RELAIS: Don't Get Lost in a Record Linkage Project

Tiziana Tuoto

**Istituto Nazionale di Statistica ISTAT via Magenta 2, 00185 Roma (Italy),
tel. 0039- 06-46733545, fax. 0039-06-46733712 tuoto@istat.it**

Nicoletta Cibella

Istituto Nazionale di Statistica ISTAT, cibella@istat.it

Marco Fortini

Istituto Nazionale di Statistica ISTAT, fortini@istat.it

Monica Scannapieco

Istituto Nazionale di Statistica ISTAT, scannapi@istat.it

Laura Tosco

Istituto Nazionale di Statistica ISTAT, tosco@istat.it

Abstract

The combined use of statistical survey and administrative data is largely widespread: joint analyses of statistical and administrative sources allow to save time and money, reducing survey costs, response burden, etc. However, data sources are sometimes hard to combine since errors or lacking information in the record identifiers may complicate the integration.

The record linkage techniques are a multidisciplinary set of methods and practices whose purpose is to identify the same real world entity, which can be differently represented in data sources. Record linkage has been subject of research for several decades: a huge number of record linkage solutions has been proposed, based on probabilistic and empirical paradigms. Being record linkage a complex process, it can be decomposed in separate phases, each requiring a specific technique and depending on given application and data requirements. To deal with such complexity and application dependency, we propose a toolkit for record linkage, called RELAIS (REcord Linkage At IStat). The toolkit is based on the idea of choosing the most appropriate technique for each phase, and of dynamically combining such techniques in order to build a *record linkage workflow*, on the basis of application constraints and data features provided as input. The modular structure of the workflow allows to omit and/or iterate some phases of the record linkage process. Moreover, the RELAIS toolkit is configured as an open source project, which allows the users community to "gather" the most appropriate techniques from the efforts of several independent experts.

A real case study, based on the Post-Enumeration Survey of the XIV Italian Population Census, validates the RELAIS idea and provides a methodological pattern for driving the design of a record linkage workflow on the basis of the requirements of a real application.

Introduction

Record linkage is a process that essentially aims to quickly and accurately identify if two (or more) records represent the same real world entity or not. A record linkage project can be performed for different purposes and the variety of the uses makes it a powerful instrument to support decisions in large commercial organizations and government institutions.

In official statistics, the field in which this work is developed, the combined use of statistical survey and administrative data is largely widespread and strongly stimulates the investigation of new methodologies and instruments to deal with record linkage projects.

Beginning from the early contributions to modern record linkage, dated back to Newcombe et al (1959) and to Fellegi and Sunter (1969), there has been a proliferation of different approaches, that make use of techniques based on data mining, machine learning, equational theory, mainly thanks to the overcoming of constraints relevant to the great development of

computer sciences. However, despite this proliferation, 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 strategy should be adopted. Specifically, record linkage can be seen as a complex process consisting of several distinct phases involving different knowledge areas, in addition for each phase, several techniques can be adopted. We consider that the choice of the most appropriate technique not only depends on the practitioner's skill but most of all it is application specific. Moreover in some applications, there are not evidences to prefer a given method to others or that different choices at some linkage stages could conduct to the same results. In addition, from the analyst's point of view, it is important to be able to experiment and modify the alternative criteria and parameters. That's 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 workflow of a given application.

In this paper we describe the RELAIS (REcord Linkage At IStat) toolkit, which relies on the above described ideas. RELAIS allows combining techniques proposed for each of the record linkage phases, so that the resulting workflow is actually built on the basis of application and data specific requirements. Moreover, the RELAIS project will include not only a toolkit of techniques, but also a library of *patterns* that, given specific data and application requirements, could support the definition of the most appropriate record linkage workflow. We start to develop the RELAIS project as an open source project. This is a choice motivated by the idea of re-using the several solutions already available for record linkage in the scientific community, and by the quite ambitious goal of providing, in the shortest possible time, a generalized toolkit for building dynamic record linkage workflows.

The major contributions of this paper can be summarized as follows. First we outline in detail the philosophy and the purposes of the RELAIS, then the phases that compose a record linkage project are explained, finally we illustrate the idea of the dynamic record linkage workflow as implemented in RELAIS. Therefore we demonstrate the RELAIS idea by means of a real case study in which a record linkage workflow is instantiated starting from data and application requirements.

RELAIS: a toolkit for building record linkage workflows

In official statistics, the combined use of statistical survey and administrative data is largely widespread and strongly stimulates the investigation of new methodologies and instruments to deal with record linkage projects. At present, many potential advantages in using administrative data for statistical purposes are known and shared by the various national statistical institutes: in fact, administrative sources usually contain larger amounts of data, possibly more accurate due to improvements over time, so in most situations the joint analysis of two or more statistical and administrative sources allows to save time and money, reducing survey costs, response burden, etc. Indeed, cooperation among different public agencies or institutes is actually based on common data sharing, that prevents from recollecting data from citizens or enterprises, if such data are already available at some of the public subjects. The various different uses of record linkage in official statistics include: *update* and *de-duplication of frame*, when multiple records referring to the same real world entity are stored within one single data source; *data integration*, across multiple data sources in order to provide a reconciled global record; *correction* across multiple data sources, performed when one source is known to have higher quality data that can be used for improving the others; *measure of a population by capture-recapture* for instance on the occasion of the post-enumeration survey of Census; *check of the confidentiality of public-use microdata*, through re-identification experiments.

However, data sources are often hard to combine since errors or lacking information in the record identifiers may complicate the integrated use of the information; in order to overcome such obstacles record linkage techniques provide multidisciplinary set of methods and practices whose purpose is to identify the same real world entity, which can be differently represented in one or more data sources.

The general and formal definition of the record linkage dates back to Fellegi and Sunter (1969). They approached the problem via a probabilistic decision model which is still widely used; recently, different approaches from the computer science field have been proposed, based on data mining, machine learning, equational theory (see Hernandez and Stolfo 1998, Monge and Elkan 1997, Ananthakrishna *et al.* 2002, Chauduri *et al.* 2005). These approaches can be classified as *empirical* in contrast with the strictly *probabilistic* record linkage procedures, which following Fellegi-Sunter, make explicitly use of probabilities. Anyway there is no evidence that a specific record linkage technique can perfectly solve all matching problems. Really, such a solution does not actually exist, whereas the linkage issues should be faced from an alternative point of view. In fact, record linkage can be seen as a complex multi-disciplinary process, composed by several distinct phases: from one hand, pre-processing, in which standardization activities are performed, and blocking, for reducing the number of comparisons, mainly involve computer science; from the other hand, the choice of a comparison function, to be used for the record comparisons, the decision rules, for coming up with a set of matched records and a set of unmatched ones, and the linkage error rate estimation involve statistics; finally linear programming methods implicate operations research. In addition for each phase, several techniques can be adopted; for instance, in order to reduce the complexity of the

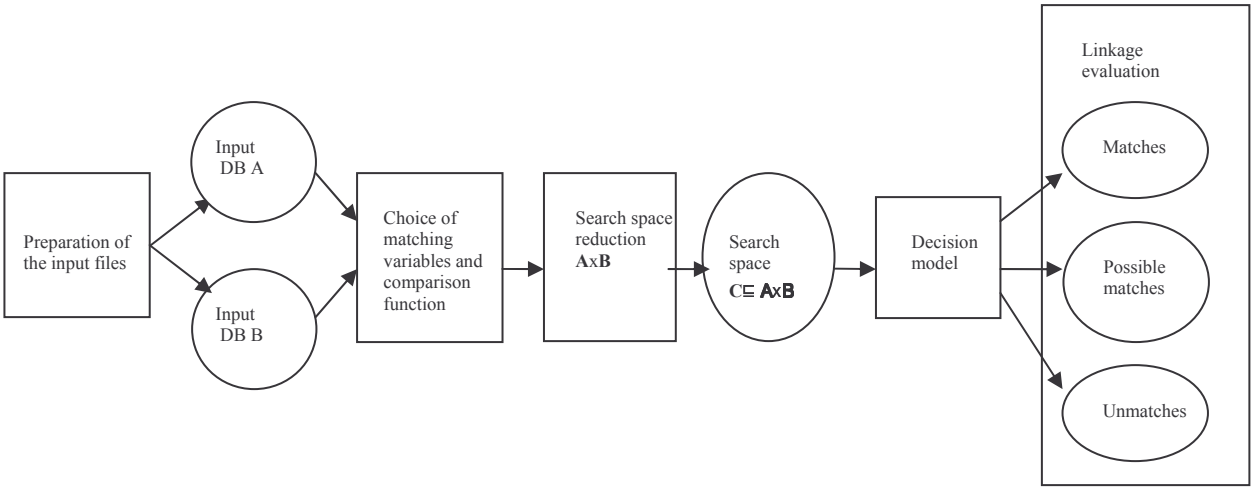
comparison stage, different blocking criteria could be implemented, or for the decision phase, the Fellegi-Sunter decision rule can be applied or in alternative it can be chosen a rule based on similarity thresholds computed on pairs of record attributes.
Due to the great attention to the integration data matters and the complexity of the 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, Tailor is an academic prototype, while others have been developed at medical-epidemiological Centres or at Universities. Some of the systems provide for the user a certain degree of flexibility, e.g. Febrl allows to choose which comparison function can be more appropriately applied. However, any of these tools provides the flexibility of multiple choices for *each* of the record linkage phase. Moreover, we want to realize the idea of dynamically building a record linkage workflow, as a result of a combination of the most appropriate technique selected at each phase. In this respect, the Tailor system is the closest one to our idea of a toolkit. However, Tailor only offers, in some of the record linkage phases, a (limited) list of methods that can be applied, without suggesting their dynamic composition based on application needs. Indeed, the purpose of Tailor is to come up with the *best* solution for record linkage, and therefore an experimental comparison was performed among techniques within each phase.
Pointing out the complexity and the modularity of the record linkage problems, its strong dependence on the handling data features and application requirements, the large number of efforts made in different fields in order to deal with the linkage issues, we propose the RELAIS (REcord Linkage At IStat) toolkit as a tool that guides in building a record linkage workflow. The inspiring principle is to allow combining the most convenient techniques for each of the record linkage phases and also to provide a library of *patterns* that could support the definition of the most appropriate workflow, in both cases taking into account the specific features of the data and the requirements of the current application. In such a way, the toolkit not only provides a set of different techniques to face each phase of the linkage problem, but also it could be seen as a compass to solve the linkage problem as better as possible given the problem constrains. In addition, RELAIS aims at joining specifically the statistical and computational essences of the matching issue. Moreover, in order to re-use the several solutions already available for record linkage in the scientific community and to gain the several experiences in different fields, we start to develop the RELAIS project as an open source project, by the quite ambitious goal of providing, in the shortest possible time, a generalized toolkit for dynamically building record linkage workflows.

The phases of a record linkage project

The complexity of the whole linking process relies on several aspects. If unique identifiers are available, in the considered data sources, then the problem is not of a high level complexity. Otherwise, the lack of unique identifiers in the datasets requires more sophisticated statistical procedures relying on “matching variables” chosen for linking data. Obviously, errors in the linking variables such as missing words, variation in spelling, use of initials, etc., may invalidate the linkage results, that’s why most of the efforts for reducing such a error needs to be made in the preparation of the input files.
The record linkage procedures is composed of some main phases (as shown in Figure 1):

1. Data cleaning - preparation of the input files (pre-processing);
2. Choice of the common identifying attributes (matching variables);
3. Comparison function;
4. Searching-Blocking;
5. Decision model;
6. Record linkage procedures evaluation.

Figure 1: Phases of record linkage



The *preparation of files* is the first phase which, according to Gill (2001), requires 75% of the whole effort for the implementation of the record linkage procedures. As a matter of fact, data can be recorded in different formats and some items may be missing or with inconsistency or errors. The key job of this phase is to convert the input data in a well defined format, resolving the inconsistencies in order to reduce errors deriving from an incorrect reported data; actually many true matches may be erroneously classified as non-matches because of a wrong comparison of the matching variables. In this phase null string are cancelled, abbreviations, punctuation marks, upper/lower cases, etc. are cleaned and any necessary transformation is carried out so as to standardize variables. Furthermore the spelling variations are replaced with standard spelling for the common words. A parsing procedure which divide a free-form field into a set of strings, could be applied and a schema reconciliation can be performed to avoid possible conflicts (i.e. description, semantic and structural conflicts) among data source schemas so as to have standardized data fields.

After the previous phase, it's important to *choose matching variables* as suitable as possible for the linking process considered. The matching attributes are generally chosen by a domain expert, hence this phase is typically not automatic but the choice can be supported by some further information that can be automatically computed. 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 the use of numeric identifiers alone. Otherwise, if unique identifiers are lacking, the choice of the common identifying attributes is more difficult and can be performed via some helpful measures, deriving from metadata description and simple statistics on the variables distribution which gives the identification power of the attributes. In this way it's possible to guide through the choice of the matching variables. Variables likes *name, surname, address, date of birth,* can be used jointly instead of using each of them alone. In such a way, one can overcome problems like the wide variations for the spelling of *name* or the changes in *surname* because of the variability of the marital status. It's evident that the more heterogeneous are the items of a variables the higher is the identification power of the variable itself; moreover, if missing cases are relevant in a field it's not useful to choose it as matching variables. For the quality measures implemented in RELAIS see the next paragraph.

The *comparison function* is used to calculate the distance between records that are compared as regards to the values of the chosen matching variables. Some of the most common comparison functions are:

a) *equality* which returns 1 if two strings fully agree on a specific characteristic, 0 otherwise;
b) *edit distance* that returns the minimum cost in terms of insertion, deletions and substitutions needed to transform a string of one record into the corresponding string of the compared record;
c) *Jaro* counts the number of common characters and the number of transpositions of characters (same character with a different position in the string) between two strings;
d) *Hamming Distance* computes the number of different digits between two numbers;
e) *Smith-Waterman* uses dynamic programming to find the minimum cost to convert one string into the corresponding string of the compared record; the parameters of this algorithm are the insertions cost, deletions cost and transposition cost;

f) *TF-IDF* is used to match strings in a document. It assigns high weights to frequent tokens in the document and low weights to tokens that are also frequent in other documents.

For a reviews of comparison functions see Koudas N. and Srivastava D. (2005).

In a linking process of two datasets, say *A* and *B*, the pairs needed to be classified as matches, nonmatches and possible matches are those in the cross product *A x B*. In case we're considering the de-duplication problem the space is *A x (A-1)/2*. When dealing with large datasets, the comparison of the matching variables is almost impracticable; as a matter of fact, while the number of possible matches increases linearly, the computational problem raises quadratically, the complexity is $\boldsymbol{O}(n^2)$(Christen and Goiser, 2005). To reduce this complexity, which is an obvious cause of problems for large databases, it is necessary to reduce the number of pairs (*a; b*), *a* belonging to *A* and *b* belonging to *B*, to be compared. Starting from this reduced search space, we can apply different decision models which define the rules used to determine whether a pair of records (*a; b*) is a match, a nonmatch or a possible match. Blocking and sorted neighbourhood are the two main methods which aim to reduce the number of comparison between records. *Blocking* sets out to remove pairs of records that are no matches; it consists of partitioning the two sets into blocks and of considering linkable only records within each block. The partition is made through blocking keys; two records belong to the same block if all the blocking keys are equal or if a hash function applied to the blocking keys of the two records gives the same result.

Sorted neighbourhood sorts the two record sets by the same variable and searches possible matching records only inside a window of a fixed dimension which slides on the two ordered record sets.

The techniques of sorting, filtering, clustering and indexing may be all used to reduce the search space.

The core of record linkage process is the *choice of decision model* which enables to classify pairs into M, the set of true matches and U, the one of the true non-matches. The decision rule can be empirical or probabilistic. A pair is a true match if it agrees completely on all the identifiers or satisfies a defined rule-base system, that is if it reaches a score which is besides a threshold when applying the comparison function. The probabilistic approach, based on the Fellegi and Sunter model, requires an estimation of the model parameters which can be performed via EM algorithm, Bayesian methods, etc.

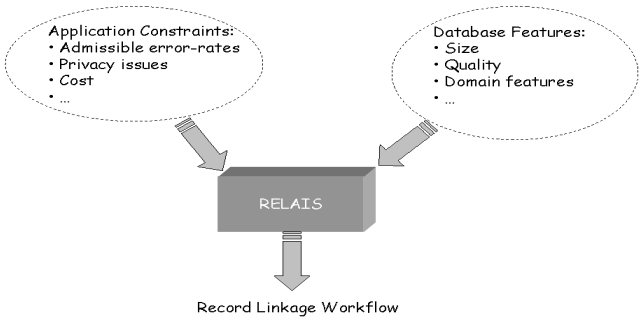
A linkage process can be also classified as one-to-one match if only one record in the A set links to only one record in B, while a many-to-one match means that a record can be matched with more than one of the compared file. Finally, the many-to-many match allows more than one records in each file to be matched with more than one record in the other. The latter two may imply the existence of duplicate records in the linkable data sources.

As not every matched records refer to the same identity, 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 nonmatches. The first type of error refers to matched records which do not represent the same identity while, the latter indicates unmatched records not correctly classified, that is truly matched entities weren't linked. Generally, false nonmatches of matching cases are the most critical ones because of the difficulty of checking and detecting them (Ding and Fienberg, 1994). The false match rate denotes the ratio between the records incorrectly matched and the whole number of matched pairs. The false nonmatch rate instead indicates the ratio between the number of incorrectly non matched records and the whole number of the correctly matched records. Generally it's not easy to find automatic procedures to estimate these two types of errors so as to evaluate the record linkage procedures quality. They can be estimated via samples of units belonging to the M and U subsets or by means of a clerical reviewed sample units or of a re-linkage procedure, assumed error-free, because performed with more accurate techniques; the bias is evaluated by means of the differences between the match and the "perfect match" results.

Description of RELAIS

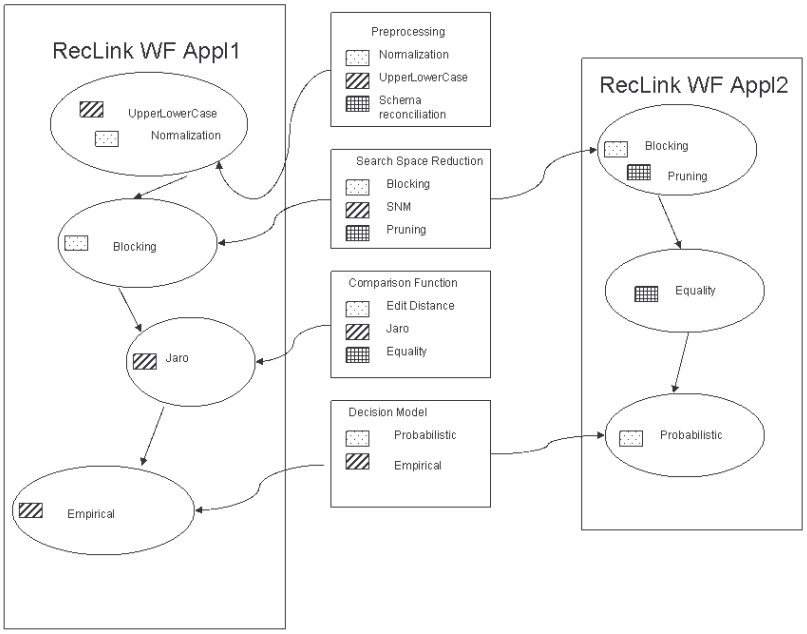
The RELAIS toolkit idea is based on the consideration that the record linkage process is application dependent. Indeed, available tools do not provide a satisfying answer to the various requirements that different applications can exhibit. As seen in the previous section, the record linkage process consists of different phases; the implementation of each phase can be performed according to a specific technique or on the basis of a specific decision model. For instance, choosing which decision model to apply is not immediate: the usage of a probabilistic decision model can be more appropriate for some applications but it can be less appropriate for others, for which an empirical decision model could prove more successful. Furthermore, even using the same decision model, in different application scenarios, a comparison function could fit better than others. Therefore, we claim that no record linkage process, deriving from the choice and combination of a specific technique for each phase, is the best for all applications.

Figure 2: The RELAIS's input-output



The RELAIS toolkit is composed by a collection of techniques for each record linkage phase that can be dynamically combined in order to build the *best record linkage workflow*, given a set of application constraints and data features provided as input (see Figure 2). As an example, if it is known that the datasets to compare have poor quality, it is suitable the usage of comparison functions ensuring high precision (e.g. Jaro distance, as defined in the previous section); as a further example, if no specific error-rates are required by the application, it can be appropriate the usage of an empirical decision model. Some phases of the record linkage process can be missing: for instance the search space reduction phase makes sense only for huge data volumes, or for applications that have time constraints. In Figure 3, examples of possible workflows that may be built with the RELAIS toolkit are shown.

Figure 3: Examples of RELAIS's workflows



In addition, to give the opportunity to the user of designing the record linkage workflow which is more appropriate for the application at hand, the RELAIS toolkit supplies a data profiling phase in which a set of quality metadata are calculated starting from real data; these metadata help the user in the critical phase of choosing the best blocking or matching variables. Moreover, in order to come towards needs of non-skilled users, the RELAIS proposes also a default set of parameters, coming from communities and manuals, to help the decision-making stages.

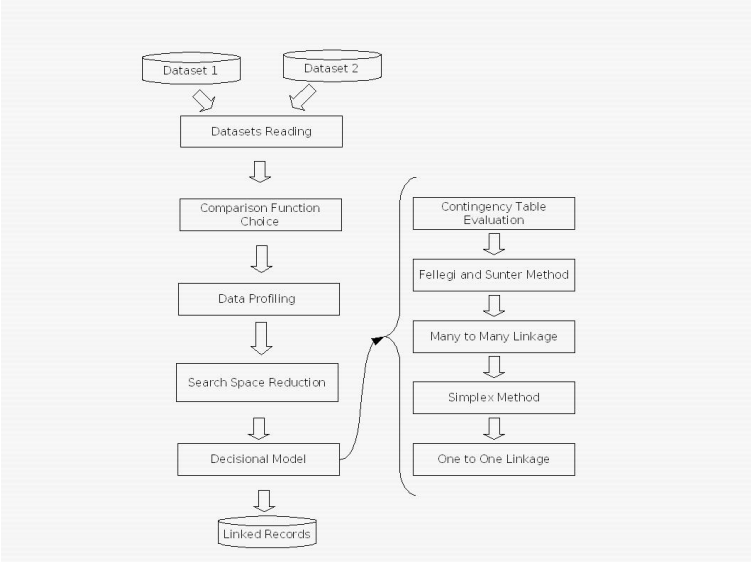
RELAIS as Open Source Project

As also remarked in the introduction, RELAIS is configured as an open source project. There are at least two reasons for this choice. First, as often highlighted above, there are many possible techniques that can be implemented for each of the record linkage phases: relying on a community of developers such set can be increased and maintained very rapidly. Second, we do believe that there have been, in the last years, several independent efforts towards the definition of a record linkage project better than the previous ones, and that such efforts have not led to the best for all solution. An open source record linkage project could instead give the possibility of gathering together the efforts already done, according to the idea described above, in order to make them available to the community for the most appropriate usage. RELAIS is mainly implemented in Java, due to the well-known features of strongly typing and platform independence; some phases are implemented in R, a free language for statistical computing (http://www.r-project.org).

Status of Implementation

RELAIS gives the opportunity to design different record linkage workflows. As shown in Figure 4, the principal phases of the record linkage process have already been implemented (with one or more techniques): (i) datasets reading, (ii) data profiling, (iii) comparison function choice, (iv) search space reduction and (v) decision model choice.

Figure 4: Status of implementation of RELAIS



In the datasets reading phase a (partial) schema reconciliation is performed: only the common variables between the two datasets are visible to the user, and only in this subset it is possible the selection of the matching and blocking variables.

In the data profiling phase, the following metadata of quality are evaluated: (i) variable completeness, (ii) identification power, (iii) accuracy, (iv) internal consistency, (v) consistency. All these metadata can be evaluated for each variable, and are merged together into a quality vector associated to the variable itself. A ranking within the quality vectors is performed in order to suggest which variable is more suitable for blocking or matching.

Some metrics are already implemented while others are still under development. We have already defined the completeness of a variable as the proportion of non-missing values for the variable on the total number of records, while the identification power is given by the ratio between the number of the different values recorded for the variables and the total number of records. The accuracy implies the comparison of the recorded values of a variable with a dictionary or a set of reference values. The measure provides the number of correct values on the overall. Finally, we have considered the consistency representing how well each item of the considered variable relates independently to the rest of the items on a scale.

More specifically, the metrics implemented for the quality metadata evaluation are listed in the following.

Given a dataset A of size N with variables $(X_1,...,X_k)$, we have:

1. **Completeness**. Let $V_i=\{v_{i1},\ldots,v_{iN}\}$ be the set of values of the variable X_i and $\underline{V_i}=\{v_{ij} \in V_i \,|\, v_{ij} \neq \text{NULL}, j \in \{1,\ldots,N\}\}$ the set of non missing values for the variable X_i, the completeness of X_i is defined as:

$$Compl(X_i) = \frac{|V_i|}{N}$$

2. **Identification power.** Let n_i be the number of different values of the variable X_i . The identification power of X_i is defined as:

$$PI(X_i) = \frac{n_i}{N}$$

3. **Accuracy.** We measure accuracy with respect to reference dictionaries of values that are known to be correct. Let $V_i^a = \{v_{ij} \in V_i \mid v_{ij}$ is labelled as "accurate"$\}$ be the set of values known to be accurate; the accuracy of X_i is defined as:

$$ACC(X_i) = \frac{|V_i^a|}{N}$$

4. **Internal consistency.** It is a specific type of accuracy computed on pairs of variables X_i and X_k, with respect to a list of paired values known to be correct. Let $V_{ik}^a = \{< v_{ij}, v_{kj} >$, with $v_{ij} \in V_i$, $v_{kj} \in V_k \mid < v_{ij}, v_{kj} >, j \in \{1,\ldots,N\}$, is labelled as "accurate"$\}$ be the set of values known to be accurate; the internal consistency of X_i and X_k is defined as:

$$Cons(X_i, X_k) = \frac{|V_{ik}^a|}{N}$$

5. **Consistency.** It takes into account the number of internal consistency relationships in which a variable X_i is involved. Being X_{ij}^c the set of variables X_j such that each X_j is involved in a consistency check with X_i, consistency is defined as:

$$Cons(X_i) = \frac{\sum_j Cons(X_i, X_{ij}^c)}{|X_{ij}^c|}$$

We refer to this set of quality metadata as quality vector q. As far as the procedure adopted to rank the quality vectors obtained by metadata evaluation is concerned, we have two distinct solutions:

1. **One-step ranking.** We take into account the fact that some elements of the quality vector associated with the X_i variable may have missing values, i.e. the evaluation of some specific metadata is not possible or not required. We use "dummy values" in place of such missing values that can be set by the user; these dummy values can be the mean, the maximum or the minimum of the metadata values computed for the other attributes different from X_i. Vectors are then compared each other by means of a Euclidean weighted norm, i.e.:

$$\| q \| = \left(\sum_{i=1}^{n} w_i q_i^2\right)^{1/2}$$

where $q=(q_1,\ldots q_n)$ is a vector of quality metadata and $w=(w_1,\ldots w_n)$ is a vector of weights. The quality vectors are then ordered on the basis of such a norm.

2. **Two-step ranking.** It is performed in two steps: (i) evaluation of the norm on the present metadata, i.e. ignoring missing ones, and sorting on the basis of such a norm; (ii) refinement of the ranking performed at step (i) by using an *insertion sort,* i.e. by comparing only the values of common metadata for each couple of vectors.

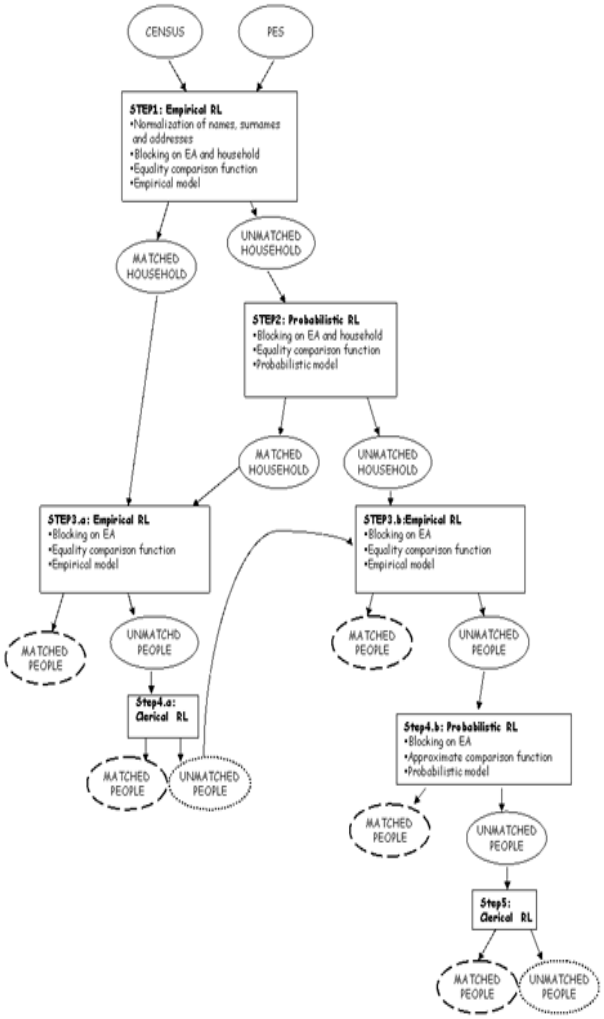
With respect to the decision model choice, we have implemented the Fellegi-Sunter probabilistic model by using the EM algorithm for the estimation of the model parameters; as detailed in Figure 4, this method takes as input a contingency table, which reports the frequencies of the agreement patterns resulting from the application of the comparison function, and the output is a many to many linkage of the datasets records. Starting from this output, we can propose to the user the clusters of matches, non-matches and possible matches. Alternatively, a subsequent phase of reduction from many to many linkage to one to one linkage can be performed by applying now the simplex method.

The Fellegi-Sunter model and the simplex method have been implemented with the R language that has a huge number of statistical packages available, thus giving us the opportunity of re-use software already developed from the scientific community, according to the open source idea of our project.

Case study: Building a Record Linkage Workflow

In this section a record linkage application concerning the Post Enumeration Survey (called *PES* in the following) of the Italian 2001 Census is described. 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 the other components of the same household was also 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.

Figure 5: The record linkage workflow of the case study



In the Figure 5, the 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 an empirical 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, Winkler 2000). In step 3.a an empirical 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 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 an empirical approach. Then, in step 4.b, for the 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 empirical 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.

Figure 6: An example of a pattern for building record linkage workflows

REQUIREMENT		CHOICE
Data requirement	Hierarchical structure	Workflow iteration: •Higher level (household) •Lower level (person)
	High quality	Equality comparison function on most of the phases
	Huge data set	Blocking Phase iteration Empirical model
Application requirement	No errors in matching process	Probabilistic model Clerical review phase

In Figure 6, a table representing the case study requirements and the corresponding choices suggested is shown. Such correspondences can be considered as a pattern useful for building record linkage workflows whereas similar application and data requirements are present.

Concluding remarks

In official statistics, data integration is of major interest as a mean of using available information more efficiently and improving the quality of statistical products, in particular, statistical indicators designed to enable sound decision and policy-making. Recently, many tools for record linkage have appeared on the market and research groups have made available to the public software packages for linkage. In this paper, we have showed the RELAIS project that aims at implementing an open source toolkit for building record linkage workflows. The idea of this project has been developed keeping in mind the complexity of a record linkage problem, that involves different techniques and sciences; the opportunity of treating the linkage with modularity, identifying several phases which can occur, even iteratively; the different suitable approaches depending on both the data features (type of data, amount of data, ...) and the application requirements (efficiency, efficacy, accuracy, …). As the practitioner has to deal with such number of situations, the toolkit wants to offer multiple techniques for record linkage, both deterministic and probabilistic, and also the possibility of building ad-hoc solution combining each modules. This approach allows to overcome the question on which method is better than others, being convinced that actually there is not a technique dominating all other across all data sets, but the suitability of each approach is dependent on the data and the given task.

In the paper, we have described a case study as a proof of concept of the inherent complexity of record linkage processes, on which the RELAIS project is based. Indeed, due to such complexity, great modularity and flexibility are necessary in order to properly build application specific record linkage workflows.

Furthermore, besides the enrichment of the set of available techniques, future work for RELAIS will include several patterns that can guide the design of record linkage workflows. Indeed, we believe that the design stage of a record linkage workflow could usefully exploit patterns extracted from previous knowledge and experiences. In this way, the toolkit can assist non-expert users in designing their specific record linkage workflows. Each technique could be characterized in terms of pre-conditions that must be respected in order to be part of a record linkage workflow. We will study the possibility of using formal languages for the specification of such preconditions, in order to check properties like consistency and completeness of a proposed workflow solution with respect to given application and data requirements.

References

Ananthakrishna R., Chaudhuri C., and Ganti V. Eliminating Fuzzy Duplicates in Data Warehouses. In *Proceedings of VLDB 2002*, Hong Kong, China, 2002.

Bertolazzi P., Santis L.D., and Scannapieco M. Automatic Record Matching in Cooperative Information Systems. In *Proceedings of the ICDT'03 International Workshop on Data Quality in Cooperative Information Systems (DQCIS'03)*, Siena, Italy, 2003.

Chauduri S., Ganti V., and Motwani R. Robust identification of fuzzy duplicates. In *Proceedings of ICDE 2005*, Tokyo, Japan, 2005.

Christen P. and Goiser K. Assessing duplication and data linkage quality: what to measure?, Proceedings of the fourth Australasian Data Mining Conference, Sydney, December 2005

Ding Y. and Fienberg S.E. Dual system estimation of Census undercount in the presence of matching error, *Survey Methodology*, 20, 149-158, 1994.

Elfeky M., Verykios V., and 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*, 2001. Washington D.C.

Febrl. http://www.sourceforge.net/projects/febrl.

Fellegi I. and Sunter A. A Theory for Record Linkage. *Journal of the American Statistical Association*, 64, 1969.

Fortini M., Liseo B., Nuccitelli A., and Scanu M. On Bayesian record linkage. *Research in Official Statistics*, 4:185-198, 2001.

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., and Rainsford C., Record linkage: Current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, Australia, April 2003

Gu L. and Baxter R. Adaptive filtering for efficient record linkage. In *Proceedings of the Fourth SIAM International Conference on Data Mining*, 2004.

Hernandez M. and Stolfo S. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Journal of Data Mining and Knowledge Discovery*, 1(2), 1998.

Jaro M. Advances in Record Linkage Methodologies as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of American Statistical Society*, 84(406):414-420, 1985.

Koudas N.and Srivastava D. Approximate joins: Concepts and techniques. In *Proceedings of VLDB 2005*, 2005.

Monge A. and Elkan C. An Efficient Domain Independent Algorithm for Detecting Approximate Duplicate Database Records. In *Proceedings of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery(DMKD'97)*, Tucson, AZ, USA, 1997.

Newcombe H, Kennedy J, Axford S and James A. Automatic Linkage of Vital Records, Science, Vol.130 pp. 954-959, 1959.

The-Link-King. http://www.the-link-king.com.

The-R-Project for Statistical Computing. http://www.r-project.org/.

Winkler W. Frequency-based matching in Fellegi-Sunter model of record linkage. Technical report, U.S. Bureau of the Census - Washington D.C., 2000. Technical Report RR/2000/06, Statistical Research Report Series.

Winkler W. Record Linkage Software and Methods for Merging Administrative Lists. Technical report, U.S. Bureau of the Census - Washington D.C., 2001. Technical Report RR/2001/03, Statistical Research Report Series.

Winkler W. Methods for Evaluating and Creating Data Quality. *Information Systems*, 29(7), 2004.

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.