# Linking Census Data based on Similarity Threshold Method

Khin Su Mon Myint
*University of Information Technology*
Yangon, Myanmar
ksmonmyint@uit.edu.mm

Thiri Haymar Kyaw
*Faculty of Computer Science*
*University of Information Technology*
Yangon, Myanmar
thirihaymarkyaw@uit.edu.mm

Win Win Naing
*Faculty of Computer Science*
*University of Information Technology*
Yangon, Myanmar
winwinnaing@uit.edu.mm

***Abstract*— Historical censuses consist of specific information about a nation's population. It provides invaluable snapshots of the state of a country. These data can be used to reconstruct important aspects of a specific period in order to trace their ancestors and families changes over time. Linking census data is a challenging task due to poor data quality, common names and household structure changes over time. During the decades, a household may split multiple households due to marriage or move to another household. In this paper, we propose an approach for data cleaning, standardization and linking of historical census data across time. This computes similarity of selected attributes with approximate string similarity matching methods such as Q-gram, String extract match, Longest common subsequences and Gaussian probability. Then, the matches and un-matches records are determined by similarity threshold method. The results of the experiment show which optimal threshold value is effective for household linkage.**

***Keywords*— historical censuses, data cleaning, data matching, record linkage, household linkage and pair-wise linkage***

## I. INTRODUCTION

The population census data provide useful information in a specific region. They play an important role in analyzing for the social, economic, education and demographic aspects of a population [4, 5, 6] in that region. These data can also be used for planning or reconstruction purposes in the country.

Censuses are taken regularly by governments every ten years. These data allow us to understand populations and their different characteristics such as population size, age structure, household compositions, occupations, and other socio-demographic aspects [11].

Historical censuses contain specific information also gives the state of the nation and facilitate the construction aspects such as birth, death, education, occupation, etc. They help organization how our ancestors of the social and demographic changes in the country.

Linking record refers to the same households from several censuses that give across the decades. It is the process of observing records that refer to the same entities from different databases. These records will greatly enhance in value. The linked results have been allowed to trace varies in the characteristics of individual households over time.

Linked information improves not only retrieval of information, but also provides new opportunities for improving the quality of the data. It can also help social scientists with dynamic character of social, economic and demographic changes [9], which helps the reconstruction of the region.

Difficulties of historical census data linkage include poor data quality due to census data collection process. Importantly, the situation of individuals in a household may vary significantly between two censuses. For example, people are born and die, get married, change occupation, or moved home. As a result, linking individuals is not reliable, and many false matches are often generated.

Due to the benefits of historical census data linkage, there are large amount of data available, automatic or semi-automatic linking methods have been developed by data mining researchers and social scientists [3, 4, 5, 6]. These methods treat historical census data linkage as a special case of record linkage, and apply string comparison methods to match individuals. Some researchers use classification algorithms to classify matches or non-matches and use group linking approach to link households based on the matched records [2].

Most of researchers aim to find households with the majority of their members matched. However, during the ten years interval between two censuses, a household may split into multiple households due to marriage or move out to another household, or servants may change jobs.

Most previous works in census household linking problem can only be matched each individual in one household to one individual in another household. Then, previous historical census matching method couldn't support the household structure changes between the decades.

This paper proposes an approach for cleaning, standardization and linking of historical census data using domain knowledge. This work considers not only each individual in one household to one individual in another household but also takes multiple household linking.

The main idea is to use household information in the cleaning and linkage steps. So, that records which contains errors and variations can be cleaned and standardized and the number of incorrectly linked records can be reduced. The proposed approach starts by detecting Household Identifiers (HHIDs). These HHIDs together with name, address, gender, and relationship to the household head attributes, are used to clean the data. Record linkage is performed on record pairs, and then the linking results are improved using similarities results.

The rest of the paper is organized as follows. Section 2 introduces related works in data cleaning and linking, as well as their application to historical census data. Section 3 introduces the problem in census data. In Section 4 gives an overview of our approach. Section 5 describes detail of

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

historical census linking process. We report on our experimental results in Section 6, and conclude this paper in Section 7 and point out future research directions.

## II. Related Work

Difficulties of historical census data linkage came from several parts. These include poor data quality and large amount of similar values in names, address and ages.

It has a more important fact that the condition of individuals in a household may vary significantly between two census periods. For example, people are born and die, get marriage, moved home or change occupation or change their full name. As a result, linking individuals is not reliable and many false matched are generated. This is also a common problem in record linkage applications.

In recent years, computer science researchers have been developed new record linkage techniques that can be used to meet the challenges presented by linking historical census data. Christen et al. [2], [4] have proposed probabilistic data cleaning techniques for names and address that outperform traditional rules-based approaches. Christen have presented an overview of both pattern matching and phonetically encoding based name matching techniques.

Zhichun Fu [3] introduced an automatic method for linking both individuals and households across several historical census datasets. They applied the proposed method using six census datasets from the United Kingdom between 1851 and 1901.

P. Christen [1] proposed a method by supervised learning and group linking methods to link historical census households across time. This approach first computes the similarity between record pairs and uses these similarities as input to Support Vector Machine (SVM) classifier, which classifies record pairs into a matched and non-matched class. They used group linking techniques to generate household linking similarities.

To improve the quality of historical census record linkage, it is very important to examine domain driven approaches. The understanding of the domain social sciences needs and combines this knowledge with the data cleaning and record linkage methods by the computer science community [4][7]. A group linking method has been applied to generate a household match score by combining similarity scores from matched individual in a household [8].

One problem in the above methods for historical census matching is that matching is performed on the majority of members in a household over a period of time. However, a household may split multiple households between two censuses due to marriage or movement of another house or may change household structure due to birth and death between two censuses. So, the previous proposed methods cannot get accurate household matching results.

This paper considers not only 1:1 household matching but also 1: many household matching using Similarity Threshold Method.

## III. Problem in Census Data

This work uses two census datasets collected in ten-year intervals from the Ireland between 1851 and 1861. The data which contain twelve attributes such as full names, ages, sexes, and their relationship to the household, occupations, places of birth and etc.

Fig.1 illustrates the structural information of a household ($H_{1851}$) from 1851 Census. Fig.2 also shows the structural information of two households ($H_{1861}$- A and $H_{1861}$- B) from 1861 Census. The individuals are associated to a single household in each dataset. A household ($H_{1851}$) in 1851 splits two households ($H_{1861}$-A and $H_{1861}$-B) in 1861 due to marriage.

To understand the changes between two considered decades, one has to find matching individuals and their changes due to marriage and address changes. So, we also need to identify which individual occurs only in one census dataset because of births, deaths, immigration and emigration.

## IV. Overview of our Approach

The proposed approach constitutes two stages as illustrated in Fig. 3. They are data preprocessing stage and record similarity stage.

There are three processes in data preprocessing stage. In data cleaning and standardization, this is solving the low quality data problem in historical data collection. The aim is to find missing values, as well as to transform the data into a standardized form. This step also provides the data quality and increases the finding of true record matches between two datasets.

The second household detection process, which assigns unique household ID (HHID) to each household. The HHIDs are future used to define the household.

The third step is blocking and indexing. In this step, datasets are subdivided into several blocks using a blocking keys (index keys), only records in the same block are compared with each other, that greatly reduces the number of record pairs which need to be compared and so speeds up the linkage process. Only record pairs which have an identical blocking key are compared with each other.

| $H_{1851}$ | SUR NAME | FIRST NAME | Relation-ship to head | Sex | Age | STREET | COUNTRY |
|---|---|---|---|---|---|---|---|
| r11 | rickard | Thomas | head | M | 40 | Aughnacur | Cavan |
| r12 | rickard | Kate | wife | F | 38 | Aughnacur | Cavan |
| r13 | rickard | William | son | M | 18 | Aughnacur | Cavan |
| r14 | rickard | James | son | M | 13 | Aughnacur | Cavan |
| r15 | rickard | kathleen | daughter | f | 20 | Aughnacur | Cavan |
| r16 | rickard | Peter | father | F | 60 | Aughnacur | Cavan |
| r17 | rickard | Bridget | mother | M | 59 | Aughnacur | Cavan |

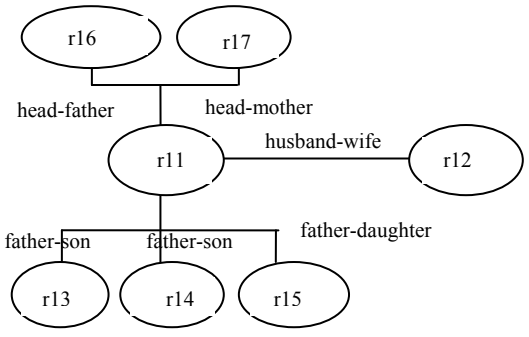
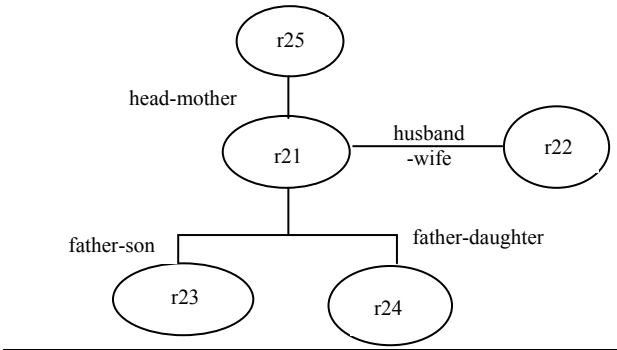



Fig. 1. An Example of a household ( $H_{1851}$) from 1851 Census

| $H_{1861}$-A | SUR NAME | FIRST NAME | Relationship to head of household | Sex | Age | STREET | COUNTRY |
|---|---|---|---|---|---|---|---|
| r21 | rickard | thomas | head | m | 50 | Aughnacur | Cavan |
| r22 | rickard | kate | wife | f | 48 | Aughnacur | Cavan |
| r23 | rickard | james | son | m | 23 | Aughnacur | Cavan |
| r24 | rickard | kathleen | daughter | f | 20 | Aughnacur | Cavan |
| r25 | rickard | bridget | mother | f | 69 | Aughnacur | Cavan |




| $H_{1861}$-B | SUR NAME | FIRST NAME | Relationship to head of household | Sex | Age | STREET | COUNTRY |
|---|---|---|---|---|---|---|---|
| r31 | rickard | william | head | m | 28 | Aghullaghy | Cavan |
| r32 | rickard | ellens | wife | f | 20 | Aghullaghy | Cavan |
| r33 | rickard | lusei | daughter | f | 3 | Aghullaghy | Cavan |

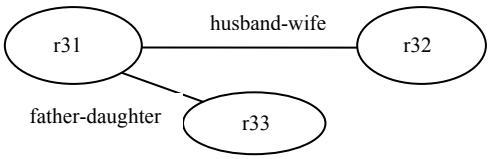



Fig. 2. An Example of two households ( $H_{1861}$-A and $H_{1861}$-B) from 1861 Census

In record similarity stage, the purpose is to compute similarities between two records. Several similarity methods have been used for this purpose. The attribute similarities were calculated by using similarity methods.

After that, attribute similarity results were summed to get record-pair similarity. Finally, candidate record pairs are classified into matches and non-matches record pairs by setting similarity threshold values.

## V. DETAIL OF HISTORICAL CENSUS LINKING PROCESS

### *A. Data Cleaning and Standardization*

The purpose of this step is to improve data quality from raw census data. It is applied for improving the quality of the data and formatting the data into a unified format. The census data return form filled by hand.

These include missing values, inconsistent values and wrong values. Errors were introduced in these stages. An example is FIRST NAME attributes with digits, letters, and other symbols which require cleaning and standardization to be applied.

The other example is the type of AGE attribute, which is mixture of digits and letters. This implies that the values were entered in different formats. Therefore, data standardization is required which improve the quality of the data and format the data to a reliable format in data cleaning and standardization step.

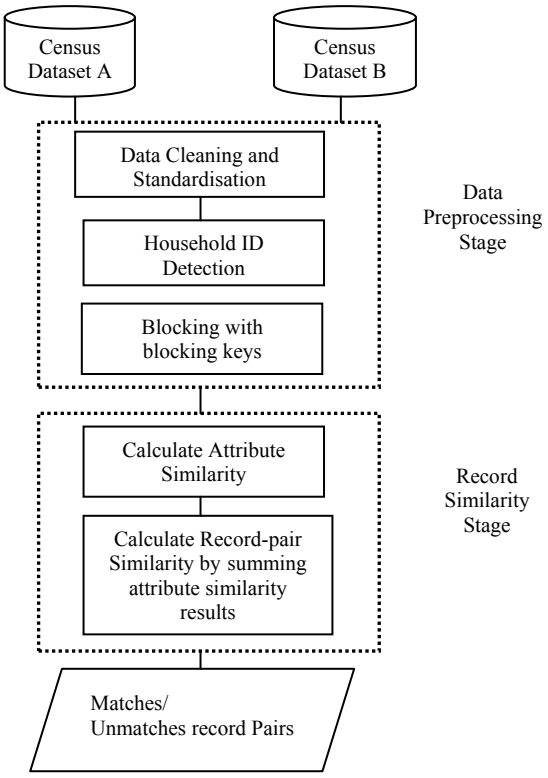



Fig. 3. Proposed Approach for historical census linkage process

This step is applied for getting the better quality of the data and structuring the data to a unified form. Many non-meaningful values are in the data. These include symbols such as "=" , "?" , and non-standard words, such as "no entry" and "not identified".

All these values are not useful. These values have been removed to improve the data quality. An example of data cleaning of gender example of data cleaning of gender values, for example, value "mm" is replaced with "m".

The standardization step formats the data into a unified form such as field names were standardized to uppercase letters and attributes values were converted to lowercase letters. It includes several operations, for example removing non-meaningful values such as "=", "?" and non-standard words, such as "no entry" and "not identified" and unifying the age format into digits-only.

The standardization step includes several operations. They are:

- All values are converted into lowercase letters
- First and middle names are split into two attributes
- The age format into a digit-only format that represent an age as number of years

### *B. Household Detection*

The purpose of household ID detection is to assign a unique household ID (HID) to each household. In each census form, the relationship to the head of household attribute always starts with the head of household. A record has a head of household role, the HID number is incremented by one, and this HID number is assigned to all following

records until another record with a head of the household role is found.

In the census table, the value for the Relationship attribute for each household should start by the head of the household. Based on the domain knowledge, possible values for the head of the household are “head”, “head of family”, “widow”, “widower” and “husband”.

We have been developed a linear algorithm to scan through census data file. If the record has a head of household role, the household ID (HHID) number is incremented by one, and this HHID is assigned to all rest records until other record with a head of household role is found. Fig.4 describes the construction of unique household ID.

*C. Blocking/ Indexing*

Before the linking process, we first applied a blocking technique to reduce the complexity of pairwise linking. This technique subdivides the datasets into several blocks, so only records in the same block are compared. When large datasets are used, the linking process is very time consuming. It is due to compare all pairs of records from both datasets.

This technique subdivides the datasets into several blocks, so only records in the same block are compared. When large datasets are used, the linking process is very time consuming. It is due to compare all pairs of records from both datasets.

We selected four key attributes and used Double Metaphone encoding algorithm to generate blocking keys. Double Metaphone phonetic algorithm allows multiple encodings for strings that have various possible pronunciations. This step greatly speeds up to the linking process.

The following three blocking keys are applied:

- first three letters of “SURNAME” attribute with “Double Metaphone” concatenated with the “SEX” attribute
- first three letters of the “FIRST_NAME” attribute with “Double Metaphone” concatenated with first four letters of the “ADDRESS” with “Double Metaphone”
- first three letters of the “FIRST_NAME” attribute with “Double Metaphone” concatenated with first four letters of the “SURNAME” with “Double Metaphone”

**Input:**
- All households in the dataset

**Output:**
- All households with unique household ID

```
1. household_ID = 0
2. for record ∈ House do
3.    Get "Relation to Head" field value in record
4. If relationHead == "head of family" ||  relationHead == "head"  ||
      relationHead == "widow" || relationHead == "widower" then
5.       household_ID = household_ID + 1
6.   End If
7.   End for
```

Fig. 4. HouseholdID Detecting Algorithm

*D. Household linkage*

When comparing the records, appropriate approximate string comparison functions have been chosen for each attribute. The list of attributes and functions used to compute the similarities between values is shown in Table I.

If the score of records are higher, the two attributes are more similar (scores of 1 indicate an exact match, 0 means no similarity).

Q-gram based approximate string comparison is applied on “SURNAME” and “FIRST NAME” attributes. Q-gram based approximate string comparison is applied on “SURNAME” and “FIRST NAME” attributes. Q-gram based approximate string comparison is to split the two input strings into short sub-strings of length q characters (called Q-grams).

In “SEX” attribute, string extract match algorithm is applied to compare two sex values. Gaussian probability is used to compare the different age values.

Longest common subsequence is used to compare “ADDRESS” attribute. This algorithm repeatedly finds and removes the longest common sub-string in the two strings compared, up to a minimum length (sets to 2 or 3).

The attribute-wise linking generates a similarity score for each attribute. A vector $Rs(r_{t,i,j}, r_{t`,i`,j`})$ can be got for each attribute. A vector $Rs(r_{t,i,j}, r_{t`,i`,j`})$ can be got for record $r_{t,i,j}$ from one dataset and $rt`,i`,j`$ from another dataset. We denoted the similarity vector as $Rs(r,r`)$.

By summing over all attribute-wise similarity scores, a total similarity score Rsim(a, b) can be calculated.For Rsim(a, b), the larger the similarity value, the more similar two records are.

We find matches and un-matches category is comparing the similarity Rsim(a,b) against a predefined threshold $\rho$. If $Rsim(a,b) \geqslant \rho$, the record pair is considered to be a match record pair.

In the experimental section, we will discuss how the value for $\rho$ is set based on the analysis of the linking results. After eliminating using threshold value $\rho$, multiple false matches for a single record can be reduced.

TABLE I. SIMILARITY METHODS USED FOR THE FIVE ATTRIBUTES

| Attribute | Method |
|---|---|
| Surname | Q-gram |
| First name | Q-gram |
| Sex | String extract match |
| Age | Gaussian probability |
| Address | Longest common subsequence |

## VI. EXPERIMENTAL RESULTS

We now provide the experiments to evaluate the record matching using the different similarity threshold values. The goal of these experiments was to get which threshold value achieves the best 1:1 or 1:many matching results for household linkage and compare their matching results.

We used two census data from Ireland historical census datasets. These data collected from the district of Aghullaghy in Cavan in Ireland for the period of 1851 and 1961. There are twelve attributes for each record, full name, age, sex, relationship to the household head, occupation and place of birth et al.

These data were standardized and cleaned before applying the household linkage step. In total, there are 96 and 97 records in the two datasets.

As mentioned previously, five attributes (SURNAME, FIRST NAME, SEX, AGE, and ADDRESS) were used in our study. After each of the attributes were cleaned, unique household ID (HHID) were identified.

Before pair-wise linking process, the datasets have been divided into many small blocks based on the three blocking keys as previous mentioned. This step tends to speed up our record comparison process.

Once the cleaned and identified household id available, pair-wise linking is started with the records 1851 datasets compared with the records 1861 datasets.

The record linkage step generated the similarity score of each selected attributes of the records, which the range of 0 and 1. The score are higher, the records are more similar. By combining all five score values, a total score $0 \leqslant Sa,b \leqslant 5$ can be calculated for each record pairs ra,b.

To define matches and un-matches record pairs, appropriate setting of the threshold value $\rho$ is very important. We evaluated the linking results with the respect value of the $\rho$. We applied five threshold values (2.5, 3.0, 3.5, 4, and 4.5) to evaluate the results as shown in Fig. 5.

The number of records in the 1851 dataset with exactly one matched records and multiple matched records in the 1911 data set, when different threshold values $\rho$ have been set.

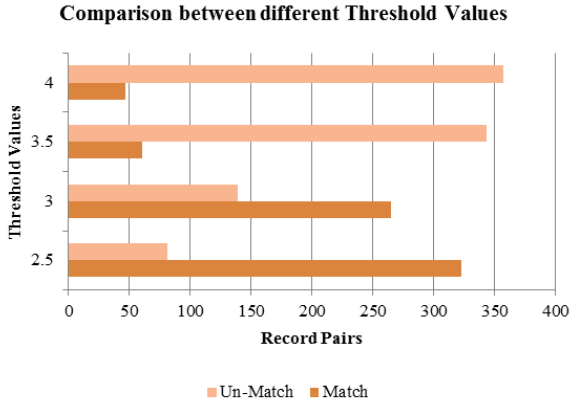



Fig. 5. Matches/Un-Matches record pairs with different threshold $\rho$

The spread of single matched records and multiple matched records are different for different $\rho$ value. The numbers of records with multiple matches have been reduced by increasing the $\rho$.

When $\rho$ value is set to 4.5, which is high considering that only selected attributes are used, there are only 22 single match record pairs. However, many true multiple record pairs are still containing in the unmatched pairs.

When $\rho$ value is set to 4, there are only 47 single (one-to-one) match record pairs but no multiple (one-to-many) matches. So, no multiple matches are found when $\rho > 4$. On the other hand, when $\rho$ is too low, a lot of multiple false matches are generated.

We evaluate the precision rate, recall rate and accuracy of the record pairs on different threshold $\rho$ values. Fig. 6 show the precision rate, recall rate and accuracy with different threshold values.

As the data shown in Fig.6 , the precision rates for the five threshold values are from 17% to 100%, the recall rates are from 44% to 100% and the accuracy are from 34% to 97%.

We found that threshold value 4 achieves the best accuracy rate of 97%, precision rate of 91% and recall rate of 86%. Although it had the best accuracy rate, it had missed many multiple (one-to-many) true links.

Threshold value 3.5 can provide 95% at accuracy, 90% at precision and 93% at recall rate. Threshold value 3 can provide 47% at accuracy, 20% at precision and 100% at recall rate. It generates many false matches.

By manually evaluating the results, threshold value 3.5 covers not only single match record but also multiple match records. It can provide 1:1 household linkage and 1:many household linkage.

This suggests that 3.5 could be an appropriate threshold value for record linking in our work. Therefore, the experiment helps us to select the most appropriate threshold value for record matching.

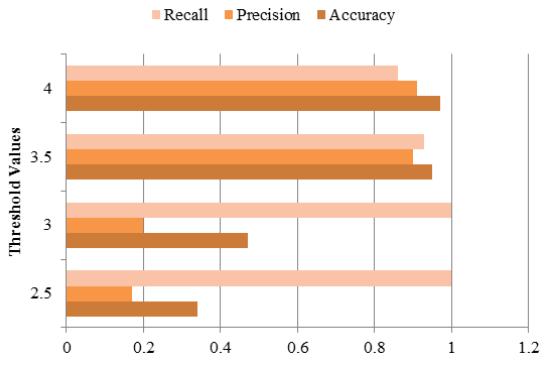



Fig. 6. Comparison of performance of record linkage with different threshold $\rho$ value

## VII. CONCLUSION

In this paper, we have introduced a data cleaning and linking approach with similarity threshold method for historical census data. The goal is to reduce ambiguous links and match households over a certain period of time. This approach uses household information, to take the record cleaning and linking steps. The record linking process is executed in two steps. The first one computes each attribute similarity scores using approximate string matching algorithms. And then a pair-wise record linkage is defined with the total similarity values. After record pairs similarities are computed, matches or un-matches are classified by setting appropriate threshold values. The experimental result shows that the matches and un-matches record pairs with different threshold values. The result also shows that ambiguous matches results exist after the threshold step. This is due to the structures of two households are very similar and family members can change substantially over time.

In the future, we will explore a classification algorithm and use graph learning methods to detect household structures due to getting married or household splitting into multiple households between the decades.

## REFERENCES


[1] Z. Fu,P. Christen, Mac Boot, "A Supervised Learning and Group Linking Method for Historical Census Household Linkage", Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, 2011 Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, 2011

[2] Z. Fu, P. Christen, Mac Boot, "Automatic cleaning and linking of historical census data using household information. In: IEEE ICDM Workshop. pp. 413–420 (2011)2.

[3] Z. Fu, H.M. Boot, Peter Christen and Jun Zhou, " Automatic Record Linkage of Individuals and Households in Historical Census Data", International Journal of Humanities and Arts Computing 8.2 , 204-225, 2014

[4] P. Christen, "Development and user experiences of an open source data cleaning, deduplication and record linkage system," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 39–48, 2009.

[5] A. Ashkpour, K. Mandemeakers and A. Meronopenuela, "The Aggregate Dutch Historical Censuse",Historical Methods", Vol 48, Number 4, Oct-Dec 2015

[6] Fure, E.: Interactive record linkage: The cumulative construction of life courses. Demographic Research 3, 11 (2000)

[7] S. Ruggles, "Linking historical censuses: a new approach," History and Computing, vol. 14, no. 1+2, pp. 213–224, 2006.

[8] Bloothooft, G.: Multi-source family reconstruction. History and Computing 7(2), 90–103 (1995)

[9] D. V. Kalashnikov and S. Mehrotra, "Domain-independent data cleaning via analysis of entity-relationship graph", ACM Transactions on Database Systems, vol. 31, no. 2, 2006

[10] B.- W. On, N. Koudas, D. Lee, and D.Srivastava, "Group linkage" ,in Proceedings of the IEEE 23rd International Conference on Data Engineering, 2007

[11] D. Quass and P. Starkey, "Record linkage for genealogical databases," in ACM KDD Workshop, Washington DC, 2003

[12] http://www.census.nationalarchives.ie/