A Qualitative Literature Review on Linkage Techniques for Data Integration

Felix Kruse
University of Oldenburg
felix.kruse@uol.de

Ahmad Pajam Hassan
University of Oldenburg
ahmad.pajam.hassan@uol.de

Jan-Philipp Awick
University of Oldenburg
jan-philipp.awick@uol.de

Jorge Marx Gómez
University of Oldenburg
jorge.marx.gomez@uol.de

Abstract

The data linkage techniques "entity linking" and "record linkage" get rising attention as they enable the integration of multiple data sources for data, web, and text mining approaches. This has resulted in the development of numerous algorithms and systems for these techniques in recent years. The goal of this publication is to provide an overview of these numerous data linkage techniques. Most papers deal with record linkage and structured data. Processing unstructured data through entity linking is rising attention with the trend Big Data. Currently, deep learning algorithms are being explored for both linkage techniques. Most publications focus their research on a single process step or on the entire process of "entity linking" or "record linkage". However, the papers have the limitation that the used approaches and techniques have always been optimized for only a few data sources.

1. Introduction

The phenomenon Big Data describes the increasing amount of data sources and its usage to generate business value. These are not only internal company data sources but also external data sources [1]. The numerous data sources can provide different, complementary or additional information [2, 3], because they have often been created for a specific task [4, 5, 6]. Data science is the discipline that uses data and information to generate relevant insights with data, text, and web mining approaches [7]. Obtaining information from different data sources is necessary for these approaches. In order to use this heterogeneous information, the data sources must be integrated. The external data sources, in particular, often do not have a common identifier (ID). This is a challenge because the data sources cannot be joined with this ID. *Record linkage* and *entity linking* can be used to integrate structured, semi-structured, and unstructured data sources without a common ID. This literature review is intended to show the state-of-the-art in the linkage techniques *record linkage* and *entity linking*.

This paper is structured as follows. First, an introduction of the concepts entity linking and record linkage is given (section 2). Section 3 describes the applied research method, the literature review, and the applied research methodology. Section 4 presents the analysis of the relevant papers. Classification categories are described descriptively and analysis results are interpreted. Lastly, a conclusion and outlook is given (section 5).

2. Record linkage and entity linking

Ma et al. (2017) define *record linkage* as "the task of identifying records that refer to the same logical entity across different data sources, especially when they may or may not share a common identifier across the data sources" [8]. Based on related literature the following 15 synonyms are used for this concept: *deduplication*, *duplicate detection*, *duplicate record elimination*, *entity identification*, *entity matching*, *entity reconciliation*, *entity resolution*, *fuzzy duplicate identification*, *identity resolution*, *object identification*, *object matching*, *reference matching*, *entity reconciliation* [9], *field matching* [10], and *object identification* [11]. In the following, the term *record linkage* will be used for this concept. The concept of *record linkage* exists since 1969 and was first based on rules. Between 2000 and 2015, the research was focused on supervised and unsupervised methods. Since 2018 the focus of the research is *record linkage* with deep learning [12].

Figure 1 illustrates a common *record linkage* workflow by *Christen (2007)*. This process contains the steps "data preparation", "blocking", "record pair comparison", "classification", and "evaluation" [13]. "Data preprocessing" is the first step in the *record linkage* process (Figure 1). In this process step, the quality of the data sources is improved by, for example, replacing missing values, resolving impossible data combinations, or handling outliers [10]. The step

URI: https://hdl.handle.net/10125/63871
978-0-9981331-3-3

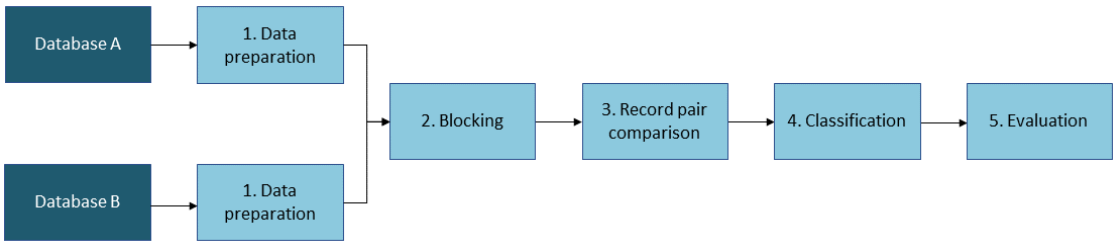

Figure 1. Standard record linkage workflow by [13]

"blocking" reduces the number of comparisons of the possible matches by generating blocking criteria on a single attribute or on combined attributes, also called blocking key. In this step, the data is divided into blocks that contain attributes with the same blocking key. The attributes in one block are possible record pairs. The step "record pair comparison" uses comparison functions that deliver a numerical similarity value on the possible record pairs. "Classification" decides if two record pairs belong together based on the similarity value calculated in the step "record pair comparison". During the "evaluation" step, the results are compared with other methods [13]. While the main task of *record linkage* is to match entities across data sets, the key focus of *entity linking* is the disambiguation of unstructured data with a knowledge base [14]. The main challenges are the ambiguity and the name variations of entity mentions in unstructured data [15, 14]. For example, the name "Obama" can refer to "Mount Obama" or "Barack Obama" [16]. *Wu et al. (2014)* define *entity linking* as "[..] the task of identifying text fragments in text which refer to an entity in a knowledge base, such as Wikipedia and Freebase. It enriches unstructured text with entities contained in knowledge base, helping people understand web pages and other documents online when they encounter unfamiliar entities" [16].

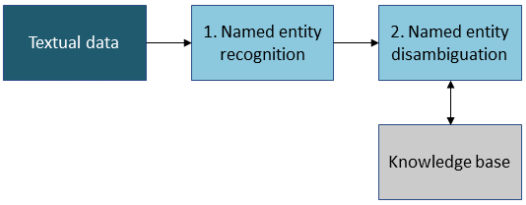

Figure 2. Standard entity linking workflow by [15, 16]

Entity linking is a subtask in natural language processing (NLP) [17]. Another synonym used for the process of *entity linking* is *named entity disambiguation* [18]. In the following, the term *entity linking* describes the concept of linking named entities in unstructured data, while *named entity recognition* and *named entity disambiguation* are the steps of this concept. The step *named entity recognition* identifies the mentions of entities in the unstructured data, while the *named entity disambiguation* step uses information from knowledge bases (e.g., Wikipedia, Freebase) to disambiguate the identified mentions. The described *entity linking* process can be represented in a standard *entity linking* process as shown in Figure 2 [15, 16].

3. State of the art data linkage techniques

Wilde and Hess (2007) define the qualitative literature review as qualitative/quantitative cross-sectional analysis [19]. In this paper, the qualitative literature review is used as the research method. The qualitative literature review is an empirical method, which allows evaluating papers under consideration of the context. In contrast, the quantitative content analysis counts the occurrence of certain words in texts [20, 21, 22]. A specific research goal is the basis of a literature review [22]. The research objective of this literature review is to provide an overview of the data, approaches, and algorithms used in *record linkage* and *entity linking*.

Jane Webster and Richard T. Watson (2002) and *Cato (2016)* describe a four-step process for a qualitative literature review. The four process steps are "definition of search strategy and parameters", "identification and selection of relevant papers", "forward and backward search", and "analysis of relevant papers" [23, 20]. The next sections describe each process step. Those were carried out in a peer review with the authors.

3.1. Definition of search strategy and parameters

The first process step defines in which databases papers are searched for and what requirements are expected of them. The research databases in which relevant literature is searched must be defined. Requirements, such as length or date of papers must be

defined. The most important requirement are the search terms used to identify the papers [23, 20].

The selected search terms for this literature review are listed in Table 1. The selection of the search words is based on the synonyms from the literature and an earlier literature review on *record linkage* [24]. All search terms have been combined with an or-operator for the database query. We chose the databases IEEE[1], ScienceDirect[2] and ACM Digital Library[3] to search for relevant papers, as listed in Table 1. Title, abstract, or keywords of identified papers must contain the search terms.

Table 1. Search strategy

Databases	Search terms
IEEE, Science Direct, ACM Digital Library	entity disambiguation; entity resolution; entity reconciliation; deduplication; duplicate detection; record linkage; redundancy elimination; object identification; reference matching; co-reference detection; non-identical duplicates; object matching; duplication detection; similarity join

3.2. Identification and selection of relevant papers

This section describes the process steps "identification and selection of relevant papers" and "forward and backward search". First, the literature search with the defined search parameters is carried out. After that, the titles and abstracts of the papers are filtered through a first screening. If necessary, the filtered papers are further filtered by a full-text screening [23, 20].

After the databases and the search terms were defined, the search has been executed. The result of the queries is described in Table 2. The query in the IEEE database returned 3073 papers from which 30 are classified as relevant. The ScienceDirect database returned 1517 papers from which 17 are classified as relevant and the ACM Digital Library returned 285 papers from which 41 are classified as relevant.

The backward search is used to search for relevant papers in the bibliography of these papers. The forward search is used to search for relevant recent papers that reference to said papers [23, 20]. The forward and backward search delivered 9 relevant papers (Table 2).

[1]https://ieeexplore.ieee.org/Xplore/home.jsp
[2]https://www.sciencedirect.com/
[3]https://dl.acm.org/dl.cfm

Table 2. Results of the conducted search strategy

Database	Results	Selected papers
IEEE	3073	30
ScienceDirect	1517	17
ACM Digital Library	285	41
Forward and backward search	/	9
Not relevant	/	5

The result reveals 97 papers. After the full-text screening, five papers were classified as irrelevant. So, the overall result reveals 92 relevant papers.

4. Analysis of the relevant papers

Mayring (2014) distinguishes between the fundamental forms of interpreting "summary", "explication", and "structuring" to analyze the papers [22]. The object of the "summary" analysis is to reduce the material in such a way that the essential contents remain, in order to create through abstraction a comprehensive overview of the base material [22]. The object of the "explication" analysis is to provide additional material on individual doubtful text components (terms, sentences...) with a view to increasing understanding, explaining, interpreting the particular passage of text [22]. The object of the "structuring" analysis is to filter out particular aspects of the material, to give a cross-section through the material according to pre-determined ordering criteria, or to assess the material according to certain criteria [22].

From these three fundamental interpretation approaches, Mayring has defined nine analysis techniques, as shown in Table 3 [22].

Table 3. Catalogue of qualitative literature review analysis Techniques

Reduction	1. Summarizing 2. Inductive category formation
Explication	3. Narrow contextual analysis 4. Broad contextual analysis
Structuring	5. Nominal deductive category assignment 6. Ordinal deductive category assignment
Mixed	7. Content structuring/theme analysis 8. Type analysis 9. Parallel forms

In this paper, the mixed method "content structuring/theme analysis" is used for the analysis of the papers. This method combines the basic procedures of inductive and deductive qualitative content analysis [22]. The approach "content structuring/theme analysis" was chosen in order not to limit the results too restrictively. The method content structuring/theme analysis contains the following process-steps [22]:

A Deductive part of the process

1. Definition of the categories (main categories and subcategories) from theory
2. Definition of the coding guidelines (definitions, anchor examples and coding rules)
3. Analysis of the papers, assigning sentences and paragraphs to categories, document anchor examples
4. Revision of the categories and coding guideline after 10-50% of the papers

B Inductive part of the process

5. Working through the beforehand categorized sentences and paragraphs to create inductive categories
6. Revision of the categories after 10-50% of the papers
7. Final working through the papers and building main categories, if useful
8. Analysis of the main and sub categories

In the first step, the categories "data sources", "data preparation", "modeling", and "research focus" were deductively defined (Table 4). These categories are based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) process steps [25]. All 92 publications have been analyzed in full text and the deductively formed categories have been assigned to the text passages. After the text passages were assigned to the categories, the inductive categories were formed from these text passages. The inductively formed categories with the relationship to the deductively formed categories are listed in Table 4. The deductive category "data source" has resulted in the categories "data set name", "application domain", and "data structure". The "similarity measures" and the "approaches and algorithms" were extracted from the category "data preparation". The text passages with the "modeling" category also provided results on "similarity measures" and "algorithms". The "research focus" category provided the focus of the respective paper. The "research focus" of the paper was assigned to the *entity linking* process, the *record linkage* process, or a process step of these. Now, the inductive categories can be evaluated quantitatively.

Table 4. Deductive categories and derived inductive categories

Deductive category	Inductive category
Data sources	Name of the data set, application domain, data structure
Data preparation	Approach or algorithm, similarity measure
Modeling	Similarity measure, algorithm, category algorithm
Research focus	Focused process (step)

4.1. Descriptive analysis of the generated data by the literature review

Figure 3 shows the number of papers published per year. In the years from 2006 to 2012, only a few papers were published constantly. From 2014, the number of papers published increased significantly. According to Google trends, the Big Data trend, which has also been rising since 2013. The relationship between Big Data and *record linkage* and *entity linking* techniques is to integrate various Big Data sources and make them available for data analysis [24, 26].
Figure 3 shows the number of published papers per year classified by *entity linking* and *record linkage*. In total, about three-quarters of the papers are assigned to *record linkage* and one-quarter of the papers to *entity linking*. The trend of an increasing number of publications since 2014 can be observed for both techniques, as shown in Figure 3.

4.1.1. Analysis of the record linkage papers We identified 68 relevant papers with *record linkage* as an object of investigation. The research focus of the publications is categorized by the *record linkage* process steps (table 5).

Most papers focus their research on the process step "classification" or the "entire process", shown in Table 5. *Kooli et al. (2018)* apply different classification algorithms to solve a *record linkage* task. For example, this paper is categorized with the research focus on the process step "classification" [27]. *Shu et al. (2012)* conduct the entire *record linkage* process with a developed framework. For example, this paper is categorized with the research focus on the "entire

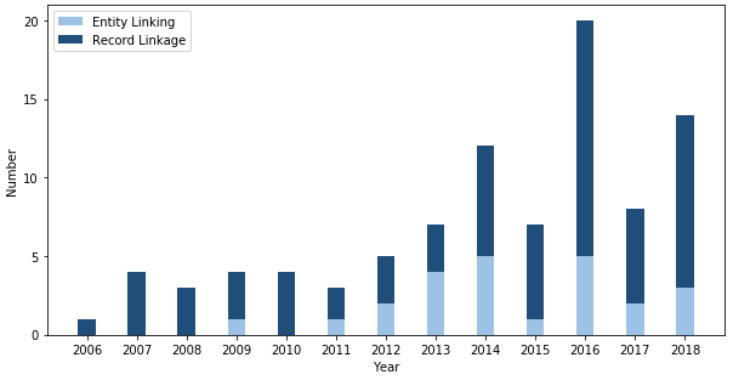

Figure 3. Number of papers per year

Table 5. Research focus of the record linkage papers

Research focus	Number
Classification	25
Entire record linkage process	23
Literature review	6
Record pair comparison	5
Blocking	5
Data preparation	3
Test record linkage software	2
Benchmarking record linkage frameworks	1
Selection of training data	1

process" [28]. *Koudas et al. (2006)*, *Elmagarmid et al. (2007)*, *Wandelt et al. (2014)*, *Enrquez et al. (2017)*, *El-Ghafar et al. (2017)*, and *Fier et al. (2018)* conduct a literature review on *record linkage* [29, 30, 31, 24, 32, 33]. Two papers focus on the special technique *similarity join* in their literature review [33, 31]. Two papers give an overview of *record linkage* but are older than ten years [30, 29]. The last two literature reviews focus on the relationship between Big Data and *record linkage* [32, 24]. *Peled et al. (2016)* apply different similarity measures to match entities across online social networks; such a paper is categorized with the research focus on "record pair comparison" [34].

Five papers focus their research on the process step "blocking" and try to improve this step [35, 36, 37, 38, 39]. Three papers are categorized with the research focus to analyze the process step "data preparation" [40, 41, 42]. For example, *Marple et al. (2017)* try to improve the *record linkage* with external data sources. To do this, their focus is on the process step "data preparation" [41]. The authors *Blanco et al. (2018)* and *Enrquez et al. (2016)* focus their research on testing *record linkage* software [4, 43]. One paper does a benchmark on *record linkage* frameworks [44] and one paper researched the selection of training data for the *record linkage* process [45].

Figure 4 shows the distribution of the algorithm categories rule-based, supervised, unsupervised, and deep learning by time. *Dong and Rekatsinas (2018)* statements regarding the historical development of *record linkage* techniques are confirmed [12]. Supervised and unsupervised learning methods have been used more often since 2015. Deep learning methods were used only in papers in 2016 and 2018. Figure 4 shows that rule-based approaches are still being used. *Dong and Rekatsinas (2018)* write about the potential of machine learning procedures like supervised and unsupervised methods to increase the automation of the *record linkage* process [12]. Despite this potential, rule-based approaches seem to be used more often than supervised and unsupervised learning, as shown in Figure 4.

Table 6 shows the use of algorithms to execute a linkage task. The table shows algorithms with more than one entry. The number of algorithms with only one entry is displayed grouped. Rule-based approaches are used in 18 papers, as for example by *Ferguson et al. (2018)*, *Jupin and Shi (2014)*, or *Kobayashi et al. (2018)* [46, 47, 48]. In 14 papers, the algorithms are not mentioned, so they are classified as not defined. The supervised learning approaches support vector machine, decision tree, and neuronal networks follow on the next places, as shown in Table 6. The paper by *Nentwig et al. (2016)* is one of the four which used clustering algorithms [49]. An example of the four graph-based entries is presented by *Liu et al. (2015)* [40]. The last three entries with three mentions are the latent dirichlet allocation, the logistic regression, and naive bayes.

Table 7 lists the similarity measures used in the *record linkage* papers. In 35 of the papers, no similarity measure is mentioned. The most commonly mentioned string similarity measures are the widely used ones

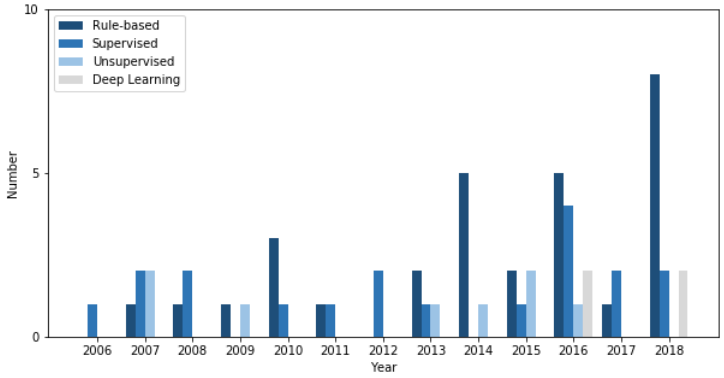

Figure 4. Number of record linkage papers per year per linkage algorithm category

Table 6. Number of mentions of algorithms for record linkage papers

Algorithm	Category	Number
Rule-based	Rule-based	18
Not defined	Not defined	14
Support Vector Machine	Supervised	13
Decision Tree	Supervised	8
Neural Network	Deep Learning	5
Clustering	Unsupervised	4
Graph-based	Rule-based	4
Latent Dirichlet Allocation	Unsupervised	3
Logistic Regression	Supervised	3
Naive Bayes	Supervised	3
Number of algorithms with one mention	-	34

Table 7. Number of mentions of similarity measures for record linkage papers

Similarity measure	Number
Not defined	35
Levenshtein	15
Jaro Winkler	10
Jaccard	10
Soundex	6
Cosine distance	5
Jaro	4
Smith-Waterman	3
NYSIIS	3
Euclidean distance	3
Metaphone	2
Longest Common Sub-String Similarity	2
Word2Vec	2
Number of similarity measures with one mention	48

like Levenshtein, Jaro Winkler, Jaccard, Soundex, Kosinus Distance, Jaro, Smith-Waterman, NYSIIS, and Euclidean Distance. Metaphone is a phonetic similarity measure like Soundex. Word2Vec is a word embedding. *Kooli et al. (2018)* use the Word2Vec embedding to consider semantic similarities of strings [27]. The embeddings GloVe and FastText are also used in the selected papers but mentioned only once [50].

Table 8 shows the mention of "data preparation" steps in the papers. As shown in Table 5, the objective of the selected paper is not focused on the *record linkage* process step "data preparation". 39 out of 68 papers do not describe their data preparation. Nine papers use the n-gram approach to split the strings in n-grams to compute the string similarity, as presented by *Elmagarmid et al. (2007)* [30]. Nine papers describe their general data cleaning approach. For example, by *Medhat et al. (2015)* who use simple rules to remove punctuation like comma, semi-colon, or colon, convert all letters into lower or uppercase, or remove spaces [51]. The tf-idf algorithm is used by eight papers to prepare the data for the next steps of the process. Four papers use word embeddings for data preparation. Three papers describe that they do feature engineering and create new features for the *record linkage* process [40, 41, 34]. *Conrad et al. (2016)* and *Gottapu et al. (2016)* use the dictionary approach to prepare data [52, 53].

Table 9 shows the data structure of the data sets used in the relevant papers. Typically, 51 out of the 68 papers use structured data sources. Fifteen papers do not describe their data sources. *Leito et al. (2007)* use semi-structured XML data [54]. The ten papers which

Table 8. Number of mentions of data preparation for record linkage papers

Data preparation	Number
Not defined	39
N-gram	9
Data Cleaning	9
Tf-idf	8
Word embedding	4
Feature Engineering	3
Dictionary	2
Number of data preparation with one mention	14

Table 9. Data structure of data sets used in the record linkage papers

Data Structure	Number
Structured	51
Not defined	15
Unstructured	10
Semi-structured	1

use unstructured data sources stand out. The use of unstructured data in the *record linkage* domain means that attributes that are not atomic but contain free text are made usable, such as a product description. The oldest paper out of the selected papers that have made product descriptions usable comes from *Köpcke et al. (2010)*. With this paper, *Köpcke et al. (2010)* provide the four benchmark data sets DBLP-ACM, DBLP-Scholar, Amazon-GoogleProducts, and Abt-Buy for *record linkage*. These four data sets are structured but contain unstructured attributes like product descriptions or short abstracts of papers [44]. Product data or bibliographical data are used in almost all 10 papers. The benchmark data sets provided by *Köpcke et al. (2010)* [44] are the most frequently used in all *record linkage* papers. Köpcke and Rahm are also the most frequent authors in the papers found with four and five authorships respectively. Besides, Köpcke's and Rahm's papers are frequently cited in the other papers.

4.1.2. Analysis of the Entity Linking papers

Twenty-four relevant papers with the technique *entity linking* were identified. Based on the *entity linking* process (see Figure 2) the research focus of the papers were classified. Most papers analyse the entire *entity linking* process and the process step *entity disambiguation*, shown in table 10. For example, the paper by *Thorne et al. (2016)* is categorized with the research focus on the "entire process". They apply general and domain-specific *entity linking* frameworks on clinical data [55]. The research focus category *entity disambiguation* is assigned if the paper tries to improve this step, for example, by *Hermansson et al. (2013)* [56]. The authors *Wu et al. (2018)* and *Lee and Hwang (2016)* did a literature review with the focus on application examples and graph-based approaches in *entity linking* [18] [57]. The paper by *Lee and Hwang (2016)* developed a system for the *named entity recognition* process step [57]. One paper shows an approach to building a knowledge base [58].

Table 10. Research focus of the entity linking papers

Research focus	Number
Entire entity linking process	12
Entity disambiguation	8
Literature review	2
Named entity recognition	1
Building a knowledge base	1

Table 11 shows that all *entity linking* paper which describe their data set use unstructured data. Most of the papers try to link news-, social media-, or bibliographical data to a knowledge base, such as *Bergamaschi et al. (2017)*, *Zwicklbauer et al. (2013)*, or *Xia et al. (2014)* [59, 60, 61]. Most of the papers use DBpedia, Wikipedia, Freebase, or YAGO as the knowledge base.

Table 11. Data structure of data sets used in the entity linking papers

Data Structure	Number
Unstructured	18
Not defined	6

Table 12 represents the distribution of the algorithm categories rule-based, supervised, unsupervised, and deep learning.

Table 12. Number of mentions of algorithms for entity linking papers

Algorithm	Category	Number
Not defined	Not defined	9
Support Vector Machine	Supervised	5
Rule-based	Rule-based	5
Vector Space Model	Unsupervised	2
Graph-based	Supervised	2
Page Rank	Rule-based	2
Latent Dirichlet Allocation	Unsupervised	2
Nave Bayes	Supervised	2
Number of algorithms with one mention	-	17

The usage of supervised methods starts in 2013, and deep learning is used since 2018. 9 out of the 24 *entity linking* papers do not describe an algorithm. Most of the papers which describe the used algorithm focus on the *entity disambiguation* process step, i.e., *Wu et al. (2018)*. They use various algorithms like the support vector machine, naive bayes, or neural networks [18]. For example, *Zwicklbauer et al. (2016)* use a graph-based approach for the *entity disambiguation* process step [60].

It is quite interesting that some papers use similarity measures from the *record linkage* domain, as shown in table 13. *Zwicklbauer et al. (2016)* or *Shen et al. (2015)* use the similarity measures Jaccard, Cosine, or Levenshtein in the *entity disambiguation* process step [60, 14].

Table 13. Number of similarity measures of algorithms for entity linking papers

Similarity measure	Number
Not defined	18
Jaccard	2
Cosine distance	2
Levenshtein	2
Number of similarity measures with one mention	10

Table 14 shows the data preparation approaches of *entity linking* papers. The approaches to data preparation are very heterogeneous. There are only two methods that have been mentioned more than once. One is the *tf-idf* approach, which is a numerical statistic that tries to reflect how important a word is to a document [14]. The other one is the *surface form identification* which attempts to standardize the various representations of words [62].

Table 14. Number of mention of data preparation for entity linking papers

Data Preparation	Number
Not defined	13
Tf-idf	3
Surface form identification	2
Number of data preparation with one mention	17

4.2. Interpretation of the descriptive analysis and derived future research

The purpose of the literature review is to show which linkage methods exist to assign different data sets to a real-world object. The results show that the importance of linkage techniques has increased in recent years. The literature review provides mainly papers dealing with *record linkage*. *Record linkage* is the classical technique used to link structured data. With the trend of Big Data, *entity linking* also becomes relevant for the research question, because unstructured data sources are to be linked with a structured knowledge base. Both linkage techniques follow the evolutionary development of rule-based algorithms, supervised and unsupervised learning, and currently deep learning algorithms (Figure 4). While *entity linking* papers use only unstructured data sources (Table 11), *record linkage* papers use both structured and unstructured data sources (Table 9). Since 2010, unstructured data sources or unstructured attributes (e.g., product descriptions) are also used to improve the *record linkage* results. Most *record linkage* papers try to optimize the process step "classification" or consider and implement the "entire record linkage process" (Table 5). The relevant *entity linking* papers also consider or implement the entire *entity linking* process or focus on the *entity disambiguation* process step (Table 10). It is noticeable that the algorithms used for classification in the *record linkage* process and in the *entity linking* process step are the same (Table 6, 12). If the unstructured data has been structured to disambiguate it, it is a classic *record linkage* problem. In this process step, the two linkage techniques overlap and can be used to connect unstructured and structured data sources. The evaluation of the similarity measures and the data preparation show a multitude of procedures in both linkage techniques. In many cases, the two process steps are not described in detail (Table 8, 14). The results of the literature review show that *entity linking* approaches - the use of unstructured data sources and attributes - are establishing in *record linkage* approaches. The focus of all papers is to optimize *record linkage* approaches on existing data sets such as the classic *record linkage* data sets from *Köpcke et al. (2010)* [44]. Furthermore, all papers focus on a single process step or the entire record linkage process (Table 5). Moreover, all papers are limited to a few data sources which are linked together.

This leads to the conclusion that no paper investigates the selection of record linkage approaches depending on the data to be integrated. None of the identified papers analyze the properties of the data sources and tries to classify the data integration problems and connect them to appropriate algorithms. The selection of an appropriate record linkage approach out of the numerous existing approaches creates much manual effort for integrating data sources. For example, if many big data sources are to be linked in order to use them in data science projects, the data scientist

must select the most suitable one for each data source from the various algorithms in the individual process steps in order to integrate the data sources. Future research should provide a linkage algorithm selection support. Many papers mention in their outlook that in future research the approaches should be applied on further data sets or domains. Moreover, *Rahm (2016)* already called for an approach to develop holistic data integration as a future reserarch topic [6]. More research should be done to describe the performance of the algorithms in relation to the data sources (research data sets or real-world data sets), data preparation, similarity measures and classification algorithms.

5. Conclusion

This paper explores the linkage approaches *record linkage* and *entity linking*. These approaches are used, especially when a common identifier is not available in different data sources. Based on a comprehensive qualitative literature review, an overview of the state-of-the-art and the development of recent years is given. A total of 92 papers were selected as relevant. Of the 92 papers, 68 deal with *record linkage* and 24 with *entity linking*. To analyze the papers, the categories "research focus", "algorithms", "similarity measures", "data preparation", "data sources" and "knowledge base" were defined, and corresponding text passages were marked. Subsequently, inductive categories were formed from the marked text passages in order to analyze them descriptively. The results show the overlapping of *entity linking* and *record linkage*, as more and more unstructured data sources and attributes are used in addition to structured data sources. One result is, that a linkage algorithm selection support is missing. The literature review shows that a large number of linkage algorithms exists. The authors want to deal with this problem in the future and make a recommendation for linkage algorithms based on data integration problems.

Acknowledgements: The authors thank Stefan Wunderlich for the valuable support and the helpful advice.

References

[1] D. Blazquez and J. Domenech, "Big data sources and methods for social and economic analyses," *Technological Forecasting and Social Change*, vol. 130, pp. 99–113, 2018.

[2] Y. Lin, H. Wang, J. Li, and H. Gao, "Data source selection for information integration in big data era," 2016.

[3] M. Pershina, *Graph-Based Approaches to Resolve Entity Ambiguity*. Dissertation, New York University, New York, New York, 2016.

[4] R. Blanco, J. G. Enriquez, F. J. Dominguez-Mayo, M. J. Escalona, and J. Tuya, "Early integration testing for entity reconciliation in the context of heterogeneous data sources," *IEEE Transactions on Reliability*, pp. 1–19, 2018.

[5] X. L. Dong and D. Srivastava, "Big data integration," in *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 1245–1248, IEEE, 2013.

[6] E. Rahm, "The case for holistic data integration," in *Advances in Databases and Information Systems* (J. Pokorný, M. Ivanović, B. Thalheim, and P. Šaloun, eds.), vol. 9809 of *Lecture Notes in Computer Science*, pp. 11–27, Cham: Springer International Publishing, 2016.

[7] F. Kruse, V. Dmitriyev, and J. Marx Gómez, "Building a connection between decision maker and data-driven decision process," *Archives of Data Science, Series A (Online First)*, vol. 4, no. 1, p. 16 S. online, 2018.

[8] B. Ma, T. Jiang, X. Zhou, F. Zhao, and Y. Yang, "A novel data integration framework based on unified concept model," *IEEE Access*, vol. 5, pp. 5713–5722, 2017.

[9] H. Köpcke, *Object Matching on real-world problems*. Dissertation, Universität Leipzig, Leipzig, 2014.

[10] F. Wang and H. Wang, "Record linkage using the combination of twice iterative svm training and controllable manual review," in *2016 IEEE 14th Intl Conf 2016*, pp. 31–38, 2016.

[11] S. Lee, J. Lee, and S.-w. Hwang, "Efficient entity matching using materialized lists," *Information Sciences*, vol. 261, pp. 170–184, 2014.

[12] X. L. Dong and T. Rekatsinas, "Data integration and machine learning," in *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18* (G. Das, C. Jermaine, and P. Bernstein, eds.), (New York, New York, USA), pp. 1645–1650, ACM Press, 2018.

[13] P. Christen, *A two-step classification approach to unsupervised record linkage: Proceedings of the Sixth Australasian Data Mining Conference (AusDM'07) : Gold Coast, Australia, 3-4 December, 2007*, vol. v. 70 of *Conferences in research and practice in information technology series*. [Sydney, N.S.W.]: Australian Computer Society, published in association with the ACM Digital Library, 2007.

[14] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 443–460, 2015.

[15] S. Hakimov, S. A. Oto, and E. Dogdu, "Named entity recognition and disambiguation using linked data and graph-based centrality scoring," in *Proceedings of the 4th International Workshop on Semantic Web Information Management - SWIM '12* (R. de Virgilio, F. Giunchiglia, and L. Tanca, eds.), (New York, New York, USA), pp. 1–7, ACM Press, 2012.

[16] C. Wu, W. Lu, and P. Zhou, "An optimization framework for entity recognition and disambiguation," in *Proceedings of the first international workshop on Entity recognition & disambiguation - ERD '14* (D. Carmel, M.-W. Chang, E. Gabrilovich, B.-J. P. Hsu, and K. Wang, eds.), (New York, New York, USA), pp. 105–110, ACM Press, 2014.

[17] Z. Zheng, X. Si, F. Li, E. Y. Chang, and X. Zhu, "Entity disambiguation with freebase," in *2012 IEEE/WIC/ACM International Conferences*, pp. 82–89, 2012.

[18] G. Wu, Y. He, and X. Hu, "Entity linking: An issue to extract corresponding entity with knowledge base," *IEEE Access*, vol. 6, pp. 6220–6231, 2018.

[19] T. Wilde and T. Hess, "Forschungsmethoden der wirtschaftsinformatik," *WIRTSCHAFTSINFORMATIK*, vol. 49, no. 4, pp. 280–287, 2007.

[20] Jane Webster and Richard T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS Quarterly*, vol. 26, no. 2, pp. xiii–xxiii, 2002.

[21] P. Mayring, "Qualitative content analysis," *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, vol. 1, no. 2, 2000.

[22] P. Mayring, *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. 2014.

[23] P. Cato, *Einflüsse auf den Implementierungserfolg von Big Data Systemen*. Dissertation, Verlag Dr. Kovač, 2016.

[24] J. G. Enríquez, F. J. Domínguez-Mayo, M. J. Escalona, M. Ross, and G. Staples, "Entity reconciliation in big data sources: A systematic mapping study," *Expert Systems with Applications*, vol. 80, pp. 14–27, 2017.

[25] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining." 2000.

[26] B. Golshan, A. Halevy, G. Mihaila, and W.-C. Tan, "Data integration: After the teenage years," in *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - PODS '17* (J. van den Bussche, F. Geerts, and E. Sallinger, eds.), (New York, New York, USA), pp. 101–106, ACM Press, 2017.

[27] N. Kooli, R. Allesiardo, and E. Pigneul, "Deep learning based approach for entity resolution in databases," in *Intelligent Information and Database Systems* (N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, and B. Trawiński, eds.), vol. 10752 of *Lecture Notes in Computer Science*, pp. 3–12, Cham: Springer International Publishing, 2018.

[28] L. Shu, C. Lin, W. Meng, Y. Han, C. T. Yu, and N. R. Smalheiser, "A framework for entity resolution with efficient blocking," in *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pp. 431–440, IEEE, 2012.

[29] N. Koudas, S. Sarawagi, and D. Srivastava, "Record linkage similarity measures and algorithms," 2006.

[30] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, pp. 1–16, 2007.

[31] S. Wandelt, J. Wang, U. Leser, D. Deng, S. Gerdjikov, S. Mishra, P. Mitankin, M. Patil, E. Siragusa, A. Tiskin, and W. Wang, "State-of-the-art in string similarity search and join," *ACM SIGMOD Record*, vol. 43, no. 1, pp. 64–76, 2014.

[32] R. M. A. El-Ghafar, M. H. Gheith, A. H. El-Bastawissy, and E. S. Nasr, "Record linkage approaches in big data: A state of art study," in *2017 13th International Computer Engineering Conference (ICENCO)*, pp. 224–230, IEEE, 2017.

[33] F. Fier, N. Augsten, P. Bouros, U. Leser, and J.-C. Freytag, "Set similarity joins on mapreduce," *Proceedings of the VLDB Endowment*, vol. 11, no. 10, pp. 1110–1122, 2018.

[34] O. Peled, M. Fire, L. Rokach, and Y. Elovici, "Matching entities across online social networks," *Neurocomputing*, vol. 210, pp. 91–106, 2016.

[35] J. Gómez-Bao, J.-L. Larriba-Pey, and J. Ribes Puig, "Record linkage performance for large data sets," in *Proceeding of the ACM first international workshop on Privacy and anonymity for very large databases - PAVLAD '09* (V. Muntés-Mulero and J. Nin, eds.), (New York, New York, USA), ACM Press, 2009.

[36] S. Mishra, S. Saha, and S. Mondal, "An automatic framework for entity matching in bibliographic databases," in *2016 IEEE Congress on Evolutionary Computation (CEC)*, pp. 271–278, IEEE, 2016.

[37] G. Simonini, S. Bergamaschi, and H. Jagadish, "Blast: a loosely schema-aware meta-blocking approach for entity resolution," 2016.

[38] I. van Dam, G. van Ginkel, W. Kuipers, N. Nijenhuis, D. Vandic, and F. Frasincar, "Duplicate detection in web shops using lsh to reduce the number of computations," in *Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16* (S. Ossowski, ed.), (New York, New York, USA), pp. 772–779, ACM Press, 2016.

[39] T. de Vries, H. Ke, S. Chawla, and P. Christen, "Robust record linkage blocking using suffix arrays," in *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09* (D. Cheung, I.-Y. Song, W. Chu, X. Hu, and J. Lin, eds.), (New York, New York, USA), p. 305, ACM Press, 2009.

[40] H. Liu, T. A. Kumar, and J. P. Thomas, "Cleaning framework for big data - object identification and linkage," in *2015 IEEE International Congress on Big Data*, pp. 215–221, IEEE, 2015.

[41] T. Marple, B. Desmarais, and K. L. Young, "Collapsing corporate confusion: Leveraging network structures for effective entity resolution in relational corporate data," in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 2637–2643, IEEE, 2017.

[42] T. Prabhu and C. S. Gnana Dhas, "Improved scalability in mining using ontology record linkage algorithm," *Computers & Electrical Engineering*, 2018.

[43] J. G. Enríquez, R. Blanco, F. J. Domínguez-Mayo, J. Tuya, and M. J. Escalona, "Towards an mde-based approach to test entity reconciliation applications," in *Proceedings of the 7th International Workshop on Automating Test Case Design, Selection, and Evaluation - A-TEST 2016* (T. Vos, S. Eldh, and W. Prasetya, eds.), (New York, New York, USA), pp. 74–77, ACM Press, 2016.

[44] H. Köpcke and E. Rahm, "Frameworks for entity matching: A comparison," *Data & Knowledge Engineering*, vol. 69, no. 2, pp. 197–210, 2010.

[45] H. Köpcke and E. Rahm, "Training selection for tuning entity matching," 2008.

[46] J. Ferguson, A. Hannigan, and A. Stack, "A new computationally efficient algorithm for record linkage with field dependency and missing data imputation," *International journal of medical informatics*, vol. 109, pp. 70–75, 2018.

[47] J. Jupin and J. Y. Shi, "Identity tracking in big data: Preliminary research using in-memory data graph models for record linkage and probabilistic signature hashing for approximate string matching in big health and human services databases," in *Proceedings of the 2014 International Conference on Big Data Science and Computing - BigDataScience '14* (A. Chin, J. Zhan, W. Ding, J. Wu, W. Xu, and F. Wang, eds.), (New York, New York, USA), pp. 1–8, ACM Press, 2014.

[48] F. Kobayashi, A. Eram, and J. Talburt, "Entity resolution using logistic regression as an extension to the rule-based oyster system," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 146–151, IEEE, 2018.

[49] M. Nentwig, A. GroB, and E. Rahm, "Holistic entity clustering for linked data," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp. 194–201, IEEE, 2016.

[50] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, "Deep learning for entity matching," in *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18* (G. Das, C. Jermaine, and P. Bernstein, eds.), (New York, New York, USA), pp. 19–34, ACM Press, 2018.

[51] D. Medhat, A. Hassan, and C. Salama, "A hybrid cross-language name matching technique using novel modified levenshtein distance," in *2015 Tenth International Conference on Computer Engineering & Systems (ICCES)*, pp. 204–209, IEEE, 2015.

[52] C. Conrad, N. Ali, V. Keselj, and Q. Gao, "Elm: An extended logic matching method on record linkage analysis of disparate databases for profiling data mining," in *2016 IEEE 18th Conference on Business Informatics (CBI)*, pp. 1–6, IEEE, 2016.

[53] R. D. Gottapu, C. Dagli, and B. Ali, "Entity resolution using convolutional neural network," *Procedia Computer Science*, vol. 95, pp. 153–158, 2016.

[54] L. Leitão, P. Calado, and M. Weis, "Structure-based inference of xml similarity for fuzzy duplicate detection," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07* (M. J. Silva, A. O. Falcão, A. A. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, and Ø. H. Olsen, eds.), (New York, New York, USA), ACM Press, 2007.

[55] C. Thorne, S. Faralli, and H. Stuckenschmidt, "Cross-evaluation of entity linking and disambiguation systems for clinical text annotation," in *Proceedings of the 12th International Conference on Semantic Systems - SEMANTiCS 2016* (A. Fensel, A. Zaveri, S. Hellmann, and T. Pellegrini, eds.), (New York, New York, USA), pp. 169–172, ACM Press, 2016.

[56] L. Hermansson, T. Kerola, F. Johansson, V. Jethava, and D. Dubhashi, "Entity disambiguation in anonymized graphs using graph kernels," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13* (Q. He, A. Iyengar, W. Nejdl, J. Pei, and R. Rastogi, eds.), (New York, New York, USA), pp. 1037–1046, ACM Press, 2013.

[57] T. Lee and S.-w. Hwang, "Linking, integrating, and translating entities via iterative graph matching," in *2016 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pp. 248–255, IEEE, 2016.

[58] E. Boschee, M. Freedman, S. Khanwalkar, A. Kumar, A. Srivastava, and R. Weischedel, "Researching persons & organizations: Awake: From text to an entity-centric knowledge base," in *2014 IEEE International Conference on Big Data (Big Data)*, pp. 1030–1039, IEEE, 2014.

[59] S. Bergamaschi, A. Cappelli, A. Circiello, and M. Varone, "Conditional random fields with semantic enhancement for named-entity recognition," in *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics - WIMS '17* (R. Akerkar, A. Cuzzocrea, J. Cao, and M.-S. Hacid, eds.), (New York, New York, USA), pp. 1–7, ACM Press, 2017.

[60] S. Zwicklbauer, C. Seifert, and M. Granitzer, "Do we need entity-centric knowledge bases for entity disambiguation?," in *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies - i-Know '13* (S. Lindstaedt and M. Granitzer, eds.), (New York, New York, USA), pp. 1–8, ACM Press, 2013.

[61] Y. Xia, H. Lin, R. Lau, and Y. Liu, "Leaning to train: Linking financial news articles to company short names," in *2014 IEEE 11th International Conference on e-Business Engineering*, pp. 240–245, IEEE, 2014.

[62] X. Han and J. Zhao, "Named entity disambiguation by leveraging wikipedia semantic knowledge," in *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09* (D. Cheung, I.-Y. Song, W. Chu, X. Hu, and J. Lin, eds.), (New York, New York, USA), ACM Press, 2009.