Universal Evaluation System Data Quality

María del Pilar Angeles, Francisco Javier García-Ugalde,
Carlos Ortiz, Ricardo Valencia,Eduardo Reyes,
Arturo Nava, Jhovany Pelcastre
Facultad de Ingeniería
Universidad Nacional Autónoma de México
México, D.F.
pilarang@unam.mx, fgarciau@unam.mx,
carlos.ortiz.eng@comunidad.unam.mx, ricardofdk8@hotmail.com, eduardorv07@hotmail.com,
arturo_shox@hotmail.com, j.pelcastre@hotmail.com

***Abstract*— This paper presents a work in progress regarding to the extension and improvement of the Freely Extensible Biomedical Record Linkage system. Currently, the prototype has been extended to directly connect to a number of database managements systems, calculate their database quality indicators, automatically generate a flat file from any database and execute an appropriate data matching process.**

Keywords-data matching; de-duplication; record linkage

I. INTRODUCTION

When an enterprise information system is meant to be built upon integration of their existing heterogeneous database systems, they would face the difficulty of comparing disparate schemas in order to identify syntactic and semantic heterogeneities; make these schemas correspond and match them through transformation functions; and finally, comparing data of unknown quality such as name, address from a single record against a large number of records.

Integrating data from different sources consists of three tasks [1]. The first task is concerned with identifying database tables, attributes and conceptual structures from disparate databases that contain data that correspond to the same type of information, namely schema matching [2]. The second task is concerned with the identification and match of records that correspond to the same entity, when they come from disparate data sources, called data matching. In the case of identification of records that actually refer to the same entity within a single database, is known as duplicate detection [1]. Duplicated records can be handled in different ways, providing the complete set of inconsistent answers, providing the complete set of answers, but ranked according to likelihood of being correct [3], providing a single value selected at random, providing a top value in a ranked answer, or providing a fused answer [4], which is the process of merging pairs or groups of records that have been classified as matches into a clean and consistent record that represents an entity. When applied on one database, this process is called de-duplication. We assume that the process of schema matching has already achieved.

The open issues on data matching are mainly concerned to the record comparison among databases in order to determine if a pair of records corresponds to the same entity or not, because the process grows exponentially as the databases to be matched get larger. In real-world data matching applications, the true status of two records that are matched across two databases is not known. Therefore, accurately assessing data matching quality and completeness is challenging [1].

This approach is aimed to the development of algorithms that reduce the quadratic complexity of the naive process of pair-wise comparing each record from one database with all records in the other database, and how to accurately classify the compared record pairs into matches and non-matches considering attributes dependency.

Nowadays, we are focused on the implementation of algorithms in order to measure, assess and help during the analysis of data quality process under a number of open and licensed database management system (DBMS), such as Oracle DB, MySQL, IBM DB2, SAP-Sybase Adaptive Server Enterprise, SAP-Sybase IQ, EnterpriseDB PostgreSQL.

For the process of identification, analysis and merge of duplicated records, we are still working on the extension of the Freely Available Record Linkage System (FEBRL) [5] developed by a research group of the Department of Computer Science at The Australian National University.

We are focused on the integration of the FEBRL system to any database from any DBMS by querying the native data dictionary; the research proposal is also aimed to the enhancement and addition of further standardization, indexing, and classification algorithms for data matching.

We have called our prototype as FEBRL-SEUCAD, it will support six DBMS at least. The application extracts the database schema directly from the data dictionary and measures the intrinsic quality of the data through the following indicators: coverage, density, completeness [6]. Since these measures are intrinsically computed through SQL queries, the assessed granularity levels are at database, table and column where applicable as we have done in previous research [7]. Furthermore, the prototype will implement a specific framework for the detection, classification and fusion (cleaning) of duplicate records within a number of databases (data matching and de-duplication) with no regard of the type of data source.

The present paper is organized as follows: The next section is focused on the assessment of data quality. The third section briefly explains the data matching process. Section IV describes the work we have carried out regarding

 ISBN: 978-1-61208-334-6

to the extension and enhancement of algorithms of the data matching process. The last section concludes the main topics achieved and the future work to be done.

II. ASSESSMENT OF DATA QUALITY

The strictness of quality assessment is a weak or strong characterization depending on evaluating the quality property as a percentage or as a Boolean function respectively as shown in [8]. The strong characterization of the quality metrics is useful in applications where it is not possible to admit errors at the corresponding level of granularity.

In the case of the assessment of data quality, we have considered the weak strictness to make possible the comparison of data sources for a number of data quality properties in. However, there might be alternatives where strictness could depend on the level of quality required, according to specific applications.

In order to assess data quality at different levels of granularity [7], we have utilized the measures provided at lower levels of granularity (data value, attribute) to determine aggregated scores (table, database) as we move through the levels of granularity.

Regarding completeness, we have taken the corresponding metrics of [6] and [8] for the value, attribute, and relation granularity levels, and we have incorporated completeness at the database level.

Coverage: This is the measure for the number of tuples a source stores; in other words as the probability that an entity of the world is represented in the source [6]. This is also contemplated under the Open World Assumption without nulls completeness case, at the relation level of granularity, refer to [8] for further detail.

Density of an attribute: is the measure of how well the attributes stored at a source are filled with actual (non-null) values (columns), in [6], a weak attribute completeness case under the Closed World Assumption with nulls in [8].

Density of the source d(S) is obtained by the average density over all density attributes [6].

Weak relation completeness is the number of tuples with all its attributes filled with non-null values divided by the number of tuples [6].

The completeness at database level will correspond to the average completeness of its corresponding relations.

The measurements are given by the aggregation of values at each of these levels as they are moving on. As a measurement of data quality is directly related to the level of granularity, we conclude that scores measured at lower level of granularity will provide a greater degree of accuracy than aggregated scores produced at higher levels.

The functions utilized for aggregation of scores are commonly average, maximum, and minimum. The appropriateness of an aggregation function will depend on the optimistic, conservative, or pessimistic approach taken according with the application context. It is not our intension to identify the best aggregation function, because there is not an absolute value. As long as the aggregation function reflects the user needs and it is consistently used, it should be enough for the estimation of quality and comparison purposes.

III. THE DATA MATCHING PROCESS

A. Introduction

The data matching process in general terms is focused on joining records from one data source with another that describe the same entity. This process requires the following tasks: data standardization [9]; indexing possible matching data in order to reduce the number of comparisons; data comparison and classification of pairs of records in possible match, not match and match. These steps are briefly explained during this section.

B. Standardization

The standardization process [9] refers to the conversion of input data from multiple databases into a format that allows correct and efficient record correspondence between two data sources. Within the first step, called tokenization, it is assumed that the attributes of the input databases contain values that are separated by spaces, known as tokens. The second step is concerned with the detection and correction of data values that contain typographical errors or variations already known. The third step is the segmentation of tokens in well-defined output fields for proper data mapping or identification and correction of duplicate values (known as de-duplication).

C. Indexing

A detailed process of records comparison is usually computationally expensive. That is, the complexity is quadratic according to the length of the attribute values (mostly chains) that are correlated. The comparison process is the most complex of all the data mapping steps. The indexing aims to reduce the number of pairs of records that will be compared, reducing those pairs of records that are unlikely to correspond to the same real world entity and retaining those records that probably would correspond in the same block for comparison reducing the number of record comparisons. Therefore, the definition of the locking key is very important, because it will specify how to keep similar records in the same block of comparison. The record similarity depends on the data types they contain because they can be similar phonetically, numerically or textually. Some of the methods implemented within Febrl are for instance, Soundex [10], Phonex [1], Phonix [1], NYSIIS[11], Double metaphone [12], QGrams.

D. Field and record Comparison Methods

As the comparison data might be of low quality (they may contain typographical errors or variations), establishing a binary or strict criterion for the comparison process such as (similar / dissimilar) is not possible or realistic. Therefore, the comparison methods implemented provide degrees of similarity and define thresholds depending on the semantics and data type of each field. Some of the methods implemented within Febrl are for instance, Qgram, Jaro - Winkler Distance [13], [14] Longest common substring Comparison.

 ISBN: 978-1-61208-334-6

E. Classification

The classification of pairs of records grouped and compared in the previous steps [15], [6], is mainly based on the similarity values were obtained, since it is assumed that the more similar two records are, there is more probability that these records belong to the same entity of the real world.

The records are classified into “matches”, “not matches” or “possible matches”, the classification of records can be an unsupervised or supervised process.

The unsupervised process classifies pairs or groups of records in the similarities between them without having access to more information about the characteristics of those records.

The supervised process requires training based on data identified as similar or not similar. In this case, comparison vectors with an associated value that determines whether records correspond or not are required.

In the case of potentially corresponding records, the duplicates detection may be performed manually.
Within Febrl, there are methods based on thresholds, probabilistic methods, costs based methods or rule-based methods.

The accuracy of data matching is mostly influenced by the comparison and classification steps. However, the indexing step will impact on the completeness of a data matching exercise because record pairs filtered out in the indexing step will be classified as non-matches without being compared.

The most commonly way to classify candidate record pairs is to sum the similarity values in their comparison vectors into a single total similarity value and to then apply two similarity thresholds to decide the class a candidate record pair belongs to. However, there are some dependencies between attributes. For instance, records with the same postcode will potentially have the same street name [1]. Therefore, if we assume that all similarity values are normalized between 0 and 1, all attribute similarities contribute in the same way towards the final summed similarity value. The importance of different attributes, as well as their discriminative power regard to distinguishing matches from non-matches, is not considered. Furthermore, with no regard of a weighted or an unweight approach, the detailed information contained in the individual similarity values is lost by such a simple summation approach.

For instance, the probabilistic classification approach of Fellegi and Sunter [6] is one of the most utilized nowadays because it allows the calculation of weights for corresponding and not corresponding pairs of attribute values, which leads to a better decision during records pair classification, but by assuming a conditional independence.

The present research is currently on the implementation of an enhancement to this issue.

F. Evaluation of Matching

Matching quality refers to how many of the classified matches correspond to true real-world entities, while matching completeness is concerned with how many of the real-world entities that appear in both databases were correctly matched [17].

Each of the record pair corresponds to one of the following categories [18]:

• True positives (TP). These are the record pairs that have been classified as matches and that are true matches. These are the pairs where both records refer to the same entity.

• False positives (FP). These are the record pairs that have been classified as matches, but they are not true matches. The two records in these pairs refer to two different entities. The classifier has made a wrong decision with these record pairs. These pairs are also known as false matches.

• True negative (TN). These are the record pairs that have been classified as non-matches, and they are true non-matches. The two records in pairs in this category do refer to two different real-world entities.

• False negatives FN) .These are the record pairs that have been classified as non-matches, but they are actually true matches. The two records in these pairs refer to the same entity. The classifier has made a wrong decision with these record pairs. These pairs are also known as false non-matches.

An ideal outcome of a data matching project is to correctly classify as many of the true matches as true positives, while keeping both the number of false positives and false negatives small.

Precision calculates the proportion of how many of the classified matches (TP + FP) have been correctly classified as true matches (TP). It thus measures how precise a classifier is in classifying true matches. [19]. It is calculated as: precision= TP/(TP+FP)

Recall measures how many of the actual true matching record pairs have been correctly classified as matches [19]. It is calculated as: recall= TP/(TP+FN).

At the present time, we have been focused on the enhancement of the Fellegi and Sunter probabilistic classification in terms of keeping the match weight after classification.

G. Related work

The FEBRL project has developed prototype software which undertakes data standardisation, which is an essential pre-processing phase for most record linkage projects, and which implements the "classical" approach to probabilistic record linkage model as described by Fellegi and Sunter in [16]. We are focused on the extension of the original FEBRL system to any database from any DBMS by querying the native data dictionary; the research proposal is also aimed to the enhancement and addition of further standardization, indexing, and classification algorithms for data matching.

We are currently analysing which de-duplication algorithms are suitable for incorporating to the FEBRL-SEUCAD in order to implement them and compare them to the already implemented on FEBRL.

 ISBN: 978-1-61208-334-6

IV. FEBRL-SEUCAD

This section presents the enhancement we have implemented to the original FEBRL project so far. However, as we have pointed out before, there is a long way of further work to be done.

The prototype FEBRL-SEUCAD is now able to connect to any application implemented under a number of Database Management Systems such as PostgreSQL, SAP Sybase Adaptive Server Enterprise, SAP Sybase IQ, MySQL, Oracle and IBM DB2 with only the name of the database to be analyzed as a parameter. FEBRL-SEUCAD contains code to extract the native data dictionary in order to obtain all the database objects created under such database name.

A. Activities

To develop the prototype we have undertaken the following activities.

- Extraction of database objects by the data dictionary from each DBMS.
- Implementation by SQL programming of quality metrics such as coverage, density and uniqueness, the latter considering primary key, because otherwise would be data de-duplication or data matching covered by Febrl. Such sql programming has been carried out for each DBMS in their corresponding SQL language.
- Extension of the Febrl application for connection to any database through the already mentioned DBMS.
- Extension of the Febrl application to incorporate the options concerned to compute the quality metrics at database, table, record and column.
- Extension of the Febrl application to incorporate the option of selecting a specific database object and its corresponding data matching process.

The prototype currently supports the measurement of data qualitative dimensions within a number of database management systems and the steps within the data matching process (indexing, comparison, and classification).

The application extracts the database schemas directly from the data dictionary and identifies the following indicators: coverage, density, complete and uniqueness, since they are intrinsically calculable through SQL, granularity levels are calculated at database, table, and column log for a number of Database Management Systems.

B. Operations

This section is aimed to briefly describe the operation of the FEBRL-SEUCAD prototype. In the case of Data profiling, the first step is to select the metrics option, specify the required metric, the level of granularity to compute, the Database management system and the database name. We have extended the Febrl system in order to calculate completeness and uniqueness at different levels of granularity by extracting from the data dictionary the database objects, this feature allows the prototype to be utilized on any database platform. Fig. 1 shows completeness at database level from a SAP-Sybase Adaptive Server Enterprise database.

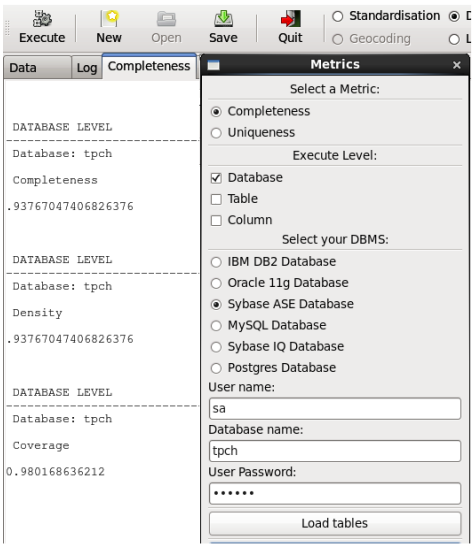
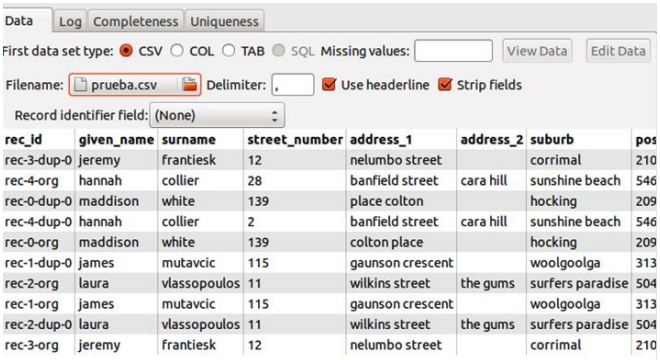


Figure 1 Completeness at database level

Once the database object has been selected, we have extended Febrl as shown in Fig. 2, in order to automatically generate the corresponding flat file in CSV, TBL formats, where originally the flat file had to be generated apart and then loaded to the application for further analysis.



rec_id	given_name	surname	street_number	address_1	address_2	suburb	pos
rec-3-dup-0	jeremy	frantiesk	12	nelumbo street		corrimal	210
rec-4-org	hannah	collier	28	banfield street	cara hill	sunshine beach	546
rec-0-dup-0	maddison	white	139	place colton		hocking	209
rec-4-dup-0	hannah	collier	2	banfield street	cara hill	sunshine beach	546
rec-0-org	maddison	white	139	colton place		hocking	209
rec-1-dup-0	james	mutavcic	115	gaunson crescent		woolgoolga	313
rec-2-org	laura	vlassopoulos	11	wilkins street	the gums	surfers paradise	504
rec-1-org	james	mutavcic	115	gaunson crescent		woolgoolga	313
rec-2-dup-0	laura	vlassopoulos	11	wilkins street	the gums	surfers paradise	504
rec-3-org	jeremy	frantiesk	12	nelumbo street		corrimal	210

Figure 2 Flat file generated from a database table

The data profiling step helps to determine the number of different data values for such attribute, the distance frequency, and the number of records with empty values. These brief data profiling allows the identification of a suitable attribute for indexing.

Indexing: The next step is the identification of attributes that would help on the execution of the indexing process along with specification of the corresponding parameters according to such indexing method. The best suited attributes for indexing are those with no missing values and uniform frequency distance. For instance, in the case of QGramIndex it is possible to specify the number of Q-grams and threshold as is shown in Fig. 3.

 ISBN: 978-1-61208-334-6

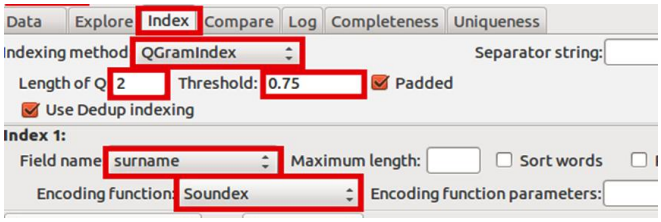


Figure 3 Indexing by QGramIndex and Soundex encoding

Comparison: The most common attributes for comparison are strings of one token, such as surname, family name, or a short number of tokens such as address, street name. Within SEUCAD-FEBRL is required to also specify the field comparison function per each comparison attribute. For instance, in Fig. 4, the comparison function Bag-Distance is used for surname and the Number-Percentage algorithm for the attribute street_number.

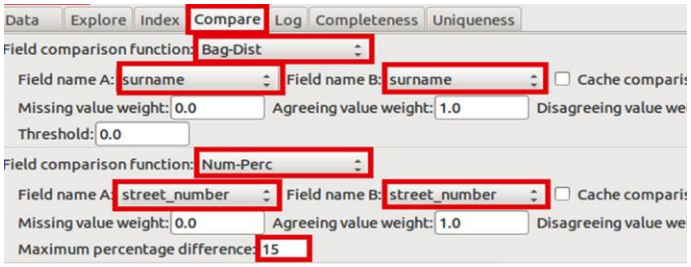


Figure 4 Comparison by Bag-Distance and Number percentage

Classification: The data type of the attributes involved for the classification process is relevant in order to specify the classification method. In Fig. 5 the classification method chosen is Fellegi Sunter with a lower threshold of 8 and a upper threshold of 1.8.

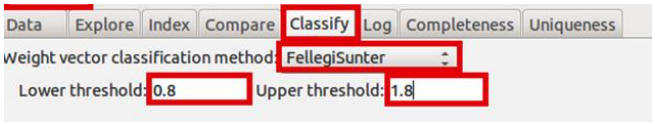


Figure 5 Classification by Fellegi and Sunter

Once identified the attributes and methods for each step, it is possible to execute the data matching process, which is the case shown in Fig. 6.

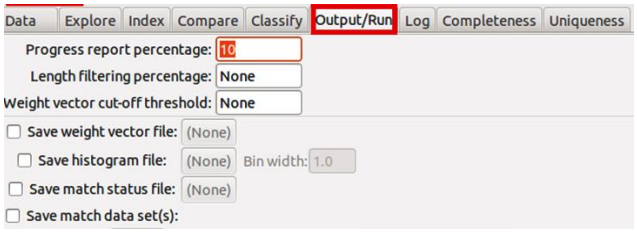


Figure 6 Execution of data matching process

Fig. 7 presents the number of matches, non-matches and possible matches as the outcome of the data matching process.

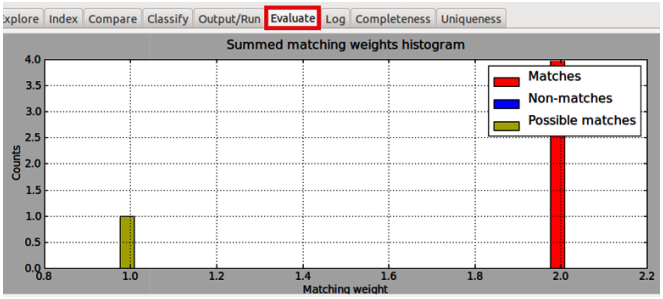


Figure 7 Outcome of matches, non matches and possible matches

The data matching outcome shows 4 matches, 1 possible match, there were not pair of records with non matches.

The evaluation of the data matching algorithms has already been coordinated by Peter Christen during the development of the FEBRL prototype [1], [5].

The data quality indicators: coverage, density, completeness and uniqueness have been identified in previous work [3], [6], [7], [8] and extended to different levels of granularity in [7]. However, as these quality indicators are an addition to the original de-duplication prototype, they have been calculated and tested at database, table, and column log for a number of Database Management Systems within FEBRL-SEUCAD.

V. CONCLUSION AND FUTURE WORK

A. *Conclusion*

We present a work in progress regarding the extension and improvement of an open software for de-duplication of records originally called FEBRL.

Currently, the FEBRL-SEUCAD prototype can directly connect to several database management systems such as Oracle DB, MySQL, IBM DB2, SAP-Sybase Adaptive Server Enterprise, SAP-Sybase IQ, EnterpriseDB PostgreSQL, extracted from the dictionary database objects of interest in order to allow data platform independency.

Our prototype is able to calculate data quality indicators allowing a better decision regarding to the identification of attributes that would help on the data matching process.

FEBRL-SEUCAD allows the flat file generation from any object database in CSV, TBL formats and then loaded to the application for further analysis.

B. *Future work*

We are planning the enhancement of some of data matching algorithms already implemented in the original FEBRL prototype, at the present time; we are focused on the enhancement of the classification process by considering the importance of different attributes through their corresponding weights.

 ISBN: 978-1-61208-334-6

The implementation of new indexing algorithms, comparison and classification methods is part of our future work.

ACKNOWLEDGMENT

This work is being supported by a grant from "Research Projects and Technology Innovation Support Program (Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica, PAPIIT", UNAM Project IN114413 named "Universal Evaluation System Data Quality (Sistema Evaluador Universal de Calidad de Datos)".

REFERENCES

[1] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Series Data-Centric Systems and Applications, Springer, 2012.

[2] E. Rahm and H.H. Do, "Data cleaning: Problems and current approaches". IEEE Data Engineering Bulletin 23(4), 2000, pp.3-13.

[3] P. Angeles and F. Garcia-Ugalde, "A Data Quality Practical Approach", para el "International Journal On Advances in Software" Vol. 2, No. 3, 2009, pp. 259-274.

[4] J. Bleiholder and F. Naumann, "Data fusion", ACM Computing Surveys 41(1), 2008, pp. 1-41.

[5] Febrl – A Freely Available Record Linkage System with a Graphical User Interface, Proceeding of the 2nd Australasian Workshop on Health Data and Knowledge Management (HDKM), Wollongong, Australia, 2008, pp.17-25.

[6] F. Naumann, J. Freytag, and U. Lesser, "Completeness of Integrated Information Sources", Workshop on Data Quality in Cooperative Information Systems (DQCIS2004), Cambridge, Mass., 2004, pp.583-615.

[7] P. Angeles and F. Garcia-Ugalde, "Assessing data quality of integrated data by quality aggregation of its ancestors", Computación y Sistemas, Centro de Investigación en Computación, Instituto Politécnico Nacional (IPN), vol. 13 No. 3, 2010, pp. 331-334, ISSN 1405-5546.

[8] M. Scannapieco and C. Batini, "Completeness in the Relational Model: A Comprehensive Framework", Research Paper, in Proceedings of the 9h International Conference on Information Quality (ICIQ-04, Cambridge, MA, USA, 2004, pp. 333-354.

[9] T. Churches, P. Christen, K. Lim, and J. X. Zhu, Preparation of name and address data for record linkage using hidden Markov models. BioMed Central Medical Informatics and Decision Making 2(9), 2002.

[10] M. Odell and R. Russell,The soundex coding system. US Patents 1261167, 1918.

[11] C. L. Borgman and S. L. Siegfried, "Getty's synonameTM and its cousins: A survey of applications of personal name-matching algorithms", Journal of the American Society for Information Science 43(7), 1992, pp. 459–476.

[12] L. Philips, "The double metaphone search algorithm", C/C++ Users J. 18, 6,pp. 38-43, 2000.

[13] M. A. Jaro, "Advances in record-linkage methodology applied to matching the 1985 Census of Tampa, Florida", Journal of the American Statistical Association 84, 1989, pp. 414–420.

[14] W. Winkler, "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage", Proceedings of the Section on Survey Research Methods, 1990, pp. 354–359, American Statistical Association.

[15] M. Neiling and H. J. Lenz, "Supplement of Information: Data Integration by Classification of Pairs of Records, Classification, Automation, and New Media Studies in Classification, Data Analysis, and Knowledge Organization", 2002, pp. 219-226, http://dx.doi.org/10.1007/978-3-642-55991-4_23 [retrieved: february, 2014], Springer Berlin Heidelberg, ISBN 978-3-540-43233-3.

[16] I. P. Fellegi and A. B. Sunter, "A theory for record linkage", Journal of the American Statistical Association 64(328), 1969, pp. 1183–1210.

[17] D. Barone, A. Maurino, F. Stella, and C. Batini, "A privacy-preserving framework for accuracy and completeness quality assessment", Emerging Paradigms in Informatics, Systems and Communication,2009, p. 83.

[18] P. Christen and K. Goiser, "Quality and complexity measures for data linkage and deduplication," In: F. Guillet, H. Hamilton (eds.) Quality Measures in Data Mining, Studies in Computational Intelligence, vol. 43, 2007, pp. 127–151, Springer.

[19] I. H. Witten, A. Moffat, and T. C. Bell, Managing Gigabytes, 2 Ed. Morgan Kaufmann, 1999.

 ISBN: 978-1-61208-334-6