Distr.
GENERAL

WP.28
16 May 2011

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2011)
(Luxembourg, 23-25 May 2011)

Topic (iv): International cooperation/collaboration

Open source software: a way to enrich local solutions

Supporting Paper

Prepared by Nicoletta Cibella, Monica Scannapieco, Tiziana Tuoto, Italian National Statistical Office, Italy

I. Introduction

1. According to the actual EUROSTAT new vision which incites to the standardization and engineering of the statistical process, the sharing of software suitable for solving specific problems among different NSIs can lead to considerable benefits in terms of cost reduction and quality improvement. Indeed, what typically happens is that each NSI develops its own software solution to address a problem that has to be faced also in other NSIs. The reuse of this solution by another institute obviously implies saving money, generally ensuring, when softwares 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. This paper describes the fruitful experiences in sharing the RELAIS software, designed by ISTAT for dealing with record linkage problem. The winning choice of the open source philosophy for developing the RELAIS tool firstly favoured spontaneous cooperation and collaboration among Institutes facing the same micro-data integration problems. Secondly, the ESSnet ISAD (Integration Survey and Administrative Data) project was able to converge these synergies in a well organized framework (Cibella et al, 2009). Finally, thanks to the DI (Data Integration) ESSnet project, a very effective way of collaboration for sharing knowledge and solutions has been found: among other initiatives, ISTAT also lead a very appreciated on-the-job training on record linkage methods. This experience was very profitable both for trainers and trainees. Among other advantages, mixing the theoretical aspects of record linkage and the solutions proposed in RELAIS allows the trainees to approach the topic and to directly test the proposed methods on their own specific problems. Some technical difficulties and possible improvements arose as well, they will be also highlighted in the paper.

II. RELAIS, a shareable tool

3. The RELAIS (REcord Linkage At IStat) system is a toolkit that, for each phase of a record linkage process, makes available a set of techniques. This approach permits a great flexibility when designing record

linkage processes. The system has a Graphical User Interface (GUI) that allows the dynamic composition and execution of record linkage processes. In particular, a RELAIS process can be run in a very flexible way. Indeed, it can be stopped and executed later in time, as the system offers a support for managing persistence of both data and metadata.

4. Istat started developing RELAIS in 2006 and the system is now at its 2.1 release, while 2.2. release is going to be published. Both design and implementation of the system were planned in order to produce a *shareable* tool:

- the design of RELAIS as a toolkit gives the possibility of adding new techniques to the system, and thus reusing solutions that are already available;
- RELAIS implementation has been carried on by making use of open source technologies, namely Java and R as programming languages and MySQL as database management system.

5. More specifically, as far as design choices, RELAIS has been developed in a modular way, with clear and defined interfaces between the various modules. The use of an object –oriented language like Java has permitted to design RELAIS by stressing software quality dimensions, like decoupling and information hiding, that have a significantly impact on reuse. Moreover, most of the record linkage phases have been designed in order to be plugged in different record linkage workflows. So, a RELAIS user can almost freely combine modules realizing the different phases in order to get the desired record linkage process.
As far as implementation choices, all the elements of RELAIS technological platform are open source. We also reused already developed software inside RELAIS, and specifically:

- most of the comparison functions are part of the Java package StringMetrics (http://www.dcs.shef.ac.uk/~sam/stringmetrics.html);
- the 1:1 reduction phase (that is one record of the first data can be matched with at most one record of the second and the other way round) is implemented by making use of the R package lpSolve (http://cran.r-project.org/web/packages/lpSolve/index.html).

6. Both source code and executables of RELAIS have been released on Istat site (http://www.istat.it/strumenti/metodi/software/analisi_dati/relais/) and on the OSOR web site (http://forge.osor.eu/projects/relais/). The system has been released under the European Union Public Licence (EUPL), a free software licence created and approved by the European Commission. It was not immediate to decide the particular licence to use for releasing the RELAIS system. Indeed, RELAIS was the first system that Istat decided to release as an open system freely downloadable from its official site, so no previous experience was available. EUPL was chosen because of its characteristic of being consistent with the copyright law in the 27 Member States of the European Union, while retaining compatibility with popular open-source software licences such as the GPL license.

III. The record linkage problem and the RELAIS solution

A. An overview on record linkage techniques

7. 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. Record linkage is also know as object identification, record matching, entity matching, entity resolution, reference reconciliation. Furthermore, 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. 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. 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 microdata files; the study of the relationship among variables reported in different sources.

8. 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. However no particular record linkage technique has emerged as the best solution for all cases. Additionally, in some applications, there is no evidence to prefer one method to others or of the fact that different choices, at a specific linkage stage, could bring to the same results. We believe that there isn't the best solution for all cases and that an alternative strategy should be adopted: it could be reasonable to dynamically select the most appropriate technique for each phase of the linkage problem and to combine the selected techniques for building an overall strategy. 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.

9. The unit identification is very hard to achieve in absence of unique identifiers or when the variables are affected by errors. So record linkage can be seen as a complex process consisting of several distinct phases involving different knowledge areas and the choice of the most appropriate technique does not depend only on the practitioner's skill but it is also application specific. The main phases in which a record linkage problem can be decomposed are: pre-processing/preparation of the input files; creation-reduction of the search space of link candidate pairs; choice of the common identifying attributes- matching variables; choice of comparison functions; choice of decision model; identification of unique links; record linkage procedures evaluation.

10. Generally speaking, the complexity of a linking process relies on several aspects. As mentioned above, if unique identifiers are available in the data sources the problem can be quite easily treated but unique identifiers are not always available and more sophisticated statistical procedures are required. Obviously, errors in the linking variables may invalidate the linkage results, thus a big effort for reducing such errors is necessary to prepare input files; such a phase, according to Gill (2001), requires 75% of the whole effort to implement a record linkage procedure. 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, nonmatches 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; blocking and sorted neighbourhood are the two main methods. After the search space creation/reduction, it is important to pay attention to the selection of matching variables. The matching variable need to be as suitable as possible for the considered linking process that is why they are generally chosen by a domain expert. Unique identifiers can be considered the best link variables; but very strict controls need to be made in case of using numeric identifiers alone. Variables like name, surname, address, can be used jointly instead of using each of them separately; in such a way, one can overcome problems like the wide variations of the name spelling. 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 it's not useful to choose the variable as a matching one. Comparison function is used to calculate the distance between records that are compared respect to the values of the selected matching variables (see for a reviews of comparison functions Koudas N. and Srivastava D. (2005)). 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, the set of non-matches or in the set of possible matches. The decision rule can be deterministic or probabilistic: the former considered a pair as a true match if it agrees completely on all the chosen matching variables or if it 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 requires an estimation of model parameters and can be performed via mixture models, Bayesian methods, etc. Additionally, a linkage process can be 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) many-to-one problem if a record in a set can be matched with more than one of the compared file, (iii) many-to-many problem allows more than one record in each file to be matched 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 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 respectively. 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. Generally, false non-matches of matching cases are the most critical ones because of the difficulty of checking and detecting them (Ding and Fienberg, 1994). In general, it's not easy to find automatic procedures to estimate these types of errors so as to evaluate the quality of record linkage procedures.

11. In the next paragraph, the main characteristics of the RELAIS toolkit are described, underlining that this software allows to dynamically select the most appropriate technique for each record linkage phase and to combine the selected techniques so that the resulting workflow is actually built on the basis of application and data specific requirements (Fortini et al (2006), Tuoto et al (2007)).

B. Procedures available in the current version of RELAIS

12. The core of the RELAIS toolkit is the idea of decomposing the record linkage process in its phases; this makes the whole process easier to manage as each phase has its own windows. Each of the phases described in the previous paragraph can be performed according to different techniques; depending on specific applications and features of the data it can be suitable to iterate and/or omit some phases, as well as it could be better to choose some techniques rather than others; in the current version, RELAIS provides some of the most widespread methods and techniques for the record linkage phases.

13. In the Relais 2.1 version is possible the:

- reading of input files both in text format and from database (mysql or oracle) tables;
- data profiling to guide the choice of matching and blocking variables;
- creation of the search space of pairs candidate to link by means of the "cross product", "blocking" method or "sorted neighborhood" method;
- choice of matching variables;
- set of comparison functions (with several string distances);
- probabilistic record linkage: estimation of the F - S model parameters via the EM algorithm;
- deterministic record linkage: both exact and rule based;
- reduction from N:M to 1:1 matching solution with optimal or greedy methods.

14. The dataset acquisition phase permits to read two input datasets from both in text format and from database (mysql or oracle) tables. 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. After 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 are used to support two critical tasks, that is blocking variables choice and matching variables selection. To give the opportunity to the user of designing the more appropriate record linkage workflow for its own application, RELAIS 2.1 supplies quality metadata, calculated starting from real data provided as input. Moreover, in order to go towards needs of less-expert users, RELAIS proposes a set quality metadata, coming from our experience 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.

15. 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 2.1 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. 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. 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 the 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.

IV. International experiences in using RELAIS

A. An overview of extra-Istat RELAIS interaction

16. Spontaneous collaboration among NSIs was favoured by the open source philosophy adopted in RELAIS and thanks to the common nature and the characteristics of micro-data integration problems faced by the different NSIs but 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 could affect the outcome of the same linkage.

17. Among the fruitful experiences in exchanging knowledge on record linkage techniques and in testing and enriching the RELAIS toolkit we would like to describe the one with the Spanish National Statistical Institute (INE). The collaboration between ISTAT and INE started during the ESSnet ISAD (Integration Survey and Administrative Data) project which 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 space search, in a context of registers with high amount of data to be compared.

18. From the Spanish tests some strengths and weaknesses of the RELAIS first release were highlighted and the collaboration with the INE group favoured an enriched planning for the 2.0 version of the software. At the same objective contributed also the continuous exchange via email of opinions, ideas and possible solutions with the statistical office of the United Kingdom, Brazil and Tunisia that also tested the RELAIS toolkit.

19. So, the initial planned 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.

B. The on the job training in ONS

20. Thanks to the experience in exchanging knowledge on record linkage and on testing RELAIS in a "real-world" tasks outside Italy, ISTAT, as coordinator of the DI (Data Integration) ESSnet project, conducted on January 2011 in U.K. a very appreciated on-the-job training on record linkage methods.

21. Istat decided to organize the training on the job according to these crucial aspects:
- the combination of the theoretical concepts of record linkage with the solutions proposed in RELAIS;
- the test of the RELAIS toolkit, during the computer session, on the specific record linkage problem faced by ONS on their own data;
- a very interactive way of conducting the lessons by the trainers.

22. Some months before the on-the-job training, ISTAT asked to ONS to test RELAIS on their computer machines in order to solve technical problems that could rise and almost all of them were solved with a very frequent exchange of emails between ISTAT and ONS experts. Unfortunately, some technical difficulties arose during the training because of different versions of the operating systems and the choice of mySql environment as relational database architecture which is not so suitable for ONS. ISTAT asked also to have the data on which ONS what to test RELAIS before the course in order to study the complexity of the ONS record linkage problem. The whole experience was very profitable both for trainers and trainees and on July a new on-the-job training will be conducted in Latvia.

23. Due to all these advantageous experiences in sharing knowledge on record linkage during the ESSNET projects and in cooperating with the other NSIs on "real-world" task by means of spontaneous cooperation, courses and on-the-job training, the RELAIS software was improved and the current version 2.1 is dramatically enriched compared to the first release and thanks to the modular approach and the open source choice, adding new techniques to the pool already available is really easy.

24. For the next future we would analyse some new functionalities for the software. The implementation of some standardized functionalities for the pre-processing phase, as character conversions, schema reconciliation, standardization, the modification of the probabilistic approach ; the interactions between matching variables; the Bayesian approach and the graphical analysis on the model fitting are some of the main important ones.

References

Christen, P. & 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.

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

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.
Koudas N. and Srivastava D. (2005), “Approximate joins: Concepts and techniques”, Proceedings of VLDB 2005.
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.
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.
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.