Journal of Statistical Science and Application 5 (2017) 39-55

A Mathematical Solution to String Matching for Big Data Linking

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

This paper describes how data records can be matched across large datasets using a technique called the Identity Correlation Approach (ICA). The ICA technique is then compared with a string matching exercise. Both the string matching exercise and the ICA technique were employed for a big data project carried out by the CSO. The project was called the SESADP (Structure of Earnings Survey Administrative Data Project) and involved linking the Irish Census dataset 2011 to a large Public Sector Dataset. The ICA technique provides a mathematical tool to link the datasets and the matching rate for an exact match can be calculated before the matching process begins. Based on the number of variables and the size of the population, the matching rate is calculated in the ICA approach from the MRUI (Matching Rate for Unique Identifier) formula, and false positives are eliminated. No string matching is used in the ICA, therefore names are not required on the dataset, making the data more secure & ensuring confidentiality. The SESADP Project was highly successful using the ICA technique. A comparison of the results using a string matching exercise for the SESADP and the ICA are discussed here.

Keyword: Big Data, Data Linking, Identity Correlation Approach, String Matching, Public Sector Datasets, Data Privacy.

Introduction

This paper presents work from the CSO's big data project called the SESADP (Structure of Earnings Survey Administrative Data Project [1] [2]. Most big data matching projects attempt to match dataset records on the basis of string variable matching. String matching is successful if the string variable records are unique and spellings of string variables are correct. However if the string variable used for matching contains data records which are similar to other data records, or are non-unique, this leads to false positives (matched records that are incorrect). Datasets which are string matched require a lot of manual checking, which is not practical with big data projects. Also false positives can be hard to quantify in the matched data records.

This paper compares the results of a mathematical solution to string matching called the Identity Correlation Approach (ICA) [3], with an exercise that involved string matching. In the ICA approach, the Matching Rate of Unique Identifier (MRUI) equation is used to calculate the matching rate for records and false positives can be eliminated. This approach provides an exact mathematical method for matching big data without string matching. Outcomes of matches are calculated mathematically before the matching process begins, as the MRUI formula quantifies the rate of exact matches beforehand and false positives can therefore be eliminated by adjusting the formula. The ICA technique is developed based on the population size and the

number of variables in the datasets to be matched with each other. There is no manual checking required for false positives. The Identity Correlation Approach (ICA) was developed in the SESADP project undertaken by the CSO to match the 2011 Irish Census dataset to a large Public Sector dataset in a hugely successful big data project to provide statistics for the Structure of Earnings Survey [4]. The SESADP replaced the CSO's annual National Employment Survey [5] [6].

Firstly, this paper will present the Identity Correlation Approach (ICA) and its application and the results from the data matching exercise. Secondly, it will detail the process and outcome of string matching in the SESADP big data project.

Identity Correlation Approach

A mathematical solution called the Identity Correlation Approach (ICA) was developed for the CSO's SESADP project to match the Census dataset to a large Public Sector dataset [3]. This innovative approach enabled 800,000 employees (50% of all employees) to be matched to their employer and other variables without the need for string matching. 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) equation.

An innovative feature of the Identity Correlation Approach (ICA) is data security and confidentiality are very strong compared with string matching [7]. 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.

Identity Correlation Approach

The Identity Correlation Approach (ICA) has been developed by the Methodology Division of the CSO as a direct response to the challenge of linking administrative data sources which do not contain Unique Identifiers at the level of the individual. Within the SESADP for 2011, the CSO's 2011 Census of Population (COP) is the administrative data source.

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. (*See Table 1)*.

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*.

(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 *2* 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 3 persons with the same DoB, gender and county.

(5) The 3 persons with the same DoB, gender & county can be further divided by NACE sector (15 categories), which provides a unique identifier of 1 person with the same DoB, gender, county and NACE sector.

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

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

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 ICS Basic Model must be modified to address the known issue of duplication.

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 2*).

Table 2

Identity Correlation Approach: Enhanced 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 (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

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).

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.

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 IDA-EM was applied to the Master Administrative Data Source (MADS) also, allowing for a Unique Identifier 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.

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 3. Once the PIK. was assigned to the Census dataset, it enabled Census data to be linked to any Public Sector Administrative Dataset.

Table 3

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	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

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.

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 Figure 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 4 illustrates the construction of each subsequent matching variable.

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.

Practical Application - Incremental Matching Process (IMP)

There are ten steps involved in the incremental matching process

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.

Table 4

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

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.

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 for false positives. 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.

False Positives

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 ent_nbr can be assigned to that individual.

Summary of ICA

In summary, the ICA method has shown to be very effective in matching the datasets. Issues around duplicates and false positives can be calculated mathematically using the MRUI formula:

The ICA approach would expect a complete match of all data records if the MRUI value is 1 or less than 1. In the exercise carried out for the SESADP, 50% of the full 1.6 million employee records were not matched possibly due to: 1) incomplete records on either the Census or the Public Sector datasets; 2) different code on either dataset. For example, the variable 'county' on the Public Sector datasets is only updated if an employee receives some payment from Dept. of Social Protection (DSP), otherwise the variable may be out of date. Also the variable for marital_status is updated only if the DSP are notified of a person's status. The variable for NACE 2-digit code may not be entirely harmonised between the Census and Business Register as the Census does not use the Business Register to code the economic activity of the employee. Census uses the description given by the employee on the Census form, which may not be compatible with the Business Register. In this case the ICA approach will not result in a match. Some Census variables may be left blank or incorrectly completed, resulting in the records being unmatched with other Public Sector datasets.

To conclude, the ICA approach resulted in 50% of records being correctly matched across the Census and Public Sector datasets where variables were coded correctly. Census records were not matched to their corresponding record on the Public Sector dataset where variables were not coded correctly. This was due to a number of issues such as variables not being updated and different coding being used for NACE economic activity. Therefore the ICA approach is also a good measure of the degree of non-harmonised coding on a dataset.

Also, the ICA approach enhances data security and confidentiality. 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 employed to replace the actual variable.

String Matching to Census Data

Data linking projects mainly involve algorithms which are based on records being matched directly (deterministic) or the probability of a match [8]. In this project string matching is based on the deterministic approach, where a record is directly matched and is uniquely identifiable [9].

A string matching exercise was carried out to link the Census 2011 dataset to the large Public Sector dataset. The purpose of the exercise was to link the employer's name on Census (variable = Business_name) with the official employer name on the CSO's Central Business Register (variable= company_name). This would enable the unique enterprise number (ent_nbr) to be added to the Census dataset for each employee. NACE industrial sector codes could then be harmonised across the datasets and the employee could be grossed up to the enterprise for earnings purposes. Details of the string matching exercise are described here.

String Matching Preparation

To prepare the Census dataset for string matching, firstly Census records with the same employer's name (variable = Business_name) were counted to identify the largest employers as stated by the individual on the Census form. A number of issues arose with this exercise as shown in Figure 1.

Census Dataset

	No. of records	Business_name
Company 1	300	CSO
	299	C.S.O.
	250	Central Statistics Office
Company 2	211	Department of Finance
	200	Dept of Finance
	198	Finance Dept
Company 3	197	High Tech Sales
	183	HTS Ltd
Company 4	169	Shoe Sales Plc
	155	Barrys Shoe Shop

Business Register Dataset

Company_name	Ent_Nbr
Central Statistics Office	EN12345678
Department of Finance	EN99349874
HTS Ltd	EN53971582
Shoe Sales Plc	EN34789632

Figure 1. An Example of List of employers in Census by no. of records, (N.B. The examples of employer names given in this paper are for illustrating concepts used in data matching. They do not reflect actual employer names on the Census dataset).

It was apparent from this step that individuals stated different versions of their employer's name on the Census form. Some individuals gave abbreviations of the employer's name, while others used the initials (e.g. CSO, C.S.O., Central Statistics Office) as shown in Figure 1. With several versions of the employer's name used for the variable 'Business_name', it was obvious that string matching would not match all employees to the official employer's name on the Central Business Register (variable = company_name). In the example given in

Figure 1, string matching only allows individual records in Census to be linked) to the Business Register if the Census employer name (Business_name) corresponds exactly to the Business Register name (company_name). Individual records with other versions of the employer name are not matched (e.g. CSO or C.S.O. are not linked to the name 'Central Statistics Office') to the Business Register.

To overcome this problem the records containing unofficial versions of the employer name were re-coded to the official version, for all large employers listed in the Census. For example, if Business_name was listed as CSO, or C.S.O. then Business_name was changed to 'Central Statistics Office'. This allowed the records for these individuals to be linked to their correct name and ent_nbr on the Business Register, as shown in Figure 2.

Census Dataset

	No. of records	Business_name	Re-coded Business_name
Company 1	300	CSO	Central Statistics Office
	299	C.S.O.	Central Statistics Office
	250	Central Statistics Office	
Company 2	211	Department of Finance	
	200	Dept of Finance	Department of Finance
	198	Finance Dept	Department of Finance

Business Register Dataset

Company_name	Ent_Nbr
Central Statistics Office	EN12345678
Department of Finance	EN99349874

Figure 2. Census employers re-coded to official company name.

False Positives – Name in different variable. On the Census form some individuals stated an internal company department as the employer name, and placed the employer name on the first address line (variable = addr_line1). This resulted in false positives occurring with the string matching exercise, due to the employer's name being placed in a different variable. An example of this is given in Figure 3 showing a false positive string match.

Census Dataset

	No. of records	Business_name	Addr_line1
Company 2	211	Department of Finance	Government Buildings
	200	Dept of Finance	University Hospital Cork
	198	Finance Dept	Government Buildings

Business Register Dataset

Company_name	Ent_Nbr
Department of Finance	EN99349874
University Hospital Cork	EN71187478

Figure 3. Census employer name inserted as address variable.

In this instance all employees who stated their employer name 'Dept. of Finance' are matched to the Government's Department of Finance. However on closer examination, of the variable addr_line1, it becomes apparent that 200 employees are actually employed by a hospital's internal Dept. of Finance. If only the variable for business_name was used for string matching then it would result in 200 employees on Census being incorrectly matched to the incorrect employer on the Business Register. Adding the variable for

addr_line1 is necessary to avoid a false positive for string matching in this instance. Therefore the variable for business_name must be populated with the information in addr_line1 for the string matching exercise. String variables must be examined carefully in this instance.

False Positives – Name used incorrectly. Another issue involving false positives arose in string matching because the individual's employer stated on Census was different from the official employer. Figure 4 demonstrates this problem using the example of the Irish Police Force (GARDA).

Census Dataset

No. of records	Business_name	occ_code
10,000	GARDA	3312, 3315, 1172
5,000	GARDA	other

Business Register Dataset

Company_name	Ent_Nbr	No. of employees
GARDA	EN58258745	10,000
Dept. Justice	EN67158746	10,001

Figure 4. Census employer stated incorrectly.

Individuals working in administrative duties for the GARDA are employed officially by the Dept. of Justice. However since the individual works in a GARDA station they state their employer as 'GARDA' in the Census variable Business_name. Linking these employees to the company_name on the Business Register would over-state the number of employees in the Police Force (GARDA). To allow only police officers to be linked to the official Business Register company_name for GARDA, an additional variable for occupation code on Census must be added to the matching process. When trying to assign employees to the ent_nbr for GARDA (Police), the occupation code for Police officers was used in addition to the employer name. Therefore only employees with employer name GARDA and occupation code for police officers were assigned the employer name GARDA. If the administrative employees stated they worked for the GARDA, then they were assigned to the Civil Service Dept. of Justice on the basis of their occupation code. See Figure 4.

Non-unique employer's names. String matching resulted in false positives where the employer's name is non-unique as shown in Figure 5a. This is an example of several different enterprises with the same name, but are completely independent of each other and in different geographical locations in the state. The example given is an enterprise called 'Murphys Pharmacy'. Individuals on Census who stated their employer as 'Murphy's Pharmacy' in the variable Business_name, will be matched to the first version of 'Murphy's Pharmacy' on the Business Register's company_name variable. In order to identifying the correct enterprise for each employee it is necessary to examine the two variables addr_line1 and addr_line2 (see Fig.5b). If addr_line1 is added then it is possible to separate out the first employer. In order to separate the second and third employer, it is necessary to use addr_line2 as the variable Business_name and addr_line1 are the same, but the town is on addr_line 2. If the address variables were not examined, then string matching of the variable business_name alone would result in false positives in the matching process.

Different Trading Name. Frequently on the Census form, individuals often state the 'Trading name' of the employer, instead of the official registered name. The ‘Trading name’ is different from the official name of the enterprise on the Business Register. Examples of this are particularly acute with small shops which may be operating as a franchise. Figure 6a gives an example of small shops on the Business Register officially registered with their own company_name (e.g. 'Jane Potter & Co. Ltd.') and have a franchise for

'SUPERPRICE' . Employees on the Census form will name their employer as 'SUPERPRICE' all over the State. This would result in several thousand employees linked to the Business Register company_name of 'SUPERPRICE' when their actual employers are numerous small shops. String matching by employer name is not possible in this example for a big data project. It is necessary to re-code the business_name to the official name on the Business Register for a string match (Fig.6b).

Census Dataset

No. of records	Business_name
105	Murphy's Pharmacy
99	Murphy's Pharmacy
50	Murphy's Pharmacy

Business Register Dataset

Company_name	Ent_Nbr
Murphy's Pharmacy	EN32198748
Murphy's Pharmacy	EN42198749
Murphy's Pharmacy	EN52198750

Figure 5a. Census employer non-unique.

No. of records	Business_name	Addr_line1	Addr_line2	Company_name	Ent_Nbr
105	Murphy's Pharmacy	Atlantic Quay	Cork	Murphy's Pharmacy	EN32198748
99	Murphy's Pharmacy	Main Street	Dublin	Murphy's Pharmacy	EN42198749
50	Murphy's Pharmacy	Main Street	Waterford	Murphy's Pharmacy	EN52198750

Figure 5b. Census employer non-unique.

Census Dataset

No. of records	Business_name
105	SUPERPRICE

Business Register Dataset

Company_name	Ent_Nbr	No. of employees
SUPERPRICE	EN11245874	10

Figure 6a. Census employers trading name (1).

Census Dataset				Business Register Dataset		
No. of records	**Business_name**	**Addr_line1**	**Re-coded Business_name**	Company_name	**Ent_Nbr**	**No. of employees**
10	SUPERPRICE	Cork		SUPERPRICE	EN11245874	10
30	SUPERPRICE	Waterford	Jane Potter & Co. Ltd	Jane Potter & Co. Ltd	EN53245874	30
25	SUPERPRICE	Tralee	L. Trainer	L. Trainer	EN91245875	25
22	SUPERPRICE	Dublin	Molloys	Molloys	EN76245876	22
19	SUPERPRICE	Athlone	Peter Jones	Peter Jones	EN55245877	19

Figure 6b. Census employers trading name (1).

This issue also occurs where there are two businesses run by the same enterprise. Figure 7 shows a business officially registered as 'Ballymore Foods' and running two businesses known as 'Traditional Foods

Ltd' and 'Bistro Garden'. Individuals on Census will state their employer as 'Traditional Foods Ltd' or 'Bistro Garden', depending on which business they work in. They will be unaware that the official name of their employer is 'Ballymore Foods'. In this instance it is impossible to link the individuals to their official enterprise name using string matching alone. In order to string match correctly, recoding was used to replace the trading name stated on Census with the official company name on the Business Register. However, this was only practical for large employers on Census due to the amount of checking required.

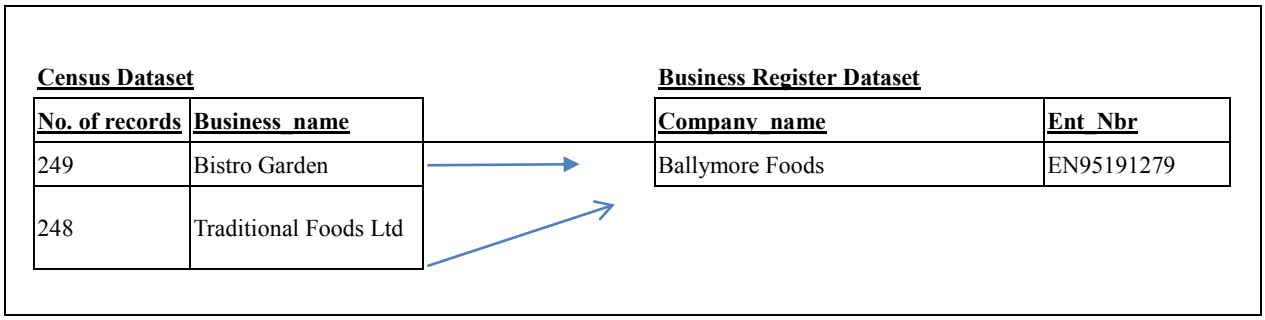

Figure 7. Census employers trading name (2).

Enterprises with different NACE economic activities. Occassionally, a group of enterprises operating in different economic activities (NACE codes) will use the same trading name. They are different distinct enterprises within the group of enterprises but use the same trading name. Figure 8 shows an example of an employer called 'MILT' with three businesses (MILT Finance, MILT Manuf, MILT R&D) in three different NACE 2-digit codes. All employees stated their employer as 'MILT' but it would not be possible to assign the employee to the correct enterprise using a string match. In order to assign the individual records to the correct enterprise on the Business Register (company_name) it is necessary to use the variable for NACE 2-digit code. This is shown in Figure 8 where both the business_name variable and the NACE 2-digit variable are used to match to the Business Register variable company_name. String matching alone is not sufficient to assign the individual to the official company name.

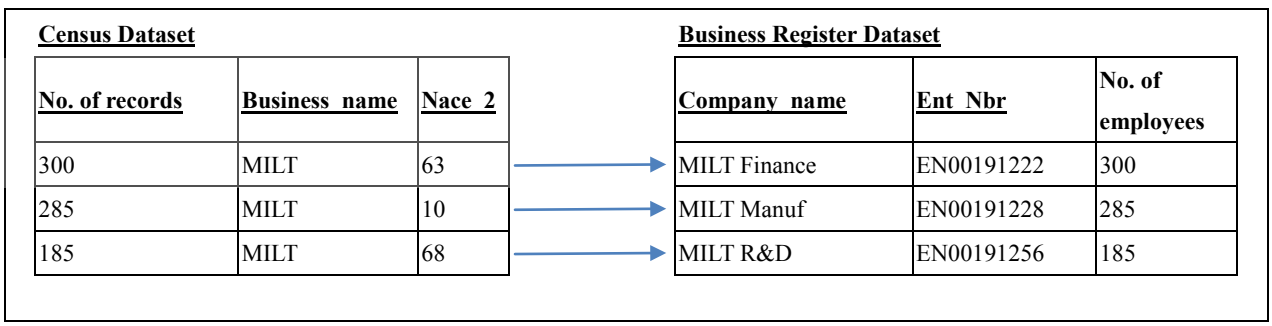

Figure 8. Census employers with different NACE activities.

Practical String Matching in SESADP

The above examples highlight the difficulty with string matching when there are issues with connecting the business name and the officially registered name. Big data projects require unique string variables, with exact spellings, for matching. String matching is ideal in big data matching where variables are obtained using drop-down menus with unique strings used. Online surveys lend themselves to this type of string matching. However several businesses with the same name (e.g. Murphys Pharmacy) would require another unique

variable to distinguish between non-unique names. In summary, string matching is useful if variables are unique and harmonised in the two datasets.

In the SESADP big data project, employees numbered approximately 1.6, million out of a total of 4.6 million records on Census. Manual checking of data to check for false positives in string matching was not practical. Focus had to be placed on getting the large organisations linked using string matching and avoiding false positives for these. String matching resulted in 130,000 records being matched. To achieve this level of matching involved recoding of some employer names to match the official name on the Business Register and employing the use of other variables. Manual checks to verify uniqueness of a business_name were also required, as described previously.

Summary of steps involved in string matching for SESADP

The variables had to be edited in the dataset before a reliable string match was obtained. The sequence in preparing the string variable for matching is summarized below:

- Uppercase all records in the Census variable for Business_name
- Count the no. of records with the same Business_name
- Sort by the largest no. of employees in each Business_name
- Focus on largest employers (greatest no. of records)
- Ensure business_name is unique and does not have other possibilities
- Employ other variables if business_name is not unique
- Recode the Business_name to the official enterprise name on the Business Register. Practically, this is only possible for large employers.
- Using string matching, match the Census Business_name to the company_name on the Business Register.

String matching in the Project resulted in linking 130,000 records to their correct employer on the Business Register, allowing the correct ent_nbr (unique business no.) to be assigned. The considerable amount of preparation of the data and hard-coding was necessary to ensure a good rate of matching for the string matching method.

Comparison of Methods

String Matching compared with ICA Method

Description	String Matching	Identity Correlation Approach
No. of records matched	130,000	800,000
Re-coding string variables	Significant	None
False positives	Difficult to quantify	Calculate match rate using formula, provided data is correctly coded
Checking records	Records have to be examined for false positives	General checking to ensure data is correctly coded
Security & Confidentiality	String variables make data less secure	Highly secure. No personal details required.
Reasons for non-matches	String variables have different names	Coding errors in datasets

Figure 9. Comparison of String Matching with ICA Method.

In comparing the results from the string matching exercise with the Identity Correlation Approach (ICA) it is obvious that the ICA is the more powerful technique (see Fig. 9). A total of 800,000 employee records were matched to their employer using the ICA approach compared with 136,000 employee records using the string matching technique. A significant amount of data preparation and re-coding was required to employ the string

matching technique. Also a knowledge of the records in the string matching variable Business_name was required in order to prepare the data for matching. With the ICA approach string variables such as names and addresses are not required. No prior knowledge of the individual records are required, apart from an overview of the different classes in each variable and the dominant class in each variable. The only preparation required for the ICA approach is the creation of a Unique Identifier (UI) by joining the existing variables on each dataset.

Data security and confidentiality are greatly enhanced using the ICA approach as the string variables are not required to match the datasets, and are therefore not retained on the datasets.

Conclusion

Results for matching Census records to the Public Sector Administrative Dataset are given in Figure 10, classified by NACE industrial sector. The lower rate of matching in some sectors (e.g. Construction) can be attributed to records not being updated for certain variables. If the theoretical MRUI value indicates a perfect match, but this is not reflected in practice, then there are issues with coding or with records not being updated.

Nace Economic Sector	No. Employees
	Total
B-E Industry	55
F Construction	26
G Wholesale and retail trade	44
H Transporation and Storage	51
I Accommodation and Food Services	31
J Information and communication	52
K-L Financial, insurance, etc.	61
M Professional, scientific & technical	39
N Administrative and support services	26
O Public administration & defence	69
P Education	63
Q Health & social work	46
R-S Arts, entertainment, other services	41
Total	47

Figure 10. Employee Population Coverage 2011 - Census & Administrative Datasets Matched.

Impact of ICA Data Matching

The ICA (Identity Correlation Approach) enabled 50% of the employee population in Ireland (800,000 of 1.6 million employees) to be matched to the Census 2011 dataset, as part of the SESADP. This enabled the Census dataset to provide variables for the SES (e.g. education level and occupation). Outputs from the SESADP produced the SES data for 2011 to 2014 and avoided having to do an expensive business Survey. IT and statistical infrastructure are now in place to produce the SES on an annual basis going forward, reducing costs from €1.6million annually to €0.1million. A Cost/Benefit Analysis of the SESADP is given in Figure 11.

Data Quality

An analysis of the data was undertaken to determine if the ICA matched the Census correctly to the other administrative data sources (ADS). There was a 90% correlation with the individual's *employer name* on Census with the *employer name* on the Business Register. The 10% with a different business name were eliminated as false positives. In Figure 12 a distribution of employees by age group and NACE in the Census dataset shows a

very good comparison with the SESADP (similar results were obtained for education, occupation & gender comparisons).

	Business Survey former NES	SESADP Project	SESADP (Annual)
Survey Type	Annual Survey	Data for 4 years	Annual basis - going forward
Reference period	Years 2002 to 2009	2011 to 2014	2015 onwards
Cost	€ 1.6 million p.a.	€0.4m	€0.1m p.a.
Timeliness	T+ 18	2 years to develop	T+ 10
Data edits	Data Edits	No edits - Revenue data	No edits - Revenue data
Sample size	65k	800k	800k+
Coverage of employee population	4%	50%	50%+
Burden	70,000 employees	None	None
Burden	5,000 enterprises	None	None
Staff Nos.	15 FTEs	4	2
Savings	-	€6.0m	€1.5m p.a.

Figure 11. Cost/Benefit Analysis of the SESADP.

NACE Economic Sector	% difference in employee nos.					
	Age Group in years					
	15-24	25-29	30-39	40-49	50-59	60 and over
	%	%	%	%	%	%
B-E Industry	-1	-1	2	1	0	-1
F Construction	-1	-1	3	0	-1	-1
G Wholesale & Retail Trade; Repair of Motor Vehicles and Motorcycles	0	-1	1	1	0	-1
H Transportation and Storage	-1	-1	0	2	1	-1
I Accomodation and Food Service Activities	-1	-1	2	1	0	-1
J Information and Communication	-1	-2	3	0	1	-1
K-L Financial, Insurance and Real Estate	-1	0	3	0	0	-1
M Professional, Scientific and Technical Activities	0	1	3	-1	-2	-1
N Administrative and Support Service Activities	0	1	4	0	-2	-2
O Public Administration and Defence; Compulsory Social Security	-1	-1	2	1	-1	-1
P Education	-2	-2	2	2	0	0
Q Human Health and Social Activities	0	-1	1	1	-1	-1
R-S Arts, entertainment, recreation and other service activities	0	1	1	1	-1	-2

Figure 12. Employees Nos. in SESADP compared to Census dataset by Nace Sector and Age Group 2011.

Annex A: Mathematical Representation of Identity Correlation Approach (ICA)

(I) Matching Rate for Unique Identifier (MRUI)

$$N \times \frac{1}{V1_{ui}} \times \frac{1}{V2_{ui}} \times \frac{1}{V3_{ui}} \times \frac{1}{V4_{ui}} \times \ldots\ldots\ldots \times \frac{1}{VX_{ui}} = MRUI \qquad \text{eqn.1}$$

(Assumes records are distributed evenly across all classes)

Where:

N = Population Size

V = Variable

X = No. of variables

ui = Uniqueness Factor = no. of classes in the variable (If records are distributed evenly across all classes).

If ui= N, then each record in the dataset can be directly matched (i.e. a unique identifier variable for each record).

$$\prod_{ui>1} VX_{ui} \to N$$

then, MRUI $\to 1$

$$\prod_{ui>1} VX_{ui} \to \infty$$

then, MRUI $\to 0$

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

(II) Adjusted Matching Rate for Unique Identifier (aMRUI)

$$N \times \frac{1}{\mathrm{V1_{di}}} \times \frac{1}{\mathrm{V2_{di}}} \times \frac{1}{\mathrm{V3_{di}}} \times \frac{1}{\mathrm{V4_{di}}} \times \dots\dots\dots \times \frac{1}{\mathrm{VX_{di}}} = MRUI \qquad \text{eqn.2}$$

(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).

If MRUI >1, there is no unique identifier for records in the largest class, however, there may exist a unique identifier for records in the smaller classes.

Annex B: Application of aMRUI Equation

Example 1:

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

Variable Name	Symbol	No. Classes	Proportion of Records in Dominant Class	Description of Classes
V1 = DoB	$V1_{di}$	365	$\frac{1}{365}$	di = 363 (days of the year) Classes evenly distributed
V2 = Gender	$V2_{di}$	2	$\frac{1}{2}$	2 genders (approx. 50% split) Classes evenly distributed
V3 = NACE 2 digit code	$V3_{di}$	50	$\frac{1}{10}$	50 different NACE2 digit codes, but approx. 1/10 of pop. in dominant nace2
V4 = marital status	$V4_{di}$	6	$\frac{1}{2}$	6 Marital status codes, Approx. 1/2 of people married, other classes = 50% (e.g. partner, single, divorced, separated, widowed)
V5 = county	$V5_{di}$	20	$\frac{1}{5}$	20 counties, but one dominant county with 1/5 of the population

Using aMRUI equation (eqn. 2):

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

MRUI = 0.89

In this example the MRUI < 1 then 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.

Example 2:

In the above example, if the records were evenly distributed across classes, then MRUI equation is used:

Using MRUI equation (eqn. 1):

$$65{,}000 \times \frac{1}{365} \times \frac{1}{2} \times \frac{1}{50} \times \frac{1}{6} \times \frac{1}{20} = \frac{65{,}000}{73{,}000} = 0.015$$

(Assuming records were evenly distributed across classes)

MRUI = 0.015

In this example the MRUI < 1 then there is a unique identifier for the individual employee in the dataset.

References

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

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.*

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

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.*

National Employment Survey 2008 and 2009 (2011), *Central Statistics Office, Ireland.*

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*

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*

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

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/