Record linkage in organisations: a review and directions for future research

Tengku Adil Tengku Izhar*

Faculty of Information Management,
Universiti Teknologi MARA,
Selangor, Malaysia
Email: tengkuadil@yahoo.co.uk
*Corresponding author

Torab Torabi

Department of Computer Science and Information Technology,
La Trobe University,
Victoria, 3086, Australia
Email: t.torabi@latrobe.edu.au

M. Ishaq Bhatti

La Trobe Business School,
La Trobe University,
Victoria, 3086, Australia
Email: i.bhatti@latrobe.edu.au

Abstract: Record linkage is a task of identifying data from large datasets across different data sources. Although record linkage approach has been applied in many areas, there is limited discussion on the literature that gives an overview on recent development that addressed record linkage in the scope of the organisational goals. This paper is classified according to the recent development on record linkage as an approach to drive the understanding of the dependencies of organisational data in relation to the organisational goals. We observed recent literature based on this classification to identify recent development on record linkage. The results show that there is no study in evaluating record linkage in the scope of organisational data that relate to the organisational goals. The contribution of this paper will serve as a first step to develop the dependency relationship between organisational data and organisational goals.

Keywords: record linkage; data goal dependency; data linkage; organisational goals; literature review.

Reference to this paper should be made as follows: Izhar, T.A.T., Torabi, T. and Bhatti, M.I. (2017) 'Record linkage in organisations: a review and directions for future research', *Int. J. Data Science*, Vol. 2, No. 4, pp.325–351.

Biographical notes: Tengku Adil Tengku Izhar is a Lecturer at the Faculty of Information Management, Universiti Teknologi MARA, Malaysia. He received his PhD from La Trobe University, Australia. His research areas include information management, information entrepreneurship, organisation modelling and ontology modelling.

Torab Torabi is a Senior Lecturer at Department of Computer Science and Computer Engineering. He is the Head of Software Engineering Interest Group organising the Pervasive Computing Interest Group in La Trobe University. He has (co)authored more than 40 journal and conference papers in Software Engineering and Mobile Computing. His research interests include software engineering, CASE tools, process modelling, software quality, XML and metadata, location based services, context-aware mobile services, integration of mobile services, model driven specification, component-based simulation.

M. Ishaq Bhatti is an Associate Professor and the Founding Director of Islamic Banking and Finance Programme at La Trobe University (LTU); the first ever in Australasian region. Previously, he has taught at Monash, Griffith, International Islamic University, University of Alberta and visited Rider, Magberg, Hitotsubahi, Auckland and Middle Eastern Universities. He is an author of more than 75 papers, three books and a member of the editorial board of various journals. His major areas of research, scholarship and teaching are in quantitative finance, Islamic finance, applied econometrics and statistics.

1 Introduction

Organisations rely on resources such as data, information and knowledge to assist the development of the business plan, business strategies and decision-making. Organisational resources such as data must be relevant to assist the decision-making process in relation to the organisational goals. Organisations should have the ability to manage their organisational resources (Omerzel and Antoncic, 2008; Schalenkamp and Smith, 2008; Smith et al., 2007). However, the growth in the amount of the organisational resources available nowadays poses major difficulties as well as challenges to decision-making (Mikroyannidis and Theodoulidis, 2010). Data is the most important asset to assist the decision-making process and to help in achieving the organisational goals. However, the trustworthiness of organisational data in relation to the organisational goals is often questionable because of the vast amount. This is because organisations create new resources such as data, information, knowledge and tools every day. Some of this data are not relevant to the organisational goals. It is difficult to identify the relevance of data and even though data analysts are trained to manage this data, the increased amount of organisational data has become a major problem in applying this data.

A large body of research has been conducted on the relationship between the organisational resources, such as information and information systems in relation to business performance, business processes and decision-making (Barjis, 2008; Elbashir et al., 2008; Harmancioglu et al., 2010; Jonsson and Lindbergh, 2010; Trkman, 2010; Xiaoying et al., 2008). To our knowledge, no study has considered the evaluation of organisational data to assist the decision-making process in relation to the organisational goals. Modelling the organisational goals has been limited to the business process and the

organisational process (Fox et al., 1996, 1998; Mansingh et al., 2009; Rao et al., 2012; Sharma and Osei-Bryson, 2008). Therefore, it is important to identify the linkage between the organisational data and organisational goals (Izhar et al., 2012).

While many studies have been carried out on data linkage in diverse areas such as data privacy, medical data, software and databases (Durham et al., 2012; Freire et al., 2012; Meray et al., 2007), limited research has been conducted on evaluating the dependency organisational data that matches to the organisational goals (Izhar et al., 2012). It is important to identify the relationship between organisational data and organisational goals to identify the relevance of the organisational data from the datasets. The first step in evaluating the relevance of organisational data is to recognise the relationship between the organisational data and the achievement of the organisational goals.

The aim of this paper is to review a number of recent developments on record linkage prior to examining the literature on organisational data that relate to the organisational goals. It is important to identify the value of analysed organisational data that match to the organisational goals so this data can be considered relevant. In contrast to the previous studies on data linkage (Durham et al., 2012; Freire et al., 2012; Meray et al., 2007), we adapt data linkage to address the dependency relationship of organisational data that relate to the organisational goals. This is because organisational data is a major resource in every organisation and it is important to identify the relevance of this data in the achievement of the organisational goals.

We define data linkage as a task to identify the relevant organisational data in the organisational datasets in relation to the organisational goals. We suggest it is important to identify the organisational data, which is relevant to the organisational goals.

The main contribution of this paper is to address the issues relating to organisational resources, respectively, taking into account both the evaluation of the organisational data and the organisational goals. In this survey, we review recent development on record linkage as an effort to advance the understanding of data dependency that match to the organisational goals.

We conducted a literature search within the scope of the organisational modelling using research sources scholarly in scientific journals. Using subject area of business information system, a number of databases have been reviewed such as ScienceDirect (Elsevier), Emerald Fulltext (Emerald) and ProQuest Business. Most of the papers reviewed in this survey are published in journals related to the information science, information system and organisational modelling. It was observed that although the papers have been reviewed since 2007 until 2012, we conclude the discussion is based on the past literatures that we reviewed and some of the past existing literatures that might not been reviewed in this survey. The papers have been classified according to record linkage in the scope of the organisation. However, whether all record linkage approaches are applicable or not to develop the data goal dependency is beyond the scope of this paper because other approaches and issues might exist and have not been identified in this survey.

The remainder of this paper is organised as follows. In Section 2, we discussed recent definition on record linkage. Section 3 discusses recent studies on record linkage activities and different approaches on record linkage in Section 4. Summary and future researches are discussed in Sections 5 and 6, respectively. Section 7 gives the conclusion.

2 Record linkage definition

There has been substantial growth in data linkage activities in recent years. Most studies use terms such as data linkage, data matching, record matching and record linkage as a definition (Abril et al., 2012; Christen, 2012; Ferrante and Boyd, 2012; Scannapieco et al., 2007; Su et al., 2010; Yakout et al., 2010). To identify the gap in the previous studies, it is important to define data linkage so that we can compare previous studies with our study. Basically, data linkage is defined as a task to find data from vast datasets from different sources.

Data linkage was first studied in 1969 by Fellegi and Sunter (1969), who defined a theory for record linkage. We use this definition to identify the gaps in the previous studies. Scannapieco et al. (2007) used the term record matching, defining this as the process of identifying whether two or more records represent the same real-world entity or not. In their study, record matching is performed across different data sources with the aim of identifying common information shared among these sources.

In 2010, Yakout et al. (2010) present a new definition for record linkage, defining it as the process of identifying records that refer to the same entity. This new approach uses entity behaviour to decide if potentially different entities are in fact the same. They also defined record linkage as the computation of the associations among records of multiple databases. It arises in contexts such as the integration of such databases, online interactions and negotiations. It is the process of identifying similar records that represent the same real-world entity. In this study, the authors state that the problem of matching records among sources that are autonomous and unwilling to share data is known as private record linkage. Meanwhile, Su et al. (2010) defined record matching as identifying the records that represent the same real-world entity as an important step for data integration. Arasu et al. (2010) defined record matching as the problem of identifying matching or duplicate records, and records that correspond to the same real-world entity.

In more recent studies, Abril et al. (2012) defined record linkage as an estimator of the disclosure risk of protected data and was initially introduced for database integration. Durham et al. (2012) defined record linkage as the task of identifying records from disparate data sources that refer to the same entity. It is an integral component of data processing in a distributed setting, where the integration of information from multiple sources can prevent duplication and enrich overall data quality, thus enabling more detailed and correct analysis.

Another definition is proposed by Ferrante and Boyd (2012) who defined data linkage in the healthcare environment, stating that it is a method that is being used increasingly in the health and human services research sector as it brings together administrative data from disparate sources and links them through various approaches, such as probabilistic, deterministic or fuzzy logic methods.

Christen (2012) defined record linkage as the process of matching records from several databases that refer to the same entities. This study reports that matched data are becoming important in many application areas because they can contain information that is not available otherwise or that is too costly to acquire.

Data linkage has been defined in diverse areas such as databases and systems as shown in Table 1. In contrast to the results shown in this table, we attempt to define data linkage as a process to identify data from datasets in an effort to identify organisational data that relate to the organisational goals. Therefore, it is important to identify previous

approaches to data linkage to identify the gaps in the existing literature on the data linkage process in the context of organisational goals.

Table 1 Summary of record linkage definition

	Scope			
Authors	*Privacy*	*De-duplication*	*Quality*	*Integration*
Scannapieco et al. (2007)	✓			
Yakout et al. (2010)			✓	
Su et al. (2010)		✓		✓
Abril et al. (2012)	✓			
Durham et al. (2012)	✓	✓		
Ferrante and Boyd (2012)			✓	
Christen (2012)		✓		

Even though the term is defined as record linkage, data linkage, record matching, data matching (Abril et al., 2012; Christen, 2012; Ferrante and Boyd, 2012; Scannapieco et al., 2007; Su et al., 2010; Yakout et al., 2010), we use the term data dependency as an effort to identify the dependency relationship between organisational data and organisational goals because we attempt to identify the dependency for every organisational data that relate to the organisational goals. In this section, we look at the gap on record linkage defined in the previous studies as shown in Table 1. As a result, we identify the dependency relationship between organisational data and organisational goals as data goals dependency based on the organisational goals ontology (Izhar et al., 2012, 2013). We define data goals dependency as a process to identify the existing organisational data from the organisational datasets in relation to the organisational goals. In the rest of this paper, we use the term data dependency unless the terms are defined or cited from past studies as record linkage, record matching, data linkage or data matching.

3 Record linkage activities

Recently, there has been substantial growth in record linkage activities (Durham et al., 2012; Freire et al., 2012; Meray et al., 2007). Most of these studies focused on the task of identifying data from datasets to prevent any redundancy. To our knowledge, no study has been carried out on the development of organisational data in relation to the organisational goals. Even though studies on organisational goals have been carried out, most focus on the modelling concept (Fox et al., 1998; Rao et al., 2012; Sharma and Osei-Bryson, 2008). Therefore, it is important to identify the relationship between organisational data and organisational goals as we suggest this data should be relevant to assist decision-making in relation to the achievement of the organisational goals (Izhar et al., 2012).

3.1 Medical record

Meray et al. (2007) focused on medical records. They applied probabilistic record linkage to combine databases without a patient identification number. The authors consider the probabilistic record linkage technique to be a valid tool to combine the patient databases. The technique allows the creation of a high-quality linked database from different datasets. In this study, the authors suggest the technique is useful for the linkage of any anonymous registries in the absence of personal identifiers and goal standards.

Another example of record linkage in medical records is Freire et al. (2012). The authors look at the record linkage process in screening patient databases because the patient information is not uniquely identified. Record linkage is presented to integrate the files in a database called Brazilian Cervical Cancer Information System (SISCOLO). The authors show that record linkage integrates SISCOLO to produce indicators for the evaluation of the cervical cancer screening program, taking the patient as a unit of observation. Record linkage assesses the effectiveness and quality of data as a way to contribute to a more efficient use of SISCOLO in the planning of health actions. Jutte et al. (2011) studied administrative record linkage as a tool for public health research. Linked administrative databases are powerful resources that provide longitudinal health and social data on large populations for the flexible and relatively low-cost investigation of pressing public health concerns.

3.2 Data privacy

Data privacy is a serious issue when implementing data linkage (Karakasidis and Verykios, 2011). The authors propose a technique to address the problem of efficient privacy preserving approximate record linkage. The technique is important to combine the speed of private blocking with the increased accuracy of approximate secure matching of data. In this study, the technique did not apply any data that related to any organisational goals; rather, it focused on data retrieval speed.

Abril et al. (2012) also studied record linkage in the context of data privacy. The authors suggest that record linkage can be used as an estimator of the disclosure risk of protected data. During the linkage process, the authors introduce a parameterisation of record linkage to improve linkage compared with standard distance-based record linkage and to identify the key attributes for record linkage. Therefore, the authors determine the weight identification for every linkage process, which expresses the importance of each variable in the linkage process. Thus, in data privacy, record linkage is used as a disclosure risk estimation of the protected data.

Another example is Durham et al. (2012). In this study, the authors aim to quantify the correctness, computational complexity and security of privacy-preserving string comparators for record linkage. In this study, privacy-preserving record linkage is a variant of the task in which data owners wish to perform linkage without revealing the identifiers associated with the records. It ensures the privacy and security of the customers' data.

Inan et al. (2008) studied data privacy and proposed a hybrid approach to private record linkage. They stated that private record linkage is the first and possibly the most important step towards the utilisation of private information. The process of identifying and linking different representations of the same real-world entity across multiple data

sources is known as the record linkage problem. Since it is a key component of data integration methodologies, record linkage has been investigated extensively.

Recently, Inan et al. (2010) studied private record matching using differential privacy. Private matching between datasets owned by distinct parties is a challenging problem with several applications. Private matching allows two parties to identify records that are close to each other according to some distance functions, such that no additional information other than the join result is disclosed to any party. Private matching can be solved securely and accurately using secure multi-party computation (SMC) techniques, but such an approach is prohibitively expensive in practice.

Yakout et al. (2010) studied efficient private record linkage. Record linkage is the computation of the associations among records of multiple databases. It arises in contexts such as the integration of such databases, online interactions and negotiations, and many others. The autonomous entities who wish to carry out the record matching computation are often reluctant to fully share their data. In such a framework where the entities are unwilling to share data with each other, the problem of carrying out the linkage computation without full data exchange has been called private record linkage.

3.3 Database and software

Holman et al. (2008) studied data linkage software and discussed design, steps to full implementation and outcomes achieved by the Western Australian Data Linkage System (WADLS). The authors conclude that the creation of a data linkage system is a challenge in a social organisation. It demands leadership, interagency and inter-sectoral cooperation, and a dedicated group of users who drive reforms with perseverance.

Cuzzocrea and Puglisi (2011) studied data linkage in a data warehouse. The authors argued that the problem of effective record linkage in a data warehouse is an issue even though many studies have been carried out on databases and software (Ferrante and Boyd, 2012; Holman et al., 2008; Su et al., 2010). A major research challenge in data warehouse research is data quality, as the quality of data stored in a data warehouse can have a significant effect and cost implications on a system that only relies on information to make decisions and predictions on business organisations and activities.

Another example of data linkage in software is Trepetin (2008). In this study, the author also looks at the aspect of privacy (Karakasidis and Verykios, 2011) and compares the existing software techniques for computing the similarity of linkage identifiers to improve linkage in a privacy-preserving fashion. An organisation using this record linkage system will be less likely to suffer a security violation but yet record linkage effectiveness will not be undermined.

Ferrante and Boyd (2012) proposed a transparent and transportable methodology for the evaluation of data linkage software. The methodology is used to evaluate data linkage software to improve the quality of the linkage. The authors evaluate a large number of packages that involve the use of synthetic data using a pre-defined linkage strategy and the use of standard linkage quality metrics to assess the performance. They suggest that this methodology provides a unique opportunity to benchmark the quality of linkage in different operational environments.

Christen (2012) looked at the indexing techniques for scalable record linkage and the de-duplication of data. The author argued that matched data are becoming important in many application areas because they can contain information that is not available otherwise, or that is too costly to acquire. Removing duplicate records in a single

database is a crucial step in the data cleaning process because duplicates can severely influence the outcomes of any subsequent data processing or data mining. In contrast, this study is important, as we attempt to prevent the inclusion of irrelevant data for data analysis in relation to the organisational goals. Previously, Christen (2008) studied an automatic record linkage using seeded nearest neighbour and support vector machine classification that aimed at automating the record linkage process.

Varghese and Sundar (2011) investigated how to improve performance in classification using data matching. They discussed that record matching cannot solve the duplication detection problem as this is an important process in data integration. In this study, the authors proposed an efficient approach to detect duplicates in a web database scenario. They presented a fast duplication detection (FDD) method that can effectively identify duplication in multiple web databases. The results in this study focused on data integration from datasets, as duplicate detection is an important step in data integration and most state-of-the-art methods are based on offline learning techniques. The general web database scenario is that the records to match are greatly query-dependent; a pre-trained approach is not applicable as the set of records in each query's results is a biased subset of the full dataset.

Su et al. (2010) discussed that record matching, which identifies the records that represent the same real-world entity, is an important step for data integration. The authors looked at this issue in relation to multiple web databases and proposed a matching method called unsupervised duplicate detection (UDD). This method works well for a web database scenario where existing supervised methods do not apply.

Adly (2009) looked at efficient record linkage using a double embedding scheme. This paper introduced a novel scheme for record linkage based on double embedding the data, aiming at improving efficiency. A two-level matching is proposed, with the first level performing fast and inaccurate matching, ensuring high recall while the second level performed a more expensive matching on a smaller set of pairs to improve accuracy. Record linkage is the problem of identifying similar records across different data sources. The similarity between two records is defined based on domain-specific similarity functions over several attributes.

Arasu et al. (2008) studied a transformation-based framework for record matching. Today's record matching infrastructure does not provide a flexible way to account for synonyms. The authors proposed a programmatic framework of record matching that takes user-defined string transformations as input. To the best of the authors' knowledge, this is the first proposal for such a framework.

Kan and Tan (2008) studied record matching in digital library metadata using evidence from external sources to create more accurately matching systems. In its most basic form, record matching can be simplified as string matching, which decides whether a pair of observed strings refers to the same underlying item. String similarity measures are usually weighted differently per column. Certain data types have been studied in depth. In fact, the need to consolidate records of names and addresses in government and industry pioneered research to find reliable rules and weights for record matching.

Elmagarmid et al. (2007) studied the comparison of past metrics used and probabilistic matching data. In this paper, the authors present a thorough analysis of the literature on duplicate record detection. They cover similarity metrics that are commonly used to detect similar field entries, and present an extensive set of duplicate detection algorithms that can detect approximately duplicate records in a database. They also cover multiple techniques for improving the efficiency and scalability of approximate duplicate

detection algorithms. They conclude with a coverage of existing tools and with a brief discussion of the big open problems in the area. However, it is currently unclear which metrics and techniques are the current state-of-the-art. The lack of standardised, large-scale benchmarking datasets can be a significant obstacle for the further development of the field as it is almost impossible to convincingly compare new techniques with existing ones.

In summary, we can see that most of these studies focus on databases in an effort to identify the link between data from databases. To our knowledge, no study has focused on organisational data that relate to the organisational goals as shown in Table 2. Therefore, it is important to identify the gaps in previous data approaches and data linkage to develop an approach for organisational data that relate to the organisational goals.

Table 2 Record linkage activities

	Scope			
Authors	*Privacy*	*Integration*	*Software/Database*	*Metrics/Weight*
Meray et al. (2007)				✓
Elmagarmid et al. (2007)				✓
Holman et al. (2008)			✓	
Trepetin (2008)	✓			
Christen (2008)				✓
Inan et al. (2008)	✓			
Kan and Tan (2008)			✓	
Arasu et al. (2008)				✓
Yakout et al. (2010)	✓			
Adly (2009)			✓	
Su et al. (2010)		✓		✓
Inan et al. (2010)	✓			
Yakout et al. (2010)	✓			
Arasu et al. (2010)			✓	
Jutte et al. (2011)			✓	
Varghese and Sundar (2011)		✓		✓
Cuzzocrea and Puglisi (2011)			✓	
Karakasidis and Verykios (2011)	✓			
Freire et al. (2012)				✓
Abril et al. (2012)	✓			✓
Ferrante and Boyd (2012)				✓
Durham et al. (2012)	✓			✓
Christen (2012)				✓

As indicated in Table 2, there are a few studies that have been carried out on metrics and weight. However, the studies look at the weight of the linkage, and based on our observations, no study has focused on metrics development that link organisational data to the organisational goals.

4 Different approaches on record linkage

In our research, the dependency relationship between organisational data and organisational goals can specify to what extent the organisational goals are achieved by evaluating organisational data that relate to the organisational goals. Therefore, it is important to identify different approaches on record linkage to evaluate this organisational data.

4.1 Metrics

Durham et al. (2012) evaluate the string comparator of data linkage using metrics. The authors evaluate each string comparator based on correctness in record linkage, computational complexity and security. Elmagarmid et al. (2007) developed a field-matching technique as a duplicate record detection approach. It is one of the common sources of mismatches in databases. The techniques involve edit distance, affine gap distance, Smith-Waterman distance, the Jaro distance metric and Q-gram distance. The authors looked at token-based similarity metrics, as this metrics works well for typographical errors. The authors also looked at phonetic similarity and numeric similarity metrics.

4.1.1 Weight

Ferrante and Boyd (2012) developed a methodology for data linkage software by creating and using synthetic datasets. However, the authors did not define a specific weight to be used in the linkage as there are considerable variations in the implementation of weighting by various software packages. Abril et al. (2012) determined the optimum weight for the linkage process for their metrics. The authors introduced a parameterisation of record linkage in terms of a weighted mean and its weight. Arasu et al. (2010) looked at a linear classifier to observe record matching packages. The authors developed an algorithm for a linear classifier and the conjunction of similarity thresholds to look at record matching for active learning. Active learning is important for record matching since manually identifying a suitable set of labelled examples is difficult.

Su et al. (2010) looked at the weight component for the similarity classifier for record matching. The authors assign a weight to a component to indicate the importance of its corresponding field under the condition that the sum of all component weights is equal to 1.

Varghese and Sundar (2011) proposed a method for record matching to solve duplication detection. The authors identify weight as a dynamic allocation to different fields of record. Christen (2008) identified the matching weight that was usually normalised such that 1.0 corresponds to exact similarity and 0.0 to total dissimilarity with

attribute values that are somewhat similar, having a matching weight somewhere in between 0 and 1.

Arasu et al. (2008) looked at string similarity as typically captured via a similarity function that given a pair of strings returns a number between 0 and 1 as a higher value, indicating a greater degree of similarity with the value 1 corresponding to equality.

4.2 Probabilistic record linkage

Abril et al. (2012) discussed distance-based record linkage by looking at probabilistic record linkage. In this study, the matching algorithm uses the linear sum assignment model to choose which pairs of the original and protected record must be matched. The approach is applied to evaluate the disclosure risk of protected data.

Meray et al. (2007) looked at probabilistic record linkage as a tool to combine databases in medical records. The authors defined the linkage weight for every linkage variable for all the possible record pairs. The linkage weight is assigned to each linkage variable and is summed over all variables, where the sum reflects the probability that these two records belong to the same unit (e.g., high weight-high probability; low weight-low probability). Elmagarmid et al. (2007) discussed probabilistic matching models based on Bayes decision rule for minimum error.

In Table 3, we can see that the scope of the previous research focused on metrics, weight and probabilistic development for data linkage. In contrast, we develop a metrics to evaluate organisational data that match the organisational goals (Izhar et al., 2012, 2013). In our approach, we consider the value of weight for this dependency relationship is important to consider organisational data is relevant to the organisational goals. In this paper, we looked at data linkage as an approach to develop the dependency relationship between organisational data and organisational goals.

Table 3 Record linkage approaches

	Scope	
Authors	*Metrics/ Weight*	*Probabilistic*
Elmagarmid et al. (2007)	✓	✓
Meray et al. (2007)		✓
Christen (2008)	✓	
Arasu et al. (2008)	✓	
Arasu et al. (2010)	✓	
Su et al. (2010)	✓	
Varghese and Sundar (2011)	✓	
Abril et al. (2012)	✓	
Ferrante and Boyd (2012)	✓	
Durham et al. (2012)	✓	

5 Summary

Although many studies have been carried out in the context of data processes, such as data mining (Kum et al., 2009; Liao et al., 2008), a limited number of studies have been conducted on evaluating organisational data in relation to the organisational goals (Izhar et al., 2012). Therefore, it is important to identify the relationship between organisational data and organisational goals as this is important in identifying the relevance of organisational data in the organisational datasets.

The data linkage approach is adapted to identify the possible dependency relationship between organisational data and organisational goals. Data linkage is commonly used to identify data that are linked, so all datasets under consideration should ideally undergo a matching process prior to data linkage (Durham et al., 2012). Studies have been carried out in relation to various issues such as software (Freire et al., 2012), privacy (Abril et al., 2012; Karakasidis and Verykios, 2011) and security (Durham et al., 2012).

In contrast to these studies, we suggest it is important to develop a standard set of models to show the relationship between data, attributes and organisational goals. Therefore, data goals dependency is presented based on the organisational goals ontology. Data goals linkage is defined as a task to identify the existing organisational data and attributes from the organisational datasets in relation to the organisational goals. We suggest it is important to identify the dependency relationship between organisational data that relate to the organisational goals.

Data linkage for the organisational goals ontology aims to show the dependency relationship between organisational data that relate to the organisational goals. In the future, we suggest data linkage is an important approach to identify the linkage between organisational data and organisational goals.

We extend the definition of data linkage in the context of organisational goals as a task of identifying relevant organisational data from datasets that relate to the organisational goals. Therefore, we apply data linkage elements in our model. We understand that organisations store their data as datasets. This data can be redundant and, therefore, not relevant. Thus, it is important to pre-process this data to identify the existing data and attributes.

In contrast to previous studies, we suggest the process of identifying data and attributes is important because:

- if we identify data and attributes that relate to the same organisational goals, we do not need to apply the same data and attributes again
- it is necessary to evaluate this data and the attributes that relate to the organisational goals.

6 Future research

In the future, we suggest data linkage can assist the process of data evaluation in relation to the organisational goals. The evaluation measures the weight of organisational data dependency to the organisational goals. Future research will implement on how the scope of data linkage in this paper will be implemented to identify the dependency relationship between organisational data and organisational goals. At the same time, future work will

cover the extension development of the organisational goals ontology by looking at the dependency relationship of organisational data that match to the organisational goals (Izhar et al., 2012, 2013).

6.1 Data as an important organisation resource to the organisational goals

Organisational goal is a main target for every organisation. To perform the organisational daily business activity in relation to the organisational goals, it relies on the organisational data. In organisations, data is important as a strategic asset that can be leveraged into a competitive advantage. This is because many organisations recognise the value and the need of the data within the organisation as an effort to assist decision-making (Karim and Hussein, 2008). However, such data is too large though not all of this data is relevant to the organisational goals. The difficulty in identifying the relevance of organisational data becomes an issue. Even though information system and technology have been recognised to manage such data, the relevance of organisation data in relation to the organisational goals is always questionable.

Data provide detailed information about specific needs while service or system such as databases executes processes involving data and returning an informative result. In the knowledge society as today, user needs both data and system to systematically provide a value to all users. To make decision, organisations depend on the amount of data to make a fact-based decision. Organisations use this data for analysis. This is noted that, even though we live in organisational business information but very little process has been work on data.

In organisations, the usage of data is very important for manager to share. It is important for the manager to receive the relevance data in relation to the organisational goals. Recently, Simsek et al. (2009) discussed that sharing of importance of data and information can develop knowledge to successful decision-making. It is very crucial for the organisation to create and generate new data and evaluate it for a better decision-making. Differences in generating new idea, information and knowledge will help in term of the decision-making and will enable such team within the organisation to use relevance of data for the organisational goals success. Data is presented in many forms such as documents and statistics and the most important resource in relation to the organisational goals. In this research, we define this data as organisational data and we use the term organisational data in most sections of this paper. It is important to further the discussion based on the relationship between organisational data and organisational goals and define this relationship to identify the relevance of organisational data in relation to the organisational goals.

6.2 Dependency relationship between the organisational data and organisational goals

In this paper, we attempt to identify the dependency relationship of organisational data in relation to the organisational goals. Even though many studies have been carried out in the context of the data process (Kum et al., 2009; Liao et al., 2008), limited study has been observed in evaluating organisational data in relation to the organisational goals (Izhar et al., 2012, 2013). Therefore, it is important to identify the dependency relationship between organisational data and organisational goals. The dependency relationship is important to identify the relevance of organisational data from datasets that

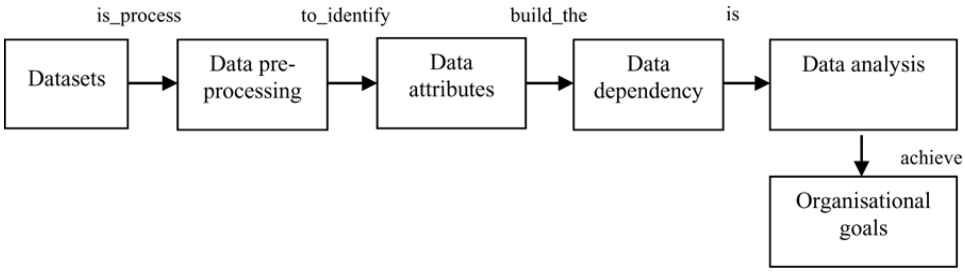
match to the organisational goals. However, organisations have a huge set of the organisational data that might be relevant to the organisational goals. This large set of the organisational data might not be relevant with respect to the organisational goals. The first step to identify the relevance of organisational data is to recognise the matching set of organisational data for the achievement of the organisational goals.

Record linkage is highly used to identify data that is being linked, so all datasets under consideration should ideally undergo a matching process prior to the data linkage (Durham et al., 2012). Even though there are studies carried out in various issues such as software (Freire et al., 2012; Jutte et al., 2011), privacy (Abril et al., 2012; Karakasidis and Verykios, 2011) and security (Durham et al., 2012), we suggest it is important to develop a standard set of approach to show the dependency relationship of organisational data, attributes and organisational goals.

6.3 Data dependency to the organisational goals

In contrast to Christen (2012), we elaborate the process of data linkage in the context of the organisational goals as shown in Figure 1. In Figure 1, we show the elements of data linkage (Christen, 2012). We consider these elements as terms used in previous studies such as datasets (Durham et al., 2012; Su et al., 2010), data pre-processing (Trepetin, 2008), data linkage (Christen, 2012) and data matching (Christen, 2012; Su et al., 2010). However, previous discussion covered the linkage for domain data such as systems and databases (Ferrante and Boyd, 2012; Trepetin, 2008). In this paper, we observe the dependency relationship between organisational data that match to the organisational goals. At the same time, we observed that there is no discussion that covers the element of organisational data and attributes. Past studies observed data linkage as a task of identifying data corresponding to the same entity from one or more data sources (Christen, 2012). Past studies only evaluate data that match to the same entities such as databases. In respond to this gap, we observe the existing organisational data and attributes from organisational datasets to evaluate the value of every organisational data and attributes that match to the organisational goals. Thus, we consider this organisational data is relevant to the organisational goals.

Figure 1 Data goal dependency



Source: Adapted from record linkage in Christen (2012)

Previous research investigated data linkage using terms such as datasets (Durham et al., 2012; Su et al., 2010), data pre-processing (Trepetin, 2008), data linkage (Christen, 2012) and data matching (Christen, 2012; Su et al., 2010). However, previous discussion covered the linkage for systems and databases (Ferrante and Boyd, 2012; Trepetin, 2008).

In contrast to previous studies (Fox et al., 1998; Rao et al., 2012; Sharma and Osei-Bryson, 2008), we included the dependency relationship of organisational data that relate to the organisational goals. However, in Figure 1, we added data dependency instead of data linkage and data matching (Christen, 2012). This is because we want to look at the dependency relationship of organisational data that relate to the organisational goals.

We also observe that there is no discussion in the existing research on organisational data and attributes in relation to the organisational goals. Past studies noted data linkage as a task to identify data corresponding to the same entity from one or more data sources (Christen, 2012). Therefore, past studies only evaluate data that match to the same entities such as databases.

In response to this gap in the existing research, we investigated the existing organisational data and attributes from organisational datasets to evaluate the value of the dependency organisational data that relate to the organisational goals. Thus, we consider these organisational data are relevant to the organisational goals.

- *Datasets*

Datasets are a collection of data that have been stored in different datasets to represent the different types of data. These datasets are important as a reference for any decision-making evaluation. However, this data can be very large and it is a challenge as to how to identify relevant data from the huge number of datasets. Therefore, it is important to discuss the creation of datasets to perform an evaluation of data from the databases (Durham et al., 2012; Su et al., 2010).

In contrast to these studies (Durham et al., 2012; Su et al., 2010), we attempt to perform the evaluation of organisational data in relation to the organisational goals. The process includes how organisational data is selected from organisational datasets using the framework and we evaluate the organisational data that relate to the organisational goals.

- *Data pre-processing*

Organisations create a vast amount of data every day but some of these data are not relevant for decision-making (Izhar et al., 2012). Christen (2012) defined data pre-processing as a process to identify quality data. In contrast to the previous study (2012), we suggest data pre-processing is important to eliminate any redundant organisational data in an effort to identify relevant organisational data in relation to the organisational goals. It is also important to identify matching data from various datasets (Trepetin, 2008).

- *Data and attributes*

In this paper, the process of identifying data and attributes from datasets is important in an effort to identify relevant organisational data. In contrast to previous studies, several studies (Ferrante and Boyd, 2012; Trepetin, 2008) discussed the linkage of domain data such as databases but no observation has been carried out in the context of organisational data in relation to organisational goals. In the future, the identification of data and attributes is important to develop the dependency relationship between organisational data and attributes that relate to the organisational goals.

- *Data dependency*

Data linkage is the task of quickly and accurately identifying records corresponding to the same entity from one or more data sources (Christen, 2012). As discussed previously, most studies on data linkage cover the linkage of domain data such as databases. Data matching is an important step in identifying data from datasets (Christen, 2012; Su et al., 2010). The approach suggests data matching is the process of bringing together data from different sources and comparing it.

In contrast to previous studies (Christen, 2012; Su et al., 2010), we refer to data linkage and data matching as data dependency to search for organisational data in the datasets that refer to the same organisational goals from different data sources. However, the term data dependency is usually defined in the context of computer science to analyse data to discover new information and knowledge.

In this paper, we use the term data dependency to show which organisational data relate to the organisational goals. We suggest data dependency is important to identify similar data from large datasets to avoid any irrelevant organisational data during the decision-making process in relation to the organisational goals. It involves analysing the dependency organisational data that relate to the organisational goals.

- *Analysis*

The main objective of data analysis is to evaluate organisational data from the vast amount of organisational datasets in relation to the organisational goals. We suggest that data analysis is important to identify the value of the organisational data that are relevant to the organisational goals.

The increased amount of other organisational resources, such as information, knowledge and tools, also makes it difficult for the decision-maker to identify the most relevant organisational data, as data might not be relevant to the organisational goals. Therefore, it is important to apply an analysis approach that can identify the organisational data that is relevant to the organisational goals.

Metrics is proposed as a measurement tool to analyse the dependencies of organisational data. In the literature, metrics is proposed to evaluate organisation resources such as information and knowledge for business process (Rao et al., 2012). In the future, metrics will be defined based on the dependency relationship of the organisational goals ontology (Izhar et al., 2012, 2013). The aim of the analysis approach discussed in the paper is to identify the value of organisational data that relate to the organisational goals and how the metrics is defined based on the dependency relationship between organisational data and organisational goals.

- *Organisational goals*

Organisational goals are the higher and important achievement target in every organisation and it consists of the process of identifying the aim of the organisation. In Figure 1, we show the dependency relationship between organisational data and organisational goals. The process in Figure 1 shows how data will be selected from datasets to evaluate this data in relation to the organisational goals. In this paper, we define a metrics as a measurement approach to evaluate this data. Therefore, this organisational data can be considered relevant for domain experts and entrepreneurs to assist their decision-making process in relation to the organisational goals.

6.4 Metrics to measure the level of the organisational goals achievement

Metrics is defined to evaluate the extent to which the organisational goals have been achieved by measuring the dependency organisational data that relate to the organisational goals. It is important to identify the value of the organisational data that relate to the organisational goals to support decision-making. In the metrics, the factors may include:

- frequency
- percentage
- rank.

The weight to analyse the dependencies of organisational data can be defined in many ways, such as percentage and frequency based on different situations. For example, domain experts and entrepreneurs might want to identify the percentage of data that relate to organisational goals. After we identify this value, it can be presented on the dashboard to show a graphical presentation of value. The comparison of this value can be presented to support the decision-making process in relation to the organisational goals.

Varghese and Sundar (2011) evaluated the weight of the matching data for metrics evaluation that used various types of databases (Herzog et al., 2007). For example, we assumed two organisational data, data a and data b, from two organisational datasets, dataset A and dataset B, relate to the same organisational goals, $a \in A$ and $b \in B$.

For example, if both organisational data a and b relate to the same organisational goals, then $M = \{(a, b): a \in A, b \in B, (a, b)\}$ is matching organisational data between data a and b to the same organisational goals.

Even though values, such as percentage and frequency, are discussed in this section, the aim of this paper is to show future direction on data linkage to develop a framework to represent systematic steps for domain experts and entrepreneurs to follow in an effort to identify relevant organisational data in relation to the organisational goals. Therefore, how domain experts and entrepreneurs want to define their own metrics is not the main objective of the research. This is because they might want to evaluate the organisational data and define the metrics in different ways.

7 Conclusion

The aim of this paper is to review a number of the recent developments in data linkage prior to the organisational data that relate to the organisational goals. In contrast to the previous studies on data linkage (Durham et al., 2012; Freire et al., 2012; Meray et al., 2007), we propose data linkage that addresses the relationship between the organisational data that relate to the organisational goals. Data is an important and valuable resource that supports managerial decision-making in daily business activities in relation to the organisational goals.

In organisations, the amount of data continues to grow and information technology has also grown beyond storage, transmission and processing (Seng and Chen, 2010). This amount of data has become vast. Therefore, it is difficult to identify the relevance of data in an effort to assist the decision-making process in relation to the achievement of the organisational goals.

In the first part of this paper, we discussed previous definitions of data linkage. We attempted to compare previous definitions and identify the gap in the existing research in the context of organisational goals. In this section, we summarised the recent studies that have been conducted on organisational goals.

In the second part of this paper, we presented a summary of the recent approaches to data linkage. We also discussed the limitations of these approaches and then discussed the findings resulting from these limitations. The results show that no studies have been conducted on the linkage between organisational data and organisational goals because most studies on data linkage focus on various issues, such as software (Freire et al., 2012; Jutte et al., 2011), privacy (Abril et al., 2012; Karakasidis and Verykios, 2011) and security (Durham et al., 2012). We suggest data linkage is important to identify the relevance of the organisational data that relate to the organisational goals. The contribution of this paper will serve as a first step in enhancing the understanding of the approach for the evaluation of dependency organisational data in relation to the organisational goals.

References

Abril, D., Navarro-Arribas, G. and Torra, V. (2012) 'Improving record linkage with supervised learning for disclosure risk assessment', *Information Fusion*, Vol. 13, No. 4, pp.274–284.

Adly, N. (2009) 'Efficient record linkage using a double embedding scheme', *International Conference on Data Mining*, 13–16 July, Nevada, USA, pp.274–281.

Arasu, A., Chaudhuri, S. and Kaushik, R. (2008) 'Transformation-based framework for record matching', *IEEE 24th International Conference*, 7–12 April, Cancun, Mexico, pp.40–49.

Arasu, A., Gotz, M. and Kaushik, R. (2010) *On Active Learning of Record Matching Packages*, translated by Indiana, USA, pp.783–794.

Barjis, J. (2008) 'The importance of business process modeling in software systems design', *Science of Computer Programming*, Vol. 71, No. 1, pp.73–87.

Christen, P. (2008) *Automatic Record Linkage using Seeded Neareast Neighbour and Support Vector Machine Classification*, translated by Las Vegas, Nevada, USA, pp.151–159.

Christen, P. (2012) 'A survey of indexing techniques for scalable record linkage and deduplication', *IEEE Transaction on Knowledge and Data Engineering*, Vol. 24, No. 9, pp.1537–1555.

Cuzzocrea, A. and Puglisi, L. (2011) 'Record linkage in data warehousing: state-of-art analysis and research perspectives', *22th International Workshop on Database and Expert Systems Applications*, 29 August–2 September, Toulouse, France, pp.121–125.

Durham, E., Xue, Y., Kantarcioglu, M. and Malin, B. (2012) 'Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage', *Information Fusion*, Vol. 13, No. 4, pp.245–259.

Elbashir, M.Z., Collier, P.A. and Davern, M.J. (2008) 'Measuring the effects of business intelligence systems: the relationship between business process and organizational performance', *International Journal of Accounting Information Systems*, Vol. 9, No. 3, pp.135–153.

Elmagarmid, A.K., Ipeirotis, P.G. and Verykios, V.S. (2007) 'Duplicate record detection: a survey', *IEEE Transaction on Knowledge and Data Engineering*, Vol. 19, No. 1, pp.1–16.

Fellegi, I.P. and Sunter, A.B. (1969) 'A theory for record linkage', *Journal of American Statistical Association*, Vol. 64, No. 328, pp.1183–1210.

Ferrante, A. and Boyd, K. (2012) 'A transparent and transportable methodology for evaluating Data Linkage software', *Journal of Biomedical Informatics*, Vol. 45, No. 1, pp.165–172.

Fox, M.S., Barbuceanu, M. and Gruninger, M. (1996) 'An organisation ontology for enterprise modeling: preliminary concepts for linking structure and behaviour', *Computers in Industry*, Vol. 29, Nos. 1–2, pp.123–134.

Fox, M.S., Barbuceanu, M., Gruninger, M. and Lin, J. (1998) *An Organization Ontology for Enterprise Modelling*, Simulation Organizations: Computational Models of Institutions and Groups, AAAI/MIT Press, Ontario, Canada, pp.131–152.

Freire, S.M., Almeida, R.T.d., Cabral, M.D.B., Bastos, E.d.A., Souza, R.C. and Silva, M.G.P.d. (2012) 'A record linkage process of a cervical cancer screening database', *Computer Method and Program in Biomedecine*, Vol. 108, No. 1, pp.90–101.

Harmancioglu, N., Grinstein, A. and Goldman, A. (2010) 'Innovation and performance outcomes of market information collection efforts: the role of top management team involvement', *International Journal of Research in Marketing*, Vol. 27, No. 1, pp.33–43.

Herzog, T.N., Sheuren, F.J. and Winkler, W.E. (2007) *Data Quality and Record Linkage Technique*, Spinger, Washington, USA.

Holman, C.D.A.J., Bass, A.J., Rosman, D.L., Smith, M.B., Semmens, J.B., Glasson, E.J., Brook, E.L., Trutwein, B., Rouse, I.L., Watson, C.R., Klerk, N.H.d. and Stanley, F.J. (2008) 'A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system', *Australia Health Review*, Vol. 32, No. 4, pp.766–777.

Inan, A., Kantarcioglu, M., Bertino, E. and Scannapieco, M. (2008) *A Hybrid Approach to Private Record Linkage*, translated by Cancun, Mexico, pp.496–505.

Inan, A., Kantarcioglu, M., Ghinita, G. and Bertino, E. (2010) *Private Record Matching using Differential Privacy*, translated by Lausanne, Switzerland, pp.123–134.

Izhar, T.A.T., Torabi, T., Bhatti, I. and Liu, F. (2012) *Analytical Dependency Between Organisational Goals and Actions: Modelling Concept*, translated by Chengdu, IACSIT Press, China.

Izhar, T.A.T., Torabi, T., Bhatti, M.I. and Liu, F. (2013) 'Recent developments in the organization goals conformance using ontology', *Expert Systems with Applications*, Vol. 40, No. 10, pp.4252–4267.

Jonsson, S. and Lindbergh, J. (2010) 'The impact of institutional impediments and information and knowledge exchange on SMEs' investments in international business relationships', *International Business Review*, Vol. 19, No. 6, pp.548–561.

Jutte, D.P., Roos, L.L. and Brownell, M.D. (2011) 'Administrative record linkage as a tool for public health research', *Annual Review Public Health*, Vol. 32, pp.91–108.

Kan, M-Y. and Tan, Y.F. (2008) 'Record matching in digital library metadata', *Communications of the ACM*, Vol. 51, No. 2, pp.91–94.

Karakasidis, A. and Verykios, V.S. (2011) 'Secure blocking+secure matching= secure record linkage', *Journal of Computing Science and Engineering*, Vol. 5, No. 3, pp.223–235.

Karim, N.S.A. and Hussein, R. (2008) 'Manager's perception of information management and role of information and knowledge manager: the Malaysian perspectives', *International Journal of Information Management*, Vol. 28, pp.114–127.

Kum, H-C., Duncan, D.F. and Stewart, C.J. (2009) 'Supporting self-evaluation in local government via knowledge discovery and data mining', *Government Information Quarterly*, Vol. 26, No. 2, pp.295–304.

Liao, S-H., Chang, W-J. and Lee, C-C. (2008) 'Mining marketing maps for business alliances', *Expert Systems with Applications*, Vol. 35, No. 3, pp.1338–1350.

Mansingh, G., Osei-Bryson, K-M. and Reichgelt, H. (2009) 'Building ontology-based knowledge maps to assist knowledge process outsourcing decisions', *Knowledge Management Research and Practice*, Vol. 7, pp.37–51.

Meray, N., Reitsma, J.B., Ravelli, A.C.J. and Bonsel, G.J. (2007) 'Probabilistic record linkage is a valid and transparent tool to combine databased without a patient identification number', *Journal of Clinical Epideminology*, Vol. 60, No. 9, pp.883–891.

Mikroyannidis, A. and Theodoulidis, B. (2010) 'Ontology management and evolution for businee intelligence', *International Journal of Information Management*, Vol. 30, No. 6, pp.559–566.

Omerzel, D.G. and Antoncic, B. (2008) 'Critical entrepreneurs knowledge dimensions for the SME performance', *Industrial Management and Data System*, Vol. 8, No. 9, pp.1182–1199.

Rao, L., Mansingh, G. and Osei-Bryson, K-M. (2012) 'Building ontology based knowledge maps to assist business process re-engineering', *Decision Support Systems*, Vol. 52, No. 3, pp.577–589.

Scannapieco, M., Figotin, I., Bertino, E. and Elmagarmid, A. (2007) *Privacy Preserving Schema and Data Matching*, translated by Beijing, China, pp.653–664.

Schalenkamp, K. and Smith, W.L. (2008) 'Entrepreneurial skills assessment: the perspective of SBDC directors', *International Journal of Management and Entreprise Development*, Vol. 5, No. 1, pp.18–29.

Seng, J-L. and Chen, T.C. (2010) 'An analytic approach to select data mining for business decision', *Expert Systems with Applications*, Vol. 37, No. 12, pp.8042–8057.

Sharma, S. and Osei-Bryson, K-M. (2008) *Organization-Ontology based Framework for Implementing the Business Understanding Phase of Data Mining Projects*, translated by Hawaii, p.27.

Simsek, Z., Lubatkin, M.H., Veiga, J.F. and Dino, R.N. (2009) 'The role of an entrepreneurially alert information system in promoting corporate entrepreneurship', *Journal of Business Research*, Vol. 62, No. 8, pp.810–817.

Smith, W.L., Schalenkamp, K. and Eicholz, D.E. (2007) 'Entrepreneurial skills assessment: an exploratory study', *International Journal of Management*, Vol. 4, No. 2, pp.179–201.

Su, W., Wang, J. and Lochovsky, F.H. (2010) 'Record matching over query results from multiple web databases', *IEEE Transaction on Knowledge and Data Engineering*, Vol. 22, No. 4, pp.578–589.

Trepetin, S. (2008) 'Privacy-preserving string comparisons in record linkage systems: a review', *Information Security Journal: A Global Perspective*, Vol. 17, Nos. 5–6, pp.253–266.

Trkman, P. (2010) 'The critical success factors of business process management', *International Journal of Information Management*, Vol. 30, No. 2, pp.125–134.

Varghese, C.E. and Sundar, G.N. (2011) 'Record matching: Improving performance in classification', *International Journal on Computer Science and Engineering*, Vol. 3, No.3, pp.1207–1212.

Xiaoying, D., Qianqian, L. and Dezhi, Y. (2008) 'Business performance, business strategy, and information system strategic alignment: an empirical study on Chinese firms', *Tsinghua Science and Technology*, Vol. 13, No. 3, pp.348–354.

Yakout, M., Atallah, M.J. and Elmagarmid, A. (2009) 'Efficient private record linkage', *IEEE 25th International Conference on Data Engineering*, Shanghai, China, 29 March–2 April, 2009, pp.1283–1286.

Yakout, M., Elmagarmid, A.K., Elmeleegy, H. and Ouzzanil, M. (2010) *Behavior based Record Linkage*, translated by Singapore, pp.439–448.

Appendix

Table A1 Existing literature review on record linkage (2007–2012)

Authors	*Objective and issues*	*Summary*
Meray et al. (2007)	To describe the technical approach and subsequent validation of the probabilistic linkage of the three anonymous, population based Dutch perinatal registries 1 LVR1 of midwives 2 LVR2 of obstetricians 3 LNR of paediatricians/ neonatologists)	A combination of probabilistic and deterministic record linkage techniques were applied using information about the mother, delivery, and child(ren) to link three known registries. Rewards for agreement and penalties for disagreement between corresponding variables were calculated based on the observed patterns of agreement and disagreements using maximum likelihood estimation Special measures were developed to overcome linking difficulties in twins. A subsample of linked and non-linked pairs was validated
Elmagarmid et al. (2007)	Comparison of past metrics used and probabilistic matching data Duplicate records do not share a common key and/or they contain errors that make duplicate matching a difficult task. Errors are introduced as the result of transcription errors, incomplete information, lack of standard formats, or any combination of these factors	In this paper, authors present a thorough analysis of the literature on duplicate record detection. They cover similarity metrics that are commonly used to detect similar field entries, and we present an extensive set of duplicate detection algorithms that can detect approximately duplicate records in a database. They also cover multiple techniques for improving the efficiency and scalability of approximate duplicate detection algorithms. They conclude with coverage of existing tools and with a brief discussion of the big open problems in the area
Scannapieco et al. (2007)	Studied on privacy preserving schema and data matching In many business scenarios, record matching is performed across different data sources with the aim of identifying common information shared among these sources. However such need is often in contrast with privacy requirements concerning the data stored by the sources	Authors propose a privacy-preserving protocol to perform record matching across two data sources, which is more efficient than protocols based on cryptographic techniques such as privacy-preserving set intersection. The protocol has an interesting new feature that concerns the privacy preservation of database schemas. Whereas the problem of privacy protection of data has been investigated, much less attention has been devoted to schema level information
Holman et al. (2008)	The paper describes the strategic design, steps to full implementation and outcomes achieved by the Western Australian Data Linkage System (WADLS)	The paper instigated in 1995 to link up to 40 years of data from over 30 collections for an historical population of 3.7 million. Staged development has seen its expansion, initially from a linkage key to local health datasets, to encompass links to national and local health and welfare datasets, genealogical links and spatial references for mapping applications

Appendix

Table A1 Existing literature review on record linkage (2007–2012) (continued)

Authors	*Objective and issues*	*Summary*
Trepetin (2008)	Comparing record linkage system the security of data	This paper reviews existing proposals for how such anonymised string comparisons might be accomplished, but demonstrates that existing methods have various operational deficiencies. It therefore argues that new, more capable methods are needed
	The organisation using the record linkage system would be less likely to suffer a security breach, yet record linkage effectiveness would not be undermined. This paper reviews the existing software techniques for computing similarity of linkage identifiers to improve linkage in a privacy-preserving fashion	Current software techniques, however, fall short in performing secure character-level analysis
Christen (2008)	Automatic record linkage using seeded nearest neighbour and support vector machine classification	In this paper, two variations of this approach are presented. The first is based on a nearest neighbour classifier, while the second improves a SVM classifier by iteratively adding more examples into the training sets. Experimental results show that this two-step approach can achieve better classification results than other unsupervised approaches
	This paper presented a novel unsupervised two-step approach to record pair classification that is aimed at automating the record linkage process	In this paper, matching weight is usually normalised, such that 1.0 corresponds to exact similarity and 0.0 to total dissimilarity, with attribute values that are somewhat similar having a matching weight somewhere in between 0 and 1
Arasu et al. (2010)	Transformation-based framework for record matching	Authors propose a programmatic framework of record matching that takes such user-defined string transformations as input. To the best of authors knowledge, this is the first proposal for such a framework. This transformational framework, while expressive, poses significant computational challenges which we address
	Today's record matching infrastructure does not allow a flexible way to account for synonyms such as ‘Robert’ and ‘Bob’ which refer to the same name, and more general forms of string transformations such as abbreviations	Propose a simple programmatic framework based on transformation rules to capture non-syntactic notions of string similarity. Informally, transformation rules generate a set of new strings for any given string; two strings are considered similar if some pair of strings generated from the original strings is similar

Appendix

Table A1 Existing literature review on record linkage (2007–2012) (continued)

Authors	*Objective and issues*	*Summary*
Inan et al. (2008)	A hybrid approach to private record linkage for privacy data Private record linkage is the first and possibly the most important step towards utilisation of private information. The process of identifying and linking different representations of the same real-world entity across multiple data sources is known as the record linkage problem. Since it is a key component of data integration methodologies, record linkage has been investigated extensively	Experiments conducted on real datasets show that authors method has significantly lower costs than cryptographic techniques and yields much more accurate matching results compared to sanitisation techniques, even when the datasets are perturbed extensively
Adly (2009)	Looked at efficient record linkage using a double embedding scheme. This paper introduced a novel scheme for record linkage based on double embedding of the data, aiming at improving the efficiency. A two level matching is proposed, with the first level performing a fast and inaccurate matching, ensuring high recall while the second level performs a more expensive matching, on a smaller set of pairs, to improve the accuracy Record linkage is the problem of identifying similar records across different data sources. The similarity between two records is defined based on domain-specific similarity	Experimental evaluation on real datasets revealed that, by using contractive embedding techniques that preserve the distance between records values, the suggested scheme outperforms the single embedding scheme achieving gains in time performance ranging from 30% to 60%, while achieving the same level of recall and accuracy. Future work will address scenarios with more than two parties and different data types such as DNA sequence, etc.
Yakout et al. (2009)	Studied on the efficient private record linkage Record linkage is the computation of the associations among records of multiple databases. It arises in contexts like the integration of such databases, online interactions and negotiations, and many others. The autonomous entities who wish to carry out the record matching computation are often reluctant to fully share their data	Provide efficient techniques for private record linkage that improve on previous work in that (i) they make no use of a third party; (ii) they achieve much better performance than that of previous schemes in terms of execution time and quality of output (i.e., practically without false negatives and minimal false positives). The software implementation provides experimental validation of authors approach and the above claims Illustrating the accuracy of the data matching. Authors analysed the effect of the likely-linked pairs' computation on data matching in terms of the classical precision and recall metrics

Appendix

Table A1 Existing literature review on record linkage (2007–2012) (continued)

Authors	*Objective and issues*	*Summary*
Inan et al. (2010)	Private record matching using differential privacy Private matching between datasets owned by distinct parties is a challenging problem with several applications. Private matching allows two parties to identify the records that are close to each other according to some distance functions, such that no additional information other than the join result is disclosed to any party. Private matching can be solved securely and accurately using secure multi-party computation (SMC) techniques, but such an approach is prohibitively expensive in practice	Propose an alternative design centred on differential privacy, a novel paradigm that provides strong privacy guarantees. The realisation of the new model presents difficult challenges, such as the evaluation of distance-based matching conditions with the help of only a statistical queries interface
Yakout et al. (2010)	Studied on behaviour based record linkage Present a new record linkage approach that uses entity behaviour to decide if potentially different entities are in fact the same. An entity's behaviour is extracted from a transaction log that records the actions of this entity with respect to a given data source	Present the necessary algorithms to model entities' behaviour and compute a matching score for them. To improve the computational efficiency of our approach, authors precede the actual matching phase with a fast candidate generation that uses a 'quick and dirty' matching method. Extensive experiments on real data show that the approach can significantly enhance record linkage quality while being practical for large transaction logs The key to the proposed strategy is that we merge the behaviour information for each candidate pair of entities to be matched. If the two behaviours seem to complete one another, in the sense that stronger behavioural patterns become detectable after the merge, then this will be a strong indication that the two entities are, in fact, the same
Arasu et al. (2010)	Studied on active learning of record matching packages Consider the problem of learning a record matching package (classifier) in an active learning setting. In active learning, the learning algorithm picks the set of examples to be labelled, unlike more traditional passive learning setting where a user selects the labelled examples	The standard approach to record matching is to use textual similarity between the records to determine whether or not two records are matches Present new algorithms for this problem that overcome these limitations. The algorithms are fundamentally different from traditional active learning approaches, and are designed ground up to exploit problem characteristics specific to record matching. Authors include a detailed experimental evaluation on real world data demonstrating the effectiveness of the algorithms

Appendix

Table A1 Existing literature review on record linkage (2007–2012) (continued)

Authors	*Objective and issues*	*Summary*
Su et al. (2010)	Record matching over query results from multiple web databases Propose a new record matching method unsupervised duplicate detection (UDD)	Present an unsupervised, online record matching method, UDD, which, for a given query, can effectively identify duplicates from the query result records of multiple web databases. After removal of the same-source duplicates, the 'presumed' nonduplicate records from the same source can be used as training examples alleviating the burden of users having to manually label training examples. Starting from the nonduplicate set, authors used two cooperating classifiers, a weighted component similarity summing classifier and an SVM classifier, to iteratively identify duplicates in the query results from multiple web databases
Jutte et al. (2011)	Study looked at administrative record linkage as a tool for public health research	Methods developed and implemented in several jurisdictions across the globe have achieved high-quality linkages for conducting health and social research without compromising confidentiality. Key data available for linkage include health services utilisation, population registries, place of residence, family ties, educational outcomes, and use of social services. Linking events for large populations of individuals across disparate sources and over time permits a range of research possibilities, including the capacity to study low prevalence exposure-disease associations, multiple outcome domains within the same cohort of individuals, service utilisation and chronic disease patterns, and life course and transgenerational transmission of health
Cuzzocrea and Puglisi (2011)	Studied on records linkage in data warehouse	Propose a critical review of research contributions focused on record linkage in Data Warehousing, along with discussion on possible research perspectives, which is the main contribution of this paper While a wide collection of research proposals and results in the context of record linkage for relational databases exists, the problem of effectively supporting record linkage in data warehousing is still an open research challenge. Contrary to this actual trend, record linkage plays a critical role in such research context, with particular regard to the extraction-transformation-loading (ETL) layer of data warehousing platforms

Appendix

Table A1 Existing literature review on record linkage (2007–2012) (continued)

Authors	*Objective and issues*	*Summary*
Karakasidis and Verykios (2011)	Studied on blocking and matching for secure record linkage	The secure framework authors propose consists of two basic components. First, authors utilise a secure blocking component based on phonetic algorithms statistically enhanced to improve security. Second, authors use a secure matching component where actual approximate matching is performed using a novel private approach of the Levenshtein Distance algorithm
	Performing approximate data matching has always been an intriguing problem for both industry and academia. This task becomes even more challenging when the requirement of data privacy rises. In this paper, authors propose a novel technique to address the problem of efficient privacy-preserving approximate record linkage	The goal is to combine the speed of private blocking with the increased accuracy of approximate secure matching
Varghese and Sundar (2011)	To improve performance in classification using data matching. Performing data matching to solves the duplication detection problem	This paper then proposes an efficient approach to detect duplicates in a web database scenario. The characteristics of relational data are analysed from the perspective of duplicate detection. Authors define constraint rules that capture these characteristics. Since in a typical database the vast majority of randomly selected record pairs are non-duplicates, it is possible to populate the training set with negative examples based on such pairs, while filtering out likely pairs of duplicate records using similarity metrics such as vector-space cosine similarity
Abril et al. (2012)	To improve record linkage with supervised learning for disclosure risk assessment	Introduce a parameterisation of record linkage in terms of a weighted mean and its weights, and provide a supervised learning method to determine the optimum weights for the linkage process. That is, the parameters yielding a maximal record linkage between the protected and original data. Authors compare their method to standard record linkage with data from several protection methods widely used in statistical disclosure control, and study the results taking into account the performance in the linkage process, and its computational effort
	To improve the linkage as compared to standard distance-based record linkage	
	To identify key-attributes for record linkage	
Friere et al. (2012)	This paper aims at to present the integration of the files of the Brazilian Cervical Cancer Information System (SISCOLO) in order to identify all women in the system. SISCOLO has the exam as the unit of observation and the women are not uniquely identified	In this study, data from June 2006 to December 2009 were used. Each table was linked with itself and with the other through record linkage methods
		The integration identified 6236 women in the histology table and 1,678,993 in the cytology table. 5324 women from the histology table had records in the cytology table

Appendix

Table A1 Existing literature review on record linkage (2007–2012) (continued)

Authors	*Objective and issues*	*Summary*
Ferrante and Boyd (2012)	To develop a transparent and transportable methodology for evaluating data linkage software	Evaluation methodology that overcomes a number of these difficulties. Authors approach involves the generation and use of representative synthetic data; the execution of a series of linkages using a pre-defined linkage strategy; and the use of standard linkage quality metrics to assess performance. The methodology is both transparent and transportable, producing genuinely comparable results The methodology was used by the centre for data linkage (CDL) at Curtin University in an evaluation of 10 DL software packages. It is also being used to evaluate larger linkage systems (not just packages). The methodology provides a unique opportunity to benchmark the quality of linkages in different operational environments
Durham et al. (2012)	To quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage	Provided a principled and comprehensive evaluation of the state-of-the-art privacy-preserving string comparators (PPSCs). The evaluation considered three axes critical to privacy-preserving record linkage (PPRL) applications: (1) correctness, (2) computational complexity, and (3) security. This research used a real dataset, evaluated the PPSCs on a common quantified space, and provided the information needed to support decisions in designing a PPRL protocol in the real world