Journal of Statistical Science and Application 5 (2017) 81-90

Privacy Protection for Big Data Linking using the Identity Correlation Approach

Kevin McCormack, Mary Smyth
Central Statistics Office, Cork, Ireland

Privacy protection for big data linking is discussed here in relation to the Central Statistics Office (CSO), Ireland's, big data linking project titled the 'Structure of Earnings Survey – Administrative Data Project' (SESADP). The result of the project was the creation of datasets and statistical outputs for the years 2011 to 2014 to meet Eurostat's annual earnings statistics requirements and the Structure of Earnings Survey (SES) Regulation. Record linking across the Census and various public sector datasets enabled the necessary information to be acquired to meet the Eurostat earnings requirements. However, the risk of statistical disclosure (i.e. identifying an individual on the dataset) is high unless privacy and confidentiality safe-guards are built into the data matching process. This paper looks at the three methods of linking records on big datasets employed on the SESADP, and how to anonymise the data to protect the identity of the individuals, where potentially disclosive variables exist.

Keywords: Big Data Linking, Data Matching, Data Privacy, Data Confidentiality, Identity Correlation Approach, Data Disclosure, Data Mining

Introduction

Work presented in this paper is based on a big data project carried out by the Central Statistics Office (CSO), Ireland, titled the 'Structure of Earnings Survey – Administrative Data Project' (SESADP) [1], [2]. The SESADP was designed to meet the Eurostat SES Regulation [3], replacing the National Employment Survey (NES) [4], [5]. This paper looks at three methods of data linking employed in the SESADP project. It then focuses on methods of anonymising the data to protect the confidentiality of the individuals. The three methods of data linking are:

- Method 1 - Data linking with a Unique Identifier
- Method 2 - Data linking by String matching
- Method 3 - Data linking with the Identity Correlation Approach

In summary, the SESADP linked public sector datasets using the Unique Identifier of the social security number (called the PPS No.[1]) to create the master administrative dataset (ADS). This master ADS dataset was then linked to the Census dataset using the Identity Correlation Approach (ICA) (developed by McCormack and Smyth, 2015) [6], [7]. A string matching exercise was also compared with the Identity Correlation Approach (ICA), to demonstrate the power of the ICA in big data linking.

[1] The Personal Public Service Number (PPSN) is a unique reference number that allows individuals access social welfare benefits, public services and information in Ireland. State agencies that use PPSNs to identify individuals include the Department of Social Protection, the Revenue Commissioners and the Health Service Executive (HSE).

Methods for anonymisation of the datasets to protect the individual's confidentiality are detailed in this paper for the three methods used in the data linking process[2]. The analysts linking the datasets did not have access to individual names and addresses, nor potentially disclosive variables such as date of birth. This SESADP project demonstrated the power of the ICA for big data linking and the ICA's unique feature in enabling variables to be encrypted without compromising the power of the linking process.

Method 1- Data linking with a Unique Identifier

An example of records being matched using a unique identifier (UI) is shown in Fig. 1a. Records on each of the datasets linked contained the unique identifier (UI) of the Social Security Number (PPS No.). This linking process is very simple. The records are simply linked by the PPS No. There are no duplicates as each record has a unique identifier (UI). Also, there are no 'false positives' (incorrectly matched records) in the data linking process as the UIs match precisely, with no possibility of error. The records on each dataset can be linked to any number of datasets to acquire additional information about the record, provided the UI is a variable on each of the datasets. Fig. 1a shows the person's record on Dataset A with the variable for income can be linked to Dataset B (using the PPS No.). Linking to Dataset B adds the additional variable date of birth to each record on Dataset A. Again by further linking to Dataset C, another two variables (gender, education level) is added to each record. This allows the analyst to look at income distributions by age, gender and education level, thus adding greater insight to the information on persons' incomes.

Dataset A		
PPS No.	Name	Income
ABC1234568	Fiona Beale	40,000
SYZ1234569	Jane Eyre	71,000
LMN1234570	Tom Sawyer	39,000
QAT1234571	JK Rowling	85,000

Dataset B	
PPS No.	Dob
ABC1234568	30/09/1955
SYZ1234569	11/10/1985
LMN1234570	17/02/1991
QAT1234571	25/08/1961

Dataset C		
PPS No.	Gender	Education level
ABC1234568	F	6
SYZ1234569	F	9
LMN1234570	M	7
QAT1234571	F	8

Figure 1a. Individual Datasets.

Dataset A		
PPS No.	DPK (Encrypted PPS No.)	Income
ABC1234568	101951	40,000
SYZ1234569	135673	71,000
LMN1234570	581001	39,000
QAT1234571	701481	85,000

Dataset B	
DPK (Encrypted PPS No.)	Dob
101951	30/09/1955
135673	11/10/1985
581001	17/02/1991
701481	25/08/1961

Dataset C		
DPK (Encrypted PPS No.)	Gender	Education level
101951	F	6
135673	F	9
581001	M	7
701481	F	8

Link records using DPK

Datasets Linked				
DPK (Encrypted PPS No.)	Income	Dob	Gender	Education level
101951	40,000	30/09/1955	F	6
135673	71,000	11/10/1985	F	9
581001	39,000	17/02/1991	M	7
701481	85,000	25/08/1961	F	8

Figure 1b. Datasets encrypted to prevent PPS No. being disclosed before linking process.

Dataset A		
PPS No.	DPK (Encrypted PPS No.)	Gender
ABC1234568	123αβγδεζθ	F

key

1=α, 2=β, 3=γ, 4=δ, 5=ε, 6=ζ, 7=η, 8=θ, 9=ι.

A=1, B=2, C=3, etc.

Figure 1c. Encrypting a variable.

[2] Synthetic data is used to demonstrate the SESADP concepts in this paper.

Method 1 - Data Protection for UIs

In the example given in Fig. 1a, it is necessary for data protection and confidentiality reasons, to remove any potentially disclosive variables to safe-guard information which could potentially identify the person. Therefore the variable with the person's name is removed in the matching process in Fig. 1b. Confidentially is further enhanced by encrypting the PPS No. Data linking was then carried out using the encrypted PPS No. known as the DPK (Data Protected Key), by the analyst, without the analyst having access to names nor the actual PPS No. Removing all the sensitive information, while still allowing the analyst to link and analyse the data enabled the master administrative dataset (ADS) to be created.

Data matching using a UI is easily protected. Firstly, the person's name is not a required variable for analysis nor data linking. Therefore the variable NAME is dropped. Secondly, since the PPS No. is required to link the datasets, the PPS No. can be encrypted as shown in Fig. 1b. The only requirement for encrypting the PPS No. is to use a consistent method which preserves the uniqueness of each PPS No.

In order to protect the PPS No. from being disclosed, the PPS No. is encrypted using a Data Protected KEY (DPK). The DPK is simply the PPS No. encrypted so that analysts can link the data and perform analysis without having access to the actual PPS No. The simplest method is to give each unique PPS No. a unique random number. In Fig. 1b each of the three original datasets contain records with a unique PPS No. The PPS No. on each dataset is encrypted with a Data Protected KEY. In Fig. 1b the DPK is a unique random number. The PPS No. is then removed from all three datasets and the DPK is used as the unique identifier for each record. The records are linked using only the DPK therefore the analyst does not have access to an individual's PPS No. This permits data mining without revealing too much sensitive information [8].

Another method of encryption shown in Fig. 1c, is to assign a value for each of the characters in the PPS No. string (e.g. A=1, B=2, C=3, 1=α, 2=β, 3=γ, 4=δ, 5=ε, 6=ζ, 7=η, 8=θ, 9=ι). Although this method has a high degree of security and is adequate for most research projects, it is not as secure as assigning a random number as the encryption method follows a pattern. Once the pattern is identified, then all records can be decrypted. In both encryption methods the key for encrypting the UI is held by the data custodian, who encrypts all datasets that contain the UI. The encryption method is not disclosed to the analyst so the analyst cannot perform unauthorised data linking exercises with datasets. The data custodian is not permitted to link datasets, therefore confidential information is not disclosed.

Method 2 - Data linking by String matching

One of the commonest methods of data linking for large datasets is matching a character variable common to both datasets. This is known as 'String matching'. Records are matched by linking the character strings of a variable, such as a person's name, on one dataset with the corresponding name on another dataset [9]. This is demonstrated in Fig. 2a, where three datasets are matched using the person's name[3]. As the variable 'name' contains all records which are unique, the matching process is very simple. Each record is matched to the exact same spelling of the character variable 'name', which is the person's name.

In situations where the names are non-unique (e.g. Fig. 2b where there are two people with the same name) then this compromises the record linking process. If the persons with the same name are incorrectly linked then this leads to false positives i.e. the person is not linked to their correct record on the second dataset and

[3] The string variable used in the SESADP was the enterprise name[7]. This example of a person's name is used to demonstrate the concept.

therefore are assigned incorrect values for variables such as date of birth and income. False positives in the data linking process are shown with the red dashed arrows in Fig. 2b. The blue arrows indicate the correct string match. Methods for data linking using string variables are discussed in the literature, and various methods of avoiding false positives are discussed [10], [11].

Data Social Security			Dataset Census			Dataset Tax returns		Link records	Datasets Linked				
DPK (Encrypted PPS No.)	Name	Gender	Name	Education level	Occupation	Name	Income		Name	Gender	Occupation	Education level	Income
10195	Fiona Beale	F	Fiona Beale	6	Actor	Fiona Beale	40,000		Fiona Beale	F	Publisher	6	40,000
135673	Jane Eyre	F	Jane Eyre	9	Publisher	Jane Eyre	71,000		Jane Eyre	F	Actor	9	71,000
58100	Tom Sawyer	M	Tom Sawyer	7	Politician	Tom Sawyer	39,000		Tom Sawyer	M	Politician	7	39,000
70148	JK Rowling	F	JK Rowling	8	Author	JK Rowling	85,000		JK Rowling	F	Author	8	85,000

Figure 2a. String matching to link datasets (unique).

Data Social Security			Dataset Census			Dataset Tax returns		Datasets Linked				
DPK (Encrypted PPS No.)	Name	Gender	Name	Education level	Occupation	Name	Income	Name	Gender	Occupation	Education level	Income
10195	Fiona Beale	F	Fiona Beale	6	Actor	Fiona Beale	40,000	Fiona Beale	F	Actor	6	40,000
135673	Jane Eyre	F	Jane Eyre	9	Publisher	Jane Eyre	71,000	Jane Eyre	F	Publisher	9	71,000
58100	T. Sawyer	M	Tom Sawyer	7	Politician	T. Sawyer	39,000	T. Sawyer	M	Politician	7	39,000
379154	T. Sawyer	F	Teresa Sawyer	8	Nurse	T. Sawyer	44,000	T. Sawyer	F	Nurse	8	44,000
70148	JK Rowling	F	Joanne Rowling	8	Author	Joanne Rowling	85,000	JK Rowling	F	Author	8	85,000
513456	JK Rowling	M	John Rowling	7	Mechanic	JK Rowling	42,000	JK Rowling	M	Mechanic	7	42,000

Figure 2b. String matching to link datasets (non-unique).

Method 2 - Data Protection for string matching

Variables with character strings such as a person's name or address are highly sensitive and must be handled securely to protect the person's privacy, and to comply with data protection laws [12]. In order to maintain data confidentiality and ensure the privacy of the individual, names and addresses should only be maintained on datasets where it is absolutely necessary. The ideal solution to data privacy is to remove character strings such as a person's name from the dataset before permitting data matching by analysts, while preserving linkability.

A solution to replacing string characters (e.g. a person's name) is to replace the name with a random number, as discussed in Method 1 above for UIs. This would enable all names to be removed from the datasets before the matching process begins. This method is dependent on the individuals' names being unique i.e. two individual's do not have the same name. The variable for individuals' names is encrypted on all datasets by the custodian with a unique random number. Then the analyst does not have access to individuals' names in the data linking process. If non-unique names exist (two people with the same name) then the individuals are identified by another variable such as Date of birth and given a unique name e.g. JK Rowling1 and JK Rowling2. As these two character string records are now different they are given a different random number in the encryption process. Linking can then proceed with the unique random numbers, when the names are removed from the datasets. Large datasets which contain multiple non-unique records for the string variable 'name' may render it difficult to identify all of the records with the same names. In this case the non-unique names would get the same random number in the encryption process, even though the individual records are different people. In this case it would be necessary to separate the datasets into records which had unique names and those that contained non-unique names. Encryption of the unique names could proceed, however it would be difficult to encrypt the non-unique names without another method such as 'Blocking'[11].

Method 3 – Data linking with the Identity Correlation Approach

The Identity Correlation Approach (ICA) is a mathematical solution for big data linking developed by McCormack and Smyth[6], [7], in the absence of a direct linkable UI. The Identity Correlation Approach (ICA) allows for record linking across datasets without the need for string variables, where a unique identifier does not exist. False positives are eliminated as a design feature of the method, and the probability of direct matches can be calculated beforehand using the Matching Rate for Unique Identifier (MRUI) formula. An innovative feature of the Identity Correlation Approach (ICA) is data security and confidentiality are very strong compared with string matching. The ICA approach does not require names nor addresses to be held on big datasets. Individual identification variables such as 'Date of Birth' can be replaced with a Data Protected Key' (DPK) for matching purposes.

Identity Correlation Approach

The Identity Correlation Approach (ICA) was developed as part of the SESADP big data matching project carried out by the Central Statistics office, Ireland [6], [7]. The aim of the SESADP was to produce data to meet the EU SES 2014 Regulation [3], from administrative data sources, and to meet annual earnings statistics requirements for 2011 to 2014. This replaced an expensive business survey, called the National Employment Survey (NES) conducted each year [4], [5].

The ICA involves combining a number of individual variables for each person until a unique identifier is arrived at. An example of this is combining the individual characteristics of each person on the 2011 Census Dataset for Ireland. Beginning with the variable for *date of birth*, then combine it with the variable *gender*, then adding variable for *county,& marital status, etc.* until a unique identifier is arrived at for each person. This is illustrated in Table 1 below.

Theoretical Application

There were 1.6m employees in Ireland in 2011. Of these, an average of 65,000 persons were born in the same year (years 1946 to 1995), as illustrated in Table 1.

Table 1

Identity Correlation Approach: Simple Model - Combining Variables

Operation	Variable	No. of Records
	Approx. No. of births each year	65,000
Divide by:	No. days in the year	365
	No. Persons with same DoB	178
Divide by:	Gender	2
	No. Persons with same DoB and gender	89
Divide by:	No. Counties	26
	No. Persons with same DoB, Gender, County	3
Divide by:	Marital Status (married & other)	2
	No. Persons with same DoB, Gender, County, marital status	1

The 65,000 persons born in the same year are divided by 365 days in the year to give approximately 178 persons with the same date of birth. The 178 persons with the same date of birth (DoB) can be divided by 2 for gender, to give 89 persons with the same DoB and gender. Dividing by the no. of counties a person lives in (89 divided by 26) results in 3 persons with the same DoB, gender & county. The 3 persons can be further

subdivided by marital status resulting in 1 person with the same DoB, gender, county and marital status. Other variables used to further breakdown the data are industrial sector (NACE code), no. of dependent kids, etc. A unique combination of variables for each person allows a person to be uniquely identified. This method is termed the Identity Correlation Approach (ICA).

Practical Application - Census 2011 data

The identity Correlation approach was applied to the Irish Census Data 2011, as described above. This allowed for a Unique Identifier (UI) to be applied to each individual by combining their personal characteristics (i.e. DoB, gender, county residence, etc.). The unique identifier is called the matching variable (matchvar) which is used to link each person's record to other datasets (see Table 2).

Table 2

Applying Identity Correlation Approach to Create Unique Identifier (Matchvar)

Date_of Birth	Gender	County	NACE	Marital Status	No.Child	Matchvar (all variables)
15031949	M	CORK	42	M	0	15031949MCORK42M 0
11021945	F	LIMERICK	31	S	1	11021945FLIMERICK31S1
21111954	M	DUBLIN	25	D	2	21111954MDUBLIN25D2
19051964	M	CARLOW	55	O	2	19051964MCARLOW55O2
22091966	M	GALWAY	82	M	3	22091966MGALWAY82M3
24031971	F	CAVAN	84	M	0	24031971FCAVAN84M0

Public Sector Administrative Datasets (ADS)

A single master Administrative Dataset (ADS) was created by linking a number of Public Sector Administrative Datasets (Revenue Commissioners Tax data, Social Security Administrative Datasets and the CSO's Administrative Datasets (e.g. Central Business Register (CBR), Earnings and Labour Force Survey)). These datasets were combined using the PPS No. for each individual and the CBR Enterprise No. for Establishment Surveys. The Identity Correlation Approach was applied to the master Administrative Dataset (ADS) also, Allowing for a Unique Identifier (UI) to be applied to each individual by combining their personal characteristics (i.e. DoB, gender, county residence, etc.). This Unique Identifier known as the match variable (matchvar) was then used to link to the UI (matchvar) in Census.

Linking Census to ADS

Variables common to both the Census dataset and the master Administrative Data Source (ADS) were identified (e.g. DoB, gender, etc.). These common variables were joined to each other to create a Unique Identifier on each dataset using the Identity Correlation Approach. By linking the two datasets using the Unique Identifier, a PPS No. could be applied to each individual person in the Census. This is shown in Table 2. Once the PPS No. was assigned to the Census dataset, it enabled Census data to be linked to any Public Sector Administrative Dataset.

Dataset Matching - Census dataset

A total of 1.6 million employee records were identified from the 4.6 million persons on Census Records. Approximately 700,000 records had a unique Business No. identifier attached (CBR No.). The first matching variable (Matchvar1) created for Census used the following variables combined: CBR No., Dob, gender, county, NACE 2, marital status, No. children. A second matching variable was created (Matchvar2) excluding NACE2 (see Table 3). Up to ten matching variables (Matchvar1 – Matchvar10) were created. Each matching variable is

similar to the previous one with one change to the composition variables for each subsequent matching variable created. Table 3 illustrates the construction of each subsequent matching variable.

Table 3

Matching Variables

Date_of Birth	Gender	County	NACE	Ent No.	Marital Status	No.Child	Match Var 1	Match Var 2
15031949	M	CORK	42	EN12345678	M	0	15031949MCORK42EN12345678M0	15031949MCORK42EN12345678M
11021945	F	LIMERICK	31	EN52345679	S	1	11021945FLIMERICK31EN523456791	11021945FLIMERICK31EN52345679S
21111954	M	DUBLIN	25	EN52795680	O	2	21111954MDUBLIN25EN527956802	21111954MDUBLIN25EN52795680O
19051964	M	CARLOW	55	EN32795681	D	2	19051964MCARLOW55EN327956812	19051964MCARLOW55EN32795681D
22091966	M	GALWAY	82	EN22795682	M	3	22091966MGALWAY82EN227956823	22091966MGALWAY82EN22795682M
24031971	F	CAVAN	84	EN52795683	M	0	24031971FCAVAN84EN527956830	24031971FCAVAN84EN52795683M
28021977	F	DUBLIN	71	EN84355684	S	1	28021977FDUBLIN71EN843556841	28021977FDUBLIN71EN84355684S
30061990	F	KERRY	35	EN73795687	M	1	30061990FKERRY35EN737956871	30061990FKERRY35EN73795687M

Mathematical Representation of Identity Correlation Approach (ICA)

Creating a Unique Identifier (UI) for each record using the Identity Correlation Approach (ICA) will result in a perfect match of records across two datasets, if there is a sufficient overlap of variables on both datasets. The probability of matching records across two datasets can be calculated by a formula known as the Matching Rate for Unique Identifier (MRUI). This is illustrated in Equation 1.

Eqn.1 Matching Rate for Unique Identifier (MRUI)

$$\boldsymbol{N \times \frac{1}{V1_{ui}} \times \frac{1}{V2_{ui}} \times \frac{1}{V3_{ui}} \times \frac{1}{V4_{ui}} \times \ldots\ldots\ldots \times \frac{1}{VX_{ui}} = MRUI}$$

(Assumes records are distributed evenly across all classes) (McCormack & Smyth, 2015)

where:

N = Population , V = Variable, x = no. of variables

ui = Uniqueness Factor. ui = no. of classes where variable is distributed evenly across all classes

(Classes in a variable do not contain an even distribution of records)

If records are not distributed evenly across all classes then ui is replaced by di.

where:

di = adjusted Uniqueness Factor = Proportion of records occurring within the largest class of the variable (where a variable does not have records evenly distributed across all classes).

MRUI Properties

The Matching Rate for Unique Identifier (MRUI) is the ability to identify a unique record in a dataset, given the combination of variables used to deduce the record.

Mathematically it is assumed that variables are discreet (non-dependent)

MRUI = 1 , then there exists a unique identifier variable for each record, allowing a direct match to the record in the dataset.

MRUI < 1, then there exists a unique identifier variable for each record and there are additional variables to increase the confidence in the matching process for each record.

MRUI > 1, then no unique Identifier variable exists and there will be duplicate records in the matching process

Application of MRUI Equation

In this example: N = Population = 65,000 employees born in same year

Variable	Symbol	Factor	Description
V1 = DoB	$V1_{di}$	365	di = 363 (days of the year)
V2 = Gender	$V2_{di}$	2	2 genders (approx. 50% split)
V3 = NACE 2 digit code	$V3_{di}$	10	50 different NACE2 digit codes, but approx. 1/10 of pop. in dominant NACE 2 code
V4 = marital status	$V4_{di}$	2	6 Marital status codes, Approx. 1/2 of people married, other classes = 50% (e.g. single, divorced, separated, etc.)
V5 = county	$V5_{di}$	5	20 counties, but one dominant county with 1/5 of the population

Put above values into aMRUI equation:

$$\mathbf{65,000} \times \frac{\mathbf{1}}{\mathbf{365}} \times \frac{\mathbf{1}}{\mathbf{2}} \times \frac{\mathbf{1}}{\mathbf{10}} \times \frac{\mathbf{1}}{\mathbf{2}} \times \frac{\mathbf{1}}{\mathbf{5}} = \frac{\mathbf{65,000}}{\mathbf{73,000}} = \mathbf{0.89}$$

MRUI = 0.89

In this example the **MRUI** ≤ 1, this indicates that there is a unique identifier for the individual employee in the dataset. When the **MRUI** = 1 then there is a unique identifier for the individual. Since the **MRUI** < 1 here, it means that there is added assurance from the data that an individual has been identified from the variables. This is useful to know if there is an issue around how the data is coded.

Method 3 - Data Protection for the Identity Correlation Approach (ICA)

An innovative feature of the Identity Correlation Approach (ICA) is data security and confidentiality can be built into the data linking process without compromising data matching in any way [13]. The ICA approach does not require names nor addresses to be held on big datasets. Individual identification variables such as 'Date of Birth' can be replaced with a Data Protected Key' (DPK) for matching purposes.

Fig. 3b shows that any of the variables can be encrypted. This does not affect the data linking process, as the method still results in the creation of a unique identifier, by joining all the encrypted variables.

The first variable '*Date_of_Birth'* is encrypted to give the age group (Age1, etc.) only. If more detail is required then the day and the month in the date of birth variable can also be encrypted.

Encrypting the second variable, '*gender'*, gives a value of X for males and Y for females. Any arbitrary value can be assigned to encrypt the gender.

'County of residence' is the third variable and each county is assigned an encrypted value such as Co1, Co2, etc.

The variable '*NACE economic sector'* of the enterprise is also assigned an encrypted DPK, as shown for the fourth variable

The fifth variable is the '*Enterprise no.'* on the CSO's Business Register. This can also be encrypted by giving it a random no. such as EN1, etc.

'Marital status' is similarly encrypted as shown in the sixth variable, with the value MA arbitrarily assigned to married status, MB assigned to single status, etc.

Finally, the seventh variable for '*No. of children*' is assigned a random value to protect the identity of the person.

As stated above for Unique Identifiers, the key for encrypting the value of each variable is held by the data custodian, who encrypts all datasets that contain the specific variables. The encryption method is not disclosed to the analyst so the analyst cannot perform unauthorised data linking exercises with datasets. Any variable can be encrypted or decrypted, depending on the requirements of the analyst and the necessity of the project, to protect data privacy.

	Date of Birth	Gender	County	NACE	Enterprise No.	Marital Status	No. Children	Unique Identifier
Record 1	15031949	M	CORK	42	EN12345678	M	0	15031949MCORK42EN12345678M0
Record 2	11021945	F	LIMERICK	31	EN52345679	S	1	11021945FLIMERICK31EN523456791
Record 3	21111954	M	DUBLIN	25	EN52795680	O	2	21111954MDUBLIN25EN527956802
Record 4	19051964	M	CARLOW	55	EN32795681	D	2	19051964MCARLOW55EN327956812
Record 5	22091966	M	GALWAY	82	EN22795682	M	3	22091966MGALWAY82EN227956823

Figure 3a. Dataset variables matched with ICA method.

	Date of Birth	Gender	County	NACE	Enterprise No.	Marital Status	No. Children	Unique Identifier
Record 1	Age2	X	Co1	N1	EN1	MA	A	Age2XCo1N1EN1MAA
Record 2	Age1	Y	Co2	N2	EN2	MB	B	Age1YCo2N2EN2MBB
Record 3	Age3	X	Co3	N3	EN3	MC	C	Age3XCo3N3EN3MCC
	etc.							

Figure 3b. Dataset variables encrypted and matched with ICA method

Results of ICA Project

Table 4

Results of Quality check on Census Matching of Names

Matched Surname	497
Matched Birthname	34
Matched Double Barrelled name	1
Matched Irish spelling of name	1
Minor Spelling/OCR error	5
Non-matched*	11
Blank (no census name)	1
Total no. of names checked	550

*of the 11 records non=matched only 3 looked completely wrong

The SESADP project successfully matched 50% of the employee population in Ireland to the Census. A quality check was carried out on a sample of 550 employees in the SESADP project to assess if the correct person on Census was indeed mapped to their correct records on the Public sector datasets. The data custodian was given a sample of 550 employees with two variables: (1) the encrypted PPS No. and (2) the Census Id no. Matching the PPS No. to the public sector dataset allowed the data custodian to extract the persons name. Similarly, using the Census Id No. the person's name was extracted from Census. When both names were compared the matching rate was 98%. The ICA method was precise in matching the employee, which would not have been possible using string matching as some employees used their Irish translated name on Census and some used their maiden name. The results of the quality checks are given in Table 4.

Conclusion

Data privacy and confidentiality can be protected across all three methods by encrypting the variables and only allowing the variables required for analysis to be disclosed. Unique Identifiers allow datasets to be linked without the need to disclose any confidential names or addresses. Where UIs do not exist string matching is often required for data linking. String variables such as a person's *name*, should be encrypted to prevent statistical disclosure before the data matching exercise begins. Of the three data linking methods, the ICA method is the most powerful as it enables data linking in the absence of a unique identifier. The probability of linking the data is calculated using the MRUI formula, before any time or resources are invested in the data linking project.

In terms of data confidentiality and security, the ICA method allows all variables to be encrypted without compromising the quality of the data linking process as data linking in the ICA method is based on mathematical probability. Only the variables required for the specific analysis need be disclosed to the analyst. Protected variables are not disclosed to the analyst, and only the data custodian holds the key to decrypting the variables.

References

[1]. McCormack,K (2015). "Constructing structural earnings statistics from administrative datasets: Structure of earnings survey – Administrative data project". *The Statistics Newsletter – OECD*, 62, 3-5.

[2]. McCormack,K. & Smyth,M. (2015). "Constructing structural earnings statistics from administrative datasets". *New Techniques and Technologies for Statistics (NTTS) 2015. Collaboration in Research and Methodology for Official Statistics.*

[3]. "Council Regulation (EC) No 530/1999 of 9 March 1999 concerning structural statistics on earnings and on labour costs". *OJ L 63, 12.3.1999, p. 6.*

[4]. "National Employment Survey 2008 and 2009". (2011). *Central Statistics Office, Ireland. www.CSO.ie*

[5]. McCormack,K. & Smyth,M. (2015). "Specific Analysis of the Public/Private Sector Pay Differential for National Employment Survey 2009 & 2010 Data". *Research Paper. Central Statistics Office, Ireland. www.CSO.ie*

[6]. McCormack,K. & Smyth,M. (2016). "Big Data Matching Using the Identity Correlation Approach". *Proceedings of the First International Conference on Advanced Research Methods and Analytics, CARMA 2016. Valencia, Spain.*

[7]. McCormack,K. & Smyth,M. (2017). "A Mathematical Solution to String Matching for Big Data Linking". *Journal of Statistical Science and Application (ISSN 2328-224X, USA). Vol. 5 (2017) pp. 39-55.*

[8]. Hall,R. & Fienberg,S.E. (2010). "Privacy-Preserving Record Linkage". PSD'10 Proceedings of the 2010 international conference on Privacy in statistical databases. Corfu, Greece. pp.269-283

[9]. Dusetzina SB, Tyree S, Meyer AM, et al. (2014). "Linking Data for Health Services Research: A Framework and Instructional Guide" [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); https://www.ncbi.nlm.nih.gov/books/NBK253312/

[10]. Fellegi, Ivan; Sunter, Alan (1969). "A Theory for Record Linkage". *Journal of the American Statistical Association.* ***64*** *(328): pp. 1183–1210.*

[11]. Cohen, W. W., Ravikumar, P. & Fienberg, S. E. (2003). "A Comparison of String Metrics for Matching Names and Records". Paper presented at the meeting of the SIGKDD, 2003.

[12]. Trepetin, S. (2008) "Privacy-Preserving String Comparisons in Record Linkage Systems: A Review". *Information Security Journal: A Global Perspective Vol. 17. ISS 5-6. pp. 253-266*

[13]. Vatsalan,D., Sehili,Z., Christen,P., Rahm,E. (2017). "Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges". Handbook of Big Data Technologies. Springer. https://www.semanticscholar.org/paper/Privacy-Preserving-Record-Linkage-for-Big-Data-Vatsalan-Sehili/e87f503e838ce7c2ae8a90bfa3cda7bbac8c919a#citingPapers