encoding method in this paper, it is important to place it in its proper context. In regard to 'privacy preserving record linkage' (PPRL) as a subject, the emphasis in the literature is on enabling record linkage between databases of different data custodians, whilst minimising the risk of exposure of personal details (e.g. name, address) between custodians during the linkage process. Name encoding and encryption play a major part in PPRL. See Chapter 8 of Christen (2012) for a thorough discussion of the subject. For the purposes of this paper, note that the ABS's situation differs from typical applications in PPRL in two ways:

1. ABS has committed to deleting the original plain text names from the Census (as discussed previously), but also to retaining an anonymised/encoded form of name that cannot be reversed (which impacts on eligible data linking algorithms). Typically in PPRL applications, there is no requirement for data custodians to delete useful linking variables, and there is no requirement to make encoding schemes irreversible by design (usually the emphasis is on restricting who might be able to reverse it, rather than saying it cannot be reversed at all).

2. ABS will not share encoded name information with any external party for record linkage, and it will only be used within the ABS ICT environment. Typically in PPRL applications, the emphasis is on encoding and protecting name so that it can be shared with other parties for the purposes of record linkage.

The proposed name encoding method also needs to be considered in light of ABS's broader ICT security and operational environment which includes:

- ABS's legislative framework,
- Separation principle and functional separation applied to data linking using the Census,
- Strict data access controls and auditing,
- A secured computing platform within the broader ABS ICT environment, and
- Well established track record in the secure management of highly sensitive and confidential data.

Information security experts recognise that data security cannot rely solely on one logical control method such as encryption / encoding of the data. Effective information security relies on 'Defence in Depth' whereby security is multi-layered across data, application, host and network layers (Wikipedia, 2017). This paper should be read with this context in mind.

The ABS needs to evaluate and mitigate both internal and external threats. Internally, ABS officers work with sensitive data, and must sign a security and fidelity undertaking, annually. Misuse of data is a criminal offence and can result in gaol sentences and/or fines for each breach. Additionally, access within ABS is highly restricted following the information security principles of 'least privilege' and 'separation of duties (SoD)' (functional separation in ABS lexicon).

In the case of Census data linking projects, only the few people involved in the linkage exercise will be able to gain access to the data and systems and no-one has more access rights than is absolutely needed to undertake her/his job at a particular point in time. Under the Separation Principle, names and addresses are stored separately from analysis fields, and administrative controls are in place to ensure that no individual is permitted access to both sets of fields at the same time, through the strict application of roles. The severity of legislative sanctions, restricted access, and limited access to linking fields all help mitigate the risk of an internal attack on encoded Census name. Access to linked data files as release products is restricted to authorised users in a secure data environment in which the Five Safes Framework is applied.

External attacks threaten not only the Census, but all ABS datasets. The ABS has ICT security measures in place to detect and prevent attacks on our data, and it is worth noting that these measures have been used for a considerable period of time for sensitive economic and person level datasets, including the Census plain text name and address retained during the Census processing period. However, the threat landscape is continually evolving and security controls need to be regularly re-assessed to ensure an appropriate level of protection.

To summarise, the encoded name ought not to reveal plain text name to an adversary (internal or external), but at the same time, the associated risks need to be put in the context of the other broader security measures in place.

2.3 IMPLEMENTATION

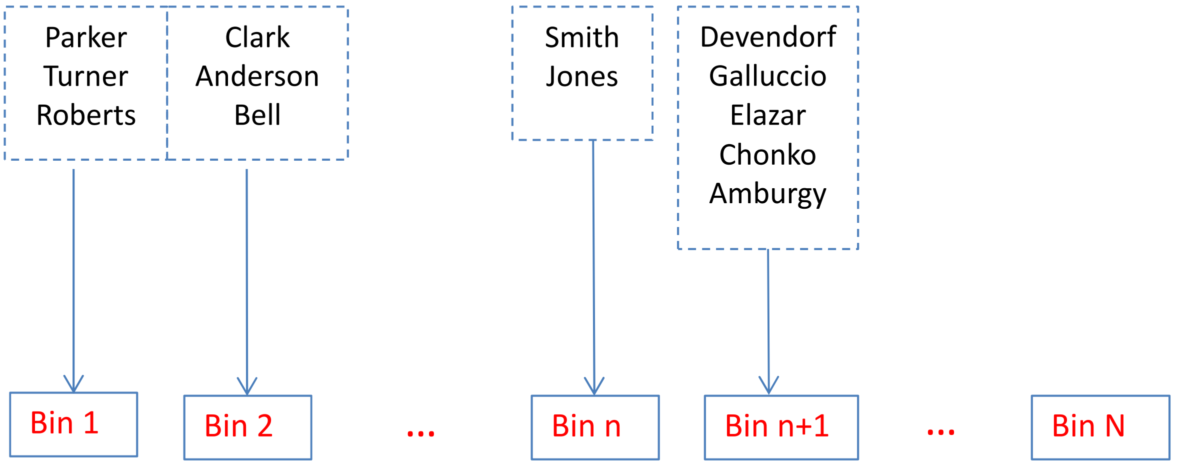
Any method for encoding Census name information needs to be fully compatible with the data integration methods used at ABS (D-MAC and probabilistic). In particular, the method needs to be sufficiently flexible, not only for linking strategies we currently use, but also those for future linkage projects that have not yet been formulated. In addition, if the adopted method requires the retention of private keys or tables, which have the potential to enable the name encoding to be reversed, they need to be kept in a highly secure manner. The method also needs to balance security with practicality, ease of implementation and computational efficiency.

3. LOSSY ENCODING

3.1 WHAT IS LOSSY ENCODING?

Lossy encoding is an algorithm whereby many input values map to a single output (encoded) value. Security is essentially provided through the loss of information. Figure 2 illustrates the basic idea of lossy encoding applied to names. Names are grouped together into a desired number of 'bins' and during data linkage, the name bin identifiers will be used instead of plain text name. The higher the number of bins, the less security protection afforded, but the greater the accuracy of linking on bin identifier. The fewer the bins, the greater the security protection afforded, but the lower the linkage accuracy. For security protection, it is desirable that bin sizes are approximately uniform and that each bin contains some minimum number of discrete names (Culnane et al. 2017).

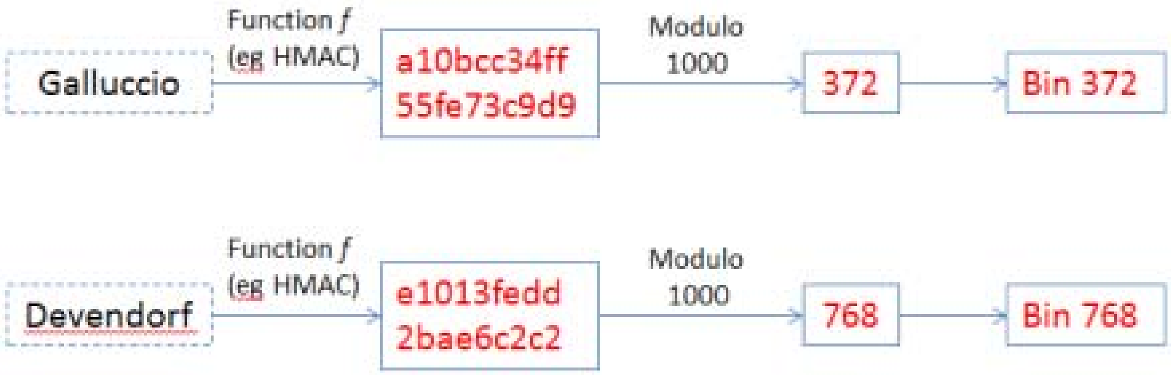
Figure 2 – Basic concept of lossy encoding applied to names
(hypothetical example for illustrative purposes only)



There are several potential methods for mapping names to bins. First, there is the method developed by the ABS, and used in previous quality studies (such as Bishop, 2007). Names were first truncated, then hashed using an in-house function and then mapped to one of a user-specified number of bins, using the modulo function.

A second, more industry standard approach, is to use a standard cryptographic hash function such as HMAC SHA 256 and apply the modulo function to the output. Figure 3 illustrates the general approach, hypothetically, for assigning names to bins. The use of HMAC here is just as a means to an end under Lossy encoding and should not be confused with the HMAC-based linkage identifiers described in Section 4.

Figure 3 – Hypothetical example of mapping plain text names to 1000 lossy bins



A third option is to manually map names to a code, via a lookup table. It is infeasible to do this for all names, but may have merit for common names, in order to smooth out any skewness in the frequency of distribution of names.

3.2 LINKAGE ACCURACY

Lossy encoding is compatible with both deterministic and probabilistic linking methodologies (both are currently used at the ABS) and sufficiently flexible with any current and future linkage strategies used with those methods. Encoded names are simply additional linking variables.

There will inevitably be some loss of linkage accuracy through Lossy encoding, however Methodology Division is currently undertaking empirical evaluations to assess the level of loss for data linkage under both deterministic and probabilistic approaches. Based on a very preliminary evaluation this loss appears to be acceptably low for linkages involving the general population.

Lastly, although fuzzy comparisons are not readily compatible with Lossy encoding, as mentioned earlier in this report, this is not a major disadvantage as an approximate string comparator is applied during name repair. Although it is possible to apply lossy encoding separately to variants of plain text name, this adds to the complexity and may not increase linkage accuracy substantially.

3.3 SECURITY

Lossy encoding ensures that the resulting encoded name is irreversible. By definition, it is impossible to explicitly reverse a many-to-one encoding. There is insufficient information to do so, and this loss of information is the baseline level of protection that this encoding method offers. Given raw names have a skewed frequency distribution, we do need to consider frequency attacks in designing the encoding method. We queried the University of Melbourne consultants about what level of uniformity of bin frequencies is required to ensure security, however there is no simple answer as it depends upon the assumed attack scenarios and other security protections put in place. Clearly any bin frequency that is slightly higher than the others may give information to a potential adversary but the risk that this information can be successfully utilised to affect an attack is more difficult to assess.

If an assignment table is required to ensure uniformity of bin frequencies for common names, this table will be kept secured in the ICT environment with restricted access using the functional separation principle.

3.4 IMPLEMENTATION

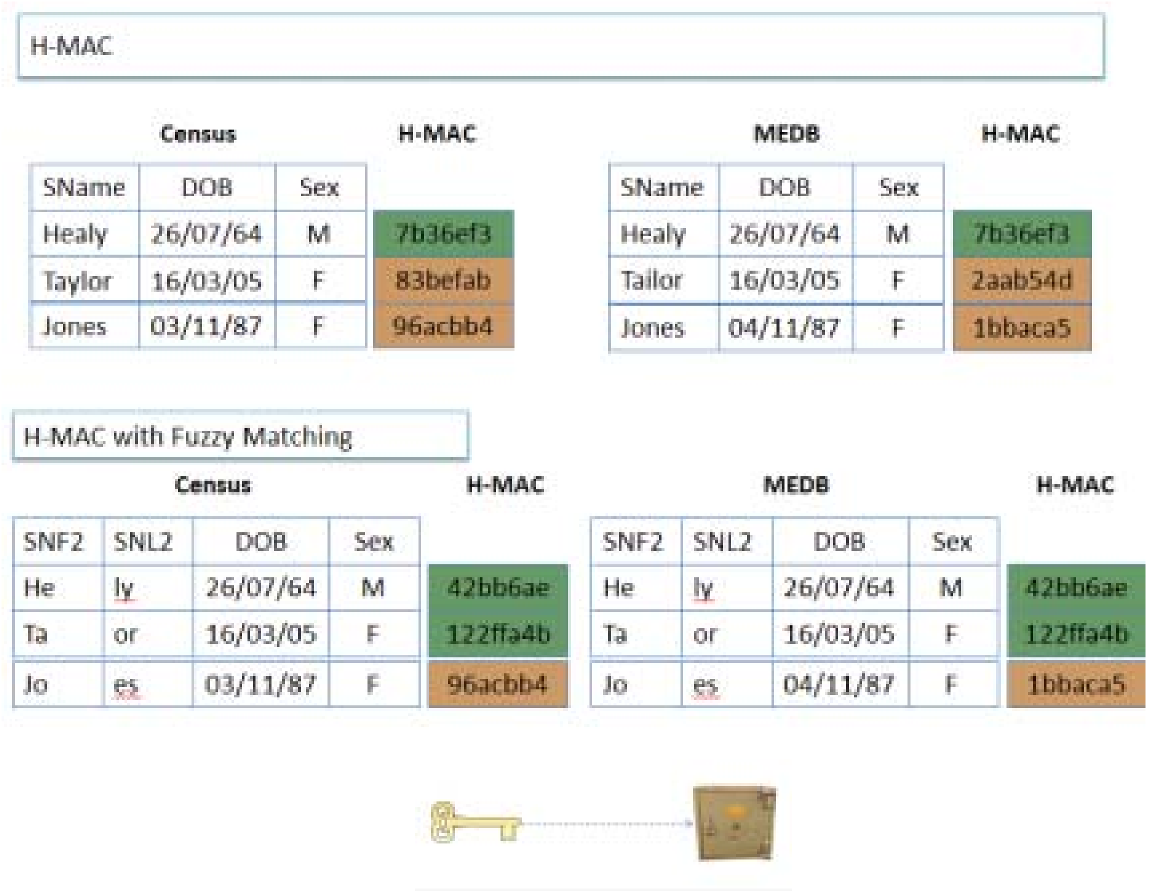
The ABS already has direct experience with lossy name encoding for record linkage, hence its implementation should be relatively straightforward. In 2007, ABS Methodology developed and used a lossy encoded name in a study exploring the feasibility of creating a longitudinal Census dataset (Bishop, 2007). We have subsequently adopted this approach in the 2011 Census for the 5% of records in the sample of the Australian Census Longitudinal Dataset (ACLD).

4. HMAC-BASED LINKAGE IDENTIFIERS

4.1 WHAT ARE HMAC-BASED LINKAGE IDENTIFIERS?

The approach suggested by Culnane et al. (2017) involves applying key-Hashed Message Authentication Code (HMAC) to a combination of linkage variables and then linking on the HMAC encrypted value. The top half of Figure 4 gives an illustration of how the method would work in its basic form.

Figure 4 – Basic concept of HMAC applied to a combination of variables (hypothetical example for illustrative purposes only)



H-MAC

Census

SName	DOB	Sex	H-MAC
Healy	26/07/64	M	7b36ef3
Taylor	16/03/05	F	83befab
Jones	03/11/87	F	96acbb4

MEDB

SName	DOB	Sex	H-MAC
Healy	26/07/64	M	7b36ef3
Tailor	16/03/05	F	2aab54d
Jones	04/11/87	F	1bbaca5

H-MAC with Fuzzy Matching

Census

SNF2	SNL2	DOB	Sex	H-MAC
He	ly	26/07/64	M	42bb6ae
Ta	or	16/03/05	F	122ffa4b
Jo	es	03/11/87	F	96acbb4

MEDB

SNF2	SNL2	DOB	Sex	H-MAC
He	ly	26/07/64	M	42bb6ae
Ta	or	16/03/05	F	122ffa4b
Jo	es	04/11/87	F	1bbaca5

A HMAC code would be formed from the combination of Surname (SName), Date of Birth (DOB) and Sex. This code would be used as a linking field along with other non-sensitive fields. We can see from the top half of Figure 4 that the slightest difference in name or any other component field would result in a non-match on HMAC code (coloured red). Agreement on HMAC code for two records, such as in the record for Healy (coloured green), requires exact agreement on each component field.

This requires knowing in advance the linking strategies for all future linkage exercises involving Census name. Each linkage strategy is composed of runs/passes involving specific combinations of linking variables. For those runs/passes involving Census name, HMAC values would need to be calculated prior to the deletion of Census names.

Another important aspect of HMAC relates to the secret key. While the secret key is primarily intended for the encoding of plain text name it could in theory be used to reverse the resulting cypher. This would require knowing or being able to guess the structure of the digest, that is, the variables used in the digest and the order in which they appear. This means that the secret key has to be kept highly secured, and even though this would be assured, there might be a public perception that the cyphers could be reversed using a dictionary attack.

An alternative form of HMAC was suggested by Culnane et al (2017) to enable some degree of fuzzy matching. This is illustrated in the bottom half of Figure 4. Essentially, plain text names can be segmented into the first two characters of Surname (SNF2) and the last two characters (SNL2) which will give agreement even where there are typographical errors in other characters.

4.2 LINKAGE ACCURACY

As discussed in Sub-section 2.1, HMAC is in a class of methods that would effectively render probabilistic linking to be no more effective in measuring the accuracy of linkages than deterministic linking. This is due to the inability to reliably calculate match scores for name in some circumstances. Disagreement on any field other than name would lead to disagreement in the HMAC cypher, regardless of whether the name field agreed or disagreed. Hence we would not be able to take into account the specific M- and U- probabilities[1] for name in order to gauge its relative strength for each record linkage. This is even more important for frequency-based matching (Herzog et al., 2007) used at the ABS. This method effectively increases the match score for rare names, compared with using a standard match score for name.

Fuzzy matching could be supported by HMAC but whether this is necessary, assuming the effectiveness of name repair, remains to be seen.

We would expect that the quality of deterministic linking would be largely uncompromised by the use of HMAC and this is supported by some preliminary empirical work.

4.3 SECURITY

HMAC would provide a high level of security for Census name information for use in data linking. However, the secure storage of the key will be critical and there may be the public perception that the HMAC cypher could be reversed at will.

4.4 IMPLEMENTATION

HMAC encryption is readily available in SAS and other software packages and would be simple to apply. The main implementation issue arises in the storage and management of a large number of potential HMAC codes for use in future linkages. As it will be almost impossible to pre-empt all future linkages HMACs for all possible combinations of Census name with other variables will have to be created and stored. The use of HMACs for fuzzy matching will greatly increase this storage and management cost.

[1] M-probability = probability that a record pair agree on a linkage variable given that the record pair is a true match
U-probability = probability that a record pair agree on a linkage variable given that the record pair is a true non-match

5. IMPLEMENTATION ISSUES FOR LOSSY ENCODING

In relation to Section 2 on general requirements of an encoding method, the following are some specific issues surrounding the implementation of lossy encoding for the Census 2016 data:

i. We desire that encoded name should have a discriminatory power that is as close as possible to that of plain name while affording security protection. To decide on the number of bins to use we examined the trade-off between linkage accuracy and number of bins. We focused on uniqueness rates (proportion of linkages that are unique) for combinations of the linking variables name, date of birth (DOB) and address/geocode available on the Medicare Enrolments Database (MEDB) for 2011. These are important variables for linking units that are inherently hard to link, such as persons who have moved, persons with name variants or persons who are more likely to have missing information. In these cases we wish to salvage whatever linkage information we can, based on at least two of these fields, giving us the best estimate of match score we can obtain. Bear in mind that obtaining a good link for hard to link units will often still depend on the name being rare and agreement on linking variables additional to these three.

ii. The encoded name has to satisfy ABS security requirements, including upholding public undertakings made by the ABS. Of particular concern is whether the frequency distribution of encoded names reveals too much information about plain text names to an attacker.

iii. Ideally we would like to be able to apply a fuzzy comparator to encoded name or at least allow for approximate agreement between records.

iv. Related to point (i), we require encoded name to support the linking of subpopulations such as Aboriginal and Torres Strait Islander Peoples and certain migrant groups. These subpopulations are known to have more data quality issues, alternative names and spelling variations, yet are often the groups of most interest to policy analysts. Analysis of these populations also benefits the most from the full population size of administrative datasets, compared to that of survey samples.

Each of these requirements will now be discussed in the following subsections.

5.1 OPTIMAL NUMBER OF BINS

For the purpose of linking, the more bins available, the more useful the encoded name will be. However, as the security of lossy encoding arises largely from the many to one mapping, there is a need to find the minimum number of bins required for high quality linkage (or alternatively, to find the maximum number of bins that will yield an encoded name which cannot be reversed).

A related factor is the number of names in each bin (as opposed to the number of people). We would aim to ensure each bin contained at least $N>1$ distinct names, but the value that N should take is yet to be decided.

The ABS first explored lossy encoding in 2007, in a data linking quality study. In the study, encoded name was a many to one mapping of full name (i.e. given name and family name combined) into one of 12,005 bins (Bishop, 2007). This encoded name was used in conjunction with other linking variables in linking the 2005 Census Dress Rehearsal (CDR) to the 2006 Census, creating what was termed a *Silver Standard* linked dataset. The same datasets were linked using plain text name (creating a *Gold Standard*).

The project demonstrated that even though a sizeable degree of name information can be lost, the linked dataset is nonetheless fit for purpose. Using Gold as a benchmark, the Silver dataset had a match rate (recall) of 91.3% and link accuracy (precision) of 96.3% (Table 7.3 of Bishop, 2007). The study also suggested that this name encoding produced a representative dataset, and one that could yield the same analytical results as analysis conducted on the Gold dataset. For instance, logistic regressions run on the Silver and Gold datasets yielded the same conclusions.

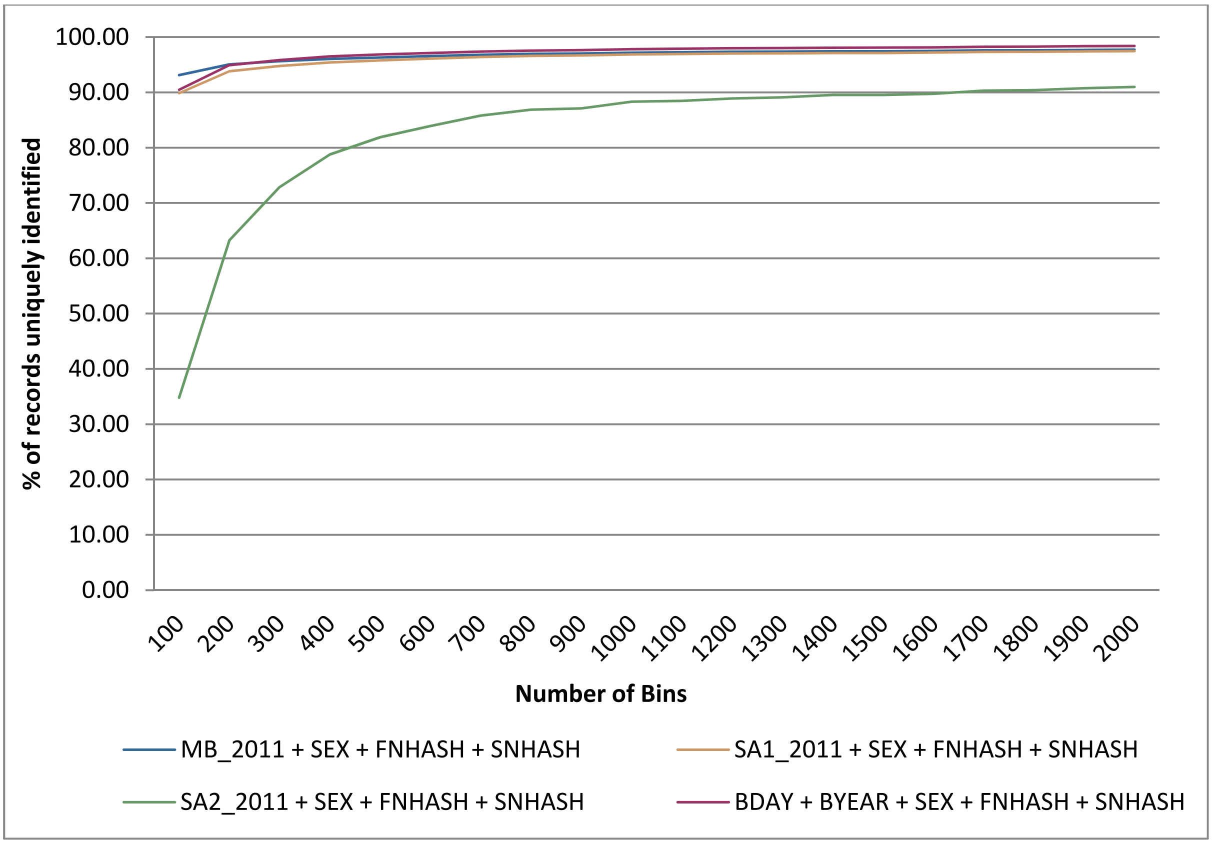
The linkage using the 12,005 full-name bins was not adopted for the following Australian Census Longitudinal Data (ACLD) linkages in which a random sample of records from each Census is linked to subsequent Census files. There was also no analysis of linkage accuracy for specific sub-populations such as Aboriginal and Torres Strait Islander Peoples and migrants.

Uniqueness analysis

To investigate optimal bin size, we used the MEDB data from 2011 to analyse uniqueness rates (proportion of MEDB records that are unique) for four different combinations of key linkage variables that include lossy encoded name, for varying numbers of encoded name bins. The quality of reported names may be higher on this dataset than from the Census, and may exclude some more uncommon names from recent migrant populations, but both datasets have a similar coverage of the Australian population. It is important to note that this uniqueness analysis is only indicative of the optimal number of bins. A final decision on the optimal number of bins for lossy encoding will be made based on a thorough quality assessment of data linkage exercises using both deterministic and probabilistic linking.

Figure 6 below shows that using the variable combinations involving Mesh Block, SA1 or DOB (Date of Birth), the vast majority of records are unique with around 700-800 bins (separately for First name and Surname). However, the choice of 2,000 bins appears to be a good compromise when considering the variable combination involving the broader geography, SA2 (green line). The uniqueness rate for the variable combination involving SA2 still continues to increase up to 2,000 bins but at a considerably slow rate. Our empirical study showed that the uniqueness rate remains at 92.08% for 10,000 through to 50,000 bins, hence given that 2,000 bins achieves 90.99 (See Table 1), there is little marginal benefit in going beyond 2,000 bins.

Figure 5: Percentage of unique Medicare Enrolments records by key linking field combinations – for lossy encoded name



MB_2011 = Mesh Block Area 2011	FNHASH = First Name Lossy Encoded
SNHASH = Surname Lossy Encoded	BDAY = Day of Birth (1-365)
BYEAR = Year of Birth	
SA1_2011 = Statistical Area 1 2011	SA2_2011 = Statistical Area 2 2011

In interpreting Figure 5, it may be useful to compare the uniqueness rates using 2,000 lossy encoded name bins with those for the same variable combinations but using plain text name. These are shown in Table 1 below, which shows that with 2,000 bins all variable combinations, except for that involving broad geography (SA2), come close to the uniqueness rate for plain text name. We can conclude from this that a 2,000 bin lossy encoding performs better when combined with a finer level geography and that under lossy encoding we may not be able to achieve quite the same linkage accuracy for persons who have changed address.

Table 1: Percentage of unique Medicare Enrolments records by linking field combinations - Plain Text Name and Lossy Encoded Name

Variable Combinations	Plain Text Name % Uniqueness	Lossy Encoded Name (2000 bins) % Uniqueness
MB_2011 + SEX + FIRST_NAME + SURNAME	99.08	97.72
SA1_2011 + SEX + FIRST_NAME + SURNAME	98.94	97.46
SA2_2011 + SEX + FIRST_NAME + SURNAME	96.18	90.99
BDAY + BYEAR + SEX + FIRST_NAME + SURNAME	99.62	98.39

5.2 THE FREQUENCY DISTRIBUTION OF THE ENCODED NAMES

Unlike one-to-one hashing, the baseline protection of lossy encoding is the fact that an original name can never be re-derived from a lossy value. The protection is in the fact that information is destroyed, not in the mathematical properties of a hashing or encryption function. Hence, while the frequency distribution of the lossy codes (i.e. the distribution of person-name instances across bins) can reveal some information, frequency is not as informative to an attacker as it would be in a one-to-one hashing algorithm. Nevertheless, it is still prudent to consider the distribution of the lossy codes and ensure we are not unnecessarily revealing information that could otherwise be easily protected.

Figure 6, below, gives the frequency distribution of 2011 Medicare Enrolments data when first names and surnames are each encoded into 500 bins (using the current ABS function approach). Figure 7 gives the frequency distribution for 2,000 bins. Given knowledge of the most common first names and last names in Australia, an attacker could reasonably infer that the bins with the highest frequency reflect the most common names. As such, we need to adopt an approach to assigning names to bins that mitigates this risk to an appropriate level.

Figure 6: Frequency distribution of encoded first name and surname, with 500 name bins

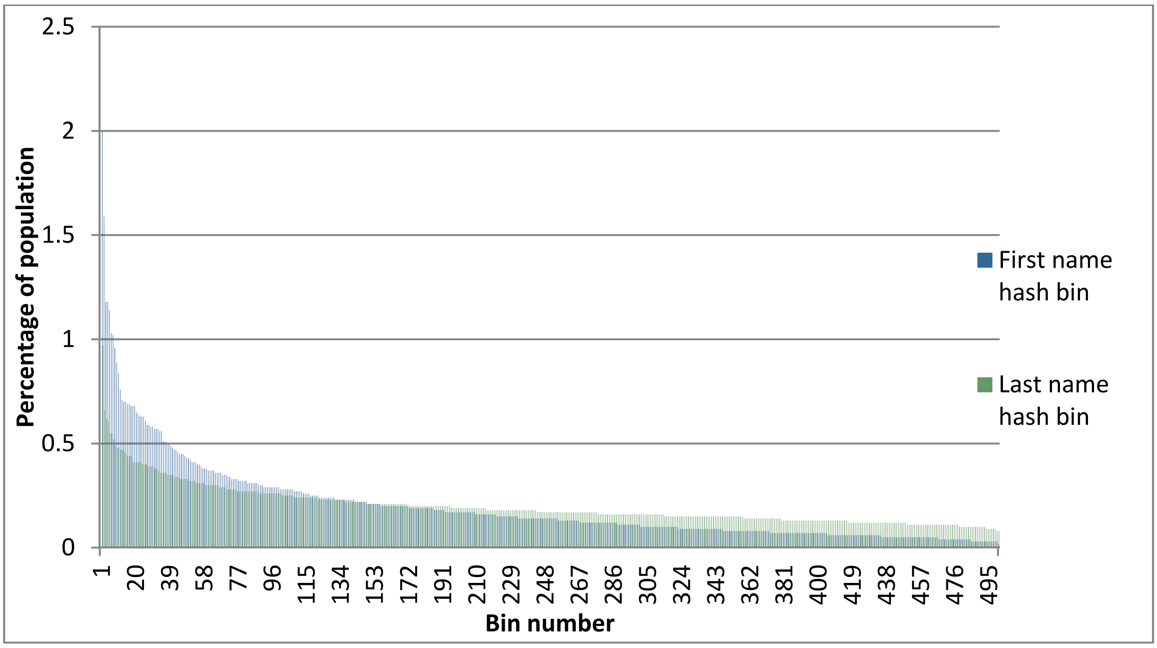
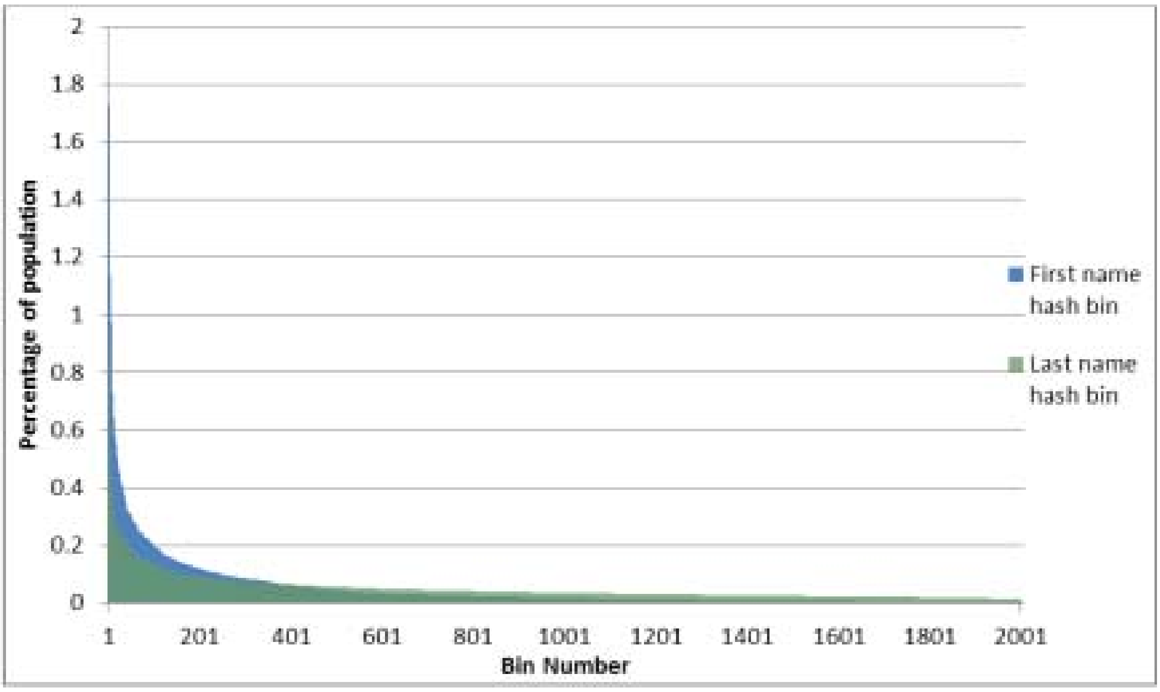


Figure 7: Frequency distribution of encoded first name and surname, with 2000 name bins



One option to mitigate this risk is to separately encode common and uncommon names, effectively segmenting the distribution into two parts, both with more uniformity in their distributions. This could be done by encoding the hundred most common names using a lookup table. It may have the added benefit that uncommon names would retain their rarity in the encoding, but this may be offset by the fact that grouping common names together would create some very large bins with much less discriminatory power.

Alternatively, we could rely on a modulo function to map all names, but hard-code a rule to create a second (and third) equally common name. For instance, prior to encoding, we could standardise the names TAYLOR and WILLIAMS to the same string so there is at least one other bin with frequency similar to the SMITH bin. This option is similar to the segmenting approach, but doesn't insist on common and uncommon name being in different bins. This option is simpler to implement.

5.3 FUZZY COMPARISONS

One of the benefits of using plain text name over encoded name is that it allows the use of fuzzy string comparisons. The ability to use fuzzy string comparisons can be helpful when, for example, there are spelling variations of uncommon names that cannot be corrected through name repair or standardisation. However, the ABS has invested resources to clean and repair Census names for data linking, and it is hoped that cleaned, repaired and standardised names do substantially improve match rates, thus reducing the need for fuzzy comparisons. Presently, we do not have any empirical evidence for how much each of these different forms of name, improve match rates.

The name repair process has produced three versions of name, in addition to the original, or plain text name. First, the *cleaned name* field has had non-alphabetic characters removed, along with prefixes (MR, MS, DR, SIR etc.), suffixes (JR, SR etc.) and nonsense name responses. The *repaired name* field is the result of an extensive automated and manual process to map the name to its closest option on a name index. The *standardised name* field takes the repair process a step further, and maps nicknames and name variants to their most common root name (e.g. Nic, Nich, Nick, Nikky are standardised to Nicholas or Nicole, conditional on gender).

A final point to consider here is whether we encode the full name, or just the first *n* characters of name. Taking, for instance, the first 4 characters of first name is a form of name standardisation ("Chris" and "Christopher" go to "Chri" prior to encoding) that may help with both nicknames and errors in the name.

5.4 APPLICATION TO CHALLENGING TO LINK SUBPOPULATIONS

Certain subpopulations, such as those with rare names, high levels of mobility or low levels of representation on linked datasets, present significant challenges in linking successfully. Nevertheless, these populations are often of keen interest to policy analysts. For example, rare names have more discriminatory power in linking, but are also more difficult to confidently repair and standardise with our validated name indexes. They are also a higher privacy risk.

Aboriginal and Torres Strait Islander Peoples and some migrant groups, often meet these criteria and we propose to give special attention to these population groups when investigating the impact of Lossy encoding on linkage accuracy. MAC's views are welcomed on whether there are potential enhancements to the lossy encoding algorithm that will help improve linkage accuracy for these groups.

Lossy encoding will result in some loss of accuracy for these population groups due to greater spelling variations in names and in some cases higher geographical mobility. The loss in accuracy will depend in part on the extent to which name repair and standardisation can mitigate for name variants. Over time, the accuracy of name repair may improve as indexes are compiled from a larger number of administrative datasets. This assumes, of course, that spelling variations are similar enough to be detected by approximate string comparators and each variant appears on these datasets.

Probabilistic linking is invariably used for the more challenging to link sub-populations as it is the method of choice when the quality of linking variables starts to deteriorate. However, as discussed, if the HMAC method were to be used, this would present some serious barriers to the reliable measurement of linkage quality.

6. SUMMARY OF PROPOSED IMPLEMENTATION

Taking into consideration the issues discussed in the previous section, the proposed approach at this stage is:

- Encode first name and surname separately. We are considering around 1,000 bins for each encoded name.
- For both first and surname, encode the repaired version of name.
- For first name, we will encode a standardised version of name. i.e. names are standardised using a nickname index prior to encoding (e.g. Jonathan, Jonathon, Jonno, are standardised to John)
- Mapping will occur primarily through a modulo function, the caveat to this being that a small subset of names will be collapsed together prior to running through the function in order to add a further level of protection to the most frequent one or two names.
- Use of the separation principle and functional separation to separate linkage variables from analysis data increases the security of the data.
- Implementing a segregated and restricted IT environment for data linking in the next couple of months.
- Encryption-at-rest is being considered in a broader organisational context, and will add protection when data is not being actively used. Regardless of the solution, encryption of linkage data and unit record data are key controls. Also, the separation of data must be robust – potentially name data (bins) and linkage data could be stored offline when not being actively used. A potential pitfall with encryption is the management of secret keys. If secret keys are not securely held, attackers can gain access or keys can be lost rendering data inaccessible.

7. CONCLUDING REMARKS

What we are proposing in this paper for Census 2016 name information is essentially an extension to the lossy encoding scheme employed for the 5% of 2011 Census records that are selected in the Australian Census Longitudinal Dataset (ACLD) sample. The extension being that we are now intending to reduce the number of names in each bin (to improve usefulness in linking), retain more encoded variables corresponding to alternative forms of name (to account for fuzzy matching), and retain encoded forms of name for every Census record, not just 5%. This will be implemented in conjunction with an increase in other forms of data protection including IT controls, a separate IT environment and the functional separation. Importantly, lossy encoding increases the protections compared to raw name and align with our public statements of irreversibility, yet still yields a potentially useful linking variable that can support higher quality record linkage.

REFERENCES

Australian Bureau of Statistics (2015) Retention of names and addresses collected in the 2016 Census of Population and Housing <http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Retention+of+names+and+addresses+collected>

Australian Bureau of Statistics (2016a) Privacy, Confidentiality & Security Statement < http://www.abs.gov.au/websitedbs/censushome.nsf/home/privacy?opendocument&navpos=130>

Australian Bureau of Statistics (2016b) Personal Income Tax and Migrants Integrated Dataset (PITMID) 2011-12 Quality Assessment, *ABS Research Papers*, cat. no. 1351.0.55.060, Canberra.

Bishop, G. (2007). Assessing the Likely Quality of the Statistical Longitudinal Census Dataset, *ABS Research Papers*, cat. no. 1351.0.55.026, ABS, Canberra.

Christen, P. (2012) Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, *Springer Science and Business Media*.

Culnane, C., Teague, V. and Rubenstein, B.I.P, (2018) Options for encoding names at the Australian Bureau of Statistics, *University of Melbourne, In press* (paper can be provided upon request).

Fellegi, I. and Sunter, A. (1969). A theory for record linkage, *Journal of the American Statistical Association*, 64 (328), 1183-1210.

Herzog, T.N, Scheuren, F.J. and Winkler, W.E. (2007) Data quality and record linkage techniques, *Springer Science and Business Media*.

Wikipedia article on Information Security, https://en.wikipedia.org/wiki/Information_security,
(Version last edited on 2 October 2017, at 07:03,
viewed on 03 October, 2017 at 9:42 AM AEST).

FOR MORE INFORMATION . . .

www.abs.gov.au the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FAX 1300 135 211

EMAIL client.services@abs.gov.au

PHONE 1300 135 070

FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

Produced by the Australian Bureau of Statistics