Journal of Mathematics and System Science 11 (2021) 18-29
doi: 10.17265/2159-5291/2021.02.002

Cyber Security — Protecting Personal Data

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

Abstract: Many organizations have datasets which contain a high volume of personal data on individuals, e.g., health data. Even without a name or address, persons can be identified based on the details (variables) on the dataset. This is an important issue for big data holders such as public sector organizations (e.g., Public Health Organizations) and social media companies. This paper looks at how individuals can be identified from big data using a mathematical approach and how to apply this mathematical solution to prevent accidental disclosure of a person's details. The mathematical concept is known as the "Identity Correlation Approach" (ICA) and demonstrates how an individual can be identified without a name or address using a unique set of characteristics (variables). Secondly, having identified the individual person, it shows how a solution can be put in place to prevent accidental disclosure of the personal details. Thirdly, how to store data such that accidental leaks of the datasets do not lead to the disclosure of the personal details to unauthorized users.

Key words: Data protection, big data, identity correlation approach, cyber security, data privacy.

1. Introduction

Identification of a person in a large dataset is possible without their name or address, if there is sufficient information about the person contained in the dataset. This paper shows firstly, how a person can be identified with a unique set of characteristics (variables), if the dataset is accidentally leaked. Exposure of personal information due to data breaches is well documented [1]. This paper presents a method for protecting individuals' personal data in big datasets to prevent accidental disclosure of confidential records. Building on the Identity Correlation Approach (ICA) [2] which was developed for matching big datasets, the concept presented here reverses the ICA process to ensure that a Unique Identifier (UI) cannot be developed for a single record. This ensures that individual records of a personal nature are kept confidential and that no confidential data is disclosed when datasets are released to researchers, made publicly available or accidentally leaked to unauthorized users. The method provides a quick and powerful solution to identify the individual's unique personal details which could potentially allow them to be identified. Encryption of variables is also demonstrated to show that the data can be protected, without inhibiting record linkage across datasets.

Corresponding author: Mary Smyth, Ph.D.; research field(s): big data, GIS, mathematical modelling. E-mail: mary.smyth@cso.ie.

2. Identity Correlation Approach

Data matching using the mathematical concept of the Identity Correlation Approach (ICA) is based on the concept of creating a Unique Identifier (UI) for each record, where a UI does not exist within a particular dataset. This UI is created by combining a number of variables. If a UI can be created for each record in a dataset, then the dataset can be matched to another dataset. This UI is created from existing variables in the dataset.

Whether or not a UI can be created from existing variables in a dataset is determined by the MRUI (Matching Rate for Unique Identifier) equation. If the MRUI equation is less than or equal to 1 then a UI can be created for each record in the dataset and the dataset can be matched based on the UI.

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 "Protected Identity Key" (PIK) for matching purposes.

2.1 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 (2015), in the absence of a direct linkable unique identifier (UI). The Identity Correlation Approach (ICA) allows for record linking across datasets without the need for string variables [3, 4], where a unique identifier does not exist. False positives are eliminated as a design feature of the method, and the success rate of direct matches can be calculated beforehand using the Matching Rate for Unique Identifier (MRUI) formula. This is different to most data linking projects which mainly involve algorithms which are based on records being matched directly (deterministic) or the probability of a match [5].

An innovative feature of the Identity Correlation Approach (ICA) is data security and confidentiality are very strong compared with string matching [6]. 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 "Protected Identity Key" (PIK) for matching purposes.

2.2 Data Linking Project

The Identity Correlation Approach (ICA) was developed as part of the SESADP big data matching project carried out by the Central Statistics office, Ireland [7, 8]. The aim of the SESADP was to produce data to meet the EU SES 2014 Regulation [9], 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) [10, 11] conducted each year.

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 of Population (COP) 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.

The Identity Correlation Approach (ICA) has been developed as a direct response to the challenge of linking administrative data sources which do not contain Unique Identifiers at the level of the individual [12].

A unique identifier is derived, within a probability environment for each individual on a data source by combining or merging in sequence a number of known demographic variables. For example, with reference to the 2011 COP, the known demographic and industrial sector variables are:

- date of birth,
- gender,
- county of residence,
- marital status,
- NACE industrial sector,

which are combined until a unique identifier is derived for each individual person (see Table 1).

2.3 ICA – Basic Model

Core to the ICA Basic Model (ICA-BM) is the determination of the probability of identifying an individual having a set of unique demographic, marital and regional characteristics.

The determination of the probabilities associated with the ICA process has a number of stages:

1) The first stage of the ICA process is the determination of the average number of individuals born in each year from 1946 to 1995 (16 years and older), which is estimated to be 65,000. The figure of 65,000 persons per annum is derived by dividing the 4.6 million

Table 1 Identity correlation approach: basic model — combining variables.

Operation	Variable	No. of Records
	Approx. No. of births each year	65,000
Divide by:	No. days in the year	365
Derived variable	No. Persons with same DoB	178
Divide by:	Gender	2
Derived variable	No. Persons with same DoB and gender	89
Divide by:	No. Counties	26
Derived variable	No. Persons with same DoB, Gender, County	3
Divide by:	NACE industrial code	15
Derived variable	No. Persons with same DoB, Gender, County and NACE	<1

population in Ireland into two categories: (1) category of "year of birth", (2) category of "employees only".

2) In the second stage, the estimated number of individuals born in a particular year (65,000) is divided by the no. of days in year (365) which provides an estimate of the number of individuals with the same date of birth (178). The assumption underlying this calculation is that the births are evenly distributed over the 365 days of the year.
3) The 178 persons with the same date of birth (DoB) are further divided into two gender groups (male and female), which provides an estimate of the number of individuals with the same DoB and gender (89).
4) It is assumed that the births in any particular year are evenly distributed by geographical location (26 county regions in the case of Ireland). The estimated number of individuals with the same DoB and gender (89) are divided by 26, which provides an estimate of three persons with the same DoB, gender and county.
5) The three persons with the same DoB, gender & county can be further divided by NACE sector (15 categories), which provides a unique identifier of one person with the same DoB, gender, county and NACE sector.

The stages involved in the ICA – Basic Model are summarized in Table 1.

2.4 MRUI Equation

An equation was developed to calculate if a UI (unique identifier) can be created from a dataset using the ICA approach. This equation is called the **Matching Rate for Unique Identifier** (**MRUI**) equation (see Eq. (1)). The **MRUI** equation (Matching Rate for Unique Identifier) is applied to a dataset. If the **MRUI** ≤ 1, then a UI can be created for each record (person) in a dataset.

Matching Rate for Unique Identifier (MRUI)

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

(Assumes records are distributed evenly across all classes) (McCormack & Smyth, 2015, 2016)

Using the MRUI equation above, input the values given in Table 1 for each variable. The value is determined by the number of classes in each variable, e.g., the variable Gender has two classes, male and female; the variable County has 26 classes. The Basic Model for the ICA assumes the records are proportionally distributed among the classes in each variable.

Table 2 Proportion of records in each class evenly distributed.

Variable Name	Symbol	No. Classes	Proportion of Records in each Class		Description of Classes
V1 = DoB	$\mathbf{V1_{ui}}$	365	0.3%	$\frac{1}{365}$	ui = 363 (days of the year) Classes evenly distributed
V2 = Gender	$\mathbf{V2_{ui}}$	2	50%	$\frac{1}{2}$	ui = 2 (genders approx. 50% split) Classes evenly distributed
V3 = county digit code	$\mathbf{V3_{ui}}$	26	3.8%	$\frac{1}{26}$	ui = 26 counties
V4 = NACE	$\mathbf{V4_{ui}}$	15	6.7%	$\frac{1}{15}$	ui =15 different NACE1 digit codes (Industrial Sector)

(Assumes records are distributed evenly across all classes). N = Population = 65,000 employees born in same year.

Using MRUI equation (Eq. (1) above):

$$65000 \times \frac{1}{365} \times \frac{1}{2} \times \frac{1}{26} \times \frac{1}{15} = \frac{65000}{284,700} = 0.23$$

In this example the **MRUI = 0.23** < 1 then there is a unique identifier for the individual employee in the dataset, derived from the variables listed above. The ICA-Basic Model assumes that the population is evenly distributed in each of the variable classes. If the population is not evenly distributed across the variable classes, then the ICA-Enhanced Model (ICA-EM) is employed, using the Adjusted MRUI equation (aMRUI Eq. (2).

2.3 ICA- Enhanced Model (ICA-EM)

It is known that the general population in Ireland is not evenly distributed by region and also that the employee population is not evenly distributed in the various NACE sectors. For example, up to a third of the working population in Ireland is located in the Dublin region; which results in a substantial number of individuals having the same DoB and gender in this region, which are referred to as *duplicates*. The ICA Basic Model must be modified to address the known issue of duplication.

If the variables are not evenly distributed the example above would be very different, as indicated in Table 3.

Table 3 Proportion of records in each class not evenly distributed.

Variable name	Symbol	No. classes	Proportion of Records in the Dominant Class		Description of Classes – one dominant class
V1 = DoB	$\mathbf{V1_{di}}$	365	0.3%	$\frac{1}{365}$	di = 363 (days of the year)
V2 = Gender	$\mathbf{V2_{di}}$	2	50%	$\frac{1}{2}$	di = 2 (genders approx. 50% split)
V3 = county	$\mathbf{V3_{di}}$	26	33%	$\frac{1}{3}$	di = 3. Dublin has one third of employees
V4 = NACE 1 digit code	$\mathbf{V4_{di}}$	6	20%	$\frac{1}{5}$	di = 5. One sector has approximately one fifth of the employees (NACE1 digit codes (Industrial Sector)

N = Population = 65,000 employees born in same year

Using MRUI equation (Eq. (1) above):

$$65000 \times \frac{1}{365} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{5} = \frac{65000}{10950} = 5.9 \approx 6$$

In this example the **MRUI = 6** > 1. Since the MRUI value is greater than 1, then we cannot derive a unique identifier (UI) for each person from the existing variables. Theoretically the MRUI value indicates the number of duplicates for each individual record. This leads to their being 6 duplicates for each person in the dominant classes for each variable. We then employ the ICA-EM.

With the ICA-EM two adjustments are made to the ICA Basic Model:

1) it is known that one fifth of the employee population are working in a dominant NACE sector, which results in 6 duplicates for individuals with the same DoB, gender, county & NACE sector. Including a marital status variable to the ICA Basic Model results in a 50% reduction in the number of duplicates, as the employee population is evenly distributed between married and non-married.
2) the inclusion of a variable representing the no. of dependent children to the ICA Basic Model allows further breakdowns of the employee population

The inclusion of these two additional variables to the ICA Basic Model, now known as the ICA Enhanced Model, allows a unique identifier for each individual to be developed. Combining or merging, in sequence, a number of the individuals known demographic, regional and industrial classification variables yields the unique identifier (see Table 4).

Eqn. 2 Adjusted Matching Rate for Unique Identifier (aMRUI)

Table 4 Identity correlation approach: Enhanced model — Combining additional variables.

Operation	Variable	No. of Records
	Approx. No. of births each year	65,000
Divide by:	No. days in the year	365
Derived variable	No. Persons with same DoB	178
Divide by:	Gender	2
Derived variable	No. Persons with same DoB and gender	89
Divide by:	No. Counties (allowing for approx. one third employees living in Dublin)	3
Derived variable	No. Persons with same DoB, Gender and County	30
Divide by:	NACE industrial code (15) - allow for one fifth employees in same NACE Sector	5
Derived variable	No. Persons with same DoB, Gender, County and NACE	6
Divide by:	Marital Status (married & other)	2
Derived variable	No. Persons with same DoB, Gender, County, NACE and marital status	3
Divide by:	No. of dependent children (3 groups)	3
Derived variable	No. Persons with same DoB, Gender, County, NACE, marital status and no. dependent children	1

Table 5 Proportion of records in the dominant class.

Variable Name	Symbol	No. Classes	Proportion of Records in the Dominant Class		Description of Classes
DoB	$\mathbf{V1_{di}}$	365	0.3%	$\frac{1}{365}$	di = 363 (days of the year). Classes evenly distributed.
Gender	$\mathbf{V2_{di}}$	2	50%	$\frac{1}{2}$	di= 2 (genders approx. 50% split).
County	$\mathbf{V3_{di}}$	26	33%	$\frac{1}{3}$	di= 3 Dublin has one third of employees
NACE 1 digit code	$\mathbf{V4_{di}}$	6	20%	$\frac{1}{5}$	di =5 One sector has approximately one fifth of the employees (NACE1 digit codes (Industrial Sector))
Marital Status	$\mathbf{V5_{di}}$	2	50%	$\frac{1}{2}$	di=2. Marital Status (married & other)
No. of dependent Children	$\mathbf{V6_{di}}$	3	33%	$\frac{1}{3}$	di=3. Divided into 3 classes

N = Population = 65,000 employees born in same year

$$N \times \frac{1}{V1_{di}} \times \frac{1}{V2_{di}} \times \frac{1}{V3_{di}} \times \frac{1}{V4_{di}} \times \ldots \times \frac{1}{VX_{di}} = \text{MRUI} \quad (2)$$

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

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).

Using the Adjusted Matching Rate for Unique Identifier (aMRUI) in Eq. (2) above, plug in the values given in Table 2 for the above example for each variable. The value is determined by the proportion of records in the largest class for each variable, e.g., the variable County has 26 classes, but one third of the employee population are concentrated in Dublin, therefore the value is 3 for the County variable.

Using aMRUI equation (Eq. (2) above):

$$65000 \times \frac{1}{365} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{5} \times \frac{1}{2} \times \frac{1}{3} = \frac{65000}{65700} = 0.99$$

In this example the **MRUI = 0.99** < 1. Therefore we can derive a unique identifier (UI) for each person from the existing variables. The ICA-EM adjusts for the records concentrated among the dominant class in each variable. This ensures that there is a UI for all records, not just the records evenly distributed in the variable class.

2.4 ICA-Basic Model for Smaller Classes

In Section 2.3 ICA-EM above, it is demonstrated that the ICA-Basic Model is not sufficient to obtain a UI if there is not enough variables to input into the equation. Therefore the ICA-EM (Enhanced Model) is employed, which examines the class sizes in proportion to the Population (N). If the proportion of records in a class is too large, then the individuals (records) in the large classes have to be broken down by additional variables. However, if the number of variables in the dataset are not sufficient for a breakdown of the larger classes, it is still possible to create a UI for the records in the variable classes with a smaller proportion of the population. This means that there will not be a UI for the entire population, but there will be a UI for the individuals in classes which are proportionally smaller. The number of records in the population with a UI may be significant enough to allow the researcher to use a representative sample to conduct analysis. This is demonstrated in Table 6 below when the values in Table 2 are plugged into the MRUI equation.

Table 6 shows that the MRUI value </ 1, when the proportion of the population in a County reaches 1/18 or smaller.

Table 6 MRUI values for proportionally smaller class sizes — Adjusting one variable (V3).

Proportion of records in variable V3 =County	N = Population	V1 = DoB	V2 = Gender	V3 = county	V4 = NACE 1 digit code		MRUI	Unique Identifier if MRUI < 1
County with 1/3 of pop.	65,000	365	2	3	5	=	5.9	>1
County with 1/4 of pop.	65,000	365	2	4	5	=	4.5	>1
County with 1/5 of pop.	65,000	365	2	5	5	=	3.6	>1
County with 1/6 of pop.	65,000	365	2	6	5	=	3.0	>1
	…	…	…	…	…	…	…	…
County with 1/18 of pop.	65,000	365	2	18	5	=	1.0	<1
County with 1/26 of pop.	65,000	365	2	26	5	=	0.7	<1

By adjusting two variables (V3, V4) in Table 7, then we can see that a UI (MRUI < 1) is created when the proportion of the variables in both NACE and County are both reduced.

Table 7 MRUI Values for proportionally smaller class sizes — Adjusting two variable (V3, V4).

Proportion of records in variable V3 = County and V4 = NACE	N = Population	V1 = DoB	V2 = Gender	V3 = county	V4 = NACE 1 digit code		MRUI	Unique Identifier if MRUI < 1
County with 1/3 of pop. NACE 1/6	65,000	365	2	3	6	=	4.9	> 1
County with 1/4 of pop. NACE 1/7	65,000	365	2	4	7	=	3.2	> 1
County with 1/5 of pop. NACE 1/8	65,000	365	2	5	8	=	2.2	> 1
County with 1/6 of pop. NACE 1/15	65,000	365	2	6	15	=	1.0	< 1

The above Tables demonstrate the significance of class size in creating a UI using the mathematical concept of the ICA. If the primary goal is to create a UI for the entire population, then the proportion of the population in the largest class size of each variable is the determining factor (where the number of variables are limited). However, in terms of Statistical Data Disclosure (SDC), if the aim is to prevent an individual from being identified then the focus must be placed on the variable class with the smallest proportion of the population.

Increasing the number of variables in the ICA equation further increases the probability of a UI for the large classes.

2.5 Application of the Identity Correlation Approach — Enhance Model (ICA-EM)

The ICA-EM was applied to the Census and Public Sector datasets for 2011 to create a unique identifier titled the matching variable (matchvar) to facilitate individual record linking across these datasets and the construction of a Master Administrative Data Source (MADS).

2.5.1 Census 2011 Dataset

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

2.5.2 Public Sector Administrative Datasets

A Master Administrative Data Source (MADS) consisting of a single dataset containing all individual characteristics (variables), was constructed from a number of Public Sector Administrative Datasets such as Revenue Commissioners Tax data, Social Security Administrative Data Sources and CSO Administrative Datasets (e.g., Central Business Register (CBR), Earnings datasets).

The MADS process consisted of combining these datasets using the PIK for each individual and the CBR Enterprise No. to link employment related data to characteristics for the individual (e.g., Dob, gender, etc.). The PIK allows data linking without revealing too much sensitive information [13].

The IDA-EM was applied to the Master Administrative Data Source (MADS) also, allowing for a Unique Identifier (UI) to be created for 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 directly associated with the person's PIK No. on the Master Administrative Data Source (MADS).

Other variables used to further breakdown the data are industrial sector in which the person works (NACE code) and no. of dependent children. In this way a unique combination of variables apply to each person allowing a person to be uniquely identified.

2.5.3 Linking Census to MADS

Variables common to both the Census dataset and the Master Administrative Data Source (MADS) 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 (ICA). By linking the two datasets using the Unique Identifier, a PIK No. could be allocated to each individual person in the 2011 Census dataset.

This is shown in Table 8. Once the PIK. was assigned to the Census dataset, it enabled Census data to be linked to any Public Sector Administrative Dataset.

Table 8 Applying identity correlation approach to dataset to create unique identifier (Matchvar).

Date of Birth	Gender	County	NACE	Marital status	No. of children	Matchvar
15031949	M	CORK	42	M	0	15031949MCORK42M0
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

2.6 Preparation of Datasets

The SESADP has a focus on employees, and this population subset must be extracted from both the census and public administrative datasets.

2.6.1 Census 2011

A total of 2.2 million records were extracted from the 4.6 million 2011 Census Records. These records consisted of employees, unemployed, students (i.e., labour force and potential participants). Approximately 200,000 of these records had a unique Business No. identifier attached (CBR No.). Another 500,000 records had a CBR No. attached using the Employer's Business name on the Census.

- The first matching variable (Matchvar1) created for Census used the following variables combined: CBR No., Dob, gender, county, NACE 2, marital status, No. of children.
- A second matching variable was created (Matchvar2) excluding NACE2 (see Fig. 4). Up to ten matching variables (Matchvar1 – Matchvar10) were created.
- Each matching variable is similar to the previous one, with a single characteristic change to the composition variables for each subsequent matching variable created.

Table 9 illustrates the construction of each subsequent matching variable.

Table 9 Matching variables.

Date_of Birth	Gender	County	NACE	Ent No.	Marital status	No. children	Match Var 1	Match Var 2	Match Var 3
15031949	M	CORK	42	EN12345678	M	0	15031949MCORK42EN12345678M0	15031949MCORK42EN12345678M	15031949MCORK42EN12345678
11021945	F	LIMERICK	31	EN52345679	S	1	11021945FLIMERICK31EN523456791	11021945FLIMERICK31EN52345679S	11021945FLIMERICK31EN52345679
21111954	M	DUBLIN	25	EN52795680	O	2	21111954MDUBLIN25EN527956802	21111954MDUBLIN25EN52795680O	21111954MDUBLIN25EN52795680
19051964	M	CARLOW	55	EN32795681	D	2	19051964MCARLOW55EN327956812	19051964MCARLOW55EN32795681D	19051964MCARLOW55EN32795681
22091966	M	GALWAY	82	EN22795682	M	3	22091966MGALWAY82EN227956823	22091966MGALWAY82EN22795682M	22091966MGALWAY82EN22795682
24031971	F	CAVAN	84	EN52795683	M	0	24031971FCAVAN84EN527956830	24031971FCAVAN84EN52795683M	24031971FCAVAN84EN52795683
28021977	F	DUBLIN	71	EN84355684	S	1	28021977FDUBLIN71EN843556841	28021977FDUBLIN71EN84355684S	28021977FDUBLIN71EN84355684
30061990	F	KERRY	35	EN73795687	M	1	30061990FKERRY35EN737956871	30061990FKERRY35EN73795687M	30061990FKERRY35EN73795687

- MADS (Public Sector Administrative Datasets)

The records in the Master Administrative Dataset contained the same set of variables used for 2011 Census data subset to create the matching variables (Matchvar1 – Matchvar10). The matching variables created in the MADS were used to match to the same variable in the Census.

2.6.2 Practical Application — Incremental Matching Process (IMP)

There are ten steps involved in the incremental matching process:

1) IMP – Step 1

The variables (Matchvar1 – Matchvar10) were used to match the 2011 Census and MADS datasets. It is known that duplicates will occur when the matching variables are created. To directly address this issue in the dataset linking process, only single occurrences of the matching variables (Matchvar1 – Matchvar10) are selected in each dataset. If there is more than one occurrence of a matching variable then the records are excluded in the matching process.

2) IMP – Step 2

In the next step, the first matching variable is chosen (matchvar1). Both datasets are matched using matchvar1. Then the second matching variable (matchvar2) is matched.

3) IMP – Steps 3 to 10

The matching process continues incrementally up to Matchvar10 until all the single occurrences of the matching variables have been matched.

Using this approach approximately 1 million records were matched between the Census and Public Sector MADS. Only 800,000 records were used in the first phase of data outputs. The reason for this was that a smaller number of variables were used for the final 200,000 records matching process. Therefore, these records would have required more time to check if they were correctly coded. Due to a tight deadline for publication of the Earnings results it was decided not to use these 200,000 records in the first phase of the publication, as they needed more time for thorough checks.

2.6.3 False Positives

False positives using the ICA approach can only occur if there is an error in the data, e.g., data is incorrectly coded.

False positives can occur in the matching process if a variable is incorrect on one of the datasets. For example, if the county variable has not been updated on the Social Welfare dataset then the county will be different on the persons record on Census. Similarly, if the NACE code is incorrect on either dataset, then it will not match a person to their correct record. False positives can be corrected using occupation codes on the Census. For example, if the occupation code refers to a police officer, then the correct NACE sector code and Enterprise_no (ent_nbr) can be assigned to that individual.

A quality check of the matching process was carried out using a sample of records with the person's name and Date_of_Birth. It showed that 99% of records matched correctly, with poor data accounting for the 1% that could not be matched.

3. Data Protection With 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. Personal data are highly sensitive and must be handled securely to protect the person's privacy, and to comply with data protection laws [14].

Although the SESADP project was carried out prior to the GDPR regulation coming into effect, the ICA approach used ensured data matching in compliance with the GDPR Regulation [15].

3.1 Statistical Disclosure Control

Section 2.4 above titled 'Basic Model for Smaller Classes', looked at how some individuals in smaller classes can be uniquely identified even though individuals in the larger classes cannot be uniquely identified. This is an efficient tool for Statistical Disclosure Control (SDC) by using it to check uniqueness of individuals. Simply take the smallest class in every variable and apply the ICA method. If the result gives a value of MRUI ≤ 1, then there is an issue with SDC and the dataset cannot be released. The next step is to merge smaller classes (in each variable) until the value of MRUI > 1. The level of SDC can be measured using the MRUI value, for example if the SDC rules require 10 or more individuals in a group to prevent disclosure, then keep merging the classes until the MRUI ≥ 10. In this way the MRUI is a measure of suppression of the data.

$$\text{MRUI} \geq X$$

where X is a measure of Statistical Disclosure Control.

3.2 Data Encryption

Individual identification variables such as "Date of Birth" can be replaced with a "Protected Identity Key" (PIK) for matching purposes. Table 10 shows the dataset variables matched and Table 11 shows the same variables encrypted.

Table 11 shows that any of the variables can be encrypted. This does not inhibit the data linking process, as the method still results in the creation of a unique identifier, by joining all the encrypted variables.

Table 10 Dataset variables matched with ICA method.

Record No.	Date of birth	Gender	County	NACE	Enterprise No.	Marital status	No. child	Unique Identifier
Record 1	15031949	M	CORK	42	EN12345678	M	0	15031949MCORK42EN12345678M0
Record 2	11021945	F	LIMERICK	31	EN52345679	S	1	11021945FLIMERICK31EN52345679S1
Record 3	21111954	M	DUBLIN	25	EN52795680	O	2	21111954MDUBLIN25EN52795680O2
Record 4	19051964	M	CARLOW	55	EN32795681	D	2	19051964MCARLOW55EN32795681D2
Record 5	22091966	M	GALWAY	82	EN22795682	M	3	22091966MGALWAY82EN22795682M3

Table 11 Dataset variables encrypted and matched with ICA method.

Record No.	Date of Birth	Gender	County	NACE	Enterprise No.	Marital Status	No. Child	Unique Identifier
Record 1	Age2	X	Co1	N1	EN1	M1	A	Age2XCo1N1EN1M1A
Record 2	Age1	Y	Co2	N2	EN2	M2	B	Age1YCo2N2EN2M2B
Record 3	Age3	X	Co3	N3	EN3	M3	C	Age3XCo3N3EN3M3C
	Etc.							

- 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 PIK, 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 unauthorized 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.

Datasets can be stored with encrypted variables as shown in Table 11. This preserves the uniqueness of each record and ensures statistical disclosure control (SDC). Therefore the data can be stored encrypted and this ensures SDC if the data is accidentally leaked. Various methods can be used to encrypt data [16].

Data on individuals is often held in different databases and shared by different organizations. Data linking across different datasets normally requires suppression of information that might directly identify an individual which inhibits the possibility of record linkage [17]. With the ICA approach data can be

suppressed by encrypting variables without inhibiting record linkage.

3.3 Data Masking

Data masking techniques such as encryption, substitution, adding dummy variables, etc. can easily be done without affecting the data matching process using the ICA approach. As long as the uniqueness of the variables are preserved, then the ICA approach will permit data matching with any data obfuscation technique.

4. Summary

In summary, the ICA approach has shown to be an effective data matching method. Data masking and encryption does not affect the ICA method's matching process. Issues around duplicates and coding errors can be calculated mathematically using the MRUI formula.

A complete match of all data records in a dataset is determined if the value MRUI ≤ 1. In this way a record is directly matched and is unique, with no possibility of duplicates.

In order to quickly evaluate the degree of SDC (Statistical Disclosure Control) in a dataset, the MRUI equation can be applied. The MRUI value indicates how many records (e.g., people) are in each variable class. For example, if the MRUI value = 10, then 10 records (people) are the smallest group size that can be determined from the dataset.

Enhanced data security and confidentiality are a feature of the ICA approach, since the method does not require string variables to be retained on datasets. In addition, encrypted versions of identifiable variables such as *date_of_birth* can be applied to replace the actual variable. This prevents accidental disclosure of the personal details on datasets. If datasets are stored encrypted in this manner then accidental leaks of the datasets prevent the disclosure of personal details to unauthorized users.

References

[1] Edwards, B., Hofmeyr, S., and Forrest, S. (2016). "Hype and Heavy Tails: A Closer Look at Data Breaches." *Journal of Cybersecurity* 2 (1): 314, doi: https://doi.org/10.1093/cybsec/tyw003.

[2] McCormack, K., and Smyth, M. (2016). "Big Data Matching Using the Identity Correlation Approach. First International Conference on Advanced Research Methods and Analytics." In: *CARMA 2016*.

[3] Fellegi, I., and Sunter, A. (1969). "A Theory for Record Linkage". *Journal of the American Statistical Association* 64 (328): 1183-1210.

[4] Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). "A Comparison of String Metrics for Matching Names and Records". In: *The Meeting of the SIGKDD*, 2003.

[5] Dusetzina, S. B., Tyree S., and Meyer A. M. et al. (2014). "Linking Data for Health Services Research: A Framework and Instructional Guide." Rockville (MD): Agency for Healthcare Research and Quality (US).

[6] Vatsalan, D., Sehili, Z., Christen, P., and Rahm, E. (2017). "Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges." In: *Handbook of Big Data Technologies*, Springer.

[7] McCormack, K. (2015). "Constructing Structural Earnings Statistics From Administrative Datasets: Structure of Earnings Survey — Administrative Data Project." *The Statistics Newsletter* (62): 3-5.

[8] McCormack, K., and 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. NTTS

[9] "Council Regulation (EC) No 530/1999 of 9 March 1999 Concerning Structural Statistics on Earnings and on Labour Costs". *OJ L* 63: 6.

[10] "National Employment Survey 2008 and 2009", 2011, Central Statistics Office, Ireland.

[11] McCormack, K., and 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.

[12] McCormack, K., and Smyth, M. (2017). "A Mathematical Solution to String Matching for Big Data Linking". *Journal of Statistical Science and Application* 5: 39-55.

[13] Hall, R., and Fienberg, S. E. (2010). "Privacy-Preserving Record Linkage." In: *PSD'10 Proceedings of the 2010 International Conference on Privacy in Statistical Databases*, Corfu, Greece, pp. 269-283.

[14] Trepetin, S. (2008). "Privacy-Preserving String Comparisons in Record Linkage Systems: A Review." *Information Security Journal: A Global Perspective* 17

(5-6): 253-266

[15] "Regulation (EU) 2016/679 of the European Parliament and of the Council". *OJ L* 119: 1-88, available online at: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[16] McCormack, K., and Smyth, M. (2019). "Privacy Protection for Big Data Linking Using the Identity Correlation Approach". *Journal of Statistical Science and Application*: 81-90.

[17] Smith, D., and Shlomo, N. (2015). "Privacy Preserving Probabilistic Record Linkage." *New Techniques and Technologies for Statistics (NTTS) 2015*, Collaboration in Research and Methodology for Official Statistics, NTTS.