Record Linkage: Current Practice and Future Directions

Lifang Gu, Rohan Baxter, Deanne Vickers, and Chris Rainsford

CSIRO Mathematical and Information Sciences
GPO Box 664, Canberra, ACT 2601, Australia
CMIS Technical Report No. 03/83
Firstname.Lastname@csiro.au
http://datamining.csiro.au

Abstract. Record linkage is the task of quickly and accurately identifying records corresponding to the same entity from one or more data sources. Record linkage is also known as data cleaning, entity reconciliation or identification and the merge/purge problem. This paper presents the "standard" probabilistic record linkage model and the associated algorithm. Recent work in information retrieval, federated database systems and data mining have proposed alternatives to key components of the standard algorithm. The impact of these alternatives on the standard approach are assessed. The key question is whether and how these new alternatives are better in terms of time, accuracy and degree of automation for a particular record linkage application.
Keywords: record linkage, data cleaning, entity identification, entity reconciliation, object isomerism, merge/purge, list washing.

1 Introduction

Record linkage is the task of quickly and accurately identifying records corresponding to the same entity from one or more data sources. Entities of interest include individuals, companies, geographic region, families, or households.

Record linkage has applications in customer systems for marketing, customer relationship management, fraud detection, data warehousing, law enforcement and government administration. These applications can be classed as 'administrative', because the record linkage is used to make decisions and take actions regarding an individual entity. Our own particular interest is record linkage for 'research purposes' where it is used to link information regarding entities to answer ethically-approved epidemiological, health system or social science research and policy questions.

In many research projects it is necessary to collate information about an entity from more than one data source. If a unique identifier or key

of the entity of interest is available in the record fields of all of the data sources to be linked, deterministic record linkage can be used. Deterministic record linkage assumes error-free identifying fields and links records that exactly match on these identifying fields. For large multiple data sources, shared error-free identifying fields are uncommon. Sources of variation in identifying fields include changes over time, robustness in minor typographical errors and lack of completeness in availability over space and time. When no error-free unique identifier is shared by all of the data sources, a probabilistic record linkage technique can be used to potentially join or merge the data sources.

Christen et al [20] note the process of linking records has different names in different research and user communities. While epidemiologists and statisticians speak of *record linkage*, the same process is called *entity heterogeneity* [32], *entity identification* [68], *object isomerism* [19], *instance identification* [106], *merge/purge* [47], *entity reconciliation* [33], *list washing* and *data cleaning* [20] by computer scientists and others. The term *record linkage* is used throughout in this paper.

This survey paper presents the current standard practice in record linkage methodology. We also consider possible improvements due to recent innovations in computer science, statistics and operations research. Another purpose of this paper is to identify likely future directions in improving the performance of current practice. Comprehensive recent surveys touching on the long history of record linkage techniques can be found elsewhere [113, 1, 42]. The current standard practice is based on probabilistic approaches [36, 52, 42]. Some recent contributions from the database and data mining research that may eventually supersede the current practice include rule-based record comparison methods, greater automation through active learning, and improved scalability through innovative data structures and search algorithms [47, 104, 34]. There is minimal cross-referencing between the statistical and database approaches.[1] A key contribution of this paper is an initial attempt at an integrated view of recent developments in the separate approaches. We are unaware of any previous attempts at this integration.

The paper is structured as follows. The definition of the record linkage problem, the formal probabilistic model and an outline of the standard practice algorithm are provided in Section 2. Section 3 presents recent proposals for alternatives to various components of the standard algorithm. Section 4 then puts the core record linkage process in a privacy and legal

[1] Exceptions being the wide referencing of the Fellegi and Sunter paper [36] and, more occasionally, Howard and Newcombe's 1959 Science [82] and 1962 CACM paper [81].

context. Issues such as record linkage over federated data sources and confidentiality protocols are briefly reviewed. Section 5 concludes with a summary of our views on the current methods and the most worthwhile research directions.

In Appendix A, government, academic and commercial record linkage systems and their known features are described. This records our general awareness of existing systems and will allow us to check on differentiators for any new proposed directions. Appendix B gives a list of research projects with a record linkage component. These studies are useful for providing greater context for understanding the detailed requirements for record linkage software for health and social science research projects.

2 The Record Linkage Problem, Model and Standard Algorithm

We first define the record linkage problem in Section 2.1. Section 2.2 then provides the formal probabilistic model of the record linkage. This probabilistic model has provided an influential framework for the practical systems forming the standard practice today. These systems largely follow the algorithm given in Section 2.3.

2.1 Record Linkage Problem Definition

Records in data sources are assumed to represent observations of entities taken from a particular population. The records are assumed to contain some attributes (fields or variables) identifying an individual entity. Examples of identifying attributes are name, address, age and gender.

Suppose source A has n_a records and source B has n_b records. Each of the n_b records in source B is a potential match for each of the n_a records in source A. So there are $n_a \times n_b$ record pairs whose match/non-match status is to be determined [43].

Two disjoint sets M and U can be defined from the cross-product of A with B, the set $A \times B$. A record pair is a member of set M if that pair represents a true match. Otherwise, it is a member of U. The record linkage process attempts to classify each record pair as belonging to either M or U.

Many matching problems are more constrained than this statement of the problem. For instance, if each record in data source B refers to a distinct entity, a record in data source A cannot be matched to two records at the same time in data source B. Cohen calls this the *constrained*

matching problem [26]. It is more generally referred to as 1-1 linkage in comparison to the alternative 1-many linkage. 1-1 linkage, since it has more constraints, is a harder optimisation problem [26].

2.2 Probabilistic Model of the Record Linkage Problem

Following Gomatam et al. [43], record pairs are labelled as:

- match, A_1.
- possible match, A_2.
- non-match, A_3.

For record a from source A and record b from source B, available information on the records is denoted $\alpha(a)$ and $\beta(b)$ respectively. A comparison or agreement vector, γ, for a record pair $(\alpha(a), \beta(b))$ represents the level of agreement between the records a and b.

When record pairs are compared on k identifying fields the γ vector has k components. $\gamma = (\gamma^1(\alpha(a), \beta(b)), ..., \gamma^k(\alpha(a), \beta(b)))$ is a function on the set of all $n_a \times n_b$ record pairs.

For an observed agreement vector γ in Γ, the space of all possible comparison vectors, $m(\gamma)$ is defined to be a conditional probability of observing γ given that the record pair is a true match. That is:

$$m(\gamma) = P(\gamma|(a, b) \in M) \tag{1}$$

Similarly,

$$u(\gamma) = P(\gamma|(a, b) \in U) \tag{2}$$

denotes the conditional probability of observing γ given that the record pair is a true non-match.

There are two kinds of possible misclassification errors: false matches (Type I error) and false non-matches (Type II error). The probability of a false match is:

$$P(A_1|U) = \sum_{\gamma \in \Gamma} u(\gamma) P(A_1|\gamma) \tag{3}$$

and the probability of a false non-match is:

$$P(A_3|M) = \sum_{\gamma \in \Gamma} m(\gamma) P(A_3|\gamma) \tag{4}$$

For fixed values of the false match rate (μ) and false non-match rate (λ), Fellegi and Sunter [36] define the optimal linkage rule on Γ at levels μ and λ, denoted by $L(\mu, \lambda, \Gamma)$ as the rule for which $P(A_1|U) =$

$\mu, P(A_3|M) = \lambda$, and $P(A_2|L) \leq P(A_2|L')$ for all other rules L'. Optimal is defined here as the rule that minimises the probability of classifying a pair as belonging to A_2, the subset of record pairs requiring manual review. Other criteria for optimality are considered in Section 3.7.

Let $\frac{m(\gamma)}{u(\gamma)}$ be ordered to be monotonically decreasing (with ties broken arbitrarily) and the associated γ be indexed $1, 2, ..., N_\Gamma$. If $\mu = \sum_{i=1}^{n} u(\gamma_i)$ and $\lambda = \sum_{i=n'}^{N_\Gamma} m(\gamma_i), n < n'$, then the optimal rule is a function of the likelihood ratio $\frac{m(\gamma)}{u(\gamma)}$ and is given by the following equations:

$$(a, b) \in A_1 \text{ if } T_\mu \leq \frac{m(\gamma)}{u(\gamma)} \tag{5}$$

$$\in A_2 \text{ if } T_\lambda < \frac{m(\gamma)}{u(\gamma)} < T_\mu \tag{6}$$

$$\in A_3 \text{ if } \frac{m(\gamma)}{u(\gamma)} \leq T_\lambda \tag{7}$$

where $T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$ and $T_\lambda = \frac{m(\gamma_{n'})}{u(\gamma_{n'})}$. Figure 1 illustrates these three regions in terms of the degree of agreement of the record pair.

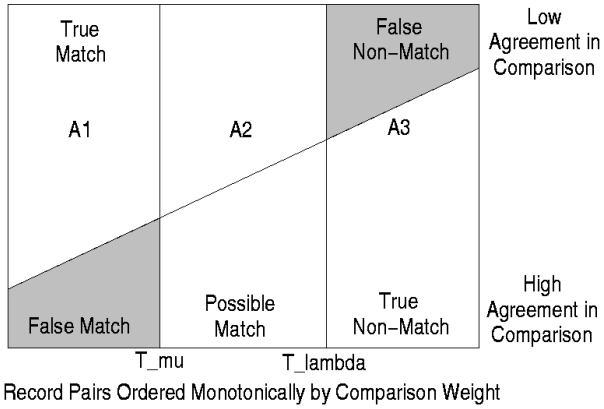

Fig. 1. The three regions of the probability model.

Under the assumption of conditional independence of the components of the γ vector, the decision rule above can be written as a function of $\log(\frac{m(\gamma)}{u(\gamma)}) = \sum_{j=1}^{k} w_j$, where the weight $w_j = \log(\frac{m(\gamma^j)}{u(\gamma^j)})$.

Intuitively, there will be many more non-matching record pairs than matching record pairs. Figure 2 is a typical histogram of record pair weights illustrating this. The non-match mode is much larger than the matching mode. The degree of separation between the modes is an indication of the level of difficulty of the linkage task and amount of Type I

and II errors that may result. This is assuming there is ground truth of matches available.

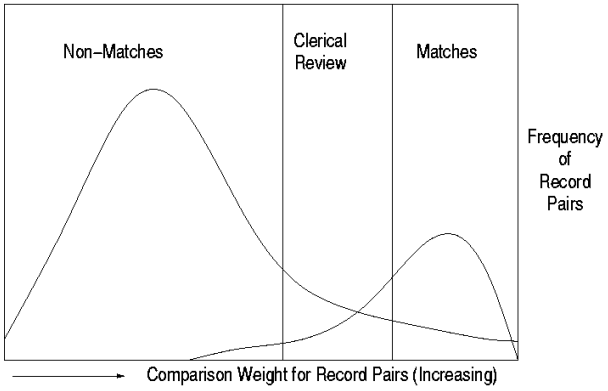

Fig. 2. Histogram of comparison weights showing a mode for matching weights and a mode for non-matching weights, and a degree of overlap where matching errors will occur.

Alternative Statistical Models This standard model represents the data probabilities directly. An alternative is to model errors in attributes explicitly [28]. Different Bayesian models of record linkage [38, 64] have also been proposed. The practicality and accuracy of these methods are not clear. Systems implementing them have not been widely used and are not widely available.

2.3 Standard Algorithm for Record Linkage

The framework of the previous section is the basis for the standard algorithm for record linkage. The operationalisation of the framework requires a method for estimating the weights, w_j, or more generally, the likelihood ratio $\frac{m(\gamma)}{u(\gamma)}$.

Jaro [51, 52] uses the expectation-maximisation (EM) algorithm [31] to estimate $m(\gamma), \gamma \in \Gamma$. The complete data vector is given by (γ, g), where g indicates the actual status (match or non-match) of the record pair. Jaro restricts the γ^j to be 0/1 values and assumes conditional independence of the γ^j. The term, g, takes a value corresponding to a match with probability p and non-match with probability $1-p$. The likelihood is

written in terms of $g, m(\gamma)$ and $u(\gamma)$. The EM algorithm uses maximum likelihood estimates of $m(\gamma)$, $u(\gamma)$ and p to estimate unobserved g.

The EM algorithm needs initial estimates of $m(\gamma)$, $u(\gamma)$ and p and then iterates. Jaro uses frequencies for estimates of $m(\gamma)$, $u(\gamma)$ and p within blocks. This results in biased $u(\gamma)$ estimates. So only the $m(\gamma)$ values from this solution are used. The $u(\gamma^j), j = 1, 2, ..., k$, probabilities are estimated as the probabilities of chance agreement of the j attribute by carrying out a frequency analysis.

The standard record linkage algorithm [51, 52] requires the following inputs:

- Initial values for $m(\gamma)$ and $u(\gamma)$ probabilities.
- Blocking attribute(s) for each blocking pass. Different blocking attributes are used in each pass. Blocking passes are described in more detail below.
- The thresholds for the weights to determine the three decision regions, A_1, A_2 and A_3.
- The type of comparison functions.

The algorithm then goes through the following steps:

- Perform a *Blocking Pass*: Select possible pairs for linking from the data sources using one or more blocking attributes. Note that there is an implicit assumption that comparisons not made due to blocking are non-match records. For example, if both data sources are sorted by postcode, the comparison pairs would consist of only records where postcodes agree. Methods for reducing errors due to blocking or for minimising blocking are described in Section 3.3.
- Estimate the matching probability, $m(\gamma)$, and non-matching probability, $u(\gamma)$, for the attributes used for matching within the stratum defined by the blocking attribute(s).
- Use the probabilities to calculate the matching weight for each attribute.
- Calculate the composite score from the weights of all the matching attributes.
- Determine whether a record pair is a match, non-match or possible match using the threshold levels on the composite score.
- Perform another blocking pass and the above steps on the remaining non-match record pairs until all blocking passes are done.

3 Record Linkage System and Components

This section describes a record linkage system design largely following the TAILOR system [34]. The basic requirements for each key system component are then described. New approaches to implementing these components are considered and compared to the standard algorithm presented in Section 2.3.

3.1 Record Linkage System Design

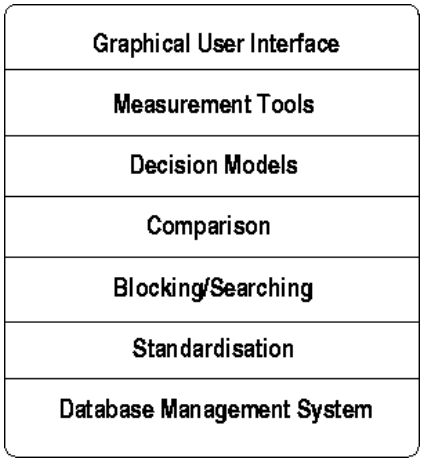

Fig. 3. Layered Design for a Record Linkage System.

Figure 3 shows a layered design of a record linkage system. The Information Flow diagram of this record linkage system, following TAILOR [34], is shown in Figure 4. We now briefly discuss the components of this record linkage system.

The Standardisation component is arguably optional, depending on the quality of the data sources to be linked. Some successful record linkage systems do not separate standardisation [97], but rather incorporate it into the Comparison component. Approaches to standardisation are described in Section 3.2.

Blocking is used to reduce the number of comparisons of record pairs and blocking strategies are described in Section 3.3.

The Comparison component performs comparison of record pairs. Comparison methods are described in Section 3.5.

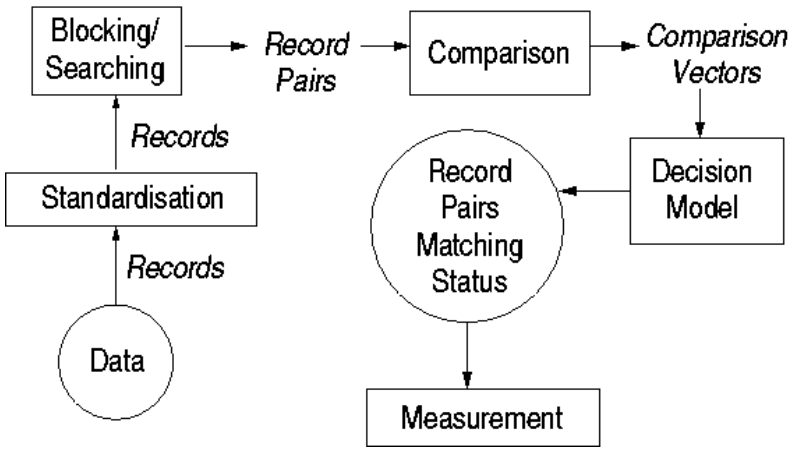

Fig. 4. Information Flow for a Record Linkage System.

The Decision component makes the decision of whether a record pair is a match, non-match or possible match. Alternative decision models are described in Section 3.6.

The Measurement component allows evaluation of the record linkage system along different performance criteria and these are described in Section 3.7.

The Graphical User Interface component should provide the following functionalities:

- Entry of the standard algorithm parameters.
- Assistance for identification and specification of common identifying attributes for use in the record linkage as discussed in Section 3.4.
- Support for the presentation and interpretation of performance measurements.
- Allowing convenient clerical review of possible matches, where the record linkage process is not fully automatic.

Alternative Record Linkage System Designs Elfekey et al. [34] present an alternative record linkage system model called the *Induction Record Linkage Model*. This model shows how training data can be incorporated in the record linkage system, where it is available.

Other alternative record linkage systems that have been proposed include AJAX [40], WHIRL [24], Intelliclean [66], Merge/Purge [48], and SchemaSQL [63]. Some of the designs and architectures for the many

available academic, government and commercial systems are found in Appendix A.

3.2 Standardisation Methods

Standardisation is also called *data cleaning* or *attribute-level reconciliation*. Without standardisation, many true matches could be wrongly designated as non-matches because the common identifying attributes do not have sufficient similarity [108].

The basic ideas of standardisation are:

- To replace many spelling variations of commonly occurring words with standard spelling.
- To standardise the representation of various attributes, to the same system of units, or to the same coding system. For example, 0/1 instead of M/F for a 'gender' attribute.
- To perform integrity checks on attribute values or combinations of attribute values.

Standardisation methods need to be specific to the population of an application and the data capture processes. For example, the most common name misspellings differ based upon the origin of the name. Therefore standardisation for Italian names optimised to handle names Italian and Latin origin will perform better than generic standardisation [30].

3.3 Searching/Blocking

Searching or *blocking* is used to reduce the number of comparisons of record pairs by bringing potentially linkable record pairs together. A good attribute variable for blocking should contain a large number of attribute values that are fairly uniformly distributed and such an attribute must have a low probability of reporting error.

Errors in the attributes used for blocking can result in failure to bring linkable record pairs together. For text attributes, various phonetic codes have been derived to avoid effects of spelling and aural errors in recording names. Common phonetic codes include Russell-Soundex and NYSIIS. These codes were optimised for specific populations of names and a specific type of English pronunciation. Some commercial systems provide tools to derive phonetic codes for specific populations worldwide [97].

Kelley [57] developed an algorithm of choosing the best blocking scheme in light of the trade-off between computation cost and false non-match (negative) rates.

Sorted Neighbourhood Method (SNM) [48] SNM involves scanning the N sorted records from sources A and B using a fixed size of window, w. Every pair of records falling within the window are compared. SNM requires $w \times N$ record comparisons. Note that the error rate induced by SNM is critically dependent on the choice of sorting keys.

Multiple passes with independent sorting keys could be used to minimise the number of errors [48]. A transitive closure over matched record pairs can be computed for combining the results of independent passes [75]. An example of the circumstances where this is useful is as follows:

```
A matches B -> Drop B
C matches D -> Drop C
E does not match A or D but would have matched
B and C
Result: A,D,E are kept separate when in reality
they are all the same entity.
```

A problem with this multi-pass approach is that the number of false positives is increased as they are propagated across each pass [79].

Priority Queue Method [76] This method is related to SNM, but sets of representative records belonging to recent clusters in the sorted record list are stored in a priority queue. Heuristics are needed to select these representative records of a cluster. The advantage is the avoidance of the need to sort the data sources for each blocking pass, which can save significant computational time for very large data sources. Winkler and Yancey [112] have implemented a similar approach in their Bigmatch system.

Blocking as Preselection [79] The idea behind preselection is based on quickly computed *rejection rules* because almost all record pairs can be classified as *non-match* through simple computation. Preselection is the application of adequate rejection rules to reduce the number of comparisons. Rejection rules can be derived from the comparison functions used and a training sample. The type of preselection comparison performed can be adjusted to control the trade-off between the time complexity of blocking and the required rate of misclassified pairs.

Canopies clustering is one of the efficient high-dimensional clustering methods which can be used for preselection [72, 10].

3.4 Selection of Attributes for Matching/Comparison

Common attributes should be selected for use in the comparison function. The issues involved in the attribute selection process include:

- Identifying which attributes are common.
- Assessing whether the common attributes have sufficient information content to support the linkage quality required for the project. Information theory measures for assessing project viability have been proposed in [90, 27].
- Selecting the optimal subset of common attributes.

Attribute characteristics that affect the selection decision include level of errors in attribute values and the number (and distribution) of attribute values, i.e. information content of the attribute.

For example, a field such as gender only has two value states and consequently could not impart enough information to identify a match uniquely. Conversely, a field such as surname imparts much more information, but it may frequently be recorded incorrectly.

One decision rule for attribute selection is to select all the available common attributes. Redundancy provided by related attributes can be useful in reducing matching errors. However, redundancy from attributes is useless if their errors are highly correlated or even functionally dependent.

The power of low quality personal identifiers can be enhanced by considering semantics of the fields such as cause of death and known co-morbidities [73].

3.5 Comparison

The probabilistic framework of Fellegi-Sunter requires the calculation of the comparison vector, γ for each record pair. Jaro [51] considers comparison vectors consisting of only match/non-match (0/1) values. However, the comparison function can be extended to include categorical and continuous-valued attribute values as well [112].

Dey et al. [33] proposed a distance-based metric for attribute comparison. They also have a probability-based metric in an earlier paper [32], but argue that probabilities are difficult to estimate accurately and so a distanced-based metric is more robust.

Text or string attributes are very commonly used as matching attributes. Several string comparators are therefore considered in the following.

String comparison in record linkage can be difficult because lexicographically "nearby" records look like "matches" when they are in fact not. For example, consider the following three strings:

```
"Apt 11,101 North Rd, Acton, 2601"
"Apt 12,101 North Rd, Acton, 2601"
"Apt 12,101 North St, Acton, 2601"
```

When transitive closure is applied to pairs of nearby records, incorrect results can often occur. This will result in the following conclusion:

```
Apt 11,101 North Rd, Acton, 2601
->
Apt 12,101 North St, Acton, 2601
```

The use of the semantics of the strings in domain-dependent comparison functions can help avoid this problem. In this example, `Apt 11` and `Apt 12` can be parsed and tagged as an apartment number and a comparison function can capture the difference of even one digit as being very significant and so the records should not be matched.

A string comparator function returns a value between 0 and 1 depending on the degree of the match of the two strings. Because pairs of strings often exhibit typographical variation (e.g., Smith and Smoth), the record linkage needs effective string comparison functions that deal with typographical variations. Hernandez and Stolfo [47] discussed three possible distance functions for measuring typographic errors; they are *edit distance*, *phonetic distance*, and *typewriter distance*. They developed an Equational Theory involving a declarative rule language to express these comparison models [47].

Some alternative string comparison methods include:

- Manual construction [39].
- Jaro [51] introduced a string comparator that accounts for insertions, deletions, and transpositions. The basic steps of this algorithm include computing the string lengths and finding the number of common characters in the two strings and the number of transpositions. Jaro's definition of "common" is that the agreeing character must be within the half of the length of the shorter string. Jaro's definition of transposition is that the character from one string is out of order with the corresponding common character from the other string. The string comparator value is given by the following formula:

$$C(s1, s2) = 1/3 * (\frac{N_{common}}{L_{s1}} + \frac{N_{common}}{L_{s2}} + 0.5 * \frac{N_{transpositions}}{N_{common}}) \quad (8)$$

Where $s1$ and $s2$ are the two strings to be compared, with lengths L_{s1} and L_{s2} respectively. N_{common} and $N_{transpositions}$ are the numbers of common characters and transpositions.

- Winkler [88] modified the original string comparator introduced by Jaro in the following three ways:
  - A weight of 0.3 is assigned to a 'similar' character when counting common characters. Winkler's model of similar characters includes those that may occur due to scanning errors ("1" versus "l") or key punch errors ("V" versus "B").
  - More weight is given to agreement at the beginning of a string. This is based on the observation that the fewest typographical errors occur at the beginning of a string and the error rate then increases monotically with character positions through the string.
  - The string comparison value is adjusted if the strings are longer than six characters and more than half the characters beyond the first four agree.
- N-gram distance [49, 44, 104]: The N-grams comparison function forms the set of all sub-strings of length n for each string. The N-grams method has been extended to Q-grams by Gravano et al [44] for computing approximate string joins efficiently.
- Edit-distance based [87, 53]: This method uses edit distance, also known as Levenshtein distance [67], to compare two strings. Edit distance, a common measure of textural similarity, is the minimum number of edit operations (insertions, deletions, and substitutions) of single characters required to transform from one string to another (i.e., make two strings equal). A dynamic programming algorithm is used to find the optimal edit distance. The time complexity of this algorithm can be an issue for large databases [53].
- Vector space representation of fields such as TF-IDF [24]: Cohen uses the Information Retrieval representation of text and heuristics for ranking document similarity as a means to do schema integration without a conceptual global model. Cohen and Richman claim that TF-IDF often outperforms edit-distance metrics and is less computationally expensive [26, 25].
- Adaptive comparator function [26]: This method learns the parameters of the comparator function using training examples. Zhu and Ungar [118] use a genetic algorithm to learn the edit operator costs for a string-edit comparator function.

3.6 Decision Models

Once matching weights of individual attributes of two records are calculated, the next step is to combine them to form a composite weight or score and then decide whether a record pair should be a match, non-match or possible match.

The simplest way to calculate the composite weight is to use the average of all the matching weights assuming each attribute contributes equally. If some knowledge on the importance of individual attributes is available, the weighted (fixed) average can be used.

If the attribute value distribution for a field is not uniform, a value-specific (frequency-based), or outcome-specific, weight [110] can be introduced. For example, surname "Smith" occurs more often than "Zabrinsky" and therefore a match of "Smith" carries less weight than a match of "Zabrinsky". The basic idea is that an agreement on rarely occurring values of an attribute is better at distinguishing matches than that on commonly occurring values of an attribute.

Dey et al. [33] propose a method of directly eliciting the attribute weights from the user. The user ranks the attributes in order of their perceived predicted power, with a rank of 1 for the most predictive attribute, a rank of 2 for the second most predictive attribute, etc. Several reasons are provided in the literature [6] in favour of rank-based surrogate weights over directly-elicited weights. The (surrogate) weights can be computed based on these ranks. The rank-centroid method was chosen for converting the ranks to weights. Although satisfactory results were achieved by getting the ranks from the user, it would be difficult to apply this method if the user does not have enough knowledge or information about the attributes.

Earlier we described the optimal rule for minimising the number of possible matches given desired Type I and Type II errors, assuming conditional independence. We also described the EM method with the conditional independence assumption. The EM method has been derived without assuming conditional independence [109, 74].

We now consider several alternative decision models for deciding whether a record pair should be a match, non-match or possible match.

Statistical Models Copas and Hilton [28] propose a matching algorithm that depends on the statistical characteristics of the errors which are likely to arise. The distribution of errors can be studied using a large training file of matched record pairs. Statistical models are fitted to a file of record pairs known to be correctly matched. These models are then used to

estimate likelihood ratios. The advantage of this approach is that the model fit can be used to consider the validity of the modelling approach.

Predictive Models Predictive models for learning the parameters (threshold values and attribute weights) have recently been proposed. Adequate training data is needed to train these models [109, 64]. Proposed models have included:

- Logistic regression [87], although this was found to not work for census data in [64].
- Support vector machines [10].
- Decision trees [103].

Active learning techniques have also been proposed to optimise efficiency in selection of training records [103].

Bayesian Decision Cost Model Verykios et al. [104] propose a Bayesian decision model for *cost* optimal record matching. Conventional models for record matching rely on decision rules that minimise the probability of error, i.e., the probability that a sample record pair is assigned to the wrong class. Because the misclassification of different samples may have different consequences, their decision model minimises the cost of making a decision rather than the probability of error in a decision.

3.7 Performance Measurement

Quality of record linkage can be measured in the following dimensions:

- The number of record pairs linked correctly (true positives) n_m.
- The number of record pairs linked incorrectly (false positives, Type I error) n_{fp}.
- The number of record pairs unlinked correctly (true negatives) n_u.
- The number of record pairs unlinked incorrectly (false negatives, Type II error) n_{fn}.

Along with the two ground truth numbers (the total number of true match record pairs, N_m, and the total number of true non-match record pairs, N_u), various measures from different perspectives can be defined from these dimensions. Several of these measures [43] are listed in the following:

- Sensitivity: n_m/N_m, the number of correctly linked record pairs divided by the total number of true match record pairs.
- Specificity: n_u/N_u, the number of correctly unlinked record pairs divided by the total number of true non-match record pairs.
- Match rate: $(n_m + n_{fp})/N_m$, the total number of linked record pairs divided by the total number of true match record pairs.
- Positive predictive value (ppv): $n_m/(n_m + n_{fp})$, the number of correctly linked record pairs divided by the total number of linked record pairs.

It can be seen that sensitivity measures the percentage of correctly classified record matches while specificity measures the percentage of correctly classified non-matches.

Two other performance measures widely used in the research field of information retrieval is *precision* and *recall*. Precision measures the purity of search results, or how well a search avoids returning results that are not relevant. Recall refers to completeness of retrieval of relevant items. For record linkage, precision can be defined, in terms of matches, as the number of correctly linked record pairs divided by the total number of linked record pairs. So precision is equivalent to the *positive predicted value* defined above. Similarly, recall is defined, in terms of matches, as the number of correctly linked record pairs divided by the total number of true match record pairs. As a result, recall is equivalent to *sensitivity* defined above. Of course, precision and recall can also be defined in terms of non-matches. Alternatively, combined measures of precision and recall can be defined in terms of overall record pairs correctly classified (matches and non-matches).

Additional performance criteria for record linkage are in terms of time and number of records requiring manual review:

- Time taken. The time complexity of a record linkage algorithm is usually dominated by the number of record comparisons performed. The time taken for sorting on a blocking key for very large data sources can also be extremely long.
- Number of records requiring clerical review. Manual review of records is time-consuming, expensive and can be error prone.

Controlling Error Rates Belin and Rubin [7] proposed a mixture model for estimating false match rates for given threshold values. This is currently the only method for automatically estimating record linkage error rates [109]. The method works well when there is good separation

between the matching weights associated with matches and non-matches, but it requires the existence of previously collected and accurate training data. For situation where there is no good separation, methods that use more information from the matching process [107, 64] can be used to estimate the error rates.

4 Wider Issues in Record Linkage

We have focussed on methodological issues for record linkage so far. In this section, we briefly mention the legal and ethical issues that are integral to any substantial record linkage project involving individuals. In particular, a number of linkage protocols for minimising risk in release of confidential information are discussed in Section 4.3.

4.1 Ethical and Legal Issues

Ethical privacy concerns and legislative obligations both need to be met in a record linkage study [80, 8, 60]. A good description of the issues, accepted processes and example documents are found in [45]. A comprehensive review of the legal issues for data linkage in Australian jurisdictions can be found in [71].

4.2 Analytic Methods for Linked Data

Understanding of the error characteristics of linked data has been flagged as a critical limiting factor in allowing widespread application of analyses to linked data [59]. Wang and Donnan [105] propose methods for adjusting for missing records in outcome studies. The methods model the missing mechanism and therefore allow the regression methods to have reduced bias and more accurate confidence intervals.

The effect of errors on registry-based follow-up studies is an important issue [17].

4.3 Linkage Protocols to Maintain Confidentiality

For research purpose linkage projects, personal identifying information is not often relevant to the research problem and can be removed from the linked dataset. One exception is where the research outcomes identify a risk reduction action for individuals in the dataset and technical and ethical issues in following up with those individuals may result.

In some projects, anonymity of individuals prior to linkage is also desirable. Methods for achieving this involve encrypting or pseudonymising identifying information in the data sources in a consistent way prior to linkage [4, 89, 12]. Record linkage on encrypted identifying attributes is likely to decrease the linkage accuracy over time [77]. For example, if an individual changes their surname, and their surname is used as part of the linkage key, then the match will not be made with new records for that individual [45].

Two key ideas in the anonymised record linkage area [22] are:

- Separate the identifying information, such as name and address, from other information, such as clinical events or diagnoses prior to linkage between data sources. This enables a separation between data for record linkage and the data for the researcher. This idea requires trust of a linkage unit or other third-party body [62, 22, 58, 45].
- Individual identifier keys should have limited scope. The scope could be limited to a single research project [58].

A protocol implementing some of these ideas has been proposed for Australian government departments and agencies [45]. Limitations of this current protocol include:

- It is designed for one-off linkage projects rather than long-term longitudinal and linked data collection with permanent storage and incremental additions over time.
- It does not cover staged multi-program linkage projects, where data can be added from new programs over time.

4.4 Linkage protocols for distributed databases

An Italian research project into data quality improvement for distributed systems has designed and is implementing a prototype for a Record Matcher and Record Improver [9, 5, 93].

It is argued that automation of the blocking key selection is needed for multiple distributed source systems [5]. Quality metadata for the sources and identification power of an attribute are used as input into the key selection algorithm.

5 Conclusions and Research Questions

Current record linkage methods perform well when the matching fields are well standardised and there are sufficient attributes for matching [109].

The key insight is that new techniques for modelling text from Information Retrieval, for performing similarity joins from Database Research and for learning optimal decision boundaries from data mining can potentially improve the record linkage process in terms of manual effort and scalability [111].

Winkler identifies a key research question regarding the selection of matching/comparison method [109]. In principle, weighting attribute matching by frequency-based weights should be more informative than simple agree/disagree matching. However, for less frequent attribute values and for noisy data sources, this is not actually the case. A method for deciding between these approaches is needed.

Adaptive learning where the comparator function is learnt has also recently been proposed [26]. Predictive models from machine learning such as bagging methods and SVMs have been suggested for learning the match/non-match decision function (Section 3.6). Other learning methods for learning the comparator functions have also been proposed (Section 3.5). A direct comparison with the Fellegi-Sunter approach has not yet been done but would be worthwhile [26].

Another area of interest is avoiding the need to sort large datasets for blocking. This can be done by using recent developments in high-dimensional similarity joins [53]. These techniques use clever data structures to store records so that good candidates for matching are stored together based on the agreed distance or probabilistic measure.

References

[1] W. Alvey and B. Jamerson, editors. *Record Linkage Techniques – 1997*. Federal Committee on Statistical Methodology, Washington, D.C., 1997.

[2] M.G. Arellano, G.R. Peterson, D.B. Petitti, and R.E. Smith. The California Mortality Linkage System (CAMLIS). *Am. J. Pub. Health*, 74:1324–30, 1984.

[3] Inc. Arkidata. Arkistra, 2003.

[4] A.L. Avins, W.J. Woods, B. Lo, and S.B. Hulley. A novel use of the link-file system for longitudinal studies of HIV infections: a practical solution to an ethical dilemma. *AIDS*, 7:109–113, 1993.

[5] R. Baldoni, C. Cappiello, C. Francalanci, B. Pernici, P. Plebani, M. Scannapieco, S.T. Piergiovanni, and A. Vigirillito. Dl4.a: Design and definition of the cooperative architecture supporting data quality, December 2001.

[6] F.H. Barron and B.E. Barrett. Decision Quality Using Ranked Attribute Weights. *Management Science*, 42(11):1515–1523, 1996.

[7] T.R. Belin and D.B. Rubin. A Method for calibrating false-match rates in record linkage. *Journal of the American Statistical Assocation*, 90(430):694–707, June 1995.

[8] G.B. Bell and A. Sethi. Matching Records in a National Medical Patient Index. *Communications of the ACM*, 44(9):83–88, 2001.

[9] P. Bertolazzi, L. DeSantis, and M. Scannapieco. Automatic Record Matching in Cooperative Information Systems. In *Int. Workshop on Data Quality in Cooperative Information Systems*, Jan 2003.

[10] M. Bilenko and R.J. Mooney. Learning to Combine Trained Distance Metrics for Duplicates Detection in Databases. Technical Report AI-02-296, University of Texas at Austin, Feb 2002.

[11] Vinayak R. Borkar, Kaustubh Deshmukh, and Sunita Sarawagi. Automatic segmentation of text into structured records. In *SIGMOD Conference*, 2001.

[12] F. Borst, F-A. Allaert, and C. Quantin. The Swiss Solution for Anonymous Chaining Patient Files. *MEDINFO 2001*, pages 1239–1241, 2001.

[13] Andrew Borthwick. Choicemaker Technologies, Inc., 2002.

[14] K.J. Brameld, C.D.J. Holman, M. Thomas, and A.J. Bass. Use of a state data bank to measure incidence and prevalence of a chronic disease: end-stage renal failure. *American Journal of Kidney Disease*, 34(6):1033–1039, 1999.

[15] K.J. Brameld, M. Thomas, C.D.J. Holman, A.J. Bass, and I.L. Rouse. Validation of linked administrative data on end-stage renal failure: application of record linkage to a 'clinical base population'. *Aust. NZ J. of Public Health*, 23:464–467, 1999.

[16] P.P. Breitfeld, T. Dale, J. Kohne, S. Hui, and W.M. Tierney. Accurate case finding using linked electronic clinical and administrative data at a children's hospital. *Journal of Clinical Epidemiology*, 54, 2001.

[17] H. Brenner, I. Schmidtmann, and C. Stemaier. Effect of record linkage errors on registry-based follow-up studies. *Statistics in Medicine*, 16:2633–43, 1997.

[18] F. Caruso, M. Cochinwala, U. Ganapathy, G. Lalk, and P. Missier. Telcordia's Database Reconciliation and Data Quality Analysis Tool. In *Proc. of the 26th Int. Conf. on Very Large Databases*, 2000.

[19] Arbee L. P. Chen, Pauray S. M. Tsai, and Jia-Ling Koh. Identifying Object Isomerism in Multidatabase Systems. *Distributed and Parallel Databases*, 4(2):143–168, 1996.

[20] P. Christen and T. Churches. *Febrl: Freely extensible biomedical record linkage*, release 0.2 edition, April 2003.

[21] P. Christen, T. Churches, et al. http://datamining.anu.edu.au/projects/linkage.html, 2003.

[22] T. Churches. A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Medical Research Methodology*, 3(1):1–13, 2003.

[23] T. Churches, P. Christen, K. Lim, and J. X. Zhu. Preparation of name and address data for record linkage using hidden Markov models. *BMC Medical Informatics and Decision Making*, 2, 2002.

[24] William W. Cohen. Integration of heterogeneous databases without common domains using queries based on textual similarity. In *Proceedings of ACM SIGMOD-98*, pages 201–212, 1998.

[25] W.W. Cohen and J. Richman. Learning to Match and Cluster Entity Names. In *Proc. of the ACM SIGIR-2001 Workshop on Mathematical/Formal Methods in Information Retrieval*, 2001.

[26] W.W. Cohen and J. Richman. Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration. In *SIGKDD'02*, 2002.

[27] L.J. Cook, L.M. Olson, and J.M. Dean. Probabilistic Record Linkage: Relationships between File Sizes, Identifiers, and Match Weights. *Methods of Information in Medicine*, 40:196–203, 2001.

[28] J.B. Copas and F.J. Hilton. Record Linkage: Statistical models for Matching Computer Records. *Journal of the Royal Statistical Society Series A*, 153:287–320, 1990.

[29] W.S. Dai, S. Xue, K. Yoo, J.K. Jones, and J. Labraico. An Investigation of the Safety of Midazolam use in Hospital. *Pharmacoepidemiology and Drug Safety*, 9, 1997.

[30] L. Dal Maso, C. Braga, and S. Franceschi. Methodology Used for Software for Automated Linkage in Italy (SALI). *Journal of Biomedical Informatics*, 34:387–395, 2001.

[31] N.M. Dempster, A.P. Laird and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:185–197, 1977.

[32] D. Dey, S. Sarkar, and P. De. A Probabilistic Decision Model for Entity Matching in Heterogeneous Databases. *Management Science*, 44(10):1379–1395, 1998.

[33] D. Dey, S. Sarkar, and P. De. A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases. *IEEE Transactions on Knowledge and Data Engineering*, 14(3), 2002.

[34] M.G. Elfeky, V.S. Verykios, and A.K. Elmagarmid. TAILOR: A Record Linkage Toolbox. In *Proc. of the 18th Int. Conf. on Data Engineering*. IEEE, 2002.

[35] M.E. Fair. Recent Developments at Statistics Canada in the linking of complex health files. In *Federal Committee on Statistical Methodology*, Washington, D.C., 2001.

[36] L. Fellegi and A. Sunter. A Theory for Record Linkage. *Journal of the American Statistical Society*, 64:1183–1210, 1969.

[37] D.R. Fletcher, M.S.T. Hobbs, P. Tan, R.L. Hockey, T.J. Pikora, M.W. Knuiman, H.J. Sheiner, A. Edis, and L.J. Valinsky. Complications following cholecystectomy - risks of the laproscopic approach and protective effects of operative cholangiography: a population based study. *Annals of Surgery*, 229:449–57, 1999.

[38] M. Fortini, B. Liseo, A. Nuccitelli, and M. Scanu. On bayesian record linkage. In *Sixth International World Meeting on Bayesian Analysis*, 2000.

[39] H. Galhardas, D. Florescu, D. Shasha, and E. Simon. AJAX: An Extensible Data Cleaning Tool. In *SIGMOD (demonstration paper)*, 2000.

[40] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C. Saita. Declarative Data Cleaning: Language, Model and Algorithms. In *Proc. of the 27th VLDB Conf.*, 2001.

[41] Helena Galhardas. http://cosmos.inesc.pt/ hig/cleaning.html, 2003.

[42] L. Gill. Methods for Automatic Record Matching and Linking and their Use in National Statistics. Technical Report National Statistics Methodological Series No. 25, National Statistics, London, 2001.

[43] S. Gomatam, R. Carter, M. Ariet, and G. Mitchell. An empirical comparison of record linkage procedures. *Statistics in Medicine*, 21:1485–1496, 2002.

[44] L. Gravano, P.G. Ipeirotis, H.V. Jagadish, N. Koudas, S. Muthukrishnan, and D. Srinivasta. Approximate string joins in a database. In *Proc. 27th Int. Conf. on Very Large Data Bases*, pages 491–500, 2001.

[45] Statistical Linkage Key Working Group. Statistical data linkage in Community Services data collections, 2002.

[46] J. Halliday, O. Griffin, A. Bankier, C. Rose, and M. Riley. Use of Record Linkage Between a Statewide Genetics Service and a Birth Defects/Congenital Malformations Register to Determine Use of Genetic Counselling Services. *American Journal of Medical Genetics*, 72, 1997.

[47] M.A. Hernandez and S.J. Stolfo. The Merge/Purge Problem for Large Databases. In *Proc. of 1995 ACT SIGMOD Conf.*, pages 127–138, 1995.
[48] M.A. Hernandez and S.J. Stolfo. Real-world data is dirty: data cleansing and the merge/purge problem. *Journal of Data Mining and Knowledge Discovery*, 1(2), 1998.
[49] J.A. Hylton. Identifying and merging related bibliographic records, 1996.
[50] Info Route Inc. ClientMatch, 2003.
[51] M. A. Jaro. Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Society*, 84(406):414–420, 1989.
[52] M.A. Jaro. Probabilistic linkage of large public health datafiles. *Statistics in Medicine*, 14:491–498, 1995.
[53] L. Jin, C. Li, and S. Mehrotra. Efficient Record Linkage in Large Data Sets. In *Int. Conf. on Database Systems for Advanced Applications (DASFAA)*, Tokyo, Japan, March 2003.
[54] C. Johansen, J.D. Boice, J.K. McLaughlin, and J.H. Olsen. Cellular telephones and cancer— a nationwide cohort study in Denmark. *J. Natl. Cancer Inst.*, 93:203–7, 2001.
[55] J.A. Johnson and S.M. Wallace. Investigating the Relationship Between β-Blocker and Antidepressant Use Through Linkage of the Administrative Databases of Saskatchewan Health. *Pharmacoepidemiology and Drug Safety*, 6, 1997.
[56] S.B. Jones. Linking Databases in Biotechnology. *Neuroimage*, 4, 1996.
[57] R.P. Kelley. Blocking considerations for record linkage under conditions of uncertainity. In *Proceedings of the Social Statistics Section, American Statistical Association*, pages 602–605, 1984.
[58] C.W. Kelman, A.J. Bass, and C.D. Holman. Research use of linked health data-a best practice protocol. *Aust N Z J Public Health*, 26:251–255, 2002.
[59] B. Kilss and W. Alvey, editors. *Record Linkage Techniques - 1985. Proceedings of the Workshop on Exact Matching Methodologies in Arlington, Virginia May 9-10.* Internval Revenue Service Publication, Washington, DC, 1985.
[60] N.R. Kingsbury et al. Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information, 2001.
[61] S. Kohli, K. Sahlen, O. Lofman, A. Sivertun, M. Foldevi, E. Trell, and O. Wigertz. Individuals living in areas with high background radon: a GIS method to identify populations at risk. *Computer Methods and Programs in Biomedicine*, 53, 1997.
[62] R.L. Kruse, B.G. Ewigman, and G.C. Tremblay. The Zipper: a method for using personal identifiers to link data while preserving confidentiality. *Child Abuse & Neglect*, pages 1241–1248, 2001.
[63] L. Lakshmanan, F. Sadri, and I. Subramanian. Schema-SQL - A Language for Interoperability in Relational Database Systems. In *Proc. of VLDB*, 1999.
[64] M.D. Larsen and D.B. Rubin. Iterative automated record linkage using mixture models. *Journal of the American Statistical Assocation*, 96(453):32–41, March 2001.
[65] D. Lawrence, C.D.J. Holman, A.V. Jablensky, and S.A. Fuller. Suicide rates in psychiatric inpatients: an application of record linkage to mental health research. *Aust. NZ J. of Public Health*, 23:468–470, 1999.
[66] M.L. Lee, T.W. Ling, and W.L. Low. Intelliclean: A knowledge-based intelligent data cleanser. In *Proc. of the Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 290–294, 2000.

[67] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

[68] E. Lim, J. Srivastava, S. Prabhakar, and J. Richardson. Entity identification in database integration. In *IEEE International Conference on Data Engineering*, pages 294–301, 1993.

[69] S. Liu and S. Wen. Development of Record Linkage of Hospital Discharge Data for the Study of Neonatal Readmission. *Chronic Diseases in Canada*, 20(3), 2000.

[70] N. Maconochie, P. Doyle, E. Roman, E. Davies, P.G. Smith, and V. Beral. The nuclear family study: linkage of occupational exposures to reproduction and child death. *British Medical Journal*, 93:203–7, 2001.

[71] R.S. Magnusson. Data Linkage, Health Research and Privacy: Regulating Data Flows in Australia's Health Information System. *Sydney Law Review*, 24(5), 2002.

[72] A. McCallum, K. Nigam, and L. Ungar. Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integation. In *Proc. of the Sixth Int. Conf. on KDD*, pages 169–170, 2000.

[73] M.C.M. Mcleod, C.A. Bray, S.W. Kendrick, and S.M. Cobbe. Enhancing the power of record linkage involving low quality personal identifiers: use of the best link principle and cause of death prior likelihoods. *Compt. Biomed. Res.*, 31:257–270, 1998.

[74] X. Meng and D.B. Rubin. Maximum Likelihood via the ECM ALgorithm: A General Framework. *Biometrika*, 80:267–278, 1993.

[75] A.E. Monge. Matching algorithm within a duplicate detection system. *IEEE Data Engineering Bulletin*, 23(4), 2000.

[76] A.E. Monge and C.P. Elkan. An efficient domain-independent for detecting approximately duplicate database records. In *Proc. of the ACM-SIGMOD Workshop on Research Issues in on Knowledge Discovery and Data Mining*, 1997.

[77] A.G. Muse, J. Mikl, and P.F. Smith. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. *Statistics in Medicine*, 14:499–509, 1995.

[78] M. Neiling, S. Jurk, H. Lenz, and F. Neumann. Object Identification Quality. In *Intl. Workshop on Data Quality in Cooperative Information Systems*, 2003.

[79] M. Neiling and R.M. Muller. The good into the Pot, the bad into the Crop. Preselection of Record Pairs for Database Fusion. In *Proc. of the First International Workshop on Database, Documents, and Information Fusion*, Magdeburg, Germany, 2001.

[80] C.I. Neutel. Privacy Issues in Research Using Record Linkage. *Pharmcoepidemiology and Drug Safety*, 6:367–369, 1997.

[81] H.B. Newcombe and J.M. Kennedy. Record Linkage. *Communications of the Association for Computing Machinery*, 5:563–566, 1962.

[82] H.B. Newcombe, J.M. Kennedy, S.J. Axford, and A.P. James. Automatic Linkage of Vital Records. *Science*, 130:954–959, 1959.

[83] T.B. Newman and A.N. Brown. Use of commercial record linkage software and vital statistics to identify patient deaths. *J. Am. Med. Inform. Assoc.*, 4:233–237, 1997.

[84] P.E. Norman, J.B. Semmens, M.M.D. Lawrence-Brown, and C.D.J. Holman. Long-term survival after surgery for abdominal aortic aneurysm in Western Australia: population based study. *British Medical Journal*, 317:852–856, 1998.

[85] The West of Scotland Coronary Prevention Study Group. Computerised Record Linkage: Compared With Traditional patient Follow-up Methods in Clinical Trials and Illustrated in a Prospective Epidemiological Study. *Journal of Clinical Epidemiology*, 48(12), 1995.

[86] K.M. Patterson, C.D.J. Holman, and D.R. English. First-time hospital admissions with illicit drug problems in indigenous and non-indigenous West Australians: an application of record linkage to public health surveillance. *Aust. NZ J. of Public Health*, 23:460–463, 1999.

[87] J.C. Pinheiro and D.X. Sun. Methods for linking and mining massive heterogeneous databases. In *Fourth Int. Conf. on Knowledge Discovery and Data Mining*, 1998.

[88] E.H. Porter and W.E. Winkler. Approximate string comparison and its effect on an advanced record linkage system. In *Proc. of an International Workshop and Exposition - Record Linkage Techniques*, Arlington, VA, USA, 1997.

[89] C. Quantin, F-A. Allaert, and L. Dussere. Anonymous statistical methods versus cryptographic methods in epidemiology. *Int. J. Med. Inf.*, 60:177–183, 2000.

[90] L.L. Roos and A. Wajda. Record Linkage Strategies. *Methods of Information in Medicine*, 30:117–123, 1991.

[91] D.L. Rosman. The Western Australian Road Injury Database (1987-1996): ten yearsof linked police,hospital and death records of road crashes and injuries. *Accident Analaysis and Prevention*, 33:81–88, 2001.

[92] Inc. Sagent. Centrus Merge/Purge, 2003.

[93] M. Scannapieco, M. Mecella, T. Cartarci, C. Cappiello, B. Pernici, F. Mazzoleni, and F. Stella. DL3: Comparative Analysis of the Proposed Methodologies for Measuring and Improving Data Quality and Description of an Integrated Proposal, 2003.

[94] SearchSoftwareAmerica. Introduction to SSA-NAME3, V2.0, 2001.

[95] SearchSoftwareAmerica. An Introduction to Identity Systems(IDS), 2002.

[96] SearchSoftwareAmerica. Introduction to the Data Clustering Engine, V2.2, 2002.

[97] SearchSoftwareAmerica. Solving the Company Name Search & Matching Problem, 2003.

[98] J.B. Semmens, P.E. Norman, M.M.D. Lawrence-Brown, and C.D.J. Holman. The influence of gender on the outcomes of ruptured abdominal aortic aneurysm. *British Journal of Surgery*, 87:191–4, 2000.

[99] J.B. Semmens, C. Platell, T. Threlfall, and C.D.J. Holman. A population-based study of the incidence, mortality and outcomes following surgery for colorectal cancer in Western Australia. *Aust. NZ J. of Surgery*, 70:11–18, 2000.

[100] J.B. Semmens, Z.S. Wisniewski, C.D.J. Holman, A.J. Bass, and I.L. Rouse. Trends in repeat prostatectomy after surgery for benign prostate disease: application of record linkage to health care outcomes. *British Journal of Urology*, 84:972–975, 1999.

[101] Group 1 Software. DataSight, 2003.

[102] Data Quality Solutions. LinkageWiz, 2002.

[103] S. Tejada, C.A. Knoblock, and S. Minton. Learning object identification rules for information integration. *Information Systems*, 26:607–633, 2001.

[104] V. Verykios, G.V. Moustakides, and M.G. Elfeky. A Bayesian decision model for cost optimal record matching. *The VLDB Journal*, 2002.

[105] J.X. Wang and P.T. Donnan. Adjusting for missing record linkage in outcome studies. *Journal of Applied Statistics*, 29(6):873–884, August 2002.

[106] Y.R. Wang and S. Madnick. The Interdatabase Instance Identification Problem in Integrating Autonomous Systems. In *Proc. Fifth Intl. Conf. Data Eng.*, pages 46–55, 1989.
[107] W.E. Winkler. Advanced Methods for Record Linkage. In *Proceedings of the Section of Survey Research Methods, American Statistical Association*, pages 467–472, 1994.
[108] W.E. Winkler. Matching and Record Linkage. In Cox et al, editor, *Business Survey Methods*. J. Wiley & Sons Inc., 1995.
[109] W.E. Winkler. The State of Record Linkage and Current Research Problems. Technical Report RR/1999/04, Statistical Research Report Series, US Bureau of the Census, Washington DC, 1999.
[110] W.E. Winkler. Frequency-Based Matching in Fellegi-Sunter Model of Record Linkage. Technical Report RR/2000/06, Statistical Research Report Series, US Bureau of the Census, Washington DC, 2000.
[111] W.E. Winkler. Machine Learning, Information Retrieval and Record Linkage. In *Proc. of the Section on Survey Research Methods, American Statistical Association*, 2000.
[112] W.E. Winkler. Quality of Very Large Databases. Technical Report RR/2001/04, Statistical Research Report Series, US Bureau of the Census, Washington DC, 2001.
[113] W.E. Winkler. Record Linkage Software and Methods for Merging Administrative Lists. Technical Report RR/2001/03, Statistical Research Report Series, US Bureau of the Census, Washington DC, 2001.
[114] www.ascentialsoftware.com. Integrity xe, 2003.
[115] www.innovativesystems.com. I/analytics, 2003.
[116] www.trilliumsoft.com. Trillium software system, 2003.
[117] W.E. Yancey and W.E. Winkler. *Advanced Methods of Record Linkage*, October 2002.
[118] J.J. Zhu and L.H. Ungar. String Edit Analysis for Merging Databases. In *KDD 2002 Workshop on Text Mining*, 2002.

A Record Linkage Software and Systems

The following list of record linkage software and systems is incomplete. The list was influenced by the excellent web resources at ANU [21] and INRIA [41].

A.1 Record Linkage Software: Government and Academic

U.S. Bureau of the Census Software (GDRIVER) This software parses name and address strings into their individual components and presents them in a standard format. Several reference files are used to assist this process. Some of these reference files contain lists of tokens that influence the way parsing is carried out. For example, if a conjunction token is found, the software must then consider producing more than one standardized record.

A pattern reference file is utilized for both name and address standardization. These files contain a pattern of tokens that may be present in the name or address strings. The documentation claims that the software can "recognize the various elements in whatever order they occur" [117], however, that order must also be present in the pattern file.

These reference files list various known characteristics of the name fields such as replacing all associated nick names with a standard name, prefixes and suffixes used for surnames and occupations. These files can be modified to incorporate characteristics of the particular data. For address standardisation, there is also a collection of files that store key address words, their variant spelling, and the various address patterns to be recognised. For example, "Ave" is an abbreviation of "Avenue" and is a street type while "RD" could be an abbreviation of "Road" and a street type or a rural route type in the US.

The reference files capture name and address standardization knowledge from the U.S. Census Bureau. Users of this software may find that their names and addresses may have features that differ from those dealt with by the Census Bureau. In this case, the reference files would need to be edited to suit specific data sources.

Febrl - Freely Extensible Biomedical Record Linkage The current publicly released version of Febrl (0.2) provides data standardisation and probabilistic record linkage with choices of methods for blocking and comparison functions. Parallelisation is also supported. Febrl's data standardisation "primarily employs a supervised machine learning approach implemented through a novel application of hidden Markov models (HMMs)" [20].

HMMs require training. Febrl requires suitable training data to be selected and the HMMs to be learnt. In the example provided in [23], a HMM for addresses took under 20 hours to train. A separate HMM is required for name standardisation. It is argued in [23] that the development of rule sets to do the same task would take "at least several person-weeks of programming time".

The approach in Febrl was influenced by Datamold, an approach developed by Borkar et al [11] for address standardisation using nested HMMs.

GRLS Statistics Canada has developed the Generalized Record Linkage System (GRLS) and Match360. Some of the features include [35]:

- Runs on UNIX and Oracle.

– Based on Fellegi-Sunter linkage methodology.
– Has a graphical interface.
– Allows multiple concurrent users.
– Allows user-defined rules which are programmed in C.
– Linked records can be grouped into 'weak' and 'strong' groups.
– Allows refinements of weights and thresholds.
– On-line help is available.
– Has NYSIIS and Soundex rules built-in.

A.2 Record Linkage Software: Commercial

Ascential Software, Trillium and Innovative Systems, Inc. have a customer marketing or customer relationship marketing (CRM) focus. SearchSoftwareAmerica Inc. is more broadly based.

Telcordia has an in-house database reconciliation and data quality analysis tool [18].

Ascential Software [114] Ascential Software bought Vality Technology. Its product is Integrity XE, whose features include:

– Probabilistic matching technology with a full spectrum of fuzzy matching capabilities.
– Thorough data investigation and analysis processes.
– Flexibility through customizability to an organization's business rules with intuitive rules definition and interactive testing.

Trillium [116] The Trillium Software System is set of tools for inspecting, identifying, standardizing and linking data [116]. Components of the tools include:

– Parser: to parse, standardize and verify data.
– Matcher.
– GeoCoder.
– DataBrowser: a GUI for analyzing data integrity issues.

Innovative Systems, Inc. [115] Innovation Systems Inc.'s data linking product is called i/Lytics. Its features includes:

– Parsing and standardizing capabilities.
– User-defined comparison fields.
– Customizable ranking methodology.

Sagent [92] Sagent's merge/purge product is called Centrus.

Choicemaker Inc. Choicemaker [13] uses many comparison functions and combines them using maximum entropy to choose weights.

Search Software America Search Software America(SSA)'s ten person research and development team is based in Canberra. SSA has 500 customers world-wide in customer integration, criminal intelligence, tax collection, criminal investigation and marketing systems [97].

SSA's three key products are:

- SSA-NAME3[94].
- Identity Systems(IDS) [95] uses SSA-NAME3 for database data. High-performance indexes are automatically maintained without changes to existing application programs. IDS is used to 'centrally index identity information from disparate source tables, databases and computers...allowing the central index to search, match, rank, group and maintain all the data simultaneously and in real-time [95].
- Data Clustering Engine(DCE) [96]. This is a stand-alone batch data grouping and investigation engine for all forms of identification data. The DCE does the transitive closure computation linking pairs of matches that did not directly match.

SSA-NAME3 [94] is the core product, providing name matching and search capabilities. This product is differentiated by the custom population databases for almost all countries and alphabets worldwide. SSA-NAME3 supports applications that need to match data using one or more of the following data types:

- Names, addresses and descriptions.
- Identification codes.
- Dates.
- Other attributes such as phone numbers, sex, eye colour, regions etc.

.

There are a few names (family or first names) that occur very frequently in a population. For instance, Smith or Williams can account for more than 1% of the population (with $50,000$ distinct names) [94]. SSA-NAME3 treats common words, codes and tokens in a different manner to uncommon values to ensure good performance at each extreme of the names distribution.

A SSA critique of text searching (wild-card, n-gram indexing, NYSIIS and Soundex phonetic algorithms) argues that no single world stabilization algorithm is suitable for all data (even within a single country). A suitable algorithm is dependent on true distribution of the errors and variation in both the population of file data and the population of the search data [97].

LinkageWiz LinkageWiz [102] was developed within the South Australian Department of Human Services. It has been applied to a data quality review of the statewide Clinical Data Repository and Enterprise Patient Master Index.

Technical features include:

- Linkage on names or Medicare numbers.
- Probabilistic matching algorithms.
- Phonetic name matching (NYSIIS and Soundex).
- Value specfic weights for attributes i.e. 'Smith' gets less weight than 'Ittak'.
- Nick name mapping.
- Identification of default values and other potential data quality problems.
- User definable fields and weights.

LinkageWiz prices range from $1,000 (for 10,000 database record limit) upwards.

Info Route Inc. Info Route Inc. [50] has the following products:

- AddressAbility.
- NamePro.
- ClientMatch.

The company is based in Oakville, Ontario, Canada and has a Canadian data focus. The product's market focus is marketing.

Arkidata Arkidata's product is called Arkistra [3]. It is designed to efficiently integrate information from multiple systems while ensuring the quality of the that information. Arkistra is an integrated set of components that provides:

- Desktop applications for managing data loading, transformation and cleaning.

- Analysis tools for optimising data cleansing and information integration using a business rule based method.
- A processing engine that optimises the data transformation process using decision tree management.

Arkidata is based in Illinois, USA.

Group 1 Software Group 1 Software [101] has a suite of products for matching data such as DataSight and Merge/Purge Plus. Group 1's headquarters is Lanham, Maryland, USA. Its products have a CRM focus. For example, they are integrated with Siebels' CRM solution.

B Research Projects Utilising Record Linkage

The case studies here are a sample of record linkage applications in epidemiological, health services and social science research. The coverage is not complete and just includes studies with which we are familiar. The case studies give an indication of the uses of record linkage in the public health and social science research domains. A comprehensive bibliography on record linkage using administrative and survey records can be found elsewhere [59, 1].

- The California Mortality Linkage System [2].
- The west of Scotland coronary prevention study [85].
- Linkage of biotechnology databases [56].
- Hospitalisations and vital statistics [83].
- Background radon risk [61].
- β-Blocker and antidepressant use [55].
- Genetic counselling and birth defects [46].
- Midazolam use in hospital safety study [29].
- Obesity, hypertension and kidney cancer [73].
- Cellular phones and cancer cohort study [54].
- Nuclear industry family study [70].
- Neonatal readmissions [69].
- Child cancer patient fever and neutropenia admission [16].
- WA Linked Data Applications
  - Incidence and prevalence of end-stage renal failure [14, 15]
  - Trends in suicide rates of psychiatric patients [65].
  - Trends in first-time admissions for illicit drug problems [86].
  - Complications following gallbladder surgery [37].
  - Outcomes following colorectal cancer surgery [99].

- Outcomes following a ruptured abdominal aortic aneurysm [98, 84].
- Trends in repeat prostatectomy [100].
- Rosman describes the linking of Western Australian police, hospital and death records to create a road injury database [91]. This linkage was done without use of names, instead using fields such as injury type, severity and treatment.

Test Datasets Testing of record linkage methods on publicly available databases is important means of enabling comparison of results between research and development teams.

– G1 data generator [47, 9].
– The Berlin-Online-Apartment-Advertisements-Database [78].
– DBGen, a public domain database generator [34, 75].
– CORA, Computer Science bibliography [26].
– RESTAURANT, New York Restaurants extracted from travel guides [103].
– DBLP, `www.informatik.uni-trier.de/ley/db/index.html` [75].
– IMDB, Internet Movie Database, `www.imdb.com` [53]
– Die Vorfahren DB, `feefhs.org/dpl/dv/indexdv.html`, [53]