Towards an Open Source Toolkit for Building Record Linkage Workflows

Marco Fortini [*] Monica Scannapieco [†] Laura Tosco[‡] Tiziana Tuoto [§]

ABSTRACT

Record linkage has been subject of research for several decades, and a huge number of record linkage solutions have been proposed, based on probabilistic and empirical paradigms. However, record linkage is a complex process, for the execution of which one single technique is often not enough; it can be seen as composed by distinct phases, each requiring a specific technique and depending on given application and data requirements. Due to such complexity and application dependency, in this paper we propose a toolkit for record linkage, called RELAIS. The toolkit is based on the idea of choosing the most appropriate technique for each phase, and of combining such techniques in a dynamically built record linkage workflow. A real case study 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.

1. INTRODUCTION

Record linkage is a process that aims to identify if two (or more) records represent the same real world entity or not. It can be performed for different purposes, including *deduplication*, when multiple records referring to the same real world entity are erroneously stored within one single source; *data integration*, across multiple data sources in order to provide a reconciled global record; *correction* across multiple data sources, performed when one source has higher quality data that can be used for improving the other sources.

Record linkage is performed because identifiers can be missing, or because, though present, such identifiers can be affected by errors. In other words, the record linkage process is not straightforward in the majority of cases, needing instead more or less complex rules for deciding the status of matches or non-matches of record pairs. In Figure 1, we show an example with two sources storing information about some shoe shops in OHIO, namely `S1` and `S2` [1]. The records `S1.1` and `S2.3` can be declared as a `match`: the names can be easily verified as equals, the addresses have some differences (the one in source `S1` is more detailed), the telephone numbers are different but we could admit that a shop can have more than one telephone number. The record `S1.2` and `S2.3` can be declared as a `non-match`: the names can be verified as equals but the addresses are very different, and the telephone numbers are also different. Similar considerations can be done for declaring as a `match` the pair `(S1.3,S2.1)` and as a `non-match` the pair `(S1.3,S2.2)`.

Due to its relevancy, record linkage has been widely investigated since the late 60s when the Fellegi and Sunter theory for record linkage was proposed [9]. This model is still widely used, and several methods for the estimation of its parameters have been proposed (see [21] for a survey). On the other hand, besides such probabilistic methods, there has been a proliferation of techniques that can be classified as *empirical*, including [12, 17, 1, 5]. However, despite such huge production, 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 including: preprocessing, in which standardization activities are performed; choice of a comparison function, to be used for the actual record comparisons; blocking, for reducing the number of comparisons; decision, for coming up with a set of matched records and a set of non-matched ones, etc. For each of these phases, several techniques can be adopted; for instance, for the decision phase, the Fellegi and 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. We claim that the choice of the most appropriate technique is application specific. Also, it is reasonable to dynamically select the most appropriate technique for each phase and to

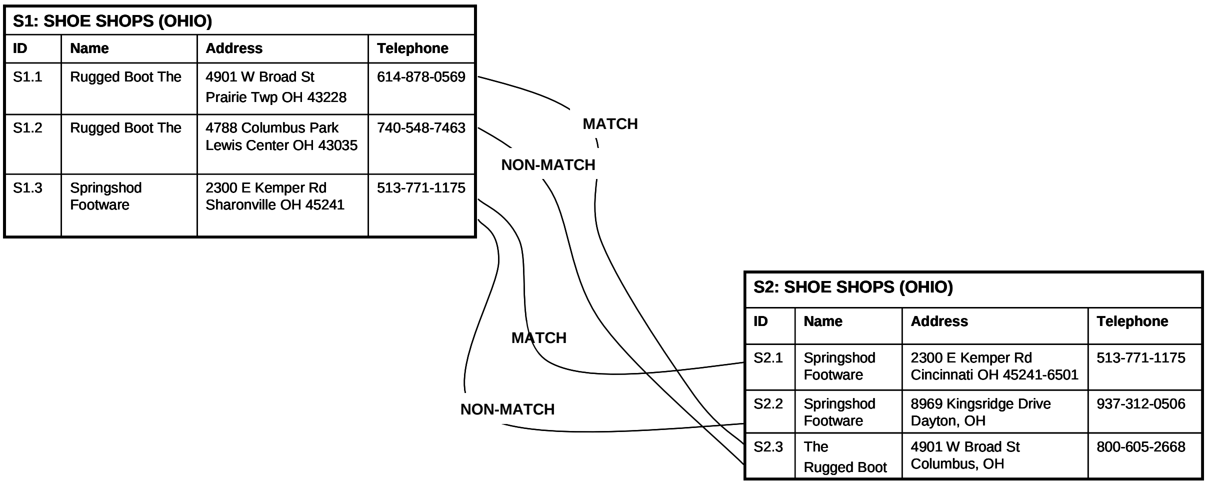
[*] Istituto Nazionale di Statistica - ISTAT, Italia; `fortini@istat.it`

[†] Istituto Nazionale di Statistica - ISTAT and Universitá degli Studi di Roma “La Sapienza”, Italia; `scannapi@istat.it,monscan@dis.uniroma1.it`

[‡] Istituto Nazionale di Statistica - ISTAT, Italia; `tosco@istat.it`

[§] Istituto Nazionale di Statistica - ISTAT, Italia; `tuoto@istat.it`

[1] The shown data are the actual result of a query posed to two yellow pages sites, namely `http://yp.yahoo.com/py/` and `http://www.yellowbook.com/`.



S1: SHOE SHOPS (OHIO)			
ID	Name	Address	Telephone
S1.1	Rugged Boot The	4901 W Broad St Prairie Twp OH 43228	614-878-0569
S1.2	Rugged Boot The	4788 Columbus Park Lewis Center OH 43035	740-548-7463
S1.3	Springshod Footware	2300 E Kemper Rd Sharonville OH 45241	513-771-1175

S2: SHOE SHOPS (OHIO)			
ID	Name	Address	Telephone
S2.1	Springshod Footware	2300 E Kemper Rd Cincinnati OH 45241-6501	513-771-1175
S2.2	Springshod Footware	8969 Kingsridge Drive Dayton, OH	937-312-0506
S2.3	The Rugged Boot	4901 W Broad St Columbus, OH	800-605-2668

Figure 1: An example of record linkage decision

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 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 some specific data and application requirements, could support the definition of the most appropriate record linkage workflow.

Several record linkage systems and tools have been proposed, in both the academic and private sectors. Such tools include Big Match [23], CANLINK [7], Febrl [8], Tailor [6] and The Link King [18]. The first two systems have been developed by the U.S. and Canadian Statistics Institutes. Some of the systems provide the user a certain degree of flexibility, e.g. Febrl allows for choosing 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, none of them relies on 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, differently from RELAIS, 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. A technological solution for composing record linkage operations is proposed in [4]; however, in this work the focus is on the performance of a service oriented architecture and record linkage is only considered as an application domain.

We will 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 dynamic record linkage workflows.

The major contributions of this paper can be summarized as follows. First, we illustrate the idea of a dynamic record linkage workflow that, to the best of our knowledge, has not been previously proposed. Second, we validate the RELAIS idea by means of a real case study in which a record linkage workflow is instantiated starting from data and application requirements. From the case study, we also abstract a pattern to be included in the RELAIS library.

2. BACKGROUND

In this section, we present a short introduction to the record linkage problem in terms of its different phases. Record linkage is a process whose purpose is to identify the same real word entity, which can be differently represented in one or more data sources. Record linkage complexity depends on several aspects, in particular: (i) the absence of keys, which forces to choose a set of attributes, called *matching attributes* to be used as keys for the linkage; (ii) keys with errors, namely keys can be affected by accuracy errors thus being jeopardized their identification power. Record linkage may give rise to two types of errors, namely: false matches, when a match is erroneously declared between two records that do not actually correspond to the same real world entity, and false non-match, when it is a non-match which is erroneously identified.

As shown in Figure 2, the record linkage process is composed of two main phases, namely: search space reduction and application of a decision model. When linking records of a set A with records of a set B, the initial search space of matching records consists of the cartesian product $(A \times B)$, therefore, given that n is the cardinality of A and of B, the complexity of an exhaustive search technique is $O(n^2)$. 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), $a \in A$ and $b \in B$, that have to be compared. Starting from this reduced search space, we can apply different decision models which define the rules used to decide if a pair of records (a, b) is a match, a non-match or a possible

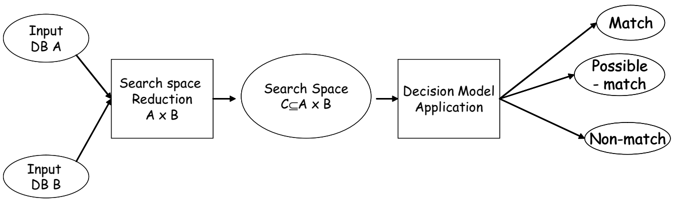
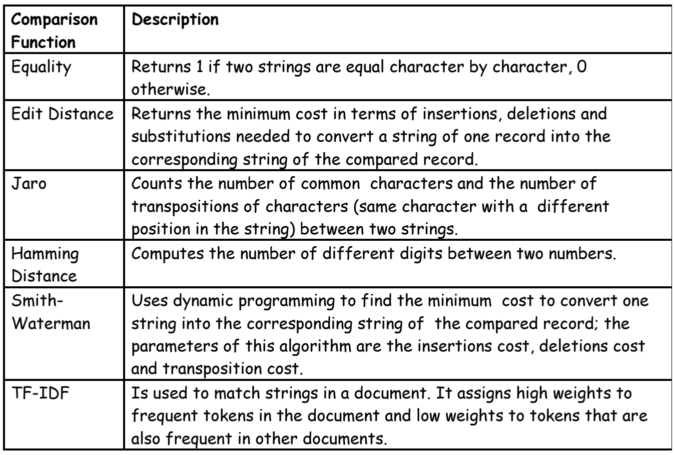


Figure 2: Phases of record linkage

match. In more detail, a record linkage process is composed of further phases, in addition to the previously described ones :

- *Preprocessing*: in this phase null strings are deleted, upper/lower cases are converted, a parsing activity can be performed for normalizing record attributes, attribute values can be standardized (e.g., the address standardization may imply that all the expressions denoting a "street" are translated into "str") and schema reconciliation is applied to avoid possible conflicts (i.e. description, semantic and structural conflicts [14]) among data source schemas.
- *Choice of the matching attributes*: in this phase the attributes to be used for linking records are chosen. A (sub)set of these attributes can be selected as a key for the following blocking and sorted neighborhood search space reduction methods. The matching attributes are typically chosen by a domain expert, hence this phase is typically not automatic. However, whereas metadata description are available on data sources to be matched, a partially automatic choice can be performed by taking into account the identification power of the attributes to select and their quality [3].
- *Choice of a comparison function*: defines the comparison function used to calculate the distance between values of the records that are compared. Some comparison functions are listed with a brief description in Figure 3; see [15] for a survey of comparison functions.
- *Search Space Reduction*: in this phase the number of comparisons necessary for finding the matching status between records is reduced. Two main methods can be used to reduce the search space, namely: blocking [11] and sorted neighborhood [12]. *Blocking* consists of partitioning the two record sets into blocks, and of searching for the the matching records only inside each block. The partition into blocks is made using blocking keys; two records belong to the same block if all the blocking keys of the two records are equal or if a hash function applied to the blocking keys of the two records gives the same result. *Sorted neighborhood* performs a sort of the two record sets using the same key, and searches possible matching records only inside a window of a fixed dimension which slides on the two ordered record sets. Occasionally, also a *pruning* step can be performed by removing all the records that certainly will not be matched with any other record. For instance, given a source that contains two disjoint partitions, e.g. women and men, and a second one containing only men, it is obvious that a pruning on the first source removing all the women should be done before starting the linkage process. Pruning rules can be more complex, and in general have the purpose to narrow a data source as down as possible towards the other(s) to be compared with.



Comparison Function	**Description**
Equality	Returns 1 if two strings are equal character by character, 0 otherwise.
Edit Distance	Returns the minimum cost in terms of insertions, deletions and substitutions needed to convert a string of one record into the corresponding string of the compared record.
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.
Hamming Distance	Computes the number of different digits between two numbers.
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.
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.

Figure 3: Comparison functions

- *Choice of decision model*: defines the method used to estimate the model parameters, and the decision rule for deciding the status of match, non-match and possible match of the compared records. Two possible approaches are:
  - the probabilistic model, based on the Fellegi and Sunter model, requires an estimation of the model parameters; such an estimation can be computed using different techniques (e.g., EM algorithm, bayesian approach, etc.);
  - the empirical model, based on identifying thresholds for comparing values of variables as well as thresholds for the classification of pairs of records (match/non-match/possible match).

On the basis of the used decision model, the whole record linkage process can be classified as *probabilistic* or *empirical.* A further classification distinguishes the following linkages: (i) one to one linkage, in which each real world entity corresponds to only one record in each data source that has to be linked; (ii) many to one linkage, in which each real world entity may correspond to two or more records in one of the involved data sources; (iii) many to many linkage, in which each real world entity may correspond to two or more records in each of the involved data sources. Note that the cases (ii)-many to one and (iii)-many to many may correspond to the existence of duplicate records in the data sources to be linked.

3. DESCRIPTION OF RELAIS

The RELAIS toolkit is composed by a collection of techniques for each record linkage phase. The toolkit idea is based on the consideration that the record linkage process is inherently very complex and existing solutions do not provide a satisfying answer to the various requirements that

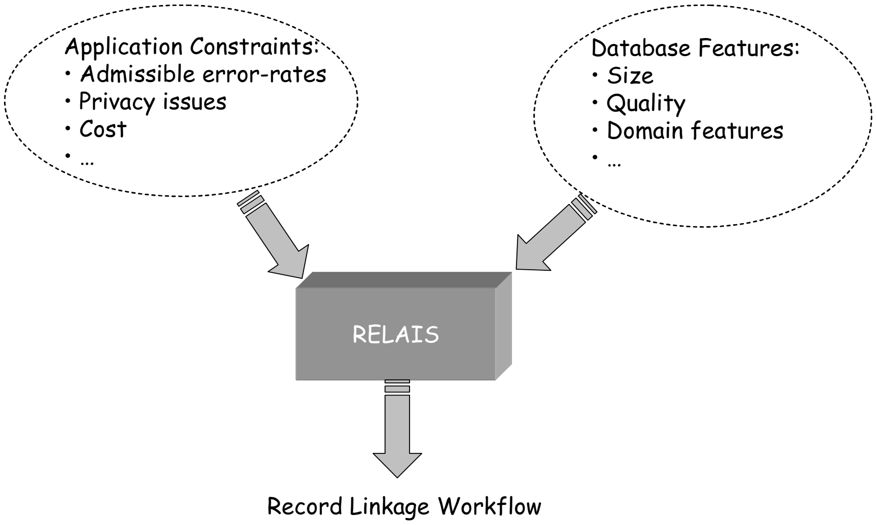


Figure 4: The RELAIS's inputs and output

different applications can exhibit. Indeed, 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 model. For instance, 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.

Therefore, we claim that no record linkage process, deriving from the combination of a specific technique for each phase, is the best for all applications. Instead, RELAIS has the purpose of offering a set of techniques that can be dynamically combined in order to build a *record linkage workflow*, given a set of application constraints and data features provided as input (see Figure 4). As an example, if it is known the databases to compare do not have high quality data, it is suggested the usage of comparison functions ensuring error tolerance (e.g. Jaro, see Figure 3), instead of the usage of an equality comparison function; as a further example, if no specific error-rates are mandatory for the application, it can be adopted an empirical decision model which can be easier to apply. Furthermore, 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 5, examples of possible workflows that may result from the RELAIS toolkit are shown.

Iterations are also possible within a record linkage process, in two forms, namely: (i) phase iteration, in which a single phase is iterated several times with different parameters in order to obtain a better result. For instance, the blocking step can be performed several times with different blocking keys; (ii) process iteration, in which the whole linkage process can be iterated by inputting to the i-th iteration the residuals of the (i-1)-th iteration. Notice that the i-th iteration should consists of a different record linkage workflow in order to obtain better results.

3.1 RELAIS as Open Source Project

As also remarked in the introduction, we intend to configure RELAIS as an open source project. There are at least two reasons for this choice. First, 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" the efforts already done in a structured way, according to the philosophy described above, and of making them available to the community for the most appropriate usage. RELAIS will be implemented in Java, due to the well-known features of strongly typing and platform independence.

4. CASE STUDY

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; it was also interesting to characterize Italian families, therefore the relationship of each enumerated person with the other component of the same household was also collected. The PES was based on the replication of the census process inside the sampled EAs and on the use of a capture-recapture model [22] for estimating the hidden amount of the population. The main objective was 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. In order to apply the capture-recapture model, after the PES enumeration of the statistical units (households and people), a

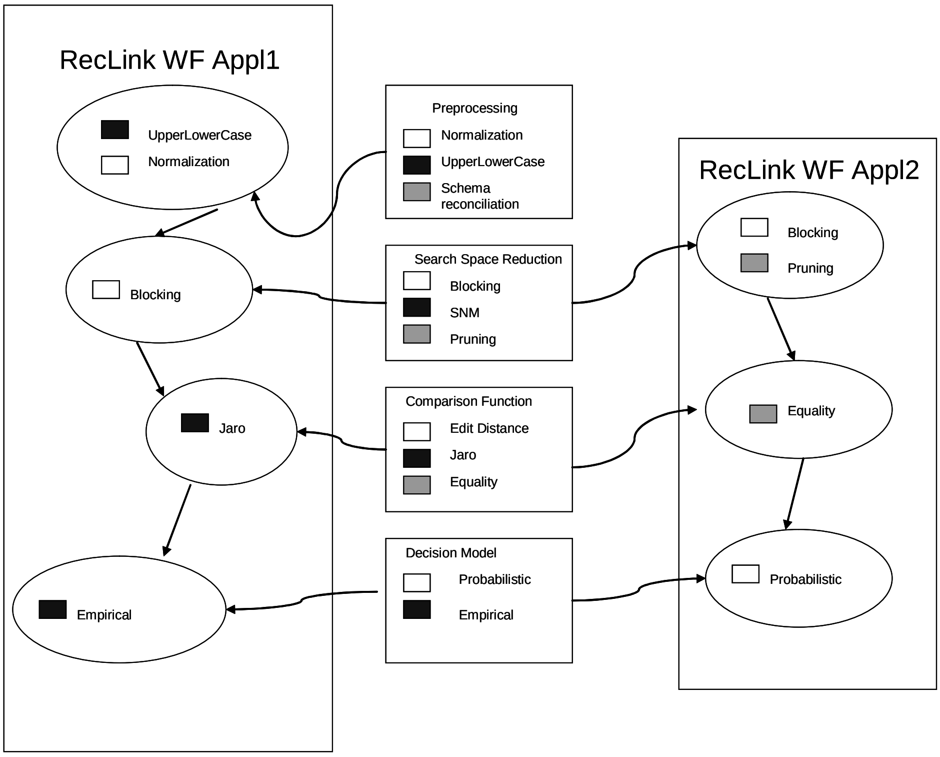


Figure 5: Examples of RELAIS's workflows

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 of the people enumerated at the census day over the hidden amount of the population, was obtained.

4.1 Record Linkage Workflow

The estimates of the census coverage rate through capture-recapture model has 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 are 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. In Figure 6, steps 1 and 2 regard two iterations of the record linkage process on households. Step 1 is performed after a preprocessing activity and it is an empirical linkage. Step 2 is a probabilistic record linkage, that can be based on the Fellegi-Sunter model [9], for which the matching weights are computed via the EM algorithm [13, 20][2]. 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 non-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 in Section 3, 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. In the case study described above, we 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

[2] We actually used a Bayesian weight estimation [10], but the detail of such usage are out of the paper's scope.

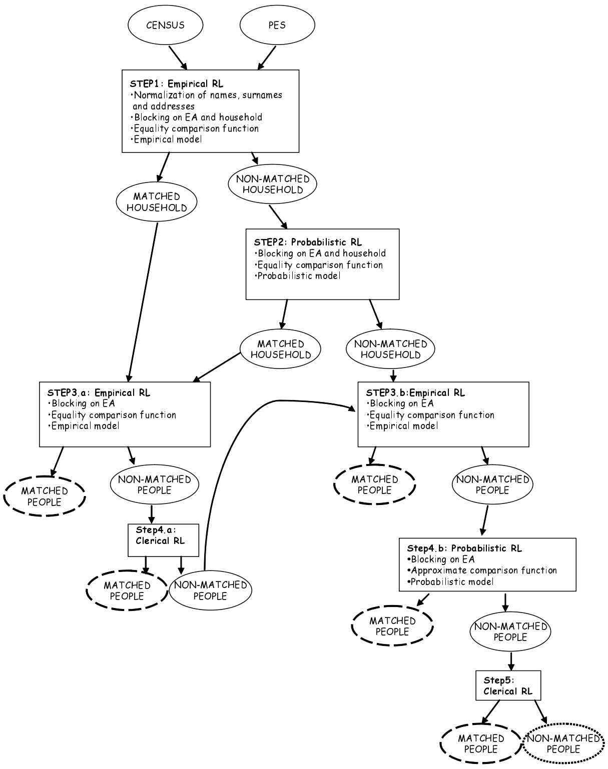
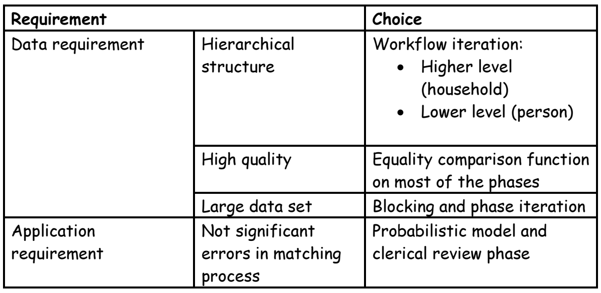


Figure 6: The record linkage workflow of the case study



Requirement		Choice
Data requirement	Hierarchical structure	Workflow iteration: • Higher level (household) • Lower level (person)
	High quality	Equality comparison function on most of the phases
	Large data set	Blocking and phase iteration
Application requirement	Not significant errors in matching process	Probabilistic model and clerical review phase

Figure 7: An example of a pattern for building record linkage workflows

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. In Figure 7, 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.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have described the RELAIS project, whose purpose is to implement an open source toolkit for building record linkage workflows. We have discussed 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 a complexity a great modularity and flexibility are necessary in order to properly build application specific record linkage workflows.

We have two main objectives in the near future for RELAIS's development. First, we intend to define the RELAIS's architecture as a service-oriented, web-accessible architecture. In order to specify each toolkit service, we are considering Semantic Web Services technologies, e.g., OWL-S [3] and WSDL-S [4]. By using semantic technologies, we can formally define input and output of the services, as well as a set of conditions that should hold prior to service invocation (preconditions) and a set of statements that should be true if the service is invoked successfully (postconditions). RELAIS's services could be thus dynamically composed in

[3] http://www.daml.org/services/owl-s/

[4] http://lsdis.cs.uga.edu/projects/meteor-s/wsdl-s/

order to form record linkage workflows. As a second step, we would like to allow the automatic generation of such record linkage workflows, on the basis of a knowledge-based reasoning on RELAIS's service specification. So far, we have informally modeled data and application requirements in the form of methodological patterns. However, we would like to formally model such knowledge in order to drive and automatic or semi-automatic workflow generation. We plan to investigate existing and current work on service composition, such as either partially automatized (e.g., [16]) or fully automatized (e.g., [2]) service composition techniques.

As remarked in the paper, we plan to carry on the implementation of the toolkit as an open source product. The core techniques can be implemented independently on the service-oriented architecture of the whole framework. Specifically, we have already started the implementation of some of the toolkit's techniques, and we plan to a have them as public available code very soon. The language used for the implementation is Java. As several methods strongly rely on statistical techniques, we plan to embed the R language [19] in Java, in order to fully use R's statistical packages.

Acknowledgments. The work of Monica Scannapieco is partially supported by MIUR, Project "ESTEEM-Emergent Semantics and co-operaTion in multi-knowledgE EnvironMents " (`http://www.dis.uniroma1.it/∼esteem/index.html`).

6. REFERENCES

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

[2] D. Berardi, D. Calvanese, G. De Giacomo, R. Hull, and M. Mecella. Automatic composition of transition-based semantic web services with messaging. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005)*, Trondheim, Norway, 2005.

[3] P. Bertolazzi, L. D. Santis, and M. Scannapieco. 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.

[4] M. Cameron, K. Taylor, and R. Baxter. Web service composition and record linking. In *Proceedings of the VLDB Workshop on Information Integration on the Web, IIWeb 2004*, Toronto, Canada,2004.

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

[6] M. Elfeky, V. Verykios, and A. K. Elmagarmid. Tailor: A Record Linkage Toolbox. In *Proceedings of the 18th International Conference on Data Engineering.* IEEE Computer Society, San Jose, CA, USA, 2002.

[7] M. Fair. Recent developments at statistics canada in the linking of complex health files. In *Federal Committee on Statistical Methodology*, Washington D.C.,2001.

[8] Febrl. http://www.sourceforge.net/projects/febrl.

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

[10] M. Fortini, B. Liseo, A. Nuccitelli, and M. Scanu. On bayesian record linkage. *Research in Official Statistics*, 4:185–198, 2001.

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

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

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

[14] W. Kim and J. Seo. Classifying schematic and data heterogeneity in multidatabase systems. *IEEE Computer*, 24(12):12–18, 1991.

[15] N. Koudas and D. Srivastava. Approximate joins: Concepts and techniques. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005)*, Trondheim, Norway, 2005.

[16] B. Medjahed, A. Bouguettaya, and A. K. Elmagarmid. Composing web services on the semantic web. *Very Large Data Base Journal*, 81(4):333–351, 2003.

[17] A. Monge and C. Elkan. 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.

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

[19] The-R-Project. http://www.r-project.org/.

[20] W. Winkler. 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.

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

[22] K. Wolter. Some coverage error models for census data. *Journal of the American Statistical Association*, 81:338–346, 1986.

[23] W. Yancey. A program for extracting probable matches from a large file for record linkage. Technical report, Statistical Research Division U.S. Bureau of the Census - Washington D.C., 2002. Research Report Series - Computing n. 2002-01.