

From theory to practice: the software RELAIS as a solution for record linkage

Tiziana Tuoto, Nicoletta Cibella, Marco Fortini, Monica Scannapieco¹

Abstract The combined use of statistical survey and administrative data is largely widespread to maximize their respective usefulness: unfortunately data sources are often hard to integrate due to errors or lacking information. Record linkage techniques are a multidisciplinary set of methods and practices aiming to identify the same real world entity, differently represented in data sources. Record linkage is a complex process but it can be decomposed in separate phases, each of them requiring a specific technique. To deal with such a problem, we propose RELAIS (REcord Linkage At IStat), an open source toolkit based on the idea of choosing the most appropriate technique for each phase and of dynamically combining them so as to build a *record linkage workflow*, given specific application constraints and input data features. The open source turned out to be a winning choice for sharing techniques and software. A case study shows the usability of RELAIS.

1 Introduction

The purpose of record linkage is to identify the same real world entity that can be differently represented in data sources, even when unique identifiers are not available or are affected by errors. In statistics, record linkage is needed for several applications, including: enriching the information stored in different datasets; de-duplicating datasets; improving the data quality of a source; measuring a population amount by capture-recapture method; checking the confidentiality of public-use microdata.

Starting from the earliest contributions, dated back to 1959 (Newcombe et al.), there has been a proliferation of different approaches based on statistics, databases,

¹ Tiziana Tuoto, Istat, tuoto@istat.it

Nicoletta Cibella, Istat, cibella@istat.it

Marco Fortini, Istat, fortini@istat.it

Monica Scannapieco, Istat, scannapi@istat.it

machine learning, knowledge representation. Despite this proliferation, however, no particular record linkage technique has emerged as the best solution for all cases. We believe that such a solution does not actually exist, and that an alternative approach should be adopted. In fact, it is suitable to see at record linkage as a complex process consisting of several phases and involving different knowledge areas; moreover, several different techniques can be adopted for each phase. The choice of the most appropriate technique not only depends on the practitioner's skill but, most of all, it is application specific. Furthermore, in some applications, there is no evidence to prefer a given method to others or of the fact that different choices, at a linkage stage, could bring to the same results. This is why it could be reasonable to dynamically select the most appropriate technique for each phase and to combine the selected techniques for building a record linkage strategy for a given application. In addition, from the analyst's point of view, it is important to have the possibility to experiment alternative criteria and parameters in the same application scenario.

In this paper we describe RELAIS (Record Linkage At Istat), a toolkit relying on these ideas (Fortini et al 2006, Tuoto et al 2007, Cibella et al 2008, Cibella et al 2009). This software is designed and developed to allow the combination of different techniques for each of the record linkage phases, so that the resulting strategy is actually built on the basis of application and data specific requirements. Moreover, this software aims to include not only a toolkit of techniques, but also a library of patterns that could support the definition of the most appropriate record linkage workflow. The toolkit has been developed as an open source project. This choice is motivated by the idea of re-using the several solutions already available for record linkage in the scientific community. It is released under the EUPL licence (European Union Public Licence).

The RELAIS project aims to provide record linkage techniques easily accessible to not-expert users. Indeed, the developed system has a GUI (Graphical User Interface) that permits to build record linkage strategy with a good flexibility and checks the execution order among the different provided techniques whereas precedence rules must be controlled.

The paper is organized as follows. In Section 2, we outline the main phases in which a record linkage process can be decomposed. In Section 3, we describe the idea, the design and the current state of implementation of RELAIS. In Section 4, a case study is described. Finally, in Section 5 some concluding remarks and direction for future works are provided.

2 Approaching record linkage problems in phases

The complexity of the whole linkage process relies on several aspects; for example the lack of unique identifiers requires sophisticated statistical procedures, the huge amount of data to process involves complex Information Technology solutions, constraints related to a specific application may require the solution of difficult linear programming problems. In order to better face with such a complexity, it can be suitable to decompose a record linkage process into main phases:

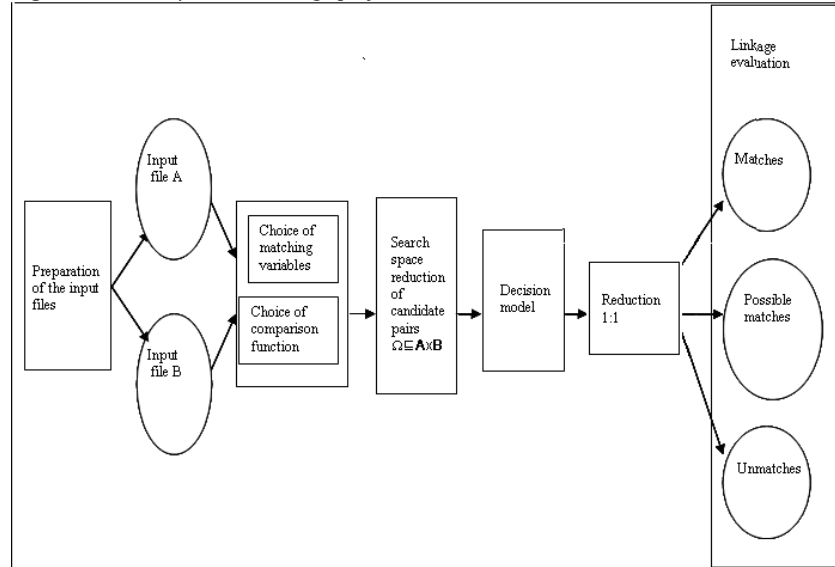
1. Pre-processing of the input files
2. Choice of the identifying attributes (matching variables)

From theory to practice: the software RELAIS as a solution for record linkage

3. Choice of the comparison function
4. Creation of the search space of link candidate pairs
5. Choice of the decision model
6. Selection of unique links
7. Record linkage evaluation

The phase decomposition (figure 1) allows to reduce the overall complexity of the linkage process, subdividing the whole problem into sub-problems.

Figure 1: Phases of a record linkage project



In this way the question on which method is better compared to the others is overcome and different methods or techniques among those proposed and available for each one of the linkage phase can be selected in order to achieve the definition of the most appropriate overall strategy on the basis of application and data specific requirements.

Generally speaking, the *preparation of input files* is the first phase which, according to Gill (2001), requires 75% of the whole effort to implement a record linkage procedure. The key job of this phase is to convert the input data in a pre-defined format, resolving the inconsistencies in order to reduce errors deriving from an incorrect reported data. In this phase null strings are cancelled, abbreviations, punctuation marks, upper/lower cases, etc. are cleaned and any necessary transformation is carried out to standardize variables. Furthermore the spelling variations are replaced with standard spelling for the common words.

After the previous phase, it is important to *choose matching variables* that are as suitable as possible for the considered linking process. The matching attributes are generally chosen by a domain expert; a set a metadata can support the users in the choice of matching attributes. If unique identifiers are available in the linkable data

sources, the easiest and most efficient way is to use these ones as link variables; but very strict controls need to be made in case of using numeric identifiers alone. It is evident that the more heterogeneous are the items of a variable, the higher is its identification power; moreover, if missing cases are relevant in a field it is not useful to choose it as a matching variable.

Comparison functions are used to compute the distance between records on the values of the chosen matching variables, depending on the kind of variables and their accuracy. For a reviews of comparison functions see Koudas and Srivastava (2005).

In a linking process of two datasets, say A and B , the pairs needed to be classified as matches, non-matches and possible matches are those in the cross product $A \times B$. If a de-duplication problem is considered the space is $A \times (A-1)/2$. When dealing with large datasets, comparing all the pairs $(a; b)$, a belonging to A and b belonging to B , in the cross product is almost impracticable: in fact while the number of possible matches increases linearly, the computational problem raises quadratic. To reduce this complexity it is necessary to reduce the number of pairs $(a; b)$ to be compared. There are many different techniques that can be applied to *reduce the search space*: techniques of sorting, filtering, clustering and indexing may be all used to reduce the search space.

Starting from the reduced search space, different *decision models* can be applied in order to classify pairs into M , the set of matches, and U , the set of non-matches, or in the set of possible matches. The decision rule can be classified as deterministic or probabilistic. While it is difficult to make a clear distinction between the two approaches, especially with respect to proposals coming from the computer science area, a difference between deterministic and probabilistic approaches is often made in research literature, where the former is associated with the use of deterministic merging rules while the latter makes an explicit use of probabilities for deciding when a given pair of records is actually a match. In practice, the deterministic and the probabilistic approaches can be combined in subsequent iteration of a record linkage process.

The probabilistic approach is the core of the statistical problem for record linkage and requires the definition of a decision model and the estimation of the model parameters. From the seminal paper of Fellegi and Sunter (1969), several models have been proposed, essentially defining latent variable for the linkage status or approaching by Bayesian methods. See Gu et al. (2003) for a review.

A linkage process can be also classified as: (i) one-to-one problem, if one record in the set A links to only one record in B and also the other way around, (ii) one-to-many problem if a record in a set can be matched with more than one of the compared file, (iii) many-to-many problem if more than one record in each file might match with more than one record in the other. The latter two problems may imply the existence of duplicate records in the linkable data sources.

Finally, as not every record matched in the linkage process refers to the same identity, in the *record linkage procedure evaluation*, it's important to establish whether a match is a "true one" or not. In other words, during a linkage project is necessary to classify records as true link or true non link, minimizing the two types of possible errors: false matches and false non-matches. The first type of error refers to matched records which do not represent the same entity, while the latter indicates unmatched records not correctly classified, that imply truly matched entities were not linked.

3 The record linkage phases as currently implemented in RELAIS

Due to the great attention on the data integration topic and the complexity of this problems, several record linkage systems and tools have been proposed, in both the academic and private sectors. Such tools include, for example, Big Match (Yancey, 2007), CANLINK (Fair, 2001), Febrl (<http://www.sourceforge.net/projects/febrl>), Link Plus (<http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm>), Tailor (Elfeky et al. 2002), The Link King (<http://www.the-link-king.com>). The first two systems have been developed at the U.S. Bureau of the Census and the Statistics Canada respectively, the other systems have been developed at medical-epidemiological centres or at universities. Some of the systems provide a certain degree of flexibility for the user: however, any of these tools provides the flexibility of multiple choices for *each* of the record linkage phase.

Keeping in mind the approach to record linkage as reported in the previous paragraph and in order to deal with both the modularity of the record linkage problem and the need of flexible choices, we proposed the RELAIS toolkit (Fortini et al 2006, Tuoto et al 2007). This toolkit is composed by a collection of techniques for each record linkage phase that can be dynamically combined in order to build the *most suitable record linkage strategy*, given a set of application constraints and data features. Some phases of the record linkage process can be missing: for instance the search space reduction phase makes sense only for large data volumes.

The strength of RELAIS consists of considering alternative techniques for the different phases that compose the record linkage process. RELAIS aims to help and guide users in defining their specific linkage strategy, supporting the practitioner's skill, due to the fact that most of the available techniques are inherently complex, thus requiring not trivial knowledge in order to be appropriately combined. RELAIS is proposed also as a toolkit for researchers: in fact, it gives the possibility to experiment alternative criteria and parameters in the same application scenario, that's really important from the analyst's point of view.

RELAIS has been designed with a modular structure. The modules implement distinct record linkage techniques and each one has a well defined interface towards other modules. In this way it is possible to have a parallel development of the different modules, and to easily include new ones in the system. Moreover, the overall record linkage process can be designed according to specific application requirements, combining the available modules. A user interface guides the design of the record linkage workflow in order to help the user in easily combining the available modules in a meaningful and controlled way.

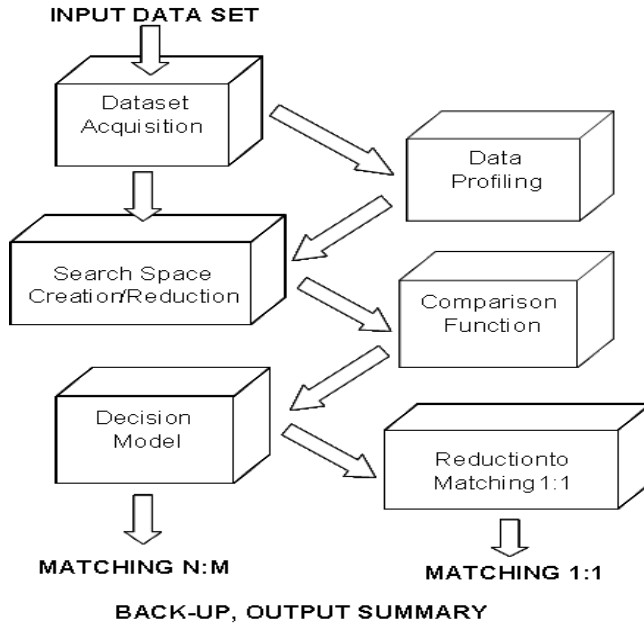
Finally, RELAIS is configured as an open source project, released under the EU Public Licence. In this way, there are many possible techniques that can be implemented in parallel for each record linkage phase: relying on a community of developers such set can be increased and maintained very rapidly. In the last years there have been several independent efforts towards the resolution of record linkage problems but such efforts have not led to the best solution. An open source project could instead give the possibility of gathering together the contributions already done in order to make them available to the community for the most appropriate usage.

From the implementation viewpoint, RELAIS is written in Java and R languages. Both languages are open source and can be used on different technological platforms.

The combined usage of these two languages permits to rely on the best features of each of them. Specifically, Java is used for the data-oriented tasks and for the development of the user interface, while R is used for the computational-oriented ones. Some R packages have been used to solve specific problems as a clear example of the possible re-use allowed by open source projects. R code is embedded in Java code, so that the calls to R software are completely transparent to the final users. While the first beta version of RELAIS (the 1.0) had a file based architecture, meaning that all data reside on text files, both input and output data as well as the intermediately produced data, the currently available version RELAIS 2.0 is evolved into a relational database architecture that permits to manage larger amounts of data more efficiently. In particular it is based on a mySql environment that is also in line with the open source philosophy of the RELAIS project.

Figure 2 shows the record linkage phases implemented in the RELAIS system.

Figure 2: *Phases implemented in RELAIS*



The *dataset acquisition* phase permits to read two input datasets from text files. The datasets must have the same names for the common variables that are the ones considered by the system in the subsequent phases. The database architecture allows both to start new project and to continue working to a previous one, saved as back-up. From the acquisition phase, it is possible to pass directly to the search space creation/reduction phase or to the data profiling phase.

The *data profiling* phase permits to characterize available variables with respect to some quality features that can be used to support two critical tasks, that is blocking variables choice and matching variables selection. To give the opportunity to the user

From theory to practice: the software RELAIS as a solution for record linkage

of designing the more appropriate record linkage workflow for its own application, RELAIS 2.0 supplies quality metadata, calculated starting from real data provided as input. Moreover, in order to go towards needs of less-expert users, RELAIS proposes also a default set of parameters, coming from communities and manuals, to help the decision-making stages. In this phase, the metadata of quality are: Completeness, Accuracy, Consistency, Entropy, Correlation and Frequency Distributions.

The *search space creation/reduction* phase allows to build the set of the candidate pairs to be linked. Besides the complete cross product of the file to link, two methods for space reduction are implemented, namely blocking and sorted neighborhood method.

A set of *comparison function* is available in order to compare strings according to an exact or an approximate procedure. The comparison function provided by RELAIS 2.0 are: Equality, Numeric Comparison, 3Grams, Dice, Jaro, Jaro-Winkler, Levenshtein, Soundex (<http://www.dcs.shef.ac.uk/~sam/stringmetrics.html>).

As far as the choice of the *decision model* to determine the matching status on candidate pairs is concerned, the current version of RELAIS implements two kind of models, *deterministic* and *probabilistic*. Deterministic approach allows two options. According to some authors, deterministic record linkage is defined as the method that individuates links if and only if there is a full agreement of unique identifiers or a set of common identifiers, i.e. the matching variables. This corresponds to the *Exact Match* option in RELAIS. Other authors backed up that in deterministic context a pair can be linked also if some specific and pre-defined criteria are satisfied. Being not exact in the strict sense this kind of linkage is assumed as almost-exact and RELAIS defines it as *Rule-based Match*. The matching rules are defined by the users throughout the selection of matching variables and related comparison function in a disjunctive format proposed by RELAIS.

The probabilistic model currently available in RELAIS consists of an implementation of the *Fellegi-Sunter decision model*, assuming latent dichotomous variable for the linkage status and conditional independence model for the manifest variables. The EM algorithm is used so to estimate the parameters.

When blocking method is performed to reduce the search space of pairs, RELAIS allows the users to choose between two different ways of applying the probabilistic model: it can be applied in a one-shot way to all the blocks or a specific block can be selected.

On the results of the model assignment, it is possible to produce an N:M matching result or a 1:1 matching result, applying a dedicated reduction phase (Jaro, 1989). The latter phase can be applied by resolving a linear programming problem on the N:M output by means of the simplex algorithm (optimal solution) or by a greedy algorithm, when the amount of data prevents from applying the simplex method due to its complexity. The reduction from a matching M:N to a matching 1:1 is available for both probabilistic and deterministic matching.

Finally, the output of the linkage process consists of several disjoint datasets: match, non-match pairs, possible match and residuals. For the Possible matches no decision is taken and they need to be processed by clerical review or by further linkage process. Also residual non-matched records resulting from the two starting files can be submitted to further analyses, that is a new record linkage process can be started by processing the residuals directly or, as an alternative, later by means of a residual back-up. Also intermediate outputs can be saved, such as blocking summary, contingency tables and parameter estimate tables.

4 A case study: building a record linkage workflow

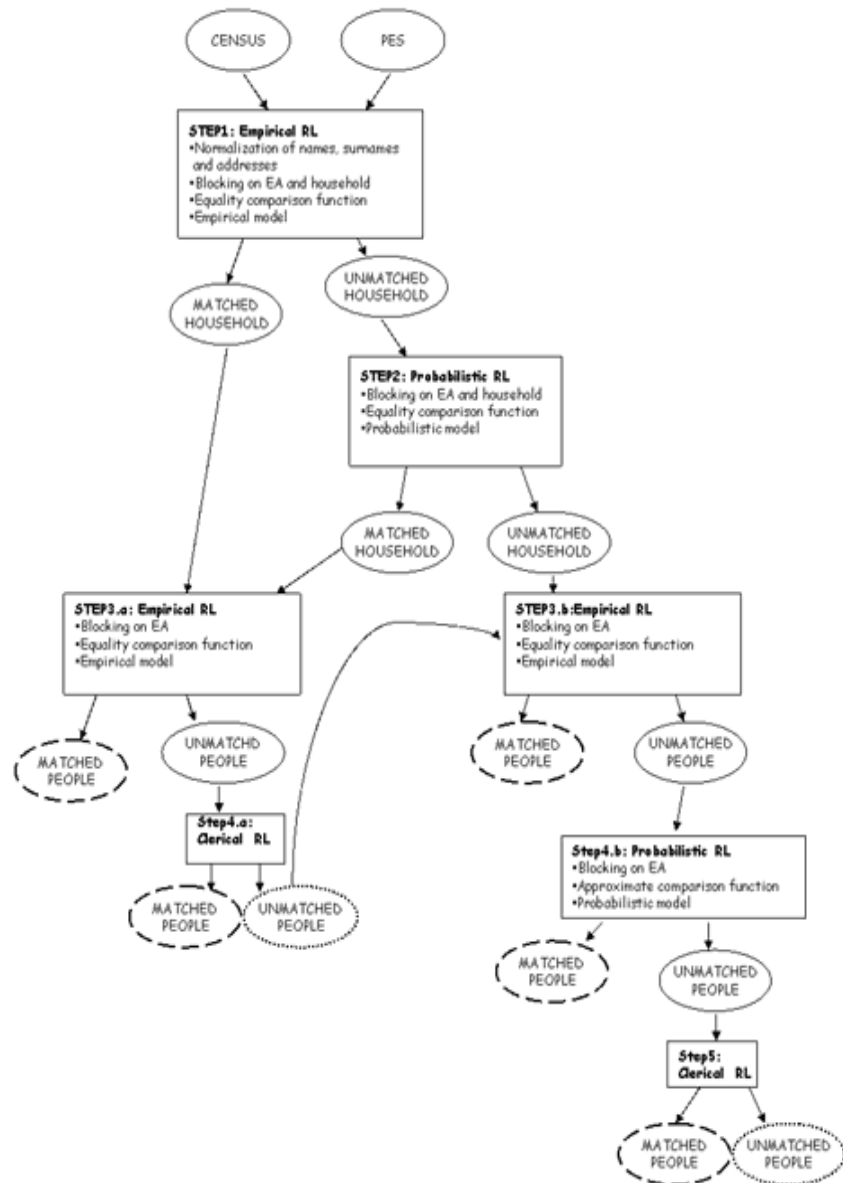
In this paragraph we briefly report the main aspects of facing a complex integration process, namely the Post Enumeration Survey (called PES in the following) of the Italian 2001 Census, according to the RELAIS idea.

The main goal of the Census was to enumerate the resident population at the Census date, the 21th of October 2001; it was also interesting to characterize Italian families, hence, the relationship of each enumerated person with his relatives was collected. The PES had the objective of estimating the coverage rate of the Census; it was carried out on a sample of enumeration areas (called *EA* in the following), which are the smallest territorial level considered by the Census. The size of the PES's sample was about 65.000 households and 170.000 people. Correspondingly, comparable amounts of households and people were selected from the Census database with respect to the same EAs. The PES was based on the replication of the Census process inside the sampled EAs and on the use of a capture-recapture model (Wolter K., 2006) for estimating the hidden amount of the population. In order to apply the capture-recapture model, after the PES enumeration of the statistical units (households and people), a record linkage between the two lists of people built up by the Census and the PES was performed. In this way the rate of coverage, consisting of the ratio between the people enumerated at the Census day and the hidden amount of the population, was obtained.

The estimates of the Census coverage rate through capture-recapture model have required to match Census and PES records, assuming no errors in matching operations. This is a strong assumption: the accuracy of the matching processes was of crucial importance because even very small matching errors could have compromised the reliability of the coverage rate estimates. To guarantee the maximum correctness of the matches between PES and Census, we had to build a structured record linkage workflow, consisting of different phases and iterations. Specifically, both empirical and probabilistic record linkage techniques were used, and also different comparison functions were selected in different phases. The resulting workflow is particular significant as a proof of concept of the RELAIS toolkit usefulness. More specifically, the first phases of the workflow identify the *easiest* matches, by means of the more straightforward computational procedures, leaving the hardest ones to the subsequent phases. The iterations of the record linkage workflow were performed on the basis of the hierarchical structure of the data, in order to take advantage of the relationships among individuals belonging to the same household. Indeed, the matching units corresponding to people can be grouped according to their households membership; this structure suggests to start by first linking households and then individuals.

From theory to practice: the software RELAIS as a solution for record linkage

Figure 3: The record linkage workflow of the case study



In Figure 3, steps 1 and 2 regard two iterations of the record linkage process on households. Step 1 is performed after a pre-processing activity and it is a deterministic linkage. Step 2 is a probabilistic record linkage, based on the Fellegi-Sunter model, for

which the matching weights are computed via the EM algorithm (Jaro, 1985). In step 3.a a deterministic linkage was performed on matched household for the purpose of identifying people. In the subsequent step 4.a, the residual individuals, not yet linked but belonging to matched households, were clerically checked. The un-matched people in the output of step 4.a were considered as input to step 3.b, together with the individuals belonging to not linked households, and were matched by means of a deterministic approach. Then, in step 4.b, for people not linked in step 3.b, a probabilistic record linkage was carried out. The residual individuals, not yet linked at the previous steps, were submitted to a final clerically linkage in step 5.

As described above, given a set of application constraints and data features, RELAIS has the purpose to suggest the best technique to choose in each record linkage phase, in order to build the best workflow for the specific application. The case study described above allows us to highlight the following requirements: (i) the data requirements include a hierarchical structure of the data sets, a quite large dimensionality and a high quality of the data; (ii) the application requirements include not significant errors in the matching process. The hierarchical structure suggests to distinguish record linkage workflow iterations at two levels, namely: we first match records at a higher level (households), and then at a lower level (persons). In this way, we take advantage of the hierarchical structure reducing the search space and, moreover, increasing the number of real matches. The dimension of the data sets implies high complexity of the linkage algorithm; this suggests to apply blocking techniques to reduce the complexity of the linkage. Moreover, due to volume of the data sets, a direct use of the probabilistic model could have been time consuming. Therefore, a first application of the deterministic model is performed with the purpose to be refined by the subsequent use of the probabilistic model. The high quality of data implies the choice of equality as comparison function in most of the phases. The requirement concerning not significant errors in the matching process suggests the adoption of a probabilistic model in the final iterations, in order to have a quantitative estimation of the errors that can be regarded as acceptable or not. Moreover, this requirement also suggests the appropriateness of a clerical review and an exact comparison function in order to achieve the desired error bounds.

5 Concluding remarks and future works

In this paper, we have illustrated the RELAIS project an open source toolkit for building record linkage workflows. The idea behind this project has been developed keeping in mind: (i) the complexity of a record linkage problem, which involves different methods and techniques; (ii) the opportunity of considering the linkage as a modular process, by identifying several phases to be carried out in their proper order, or even iteratively; and (iii) the different suitable approaches depending on both the data features (e.g. type of data, amount of data) and the application requirements (e.g. efficiency, efficacy, accuracy). The toolkit aims to offer multiple techniques for record linkage, both deterministic and probabilistic, with the possibility of building ad-hoc solution combining each modules. In fact, RELAIS proposes solutions based on a set of choices and parameters, i.e. which variables have to be used for blocking and matching respectively, the choice of the decision model, whose parameter settings is

left to the user's choices. However, RELAIS also suggests to non-experts how to take a decision and to assign values to parameters. Moreover, solutions are not fixed for each application since finding the ideal methods and parameters is not straightforward, or no such single ideal solution even exists. In fact, a variety of methods exist with different pros and cons; the same method can yield bipolar results against different applications and datasets, so dynamically selecting and combining the suitable method and parameters for the given record linkage problem and dataset allow to handle diverse scenarios and make existing solutions more flexible and applicable.

The RELAIS project have been enriched also thanks to the cooperation with many researchers and users in international context (Essnet ISAD and Essnet DI projects, Cibella et al, 2009). At the same time other remarks come from the profitable share of knowledge and solutions among different institutes and countries in dealing with 'real-world' tasks: first of all, the awareness of the common nature of the faced problems; then, the advantages in designing standardized answers to specific, though widespread, applications.

In future work, we plan to extend the current functionalities of RELAIS and to optimize its performances, especially to deal with the next Italian Censuses. As far as the Population Census is concerned, the extensive use of data integration techniques is a key challenge to support innovations as like the use of administrative registers to contact people on the field while taking into account for coverage issue.

References

- Cibella, N., Fortini, M., Scannapieco, M., Tosco, L., Tuoto, T.: Theory and practice of developing a record linkage software, In Proc. of the Combination of surveys and administrative data Workshop of the CENEX Statistical Methodology Project Area "Integration of survey and administrative data", Vienna, Austria (2008)
- Cibella, N., Fernandez, G.L., Fortini, M., Guigò, M., Hernandez, F., Scannapieco, M., Tosco, L., Tuoto, T.: Sharing Solutions for Record Linkage: the RELAIS Software and the Italian and Spanish Experiences, In Proc. of the New Techniques and Technologies for Statistics (NTTS) Conference, Bruxelles, Belgium (2009)
- Elfeky, M., Verykios, V., Elmagarmid, A.K.: Tailor: A Record Linkage Toolbox. In Proceedings of the 18th International Conference on Data Engineering. IEEE Computer Society, San Jose, CA, USA (2002)
- Fair, M.: Recent developments at Statistics Canada in the linking of complex health files. In Federal Committee on Statistical Methodology, Washington D.C., USA (2001)
- Fellegi, I.P., Sunter, A.B.: A Theory for Record Linkage. Journal of the American Statistical Association, {64}, 1183--1210 (1969)
- Fortini, M., Scannapieco, M., Tosco, L. Tuoto, T.: Towards an Open Source Toolkit for Building Record Linkage Workflows. In Proc. of SIGMOD 2006 Workshop on Information Quality in Information Systems (IQIS), Chicago, USA (2006)
- Gill, L.: Methods for Automatic Record Matching and Linkage and their Use in National Statistics. National Statistics Methodological Series no. 25, HMSO Norwich, UK (2001)
- Gu, L., Baxter, R., Vickers, D., Rainsford, C.: Record linkage: Current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, Australia (2003)
- Jaro, M.: Advances in Record Linkage Methodologies as Applied to Matching the 1985 Census of Tampa, Florida. Journal of American Statistical Society, {84}, 414--420 (1985)
- Koudas, N., Srivastava, D.: Approximate Joins: Concepts and Techniques. In Proc. of International Conference on Very Large Data Bases (VLDB), Trondheim, Norway (2005)

- Newcombe, H., Kennedy, J., Axford, S., James, A.: Automatic Linkage of Vital Records, Science, Vol. {130}, 954--959 (1959)
- Tuoto, T. , Cibella, N., Fortini, M., Scannapieco, M. Tosco, L.: RELAIS: Don't Get Lost in a Record Linkage Project, In Proc. of the Federal Committee on Statistical Methodologies (FCSM) Research Conference, Arlington, VA, USA, (2007)
- Wolter, K.: Some coverage error models for census data. Journal of the American Statistical Association, {81}, 338--346 (1986)
- Yancey, W.: BigMatch: A Program for Extracting Probable Matches from a Large File. Technical report, Statistical Research Division U.S. Bureau of the Census - Washington D.C. Research Report Series - Computing n. 2007-01 (2007)