Estadística Española • Volumen 54, número 179 / 2012, pp. 311-328

Record Linkage with RELAIS: Experiences and Challenges

Nicoletta Cibella
Italian National Statistical Office, Italy

Monica Scannapieco
Italian National Statistical Office, Italy

Laura Tosco
Italian National Statistical Office, Italy

Tiziana Tuoto
Italian National Statistical Office, Italy

Luca Valentino
Italian National Statistical Office, Italy

Abstract

The combined use of data coming from different sources 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. In this paper we describe 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*. The open source turned out to be a winning choice for sharing techniques and software and it strongly contributed to assert RELAIS in the National Statistical Institutes' community. In the paper we show the usefulness and the profitability of RELAIS in facing several challenges in linking data at micro-level, achieving a high quality of the linkage process and of the related results.

Keywords: Record linkage, Open-source software, Data Quality

AMS classification: 62-02, 68T10

El enlace de registros con RELAIS: experiencias y retos

Resumen

El uso combinado de datos provenientes de distintas fuentes se ha generalizado en gran medida para maximizar su respectiva utilidad: lamentablemente sucede que a menudo las distintas fuentes de información son difíciles de integrar debido a errores o a la falta de información. Las técnicas de enlace de registros son un conjunto multidisciplinario de métodos y prácticas que pretenden identificar la misma entidad del mundo real, diferentemente representada en las distintas fuentes de datos. El enlace de registros es un proceso complejo que puede descomponerse en fases separadas, en la que cada una de ellas requiere de una técnica específica. En este artículo se describe RELAIS (Record Linkage At IStat), un conjunto de herramientas de código abierto basado en la idea de seleccionar la técnica más adecuada para cada fase y combinarlas dinámicamente para construir el flujo de trabajo del enlace de registros. El código abierto resultó ser una opción acertada para compatir técnicas y software y contribuye fuertemente a hacer valer RELAIS entre los Institutos Nacionales de Estadística. En el artículo se muestra la utilidad y los beneficios de RELAIS a la hora de enfrentar determinados desafíos en el enlace de microdatos, alcanzando un alto nivel de calidad en el proceso de enlace y en los resultados asociados.

Palabras clave: Enlace de registros, programas de código abierto, calidad de la información

Clasificación AMS: 62-02, 68T10

1. Introduction

Nowadays data integration procedures are becoming extremely important in official statistical institutes. In particular, record linkage procedures, aiming at matching records referring to the same entities, both within a dataset and from two or more different data sources, improve the quality of information collected and make possible more detailed analyses. A large number of linkage techniques are available and commonly used; in the field of the official statistics. Moreover, the emerging Big Data Paradigm poses new challenges related to the quality of information that can be extracted from the Deep Web. Record linkage procedures can significantly help to improve the "noisy" feature of Big Data by comparing different data spread redundantly among several sources.

The record linkage has the purpose to identify, quickly and accurately, the same real world entity, which can be differently represented in one or more data sources. A record linkage project can be performed for different purposes and this richness of possible applications makes it a powerful instrument to support decisions in large commercial organizations and government institutions. In the official statistics context, 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 and to accurately identify units across different data sources.

Actually, 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 large amounts of data, often very accurate, due to improvements made over time. For this reason in most situations the joint analysis of statistical and administrative sources allows to save time and money, reduces survey costs and 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.

However, data sources are often hard to combine since errors or lacking information in the record identifiers may complicate the joint use of information. In order to overcome such difficulties, record linkage techniques provide a multidisciplinary set of methods and practices.

Record linkage procedures substantially improve the quality and the quantity of the available information by integrating different frames. Actually, identifying pairs of records coming from either the same or different data files, can help in evaluating the accuracy of the information coming from a given source, when some of the same variables are collected across different data files and can overcome lack of information; moreover the linkage enables in carrying out more detailed analysis. It needs also to be remarked how important is to evaluate the quality of the linkage output, so reducing as much as possible the matching errors, particularly when further analysis are based on previously linked data.

Starting from these analyses we designed and implemented an open source toolkit for record linkage called RE.L.A.IS. (REcord Linkage At IStat). The fist version, RELAIS 1.0, was released in February 2008 on the Istat website with the implementation of a probabilistic model based on Fellegi-Sunter theory, the EM algorithm for the parameters estimation and file architectural structure. Due to profitable collaborations we had with other NSIs, that will be described in section 2.1, the RELAIS software was improved and the version 2.2 was dramatically enriched compared to the first release.

In this paper we will also describe the features and the functionalities of the RELAIS 2.3 version, given that its release is quite immediate.

Being an open source project, both source code and executables of RELAIS have been released on Istat site (http://www.istat.it/en/tools/methods-and-software) and on the JOINUP web site (http://joinup.ec.europa.eu/software/relais/description). The system has been released under the European Union Public License (EUPL), a free software license created and approved by the European Commission.

The paper is organized as follows: in Section 2 we focus on record linkage in the official statistic and we describe the idea, the design and details of the RELAIS toolkit. In Section 3 we describe some workflows of record linkage implemented with RELAIS.

In Section 4 we describe our record linkage experiences shared with other NSIs. Finally, in Section 5 some concluding remarks and direction for future works are provided.

2. The state of the art of Record Linkage in Official Statistic and the RELAIS Toolkit

Record linkage techniques are a multidisciplinary set of methods and practices with the main purpose of accurately recognize the same real world entity at individual micro level, even when differently stored in sources of various type. A complete overview of most used techniques can be found in the Report of WP1 - State of the art on statistical methodologies for integration of surveys and administrative data of the ISAD project (2008). The overall record linkage workflow could change from user to user, due to different restrictions, such as legal and practical issues, in various fields and countries.

There are different purposes to perform a record linkage project and it has recently revealed a powerful support to decisions in large commercial organizations and government institutions. In official statistics data integration procedures are becoming extremely important due to many reasons: the cut of the costs, reduction of response burden and use of information derived from administrative data are some of the most crucial ones; this is a strong incentive to the investigation of new methodologies and instruments to deal with record linkage projects.

Several applications need record linkage techniques, including: enriching and updating the information stored in sources; the elimination of duplicates within a data frame; the creation of a sampling list; improving the data quality of a source; estimating number of units in a population amount by capture-recapture method; assessing the disclosure risk when releasing micro data files; the study of the relationship among variables reported in different sources. The combined use of statistical survey and administrative data is largely widespread in official statistics and the potential advantages in using administrative data for statistical purposes known and shared by the various national statistical institutes strongly stimulate the investigation of new methodologies and instruments to deal with record linkage projects and to accurately identify units across different data sources. The nature common to different NSIs of the record linkage problems strongly emerged during the project "ESSnet Statistical Methodology - Area ISAD" (Integration of Survey and Administrative Data, http://www.essnet-portal.eu/finished-projects/isad-finished) which last for 18 months and started in December 2006. The Essnet was composed by the NSIs of Italy (the coordinator), Austria, the Czech Republic, the Netherlands and Spain, and aimed at promoting knowledge and application of methodologies for the joint use of existing data sources (both administrative and statistical) in the production of official statistics. The NSIs in the ESS face growing duties of data supply and dissemination, while budgetary constraints and increasing public concern about response burden and data privacy on the respondent units - either individuals, households or enterprises - arise. The integration of different data sources, especially those of administrative nature, seems to be a cheap and reliable alternative. In 2009 started the two years ESSnet DI (Data Integration,

http://www.essnet-portal.eu/di/data-integration), coordinated by ISTAT. The focus was on methodologies for data integration (Record Linkage, Statistical Matching, Micro integration Processing) and on statistical aspects to be considered to make those methods concretely applicable by NSIs.

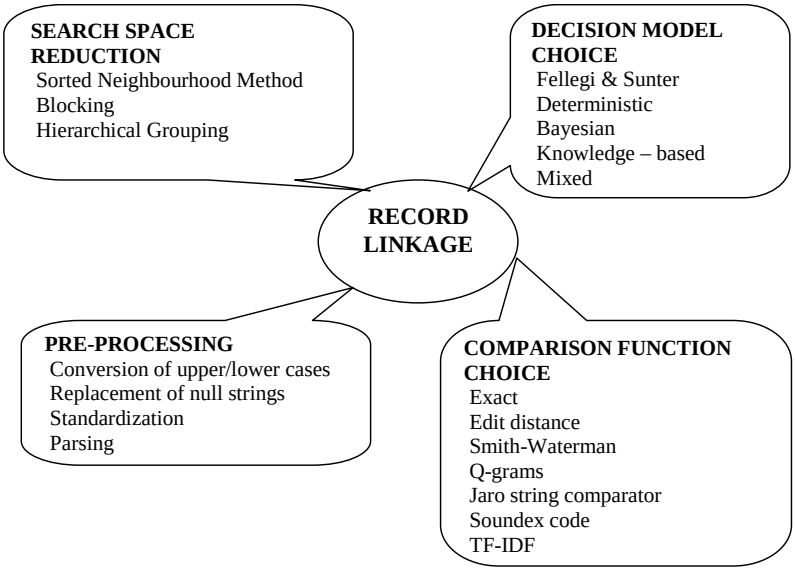
Actually, even in a statistical system with shared goals and regulations, as the European Statistical System, different constraints, for instance based on language features, may be present and affect the outcome of the same linkage. According to the actual EUROSTAT new vision which incites to the standardization and engineering of the statistical process, the sharing of solutions suitable for solving specific problems among different NSIs can lead to considerable benefits in terms of cost reduction and quality improvement. The reuse of this solutions or software by another institute obviously implies saving money, generally ensuring, when they are thought to be shareable, also very high quality results. The technical choices play however an important role: the reuse is not always possible whereas technical architectures are different among NSIs.

2.1 History and Philosophy of RELAIS

In the Italian National Institute of statistics (ISTAT) there is a wide use of record linkage techniques in different production processes; the first experiences date back to '80s but the common practice was to develop ad-hoc linkage procedures for each project, basically via deterministic approaches. Unfortunately, there was little awareness of linkage errors in further analyses of linked data and only a few official experiences with probabilistic method. However the decennial studies on the Fellegi-Sunter theory and the belief that there is no a unique solution to record linkage problem, led an Italian team of IT experts and statistical methodologists to design and implement the RELAIS system. It allows the realization of a record linkage process with a modular approach. We started the RELAIS project (Fortini et al. 2006) by the observation that since the earliest 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 also of techniques based on data mining, machine learning, equational theory but no one of these technique has emerged as the best solution for all cases. Moreover, in some applications, there is no evidence to prefer one method to others or of the fact that different choices, at a linkage stage, could bring to the same results: i.e. alternative reduction methods will lead to different pair search space and consequently different results. Finally, the choice of 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 a deterministic decision model could prove more successful. The overall complexity of the linkage process and the several suitable solution for each constituting sub-problems is represented in Figure 1.

Figure 1

The record linkage complexity



The increasing attention devoted to record linkage methodologies produced a large number of software and tools to face with linkage problem; to restrict the view only to official statistics, let consider BigMatch (Yancey, 2007), GRLS (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); but, in our opinion, any of these tools provides the flexibility of multiple choices for each of the record linkage phase.

Starting from this idea, we designed RELAIS to allow the dynamic selection of the most appropriate technique for each of the record linkage phases and the combination of the selected techniques so that the resulting workflow is built on the basis of application and data specific requirements. We believe that the choice of the most appropriate technique not only depends on the practitioner's skill but, most of all, it is application specific. 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 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 developed the RELAIS toolkit as an open source project in order to provide, in the shortest possible time, a generalized toolkit for dynamically building record linkage workflows. Moreover, a *shareable* tool make possible to:

- Share easily different solutions and experiences developed in different contest;

- Expert users can easily modify the toolkit to adapt it to their specific needs;
- Different record linkage solutions are available in a unique software;
- Easy growth and update of the software with the contribution of the scientific community.

As far as design choices, RELAIS has been developed in a modular way, with clear and defined interfaces between the various modules giving the possibility of adding new techniques to the system, and thus reusing solutions that are already available. We remark that this architectural feature, combined with the open source choice, enables skilled developers adding new techniques and possibly customizing existing ones by editing a very small number of modules.

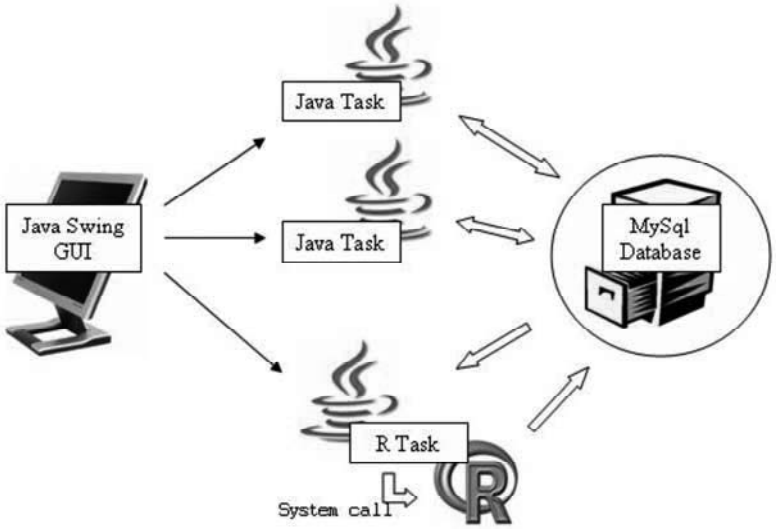
Moreover, most of the record linkage phases have been designed in order to be plugged in different record linkage workflows: a RELAIS user can almost freely combine modules realizing the different phases in order to get the desired record linkage process.

RELAIS implementation has been carried on by making use of open source technologies, namely Java and R as programming languages and MySQL as data base management system. This choice depends on our belief that a record linkage process is composed of techniques for manipulating data, for which Java is more appropriate, and of calculation-oriented techniques for which R is a preferable choice. Moreover, MySQL optimizes the performances with respect to the management of huge amount of data through the whole record linkage project (input, intermediate phase and output). Finally, the usage of an object–oriented language (Java) has permitted to design RELAIS by stressing software quality dimensions, like decoupling and information hiding, that have a significantly impact on reuse.

Being RELAIS a record linkage toolkit designed not only for expert users, we developed a Graphical User Interface (GUI) that, on the one hand, permits to build record linkage workflows with a good flexibility; on the other hand, it checks the execution order among the different provided techniques whereas precedence rules must be controlled.

Figure 2

RELAIS software architecture



Current version of RELAIS can run on several platforms (namely, Windows, Linux and Mac), because its modules are written in cross-platform languages (i.e., R and Java). With respect to application frameworks available on the IT market, RELAIS has some relationships with ETL (Extract, Transform, Load) tools. In particular, some pre-processing functionalities of RELAIS could also be implemented via an ETL tool. On the other hand, the whole RELAIS toolkit could be embedded in an ETL tool; this is made possible by the batch invocation mode of RELAIS. We are currently investigating this latter possibility, in order to have all the necessary functionalities in an integrated environment.

2.2 RELAIS in Details

As stated in the previous paragraph, RELAIS want to be a flexible toolkit to build dynamic workflow to face record linkage problems. In order to deal efficacy with such a complex process like the linkage one, the most important step is to decompose the whole procedure in its basic and more simple constituting phases. The main phases in which we chose to decompose record linkage problem are:

1) pre-processing/preparation of the input files;
2) creation-reduction of the search space of link candidate pairs;
3) choice of the matching variables (common identifying attributes);
4) choice of comparison functions;
5) choice of decision model;

6) identification of unique links;
7) linkage quality evaluation.

The RELAIS toolkit makes available, for each of these phases a set of different techniques.

The preliminary phase is the dataset acquisition that permits to read the input datasets from a textual format and from database (MySQL or Oracle) tables. In case of deduplication process only one input dataset is required. The datasets must have common variables that are the ones considered by the system in the subsequent phases; the schema reconciliation is automatically realized when the variables already have the same name, otherwise it is possible to realize a guided schema reconciliation available between the preprocessing actions. In addition to start a new project, it is also possible to continue to work to a previous one previously interrupted.

RELAIS offers some features for data preprocessing divided into two main types: (i) check features (namely check format) and (ii) conversion features (namely field standardization, fields merge, fields parse and inaccuracy repair).

After the acquisition phase, it is possible to perform a data profiling phase. Specifically, the data profiling phase permits to characterize available variables with respect to some quality features that are used to support two critical tasks: blocking variables choice and matching variables selection. The metadata supplied by RELAIS to support the user in the choice of the blocking and matching variables are quality metadata, calculated starting from real data provided as input, specifically: Completeness, Accuracy, Consistency, Entropy, Correlation and Frequency Distributions.

In a linkage of two datasets, say A and B, all pairs in the cross product A x B needed to be classified as matches, non-matches and possible matches. When dealing with large datasets, comparing all the pairs in the cross product of the two datasets is almost impracticable, in fact while the number of expected matches increases linearly, the computational problem raises quadratically (Christen and Goiser, 2005). To reduce this complexity it is necessary making use of many different techniques that can be applied to reduce the search space. RELAIS makes available three different search space reduction technologies namely: (i) blocking, (ii) sorted neighborhood and (iii) the combination of these two methods called nested blocking.

Starting from the pairs in the cross product or in the reduced search space, different decision models can be applied in order to classify them into the set of matches, non-matches or possible matches. In RELAIS both approaches deterministic and probabilistic are available.

Actually, 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. Other authors backed up that in deterministic context a pair can be linked also if some specific and pre-defined criteria are satisfied. RELAIS makes available both methods: the first case corresponds to the Exact Match option; the second case, being not exact in the strict

sense is assumed as almost-exact and corresponds to the Rule-based Match option; 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 (1969) decision model, assuming latent dichotomous variable for the linkage status and conditional independence model for the manifest variables. The EM algorithm is used to estimate the parameters. Moreover, it is possible to apply the Fellegi and Sunter approach without estimation of parameters using an input file with marginal probabilities for each matching variable.

In the decision model phase, both the selection of the matching variables and the choice of the corresponding comparison functions are very important. Thus, RELAIS makes available a set of comparison functions to compare strings according to an exact or an approximate procedure. The comparison functions provided are: Equality, Numeric Comparison, 3Grams, Dice, Jaro, Jaro-Winkler, Levenshtein, Soundex (http://sourceforge.net/projects/simmetrics/) and WindowEquality where two numbers are considered equal if their distance is minor or equal of a window dimension defined by the user.

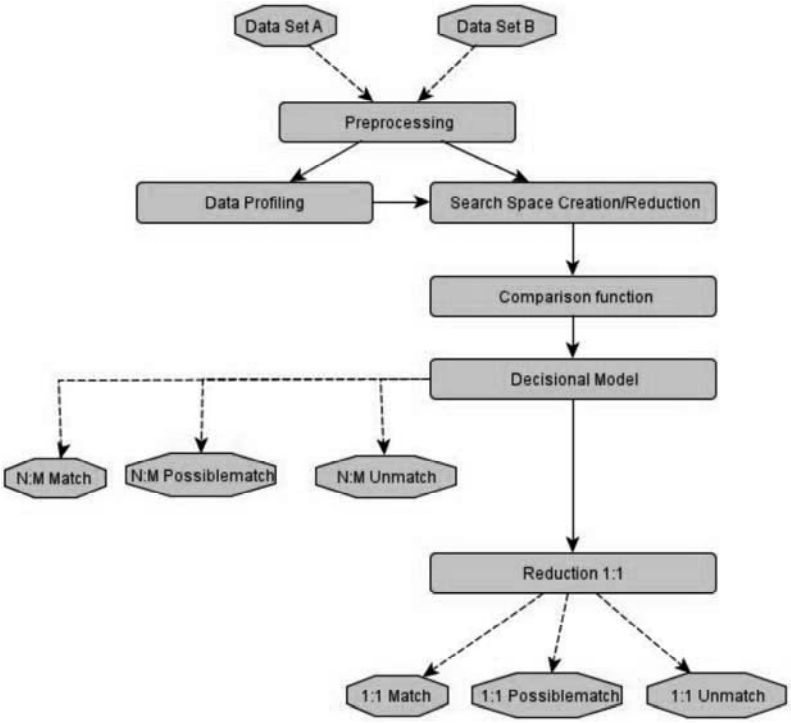
When blocking method is performed to reduce the search space of pairs, RELAIS allows the users to choose between two different ways of applying the probabilistic model: it can be applied in a one-shot way to all the blocks or to a specific block selected by the user.

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

Finally, the output of the linkage process consists of several disjoint datasets: match, non-match pairs, possible match and residuals of the starting files. For the Possible matches no decision is taken and they need to be processed by clerical review or by further linkage process. Residuals can be submitted to further analyses starting a new linkage process.

Figure 3

Principal RELAIS functionalities



To help the analysis of the linkage process results intermediate outputs can be saved, such as blocking summary, contingency tables and parameter estimate tables. Moreover, if the probabilistic approach has been applied, the user has important information about quality of the record linkage results in terms of precision and recall. This two parameters are very useful when critical quality restriction are imposed to the results.

3. Record Linkage Workflows with RELAIS

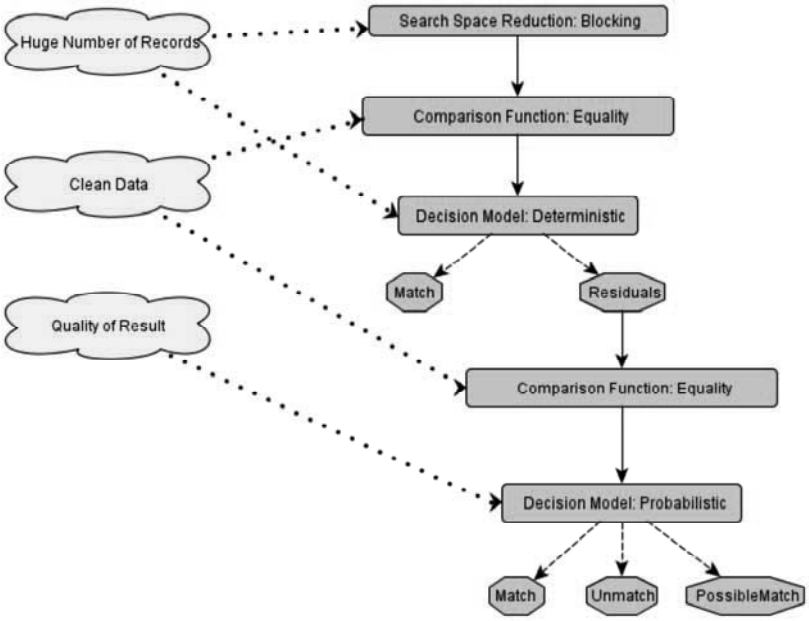
In this Section we describe some workflows of record linkage implemented with RELAIS. In particular, by using RELAIS in some concrete data integration projects, we have abstracted some workflow templates that, given some requirements, led us to choose a specific record linkage step (or sequence of steps) to become part of a record linkage process. In Figure 4.a, 4.b, 4.c, we show such abstracted templates.

Template 1 (Figure 4.a) shows how the whole set of requirements ("Huge Number of Records", "Clean Data" and "Quality of the Result") can result in the described record linkage workflow.

The Post Enumeration Survey (PES) projects are an instance of such a template. The PES projects use record linkage techniques for the evaluation of the census coverage by means of capture-record models. The quality of the Population and Agricultural Census is verified adopting the Petersen model; linkage between census and post enumeration surveys data is performed in order to estimate the coverage rate of the Census. The PESs are carried out on a sample of enumeration areas, which are the smallest territorial level considered by the Census. The PES was based on the replication of the Census process inside the sampled enumeration areas 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 for Population Census and farms for Agricultural Census, respectively), 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 units enumerated at the Census and the hidden amount of the population, was obtained.

Figure 4.a

Template 1. Example: PES



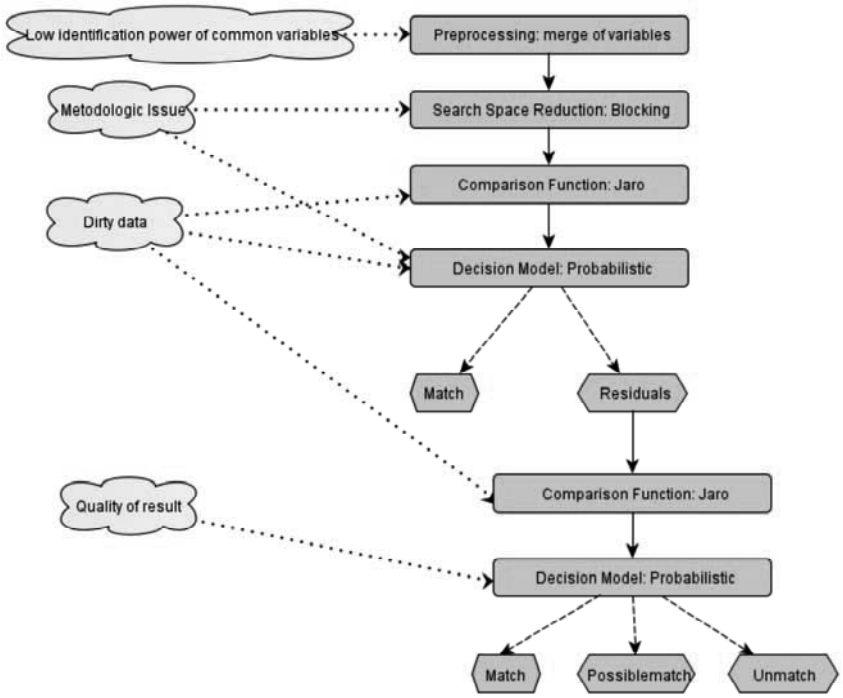
The estimates of the Census coverage rate through capture-recapture model require to match Census and PES records, assuming no errors in matching operations. This is a strong assumption: the accuracy of the matching processes is of crucial importance because even very small matching errors could have compromised the reliability of the coverage rate estimates. The PES projects have the requirements shown in Figure 4.a,

namely: (i) huge number of records; (ii) high quality of data and (iii) need to evaluate the quality of the result. 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 the quality of the result suggests the adoption of a probabilistic model, in order to have a quantitative estimation of the errors that can be regarded as acceptable or not.

Template 2 in Figure 4.b is instantiated by RELAIS record linkage procedures we applied in socio-demographic studies. For instance, a currently running project aims to build an integrated system on road accident. Here, some pre-processing steps were necessary due to the low identification power of the available variables. The blocking step was performed, this time, in order to ensure the convergence of the probabilistic method (and not due to huge volumes of data to reduce). The dirty data requirement forced to have an approximate distance as a comparison function.

Figure 4.b

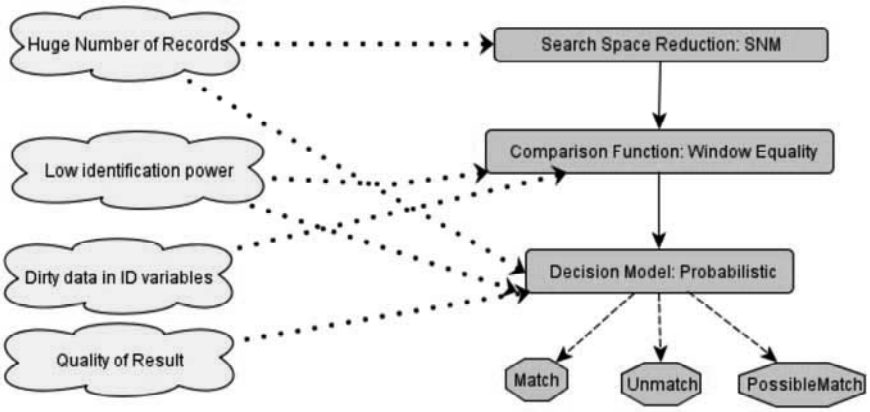
Template 2. Example: road accidents



Template 3, in Figure 4.c, is also an instantiation of a record linkage of socio-demographic data, but concerning pregnancy outcomes. Huge number of records led again to the choice of a search space reduction, but performed by Sorted Neighbourhood Method, due to data characteristics (that would have led to blocks of too much small sizes for a probabilistic approach). This choice interestingly show how the "whole" set of requirements leads to a specific workflow, that is the arrows shown in Figure 4.c between requirements and record linkage steps has not to be interpreted locally.

Figure 4.c

Template 3. Example: pregnancy outcomes



4. RELAIS: Shared Experiences Among National Statistical Institutes

The open source philosophy adopted in RELAIS appeared as a winning choice in order to favor spontaneous collaboration among different organizations. In official statistics, the advantages, in terms of quality and costs, due to the combined use of administrative data and sample surveys strongly encourage the researchers in different National Statistical Institutes (NSIs) to the investigation of new methodologies and instruments to deal with record linkage projects.

In the project "ESSnet Statistical Methodology - Area ISAD", coordinated by Italy and involving different NSIs we started a fruitful collaboration in exchanging experiences on record linkage techniques and in testing and enriching the RELAIS toolkit especially with the Spanish National Statistical Institute (INE). The collaboration between Italian National Institute for Statistics (ISTAT) and INE helped in making the synergies organized in a well-planned framework. The Spanish researcher tested the RELAIS tools with the following aims:

- assessing the capabilities of the various functionalities included in the RELAIS toolkit, e.g. the use of the EM algorithm for record linkage purposes;

- comparing the results achieved by the software with those obtained throughout some alternative ad hoc techniques;
- testing in terms of performances the blocking methods implemented in RELAIS so as to reduce the search space, in a context of registers with high amount of data to be compared.

The collaboration with the statistical office of Spain, and other collaborations with the statistical offices of United Kingdom, Brazil and Tunisia to test the RELAIS software highlighted some strengths and weaknesses of the first release. The initial planned RELAIS project has been enriched by these cooperation with many researchers and users in international context (Cibella et al, 2008) and the profitable share of knowledge and solutions among researchers coming from different institutes and countries in dealing with 'real-world' tasks point out some remarks:

- the awareness of the common nature of data integration problems faced;
- the common needs of an high quality outputs;
- the advantages in designing standardized answers to specific, though widespread, applications;
- the winning choice of the open-source solution for sharing techniques and software.

After this profitable experience, Istat in 2009 was involved, as coordinator, in two years ESSnet DI (Data Integration) which focuses on methodologies for data integration (Record Linkage, Statistical Matching, Micro integration Processing) and on statistical aspects to be considered to make those methods concretely applicable by NSIs. Thanks to the experience in exchanging knowledge on record linkage and on testing RELAIS in a "real-world" tasks outside Italy, ISTAT, conducted on January 2011 in U.K. a very appreciated on-the-job training on record linkage methods (Cibella et al, 2011). During the ESSnet DI (Data Integration) some new functionalities for the software were analyzed. In particular the implementation of some standardized functionalities for the pre-processing phase are discussed and shared among the different NSIs involved. The most common ones, as character conversions, schema reconciliation, standardization are still ready for the very next 2.3 version.

The whole experience was very profitable both for trainers and trainees thus on next July a new on-the-job training will be conducted in Latvia with the same characteristics and modality.

5. Conclusions

In this paper, we have provided an overview of the record linkage problem in official statistic institutes. We have described the Istat experiences and collaborations that helped us to improve our proposal for solving the record linkage problem, namely the RELAIS toolkit. We described RELAIS' history, philosophy and its details. Specifically, we showed how RELAIS 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 strategy, given a set of application constraints and data features provided as input. We also described how the toolkit offers multiple techniques for record linkage, both deterministic and probabilistic, with the possibility of building *ad-hoc* solution combining each module.

The RELAIS project, born from the collaboration of Istat IT expertise and statisticians researchers, have been enriched also thanks to the cooperation with many researchers and users in international context (Cibella et al, 2009). Several 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 the next future work, we have planned to extend the current functionalities of RELAIS. Specifically, we will provide:

- the implementation of a new method for optimal 1:1 reduction based on the Hungarian algorithm;
- the improvement of GUI functionalities for output management and user interactions (manual review);
- a graphical guide in the thresholds choice phase;
- new techniques for search space reduction (Mancini et al, 2012);
- new probabilistic decision model (Zardetto et al, 2010);
- some clustering techniques in the deduplication process;
- some methods for choosing cluster representative records;
- all other requirements motivated by real applications.

References

AMATO R. , BRUZZONE S., DEL MONTE V., FAGIOLO L. (2006). «Le statistiche sociali dell'ISTAT e il fenomeno degli incidenti stradali: un'esperienza di record linkage», *Istat Contributi* n. 4.

CHRISTEN, P. AND GOISER, K. (2005), «Assessing deduplication and data linkage quality: What to measure?». *Proceedings of the fourth Australasian Data Mining Conference* (AusDM 2005), Sydney.

CIBELLA, N., FORTINI, M., SCANNAPIECO, M., TOSCO, L., TUOTO, T., (2008), «Theory and practice of developing a record linkage software», *Proceedings of the Combination of surveys and administrative data Workshop of the CENEX Statistical Methodology Project Area* "Integration of survey and administrative data", Vienna, Austria.

CIBELLA, N., FERNANDEZ, G.L., FORTINI, M., GUIGÒ, M., HERNANDEZ, F., SCANNAPIECO, M., TOSCO, L., TUOTO, T. (2009), «Sharing Solutions for Record Linkage: the RELAIS

Software and the Italian and Spanish Experiences» *Proceedings of the New Techniques and Technologies for Statistics (NTTS) Conference, Bruxelles, Belgium*.

CIBELLA, N., SCANNAPIECO, M., TUOTO T. (2011) «Open source software: a way to enrich local solutions», *supporting paper at the International Meeting on the Management of statistical information systems (MSIS 2011) Luxembourg, May, 2011*

ELFEKY, M., VERYKIOS, V., ELMAGARMID, A.K.: TAILOR (2002), «A Record Linkage Toolbox», *Proceedings of the 18th International Conference on Data Engineering IEEE Computer Society, San Jose, CA, USA.*

FAIR, M. (2001), «Recent developments at Statistics Canada in the linking of complex health files», *Federal Committee on Statistical Methodology, Washington D.C., USA.*

FELLEGI, I.P., SUNTER, A.B. (1969), «A Theory for Record Linkage», *Journal of the American Statistical Association*, 64, pp. 1183-1210.

FORTINI, M., SCANNAPIECO, M., TOSCO, L. TUOTO, T. (2006), «Towards an Open Source Toolkit for Building Record Linkage Workflows», *Proceedings of SIGMOD 2006 Workshop on Information Quality in Information Systems (IQIS), Chicago, USA.*

GILL, L. (2001), «Methods for Automatic Record Matching and Linkage and their Use in National Statistics. National Statistics Methodological Series», *25, HMSO Norwich, UK.*

ISAD PROJECT (2008) «Report of WP1 - State of the art on statistical methodologies for integration of surveys and administrative data of the» (http://www.essnet-portal.eu/deliverables-wp1literature-review-isad-wp1).

JARO, M. A. (1989). «Advances in record linkage methodology as applied to the 1985 census of Tampa Florida», *Journal of the American Statistical Society*, 84 (406), pp.414–20.

MANCINI, L., VALENTINO, L., BORRELLI, F., MARCONE, L. (2012) «Record Linkage between Large Datasets: Evidence from the 15th Italian Population Census» in *Quaderni di statistica* vol. 12 (Proceedings of Conference "Methods and Models for Latent Variables" MMLV 2012), pp. 149-152.

NEWCOMBE, H., KENNEDY, J., AXFORD, S., JAMES, A. (1959), «Automatic Linkage of Vital Records», *Science*, 130, pp. 954-959.

TUOTO, T. , CIBELLA, N., FORTINI, M., SCANNAPIECO, M. TOSCO, L. (2007), "RELAIS: Don't Get Lost in a Record Linkage Project", In Proceedings of the Federal Committee on Statistical Methodologies (FCSM) Research Conference, Arlington, VA, USA.

TUOTO, T., BRUZZONE, S., VALENTINO, L., BALDASSARRE, G., CIBELLA, N., AND PAPPAGALLO M. «Towards an integrated surveillance system of road accidents» *Proceedings of 46th Scientific Meeting of the Italian Statistical Society*, Roma 20-22 giugno 2012.

YANCEY, W. (2007), «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, 2007-01.

ZARDETTO D., SCANNAPIECO M., CATARCI T. (2010) «Effective Automated Object Matching», *Proceedings of the 26th International Conference on Data Engineering* (ICDE), 1-6, IEEE.