Entity reconciliation in big data sources: A systematic mapping study

J.G. Enríquez [a, *], F.J. Domínguez-Mayo [a], M.J. Escalona [a], M. Ross [b, c], G. Staples [c]

[a] Department of Computer Languages and System, University of Seville, Seville, Spain
[b] Faculty of Maritme and Technology, Southampton Solent University, Southampton, United Kingdom
[c] BCS Quality Specialist Group, United Kingdom

Keywords:
Systematic mapping study
Entity reconciliation
Heterogeneous databases
Big data

A B S T R A C T

The entity reconciliation (ER) problem aroused much interest as a research topic in today's Big Data era, full of big and open heterogeneous data sources. This problem poses when relevant information on a topic needs to be obtained using methods based on: (i) identifying records that represent the same real world entity, and (ii) identifying those records that are similar but do not correspond to the same real-world entity. ER is an operational intelligence process, whereby organizations can unify different and heterogeneous data sources in order to relate possible matches of non-obvious entities. Besides, the complexity that the heterogeneity of data sources involves, the large number of records and differences among languages, for instance, must be added. This paper describes a Systematic Mapping Study (SMS) of journal articles, conferences and workshops published from 2010 to 2017 to solve the problem described before, first trying to understand the state-of-the-art, and then identifying any gaps in current research. Eleven digital libraries were analyzed following a systematic, semiautomatic and rigorous process that has resulted in 61 primary studies. They represent a great variety of intelligent proposals that aim to solve ER. The conclusion obtained is that most of the research is based on the operational phase as opposed to the design phase, and most studies have been tested on real-world data sources, where a lot of them are heterogeneous, but just a few apply to industry. There is a clear trend in research techniques based on clustering/blocking and graphs, although the level of automation of the proposals is hardly ever mentioned in the research work.

1. Introduction

Nowadays, we can say that information is money. When one actor has a higher level of knowledge on a topic than another, he/she has more possibilities to seize strategic opportunities. Extrapolating this concept to companies, it becomes the most important asset when they are trying to be competitive. Recent studies confirm that Big Data can generate significant financial value across sectors (McKinsey & Company, 2011), for instance, Chen, Mao, and Liu (2014) conclude that:

- 1.8ZB: is the amount of data generated in two days in 2011 (larger than the accumulated amount of data from the origin of civilization to 2003).
- 750 million: is the amount of pictures uploaded to Facebook.
- 966PB: is the storage capacity of American manufacturing industry for 2009.
- 209 billion: will be the number of RDIF tags in 2021 (12 million till 2011).
- 200+TB: is the data downloaded during a computer geek's 2450 thousand hours.
- 200PB: is the amount of data generated by a smart urban project in China.
- 800 billion dollars: will be the value for personal location data in ten years.
- 300 billion dollars: will be the value for medical expense saving by Big Data analysis in America.
- 32+ billion dollars: is the purchase of the four big companies since 2010.

Thus, the fact of having as much information as possible and making it sure that it has a very good quality, makes the value of a company grow considerably. In this context, the problem of reconciling entities gets a very important significance.

Entity reconciliation (ER) is a fundamental problem in data integration. It involves identifying entities from the digital world that refer to the same real-world entity. ER process plays a fundamental role in the context of information integration and management, aimed to infer a uniform and common structure from var-

* Corresponding author.
E-mail addresses: jose.gonzalez@iwt2.org (J.G. Enríquez), fjdominguez@us.es (F.J. Domínguez-Mayo), mjescalona@us.es (M.J. Escalona), margaret.ross@solent.ac.uk (M. Ross), geoff.staples@bcs.org.uk (G. Staples).

ious large-scale data collections, with which to suitably organize, match and consolidate the information of the individual repositories into one data set (Costa, Cuzzocrea, Manco, & Ortale, 2011). This is a complex problem, since it is not trivial to assert that two heterogeneous data instances represent the same real object. Heterogeneity can happen in data structure as well as in data values (Dorneles, Gonçalves, & dos Santos Mello, 2011). This problem can be applied to many different domains such as: e-health, citations, smart cities, meteorological predictions, manufacturing and many other different environments.

From the point of view of expert systems, ER is an operational intelligence process, whereby organizations can unify different and heterogeneous data sources in order to match non-obvious entities. This process analyzes all the information related to entities from data sources. Then, it applies probability and scoring to determine which identities can be matched and which non-obvious relationships exist among those identities.

This Systematic Mapping Study (SMS) emerges from the need to find the solution that best suits a real-world entity reconciliation problem. In this case, this problem is focused on the management of cultural heritage information in Andalusia, (a southern region of Spain). The Regional Government of Andalusia has the competence of disseminating cultural heritage information, however, it is quite simple to find nonofficial data sources such as: Wikipedia, DBPedia or Yelp, among others, that store information about this topic. Normally, the information provided by these data sources, are not the same as the one provided by the official information system for cultural heritage information, called MOSAICO (Ponce, Escalona, Gómez, Luque, & Molina, 2010). Thus, the Regional Government of Andalusia has to look for a solution that can cover all the problems.

In this context, this SMS contributes to ongoing research in the field of ER in the context of Big Data in four ways: (i) reviewing and showing all methods, techniques or tools that assist the reconciliation of entities in a Big Data context, (ii) summarizing the problems addressed during the process of reconciliation, (iii) creating a new classification for the currently known solutions to this problem, and (iv) offering directions for future research. There are four research questions that will guide this SMS: what methods, techniques or tools have been investigated for ER in the Big Data environment? What methods, techniques and tools have been used for ER in the Big Data environment? What is the nature of found methods, techniques and tools for solving ER in Big Data environment? And what are the objectives pursued in research works for solving the ER in Big Data environment?

This paper is structured as follows. After this introduction, Section 2 summarizes the related work in systematic literature reviews, systematic mapping studies, surveys or reviews about ER. Section 3 presents the method used for the systematic mapping study and its execution. Section 4 discusses the results of the previous work and finally, Section 5 states a set of conclusions and future lines of works.

2. Related work

ER is a topic that has been discussed and studied for many years. In this section, some related works such as systematic literature reviews, surveys or comparisons are presented.

Maddodi, Attigeri, and Karunakar (2010) discuss different strategies of deduplication with their pros and cons and some methods to prevent duplication in databases. This paper discusses seven techniques for detecting duplicate data (deduplication using correlated subquery, using temporary table, using derived table, by creating new tables and renaming it, using common table expression and using merge statements) and three preventive methods for SQL (the primary key, the unique key and the *IGNORE_DUP_KEY* constraint). Finally, the authors make a performance evaluation with Microsoft SQL-Server 2008 in different Data Warehouses.

Dorneles et al. (2011) divided "approximate data matching" into two basic groups: (i) those which compare data based on data values; and (ii) those which compare data based on their structure, exploiting and extracting relevant data to the comparison. They review both categories identifying different approaches and they present a comparative analysis. The authors only focus on work that relies on a similarity function when executing the data matching process. Costa et al. (2011) present an overview of research on data deduplication with the aim to provide a general assessment of useful references and ideas on this topic. Firstly, the authors describe the problem and after that, they propose two categories of techniques for deduplication: supervised (relational data, multidimensional data, data-mining/data-results, linked and XML data and streaming data approaches) and unsupervised (based on clustering, (dis)similarity-search in metric spaces and locality-sensitive hashing).

Yumusak, Dogdu, and Kodaz (2014), present a brief survey dealing with linked data ranking, classifying methods in: ontology ranking, RDF (Resource Description Framework) document ranking, graph ranking, entity ranking and document/source ranking.

Gaikwad and Bogiri (2015), present a survey analysis on duplicated detection in the domain of hierarchical data. They have oriented the paper to experts who are doing research in duplicate detection in xml data or hierarchical data.

Brizan and Tansel (2006) , divide techniques for performing ER or record linkage into: (i) establishing good match criteria between any pair of tuples, and (ii) applying these criteria to one or more relations and they described both. Otero-Cerdeira, Rodríguez-Martínez, and Gómez-Rodríguez (2015), present a literature review regarding ontology matching, with the purpose of helping in guiding new practitioners to get an idea on the state of the field and determines possible research lines based on the decade of 2005 to 2015 and classifying the papers following the framework they proposed.

Beheshti et al. (2016), present a systematic review and a comparative analysis of Cross-document Coreference Resolution methods and tools (CDCR). The authors present a systematic review of the state-of-the-art of challenges and solutions to CDCR, a taxonomy of CDCR and an identification of a set of quality attributes approaches. Papadakis, Svirsky, Gal, and Palpanas (2016), propose a comparative analysis of approximate blocking techniques for ER presenting 17 state-of-the-art blocking methods, 6 popular real datasets and 7 established synthetic datasets that range from 10,000 to 2 million entities.

A comparison of previous approaches in terms of pros and cons is described in Table 1 in order to clarify the contribution that this paper proposes.

As reflected in Table 1, the two found closest papers to our research are those presented by Beheshti et al. (2016) and Papadakis et al. (2016), respectively. Both use a specific methodology, define the number of databases that were consulted and apply a systematic process. However, their scope is specific, that is to say, a particular topic for ER, but not for ER in general. Although most of the papers propose a classification framework, they neither perform a systematic process nor specify the number of databases consulted. Finally, it is important to note that the number of databases consulted in this paper is more than the double proposed in the other papers.

3. Method

The method proposed by Kitchenham and Charters (2007) is one of the most widely accepted in the field of software engineering although it has received some critics and

Table 1
Comparison of previous approaches.

Reference	Number of databases	Apply a specific methodology	Classification framework	Systematic process	General scope	Specific scope
Maddodi et al. (2010)	Unknown	X	X	X	X	√
Dorneles et al. (2011)	Unknown	X	√	X	X	√
Costa et al. (2011)	Unknown	X	√	X	X	√
Yumusak et al. (2014)	Unknown	X	√	X	X	√
Gaikwad and Bogiri (2015)	Unknown	X	√	X	X	√
Brizan and Tansel (2006)	Unknown	X	√	X	X	√
Otero-Cerdeira et al. (2015)	5	√	√	√	X	√
Beheshti et al. (2016)	4	√	√	√	X	√
Papadakis et al., (2016)	Unknown	X	√	X	X	√
Our proposal	11	√	√	√	√	X

proposals for improvements. The proper author in Barbara Kitchenham et al. (2010), analyzes the way software systematic reviews were being used, concluding that the number of revisions had increased significantly. Despite this fact, it still could not be considered a major method for software engineering research since, although it covers a multitude of subjects, it does not often evaluate the quality of the primary studies obtained. A similar work is the one by Da Silva et al. (2011) with a very similar conclusion: most systematic reviews do not evaluate the quality of primary studies and do not provide guidelines for professionals, which reduces their possible impact on the practice of software engineering.

The authors in Zhang, Babar, and Tell (2011), focus on the search strategy, considering it as a critical step in the correct application of the systematic review methods. The authors argue that it is a time-consuming and error prone step, so it must be carefully planned and executed. This work aims at improving this step by incorporating two concepts: "quasi-gold standard" (QGS), which consists of collecting known studies, and "quasi-sensitivity", which involves evaluating the performance of the search. These same authors in (Zhang, Babar, & Ali Babar, 2013) consider the importance of systematic reviews from the empirical point of view. In this article it is shown that researchers are convinced of the value of using a rigorous and systematic methodology for literature reviews. However, they consider that a balance must be struck between methodological rigor and the necessary effort.

Wohlin and Prikladniki (2013) propose to expand the search process with the "snowballing" approach, according to which the included studies could be extended if, in addition to the work carried out with the Kitchenham method, the reference lists of these publications, which involve a backward view, as well as the citations of these publications, offering a forward view, are both taken into account.

As a result of all the criticism received, Kitchenham and Brereton (2013) investigate whether the guidelines should be modified. Several conclusions regarding the improvement of the method are: it warns to withdraw the advice to use structured questions to construct the search strings and to include the advice to use the "quasi-gold standard" concept, based on a limited manual search to build the search strings and further evaluate the search process. The authors also comment that textual analysis tools could possibly be useful for the inclusion and exclusion decision, as well as for the definition of the search chain, but a more rigorous evaluation of the search chain should be done. Besides, they consider that an independent validation of the use of existing tools for managing the systematic review process is needed. Finally, the evaluation of the quality of studies using empirical methods still remains as a major problem.

To carry out this study, two models have been taken as reference. The last one presented by Kitchenham described before and the model of Systematic Mapping Study (SMS). SMS is a form of Systematic Literature Review (SLR) that aims to identify and categorize the available research on a broad software engineering topic. SMSs (Genero, Cruz-Lemus, & Piattini, 2014) are secondary studies with a broader scope than SLRs that aim to provide an overview of an interesting topic and identify the number and type of research as well as the available related results. This allows identifying subjects that lack empirical evidence and performing more empirical studies is needed. It is very common in SMSs to calculate the frequency of publications over time to identify trends or classify the found items in a default classification scheme. SMSs typically consume less time than SLRs and are useful for researchers as a basis to do further work with high level of rigor.

The model presented by Kitchenham establishes that a review should be composed of three phases: planning, conducting and reporting.

- Planning the review. Prior to a SLR, it is necessary to confirm the necessity of the research. The most important activity is writing the research questions that define the review protocol.
- Conducting the review. It deals with executing the protocol that is defined.
- Reporting the review. It describes how the final report is elaborated.

Fosso Wamba, Akter, Edwards, Chopin, and Gnanzou (2015), Ngai and Gunasekaran (2007), Ngai, Hu, Wong, Chen, and Sun (2011a) and Ngai, Moon, Riggins, and Yi (2008) are taken as reference to create a classification framework. In these papers, the authors also propose a methodology comprising three phases: developing a classification framework, conducting the literature review and classifying relevant journal articles.

Fig. 1 shows the different phases of the selected model. It also shows all the activities that compose each of them and the time spent in running them. Details of the complete method step by step can be found below.

3.1. Planning

In this section, each of the tasks that have been made during the process the planning the SMS are presented. These are: identifying the necessity of the review, formulating research questions, defining the review protocol and validating the review protocol.

3.1.1. Identifying the necessity of the review

ER is not a new necessity. Such a need aroused since databases started to be used. Neither extracting data from a same identity nor integrating them into different databases are simple tasks, since to determine which entities are the same for each database is difficult.

In this task, the existing literature reviews concerning frameworks, methodologies or techniques that solve the ER problem in big and heterogeneous data sources have been evaluated.

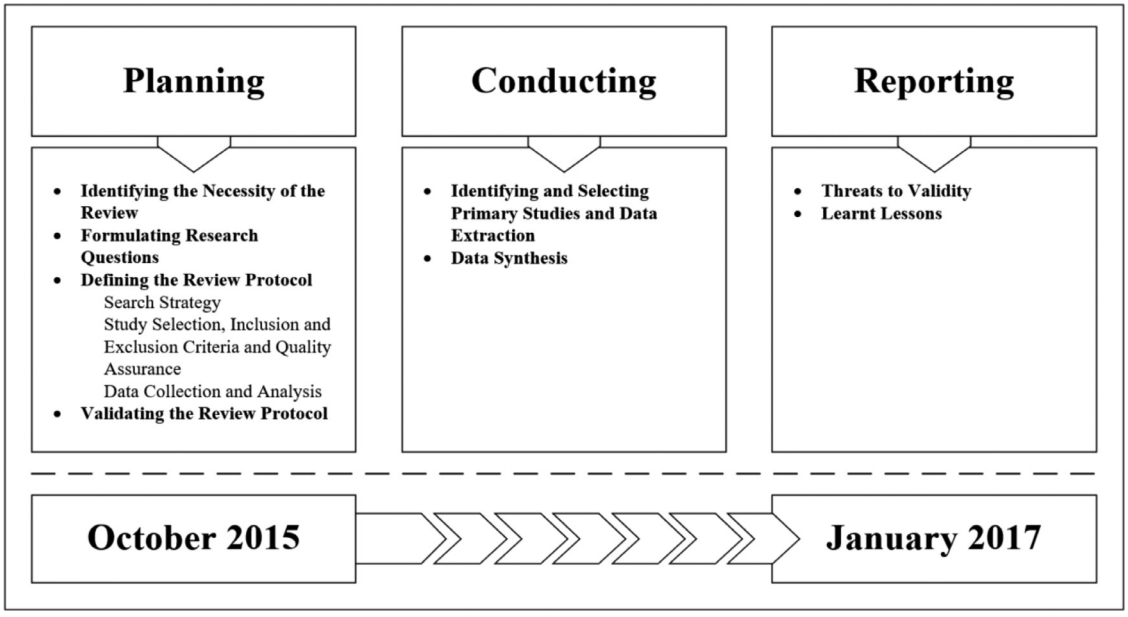

Fig. 1. Method.

Table 2
Research questions.

Research question	Motivation
RQ1. What methods, techniques or tools have been investigated for ER in Big Data environment?	Find out what methods, techniques or tools have been investigated for ER in Big Data environment.
RQ2. What methods, techniques and tools have been used for ER in Big Data environment?	Determine if the proposed research works in this field are more practical or theoretical and identify opportunities for future research works.
RQ3. What is the nature of found methods, techniques and tools for solving ER in Big Data environment?	Identify the nature of found methods, techniques and tools for ER in Big Data environment and assess the state of this field.
RQ4. What are the objectives pursued in research works for solving the ER in Big Data environment?	Point out what the major point of research interest is and which areas have been less investigated, by exploring concepts, compiling current knowledge or advancing the practice through the science of design practices.

The objective of this SMS is to identify what has been already done and what must be done in the future in the context of ER in big data sources. It is completely different to the reviews performed until today. This SMS differs from the revisions presented earlier in the related work section in four aspects: (i) the goal is different, (ii) some works are based on a specific area and others are more general not covering all the areas, (iii) this revision is broader and more systematic (considering that only one of the existing works conducts a systematic review) and the classification of primary studies is more exhaustive and (iv), we have not found recent surveys or systematic studies related to this domain that do not focus on a unique method, technology or technique. Thus, an extension of the work carried out up to date is proposed in this work.

3.1.2. Formulating research questions

Fulfilling the objective of understanding the existing research proposals within ER problem in big data sources, it was necessary to formulate some research questions (RQ). RQs will guide and center our research and they will be clearly focused on the topic, as well as they will synthesize multiple sources to present our unique argument. Table 2 lists the RQs proposed for this SMS together with their motivations.

3.1.3. Defining the review protocol

Now, once the necessity of undertaking this research work has been identified and the research questions that guide it has been formulated, we will describe each of the elements of the protocol defined for this SMS defining: search strategy, procedure for selection of studies, checklists and procedure for evaluating the quality of studies, data extraction strategy, data synthesis, dissemination strategy and project calendar.

3.1.3.1. Search strategy. This section describes the method that has let us deeply search the most relevant papers related to the topic that we are working on in the principal digital libraries. The searching strategy has been divided into three phases: pre-search, systematic search and manual search.

In the pre-search phase, keywords for the search were selected. As this selection was known to be relevant for the quality of results, general terms have been used with the aim of confirming that most of the research papers are included in the study. We have classified these terms in two main categories: problem, and technologies, tools, frameworks and concepts. The problem category is based on the ER problem, having this key as the main one and getting all the synonyms identified in the pre-search that refer to this problem. The technologies, tools, frameworks and concepts category is based on the domain where we tend to apply the category of the problem. After that, a combination between both categories and all the sets of words identified for each one was carried out. The initial list of words is shown in Table 3.

The first set of databases were taken according to the criteria presented by Ngai, Hu, Wong, Chen, and Sun (2011b) adding some other new ones proposed by the authors of this paper. These are: ABI/INFORM Database, Academic Search Premier, ACM, Business Source Premier, Emerald Full text, IEEE Xplore Digital Library,

Table 3
First set of keywords giving main terms.

Concept	Keywords
ER problem	Entity Matching. Entity Identity, Entity Name System, Entity Recognition, Entity Parsing, Entity Linking, Entity disambiguation, Entity Resolution, Entity Reconciliation, Identity Matching, Identity Management, Identity Resolution, Identity Attributes, Identity Search, Identity Linking, Duplicate Detection, Deduplication, Record Linkage, Object Identification, Reference Matching, Co-Reference Detection, Non-identical Duplicates, Redundancy elimination, Object Matching, Fuzzy Matching, Similarity join processing, Duplication Detection, Reference Reconciliation, Co-Reference Resolution, Relational Blocking
Technologies, tools, frameworks and concepts	Data Integration, Heterogeneous Data Sources, Data Sources, Data warehouse, Unstracted data, Inter-media data retrieval, ETL, Extract transform load, Extract transform and load, Big Data, Open Data, Database Management, Data quality, DSL, Domain specific language, Massive Data, Large Data, MDE, Model Driven Engineering

Science Direct, Springer-Link Journals, World Scientific Net, SCOPUS and Web Of Knowledge. Once the searching process was finished, it was realized that some databases included articles already found in others, therefore, they did not bring new value. Thus, the articles were grouped into four large databases: ACM, IEEE Xplore Digital Library, SCOPUS and Web Of Knowledge.

In the systematic search of phase two, once the relevant keywords have been found and some pilot testing was carried out, a Python script was developed for making the combination between all of them. In this context, two category files were created: one for the ER problem and another one for the technologies, tools, frameworks and concepts. Having these two files, the script was programed by taking one of the keywords of the ER problem file and combining it with all of the keywords of the second file. Besides, search queries were generated concretely for each database selected to conduct the systematic search. All kind of papers have been included such as: journal papers and presentations at conferences, congresses, tutorials and workshops. A very large number of queries have been executed and for each database they have been customized depending on: the query syntax of the relevant database, the possibilities that the database offers to make filters, year of publication and specific topics. Table 4 shows some examples of the queries that have been executed.

Once the queries for each database were created, a new specific Python script was designed for each one. Besides, Selenium, a software testing tool for Web-based applications (Selenium, 2017) was used. The process of searching a paper in each database was replicated and it was automated for getting the results based on the queries created before using this Python script and Selenium. Finally, another Python script was developed for removing the duplicate records found out during the process of search. Because the process of analysis of the results obtained was quite long over time, the searching process previously described was repeated twice.

In the last phase of manual search, papers recommended by experts in the ER problem were looked for. These papers were very important because they were very close to the topic and we could discard them because of the problem of bias.

The Web version of the application Mendeley was used for managing this amount of data. It is a reference manager tool that helps to handle papers. Mendeley is integrated into the Web browser, allowing adding directly the articles from the digital libraries to a personal document database, avoiding the duplicated ones and saving them (when possible) in PDF format (Mendeley Support Team, 2011).

3.1.3.2. Study Selection, inclusion and exclusion criteria and quality insurance. This SMS includes papers written in English that refer to ER problems and technologies, tools or frameworks that try to solve this problem, published from 2010 up to January 2017 in indexed journals, such as Journal Citation Reports (JCR) and prestigious conferences, congresses or workshops categorized in the CORE ranking (CORE Conference Ranking).

It excludes discussion or opinion papers or those that are only available in PowerPoint or abstract formats, duplicates (always considering the most completed one) and those whose main contribution is not referred to ER problems and technologies, tools or frameworks that try to solve it or just scarcely mention it.

The first filter for selecting primary studies was based on the title and abstract of the paper. If it is not relevant to the study, it is automatically excluded. After this process, the inclusion/exclusion criteria were applied when reading the abstracts of the found items. Once read, if there was still any doubt with the inclusion/exclusion criteria, the paper was completely read. The first author of the work conducted the selection of the studies and the second author randomly collected 30% of the articles to corroborate if the inclusion/exclusion criteria were applied correctly. He/she would consult the other authors in case of doubts or discrepancies.

3.1.3.3. Data collection and analysis. First, a quantitative synthesis considering the number and/or percentage of items in each category was made, illustrating them with tables and graphics, to thereby give an answer to each research question, matching each question with category. Moreover, an interpretation of retrieved results and some suggestions deduced from the synthesis are presented.

In addition, it was analyzed: (i) the number of publications per year to detect and justify trends and (ii) the number of publications by publication type to detect the journals in which more has been.

3.1.4. Validating the review protocol

The protocol was reviewed by the authors of this research work to ensure that all relevant aspects were taken into account to achieve the objectives of this SMS, also considering the recommendations provided by Kitchenham and Brereton (2013).

3.2. Conducting

Once the protocol was agreed, the proper study started. There were two main sections during the process of carrying out this SMS: (i) detect and select primary studies and data extraction, and (ii) apply the inclusion and exclusion criteria for selecting the primary studies that will be used for the work, showing the finally selected ones and the data synthesis phase, where a statistical study was conducted. It showed the main conclusions that obtained after running the previous phase.

3.2.1. Detect and select primary studies and data extraction

Papers published between 2010 and 2017 were found using the search strategy defined in the protocol. Because of the limitations that certain search sources offered (for example, not allowing the use of complex search strings), it was necessary to design specific strings for each source and manipulate the outcome of searches to

Table 4
Example of queries.

	Database	Keywords
Query 1	Scopus	2010 (TITLE("Data fusion") AND KEY("Entity Matching")) OR (TITLE("Data fusion") AND KEY("Entity Identity")) OR (TITLE("Data fusion") AND KEY("Entity Name System")) OR (TITLE("Data fusion") AND KEY("Entity Recognition")) OR (TITLE("Data fusion") AND KEY("Entity Parsing")) OR (TITLE("Data fusion") AND KEY("Entity Linking")) OR (TITLE("Data fusion") AND KEY("Entity disambiguation")) OR (TITLE("Data fusion") AND KEY("Entity Resolution")) OR (TITLE("Data fusion") AND KEY("Entity Reconciliation")) OR (TITLE("Data fusion") AND KEY("Identity Matching")) OR (TITLE("Data fusion") AND KEY("Identity Management")) OR (TITLE("Data fusion") AND KEY("Identity Resolution")) OR (TITLE("Data fusion") AND KEY("Identity Attributes")) OR (TITLE("Data fusion") AND KEY("Identity Search")) OR (TITLE("Data fusion") AND KEY("Identity Linking")) OR (TITLE("Data fusion") AND KEY("Duplicate Detection")) OR (TITLE("Data fusion") AND KEY("Deduplication")) OR (TITLE("Data fusion") AND KEY("Record Linkage")) OR (TITLE("Data fusion") AND KEY("Object Identification")) OR (TITLE("Data fusion") AND KEY("Reference Matching")) OR (TITLE("Data fusion") AND KEY("Co-Reference Detection")) OR (TITLE("Data fusion") AND KEY("Non-identical Duplicates")) OR (TITLE("Data fusion") AND KEY("Redundancy elimination")) OR (TITLE("Data fusion") AND KEY("Object Matching ")) OR (TITLE("Data fusion") AND KEY("Fuzzy Matching")) OR (TITLE("Data fusion") AND KEY("Similarity join processing")) OR (TITLE("Data fusion") AND KEY("Duplication Detection")) OR (TITLE("Data fusion") AND KEY("Reference Reconciliation")) OR (TITLE("Data fusion") AND KEY("Co-Reference Resolution")) OR (TITLE("Data fusion") AND KEY("Relational Blocking"))
Query 2	ACM	2010 ("Title":"data fusion" AND "Title":"entity matching") OR+("Title":"data fusion" AND "Title":"entity identity") OR+("Title":"data fusion" AND "Title":"entity name system") OR+("Title":"data fusion" AND "Title":"entity recognition") OR+("Title":"data fusion" AND "Title":"entity parsing") OR+("Title":"data fusion" AND "Title":"entity linking") OR+("Title":"data fusion" AND "Title":"entity disambiguation") OR+("Title":"data fusion" AND "Title":"entity resolution") OR+("Title":"data fusion" AND "Title":"entity reconciliation") OR+("Title":"data fusion" AND "Title":"identity matching") OR+("Title":"data fusion" AND "Title":"identity management") OR+("Title":"data fusion" AND "Title":"identity resolution") OR+("Title":"data fusion" AND "Title":"identity attributes") OR+("Title":"data fusion" AND "Title":"identity search") OR+("Title":"data fusion" AND "Title":"identity linking") OR+("Title":"data fusion" AND "Title":"duplicate detection") OR+("Title":"data fusion" AND "Title":"deduplication") OR+("Title":"data fusion" AND "Title":"record linkage") OR+("Title":"data fusion" AND "Title":"object identification") OR+("Title":"data fusion" AND "Title":"reference matching") OR+("Title":"data fusion" AND "Title":"co-reference detection") OR+("Title":"data fusion" AND "Title":"non-identical duplicates") OR+("Title":"data fusion" AND "Title":"redundancy elimination") OR+("Title":"data fusion" AND "Title":"object matching ") OR+("Title":"data fusion" AND "Title":"fuzzy matching") OR+("Title":"data fusion" AND "Title":"similarity join processing") OR+("Title":"data fusion" AND "Title":"duplication detection") OR+("Title":"data fusion" AND "Title":"reference reconciliation") OR+("Title":"data fusion" AND "Title":"co-reference resolution") OR+("Title":"data fusion" AND "Title":"relational blocking")
Query 3	IEEE	2010 ("Document Title":"Data fusion" AND "Document Title":"Entity Matching") OR ("Document Title":"Data fusion" AND "Document Title":"Entity Identity") OR ("Document Title":"Data fusion" AND "Document Title":"Entity Name System") OR ("Document Title":"Data fusion" AND "Document Title":"Entity Recognition") OR ("Document Title":"Data fusion" AND "Document Title":"Entity Parsing") OR ("Document Title":"Data fusion" AND "Document Title":"Entity Linking") OR ("Document Title":"Data fusion" AND "Document Title":"Entity disambiguation") OR ("Document Title":"Data fusion" AND "Document Title":"Entity Resolution") OR ("Document Title":"Data fusion" AND "Document Title":"Entity Reconciliation") OR ("Document Title":"Data fusion" AND "Document Title":"Identity Matching") OR ("Document Title":"Data fusion" AND "Document Title":"Identity Management") OR ("Document Title":"Data fusion" AND "Document Title":"Identity Resolution") OR ("Document Title":"Data fusion" AND "Document Title":"Identity Attributes") OR ("Document Title":"Data fusion" AND "Document Title":"Identity Search") OR ("Document Title":"Data fusion" AND "Document Title":"Identity Linking") OR ("Document Title":"Data fusion" AND "Document Title":"Duplicate Detection") OR ("Document Title":"Data fusion" AND "Document Title":"Deduplication") OR ("Document Title":"Data fusion" AND "Document Title":"Record Linkage") OR ("Document Title":"Data fusion" AND "Document Title":"Object Identification") OR ("Document Title":"Data fusion" AND "Document Title":"Reference Matching") OR ("Document Title":"Data fusion" AND "Document Title":"Co-Reference Detection") OR ("Document Title":"Data fusion" AND "Document Title":"Non-identical Duplicates") OR ("Document Title":"Data fusion" AND "Document Title":"Redundancy elimination") OR ("Document Title":"Data fusion" AND "Document Title":"Object Matching ") OR ("Document Title":"Data fusion" AND "Document Title":"Fuzzy Matching") OR ("Document Title":"Data fusion" AND "Document Title":"Similarity join processing") OR ("Document Title":"Data fusion" AND "Document Title":"Duplication Detection") OR ("Document Title":"Data fusion" AND "Document Title":"Reference Reconciliation") OR ("Document Title":"Data fusion" AND "Document Title":"Co-Reference Resolution") OR ("Document Title":"Data fusion" AND "Document Title":"Relational Blocking")
Query 4	Web Of Knowledge	2010 (Data fusion AND Entity Matching) OR (Data fusion AND Entity Identity) OR (Data fusion AND Entity Name System) OR (Data fusion AND Entity Recognition) OR (Data fusion AND Entity Parsing) OR (Data fusion AND Entity Linking) OR (Data fusion AND Entity disambiguation) OR (Data fusion AND Entity Resolution) OR (Data fusion AND Entity Reconciliation) OR (Data fusion AND Identity Matching) OR (Data fusion AND Identity Management) OR (Data fusion AND Identity Resolution) OR (Data fusion AND Identity Attributes) OR (Data fusion AND Identity Search) OR (Data fusion AND Identity Linking) OR (Data fusion AND Duplicate Detection) OR (Data fusion AND Deduplication) OR (Data fusion AND Record Linkage) OR (Data fusion AND Object Identification) OR (Data fusion AND Reference Matching) OR (Data fusion AND Co-Reference Detection) OR (Data fusion AND Non-identical Duplicates) OR (Data fusion AND Redundancy elimination) OR (Data fusion AND Object Matching) OR (Data fusion AND Fuzzy Matching) OR (Data fusion AND Similarity join processing) OR (Data fusion AND Duplication Detection) OR (Data fusion AND Reference Reconciliation) OR (Data fusion AND Co-Reference Resolution) OR (Data fusion AND Relational Blocking)

get the same results that may have been obtained using the original search string. The search was made on the title and abstract of the papers except for those databases that did not allow this. In that case, the search had to be performed in the full text.

Each search source stored search strings, metadata of found items (title, author, year of publication, etc.) and abstracts of the papers.

After reading their abstracts and excluding those irrelevant to the ER problem, 276 papers out of 2255 were eliminated for being duplicates. Then, according to the range of years that we have chosen, 434 papers that were written before 2010 were also eliminated.

Consequently, the inclusion/exclusion criteria were applied to the 1545 remaining items, and 1024 papers that were not classified into the computer science or information systems category were eliminated. The last filter was applied to the Big Data area and 382 papers were discarded, remaining 139 candidates. From them, 72 papers were supposed to be duplicated, remaining 67

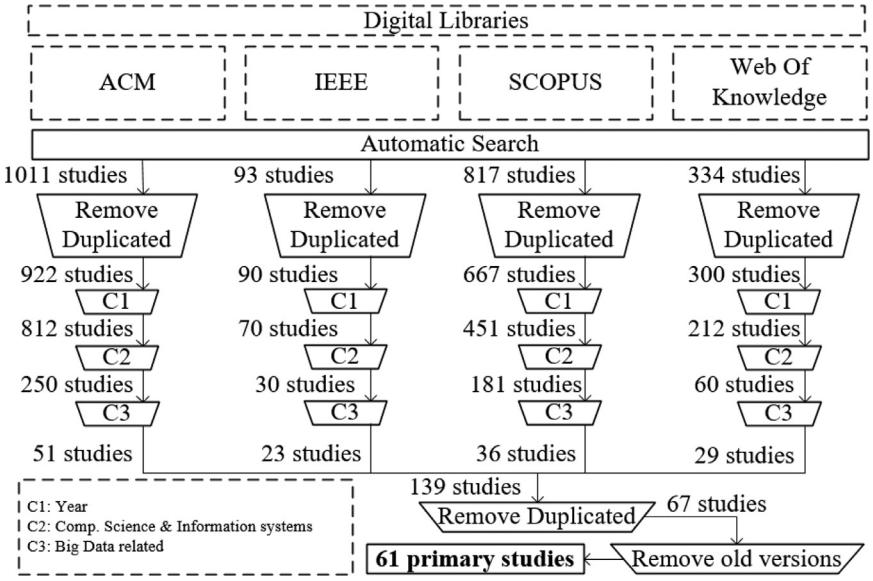

Fig. 2. Primary studies Selection Process.

papers. Then, 6 old versions were also eliminated and finally, 61 primary studies were analyzed in depth reading the full text. As it was classified in the protocol, primary studies were identified and selected by the first author of the article, while the second author chose randomly 30% to corroborate the correct choice. The doubts that arose during the selection of items were resolved among all the authors. Table 5 shows the 61 primary studies selected and Fig. 2 summarizes the process carried out.

3.2.2. Data synthesis

In this section, the information contained in the data extraction form is displayed in order to answer the research questions formulated previously. In addition to the quantitative data shown through tables and graphics, an interpretation of the results is also presented.

- RQ1. As reflected in Table 6, a classification that divides the publications in those based on algorithms as solution and those data structured-based is presented. There is a total of 68 publications (seven more than the number of primary studies because there are some papers that mention both categories). Studying data, the percentage of publications is very similar in both categories, having 48.53% the structural ones, the 50% the algorithmic ones and just 1.47% represents the others.
- RQ2. As shown in Table 7, it is interesting to note that most research works have been validated with a theoretical approach (defining a theoretical approach as that which has validated its proposals with any dataset). They represent 95.08%. Two papers (3.28%) do not present any validation and just one (1.64%) presents a validation based on a real-industry scenario.

Table 8 presents the datasets used by the authors for validating their approaches. Most of the proposals have been validated with real datasets (76.74%), followed by those which have used both real and synthetic datasets (22.95%) and finally those which have used synthetic datasets (14.75%). It is important to note that there are 77 papers because there are two papers that do not present any validation and 9 of them include the two types of datasets.

- RQ3. As shown in Table 9, most research efforts have been focused on graph-based works (26.23%), followed by those based on Clustering/Blocking (22.95%). It also highlights rule-based works (14.75%), and those based on algorithms (16.39%) and probabilistic methods (11.48%). There are two categories based on programming languages and ontologies that represent 4.92%, as well as the learning category that represents 3.28%. Finally, there are three categories based on hints, sorted neighborhood and patterns that represent 1.64%.
- RQ4. As shown in Table 10, all of the primary studies are based on the operation phase in contrast to the phase of design that only takes 4.92% (three studies present both design and operation phases). More than half of the selected studies (62.30%) apply their experiments to heterogeneous data sources. The lowest result is on multi-applications, where one study that represents 1.64% is mentioned. Finally, automation and multi-domain objectives are poorly represented with only 3.28% and 8.20%, respectively, and just 12 studies that represent 19.67%, mention the multi-relational objective.

Once the research questions have been answered and after an in-depth study of the retrieved data, some other conclusions are presented.

At it is observed in Table 11, we can conclude that the topic that we are analyzing in this SMS is arising a lot of interest. From 2010 up to date, the numbers of papers related to ER have been increasing (omitting year 2012 where just one paper less than in 2010 was published). The growth curve between 2012 and 2014 is quite large, almost tripling the number of publications. The number of publications in 2015 remains constant with respect to those published in 2014 and finally, it increases in one more publication having 15 in total. Moreover, Table 12 summarizes the evolution of the publications based on its category and the year of publication. It shows a clearest trend in this area of research is focused on graph-based methods, techniques or tools followed by those clustering/blocking-based. Those learning-based are the most scattered, finding only one publication in the beginning of the search period and another one in the end. Those sorted neighborhoods and pattern-based but in this case, they are placed at end of the search period.

Table 13 shows the result that has been retrieved from the different digital libraries. In this case, the ACM Digital Library is on top of the selected primary studies with 36.69% followed by Scopus with 25.90%, Web of Knowledge with 20.86%, and finally IEEE Xplore Digital Library with 16.55%. It is important to remark that

Table 5
Selected primary studies.

Title	Reference
Entity resolution for distributed probabilistic data	(Ayat, Akbarinia, Afsarmanesh, & Valduriez, 2013)
Incremental entity resolution on rules and data	(Whang & Garcia-Molina, 2014)
Efficient entity resolution based on subgraph cohesion	(Wang, Li, & Gao, 2015)
Domain-specific entity extraction from noisy, unstructured data using ontology-guided search	(Bratus, Rumshisky, Khrabrov, Magar, & Thompson, 2011)
Entity resolution for probabilistic data	(Ayat, Akbarinia, Afsarmanesh, & Valduriez, 2014)
Entity resolution based EM for integrating heterogeneous distributed probabilistic data	(Dharavath & Kumar, 2015)
Pay-As-You-Go Entity Resolution	(Whang, Marmaros, & Garcia-Molina, 2013)
Interaction between Record Matching and Data Repairing	(Fan, Li, Ma, Tang, & Yu, 2011)
Conflict Resolution with Data Currency and Consistency	(Fan, Geerts, Tang, & Yu, 2014)
Information Fusion for Entity Matching in Unstructured Data	(Ali & Cristianini, 2010)
Dynamic Sorted Neighborhood Indexing for Real-Time Entity Resolution	(Ramadan, Christen, & Liang, 2014)
Disambiguation of named entities in cultural heritage texts using linked data sets	(Brando, Frontini, & Ganascia, 2015)
Adaptive Connection Strength Models for Relationship-Based Entity Resolution	(Nuray-turan, Kalashnikov, & Mehrotra, 2013)
Context-based Entity Description Rule for Entity Resolution	(Li, Li, Wang, & Gao, 2011)
Efficient and Effective Duplicate Detection in Hierarchical Data	(Leitaõ, Calado, & Herschel, 2013)
Entity Disambiguation in Anonymized Graphs Using Graph Kernels	(Hermansson, Kerola, Johansson, Jethava, & Dubhashi, 2013)
HIL: A High-Level Scripting Language for Entity Integration	(Hernández & Koutrika, 2013)
A Clustering-Based Framework to Control Block Sizes for Entity Resolution	(Fisher, Christen, Wang, & Rahm, 2015)
A Probabilistic Model for Linking Named Entities in Web Text with Heterogeneous Information Networks	(Shen, Han, & Wang, 2014)
A Scalable Machine-Learning Approach for Semi-Structured Named Entity Recognition	(Irmak & Kraft, 2010)
BEAR: Block Elimination Approach for Random Walk with Restart on Large Graphs	(Shin, Jung, Lee, & Kang, 2015)
Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data	(Papadakis, Ioannou, Niederée, Palpanas, & Nejdl, 2012)
Efficient Entity Resolution for Large Heterogeneous Information Spaces	(Papadakis, Ioannou, Niederée, & Fankhauser, 2011a)
Efficient SPectrAl Neighborhood blocking for entity resolution	(Shu, Chen, Xiong, & Meng, 2011)
Entity Linking on Graph Data	(Yu, 2014)
Entity Matching across Heterogeneous Sources	(Yang, Sun, Tang, Ma, & Li, 2015)
Entity type recognition for heterogeneous semantic graphs	(Sleeman & Finin, 2013)
A load-balanced mapreduce algorithm for blocking-based entity-resolution with multiple keys	(Hsueh, Lin, & Chiu, 2014)
Web-based Graphical Querying of Databases through an Ontology: the WONDER System	(Calvanese, Keet, Nutt, Rodríguez-Muro, & Stefanoni, 2010)
Large-Scale entity resolution for semantic web data integration	(G. D. A. Costa, 2016)
Populating Entity Name Systems for Big Data Integration	(Kejriwal, 2014)
Domain-adapted named-entity linker using Linked Data	(Frontini et al., 2015)
Learning-based entity resolution with MapReduce.	(Kolb, Köpcke, Thor, Databases, & Systems, 2011)
ERGP: A combined entity resolution approach with genetic programming	(Sun, Shen, Kou, Nie, & Yu, 2014a)
Learning an accurate entity resolution model from crowdsourced labels	(Wang, Oyama, Kurihara, & Kashima, 2014)
Entity resolution for high velocity streams using semantic measures	(Priya, Prabhakar, & Vasavi, 2015)
A confidence-based entity resolution approach with incomplete information	(Gu, Zhang, Cao, Xu, & Cuzzocrea, 2014)
A framework for entity resolution with efficient blocking	(Shu et al., 2012)
Entity matching: A case study in the medical domain	(Carvalho, Laender, & Meira, 2015)
An Identification Ontology for Entity Matching	(Bortoli, Bouquet, & Bazzanella, 2014)
DS-Dedupe: A scalable, low network overhead data routing algorithm for inline cluster deduplication system	(Sun, Xiao, Liu, & Fu, 2014b)
A fast entity resolution method based on wave of records	(Liu, Wang, & Gao, 2011)
Cleaning Framework for Big Data - Object Identification and Linkage	(Liu, Kumar, & Thomas, 2015)
To compare or not to compare: making entity resolution more efficient	(Papadakis, Ioannou, Niederée, Palpanas, & Nejdl, 2011b)
Entity Resolution for High Velocity Streams Using Semantic Measures	(Priya et al., 2015)
An Ensemble Blocking Scheme for Entity Resolution of Large and Sparse Datasets	(Balaji et al., 2016)
Unsupervised Entity Resolution on Multi-type Graphs	(Zhu, Ghasemi-gol, Szekely, Galstyan, & Knoblock, 2016)
Entity Matching Across Multiple Heterogeneous Data Sources	(Kong, Gao, Xu, Quian, & Zhou, 2016)
Efficient Entity Resolution on Heterogeneous Records	(Lin, Wang, Li, & Gao, 2016)
Linked Data Entity Resolution System Enhanced by Configuration Learning Algorithm	(Nguyen & Ichise, 2016)
Linking Heterogeneous Data in the Semantic Web Using Scalable and Domain-Independent Candidate Selection	(Song, Luo, & Heflin, 2016)
Using Memetic Algorithm for Instance Coreference Resolution	(Xue & Wang, 2016)
Rule-Based Method for Entity Resolution	(Li, Li, & Gao, 2015
Entity resolution in disjoint graphs: an application on genealogical data	(Rahmani, Ranjbar-Sahraei, Weiss, & Tuyls, 2016)
Parallel Meta-blocking for Scaling Entity Resolution over Big Heterogeneous Data	(Efthymiou et al., 2016)
Minoan ER: Progressive Entity Resolution in the Web of Data	(Efthymiou, Stefanidis, & Vassilis, 2016)
Entity resolution in disjoint graphs: an application on genealogical data	(Rahmani et al., 2016)
Semantic-Aware Blocking for Entity Resolution	(Wang, Cui, & Liang, 2016)
Online entity resolution using an Oracle	(Firmani, Saha, & Srivastava, 2016)
Entity Resolution-Based Jaccard Similarity Coefficient for Heterogeneous Distributed Databases	(Dharavath & Singh, 2016)
A Blocking Scheme for Entity Resolution in the Semantic Web	(de Assis Costa & de Oliveira, 2016)

Table 6
RQ1 – Data synthesis.

Type	Number	Percentage
Structural	33	48.53%
Algorithmic	34	50%
Others	1	1.47%

Table 7
RQ2 – Data synthesis (validation).

Validation	Number	Percentage
Not Validated – Theoretical Approach	2	3.28%
Validated – Theoretical Approach	58	95.08%
Validated – Approach in Industry	1	1.64%

Table 8
RQ2 – Data synthesis (datasets).

Dataset	Number	Percentage
Real	54	76.74%
Synthetic	14	22.95%
Real + Synthetic	9	14.75%

Table 9
RQ3 – Data synthesis.

Method, Technique, Tools	Number	Percentage
Rule-based	9	14.75%
Probabilistic Method	7	11.48%
Learning-based	3	4.92%
Graph-based	16	26.23%
Programming Languages	2	3.28%
Clustering/Blocking-Based	14	22.95%
Ontology	3	4.92%
Patterns	1	1.64%
Sorted Neighborhood	1	1.64%
Algorithms	10	16.39%
Hints	1	1.64%

Table 10
RQ4 – Data synthesis.

	Objective	Number	Percentage
UML	Design	3	4.92%
	Operation	61	100%
Challenges	Automation	2	3.28%
	Multi-Relational	12	19.67%
	Multi-Domain	5	8.20%
	Multi-Applications	1	1.64%
Type of Dataset	Heterogeneous	38	62.30%
	Non-heterogeneous	23	37.70%

Table 11
RQ3 – Data synthesis.

Year	Number	Percentage
2010	3	4.92%
2011	7	11.48%
2012	2	3.28%
2013	6	9.84%
2014	14	22.95%
2015	14	22.95%
2016	15	24.59%

the amount of papers of this table is higher because the duplications among databases were not eliminated (Figs. 3–5).

3.3. Reporting

The main threats to validate this work follow the ones proposed in Shull, Singer, and Sj??berg (2008): bias in the selection of articles, inaccuracy in data extraction and errors that could be taken through the process of classification.

It is impossible to achieve full coverage of everything written on a topic. Four digital research databases were used, including journals, conferences and relevant workshops related to the ER topic. The scope of journals and conferences that has been discussed in this SMS is large enough to reach a reasonable completeness in the studied field.

Helping to ensure a fair selection process, the research questions were defined in advance and the selection of items was organized in a series of stages in which all authors of the article were involved. As discussed above, the decision to select primary studies for this SMS was made among the researchers involved in this paper and rules were rigorously enforced.

Duplication of articles is a potential threat to calculate the frequency of articles and statistical data. As it is also discussed above, an automatic process was applied to remove the duplicated publications, therefore we do not believe that any undetected duplications exist.

4. Discussion

This SMS discovered 61 primary studies classified in peer-reviewed journals, conferences and workshops. They were classified in 4 sections represented by the 4 research questions proposed in Section 3. In this section, we will discuss the obtained results.

RQ1 asked: "What methods, techniques or tools have been investigated for ER in the Big Data environment?" Following Unified Modeling Language (UML) specification (Group, 2017) that classifies diagrams in two categories: (i) structure-based diagrams, which show the static structure of the system and its parts on different abstraction and implementation levels and how they are related to each other, (ii) and behavior-based diagrams, which show the dynamic behavior of the objects in a system extrapolating it to our problem, we have organized the selected primary studies in two big groups with the aim of finding out what methods, techniques or tools have been investigated: (i) structural-based and algorithmic-based solutions, understanding structural-based as those studies that propose a solution supported by data structures, (ii) and algorithmic-based solutions, which refer to those studies in which the solution comes from applying an algorithm. Results show that the structural-based solutions are a little bit important than the algorithmic-based ones with a difference of just 1.57%. It is also relevant to highlight that there are some papers that represent both structural-based and algorithmic-based solutions. With this analysis, it is concluded that the efforts invested by the researches in both solutions is very similar.

RQ2 asked: "What methods, techniques and tools have been used for ER in Big Data environment?" We have made a classification based on solutions that have not been validated (present a theoretical approach) and solutions that have been validated (either with own experiments or in the industry), in order to determine whether the research study is more practical or theoretical. Besides, for the solutions that have been validated, we have classified the type of datasets that was used for the validation as follows: real (real-world dataset), synthetic (non-real-world dataset) and real + synthetic (those which have validated their approach with both types of datasets).

The selection criteria for the selected studies have been very strict, trying to obtain very good quality publications. Therefore, as expected in this type of publications, most studies show a validation. Only one of them presents a theoretical solution which does not mention whether the proposal has been validated or not. Some others show a theoretical validation and only one of them shows a real-case validation performed in industry. Furthermore, most of the studies have been validated using real-world datasets, under-

Table 12
RQ3 – Data synthesis.

Year	Rules	Probabilistic	Learning	Graph	Programming	Clustering blocking	Ontology	Patterns	Sorted neigh.	Algorithm	Hints
2010	0	0	1	2	0	0	0	0	0	0	0
2011	2	0	0	0	0	3	1	0	0	1	0
2012	0	0	0	0	0	2	0	0	0	0	0
2013	0	2	0	2	1	0	0	0	0	0	1
2014	3	3	1	1	1	2	1	0	0	3	0
2015	4	1	0	4	0	3	0	1	1	2	0
2016	0	1	1	7	0	4	1	0	0	4	0
Totals	9	7	3	16	2	14	3	1	1	10	1

Table 13
RQ3 – Data synthesis.

Library	Total	Percentage
ACM	51	36.69%
IEEE	23	16.55%
SCOPUS	36	25.90%
WOK	29	20.86%

standable when researchers aim to provide their proposals with a certain level of quality, and just a few studies show a validation with synthetic datasets and real-world + synthetic datasets.

RQ3 stated: "What is the nature of the methods, techniques and tools found for solving ER in Big Data environment?" A new subdivision of the classification obtained in RQ1 has been created. This subdivision presents: (i) for algorithmic-based solutions: rule-based, probabilistic-based methods, learning-based, programming languages, sorted neighborhood and others algorithms, (ii) for structure-based solutions: graph-based, clustering/blocking-based, ontology-based and pattern-based (iii) and finally, for others: hints.

Taking into account the previous classification, the ones which take more importance are graph-based and clustering/blocking-based solutions, representing 49.18% of the total. This is because when working with big datasets, the computational time is key and blocking is often used to improve efficiency and graph technology is a natural solution to treat problems related to Big Data and especially for the relationships among entities. Moreover, those solutions based on rules and probabilistic methods, often used since the beginning of the study of this problem in relational databases, are also important.

Finally, RQ4 asked: "What are the objectives pursued in research work for solving ER in Big Data environment?" . With the aim to determine where the majority of research interest is and which areas have been little investigated, we have made a classification based on: taking the extrapolation mentioned before of UML, design and operation, Big Data challenges for ER proposed in Getoor and Machanavajjhala, (2013), multi-relational, dealing with structure of entities, multi-domain, dealing with customizable methods that span across domains and multi-applications, dealing with systems that serve diverse application with different accuracy requirements, level of automation of the proposal, and finally the types of the that were used for the validation of the proposals, understanding them as heterogeneous or non-heterogeneous.

Results show that most of the proposals focus on the implementation phase that provides a solution to a problem, however only three studies refer to design. It is important to note that in the challenges proposed in Getoor and Machanavajjhala, (2013), a great research effort has not been made since the best score (19.67% of studies) obtained is related to proposals that address the multi-relational objective. Nevertheless, it is not very significant because it is clearly related to graph-based studies. Only 5 studies work with different domains and it is surprising that none of the papers found mention anything about servicing to different applications with different requirements. Moreover, it is also quite relevant that only two of the works found show some level of automation in its solutions. Finally, the heterogeneity of the datasets is acceptable as more than half of the works found (62.30%) use heterogeneous data sources.

5. Conclusions and future works

This paper emerges from the necessity of the Regional Government of Andalusia (Spain) to look for a solution that may solve the ER problem. This problem lies in the difficulty that this Government has to manage all the information available coming from nonofficial data source, that store information related to the cultural heritage patrimony, and such information does not match with the one stored in MOSAICO (official information system for managing this kind of information). To achieve this goal, this paper has presented a SMS of methods, techniques or tools that provide solutions to ER problems in Big Data sources.

From the experts' systems point of view and considering ER as an operational intelligence process, a meaningful number of proposals that provide a solution to this problem has been identified and categorized in different domains such as: rule and learning-based techniques, probabilistic methods or ontologies, among others.

This SMS is composed of three phases: (i) planning the review, where the necessity of making this research has been demonstrated taking into account that the goal of papers presented in related work is different; some were based in one topic and others in different topics, but they did not cover all the proposed fields in this work and just one paper made a systematic process similar than the presented one, (ii) conducting the review, where the protocol defined was executed, (iii) and reporting results. After managing the defined protocol, the review was conducted, where a total of 61 primary studies were selected.

A comparison table was created in order to classify the primary studies. The tags identified for classifying them were created according to the specifications of the problem provided by the Regional Government of Andalusia. Criteria such as design or operation phase of the ER problem, level of automation, heterogeneity of datasets used for validation or multi-relational characteristics, among others, have been considered.

The analysis of the results shows that research efforts in structural-based and algorithmic-based solutions are practically the same. Most of the studies have been validated using real-world datasets, but just one of them has applied its proposal in a real case in the industry. The heterogeneity of the datasets is acceptable knowing that more than a half uses heterogeneous data sources. Besides, most of the research work has been focused on the operation phase of the reconciliation and not in the design phase. Finally, the efforts made to automate the process of reconciliation have been very limited.

According to the obtained results and as future research work, it is proposed to extrapolate the solutions to the problems raised in this research area to real problems in the industry, as it has been

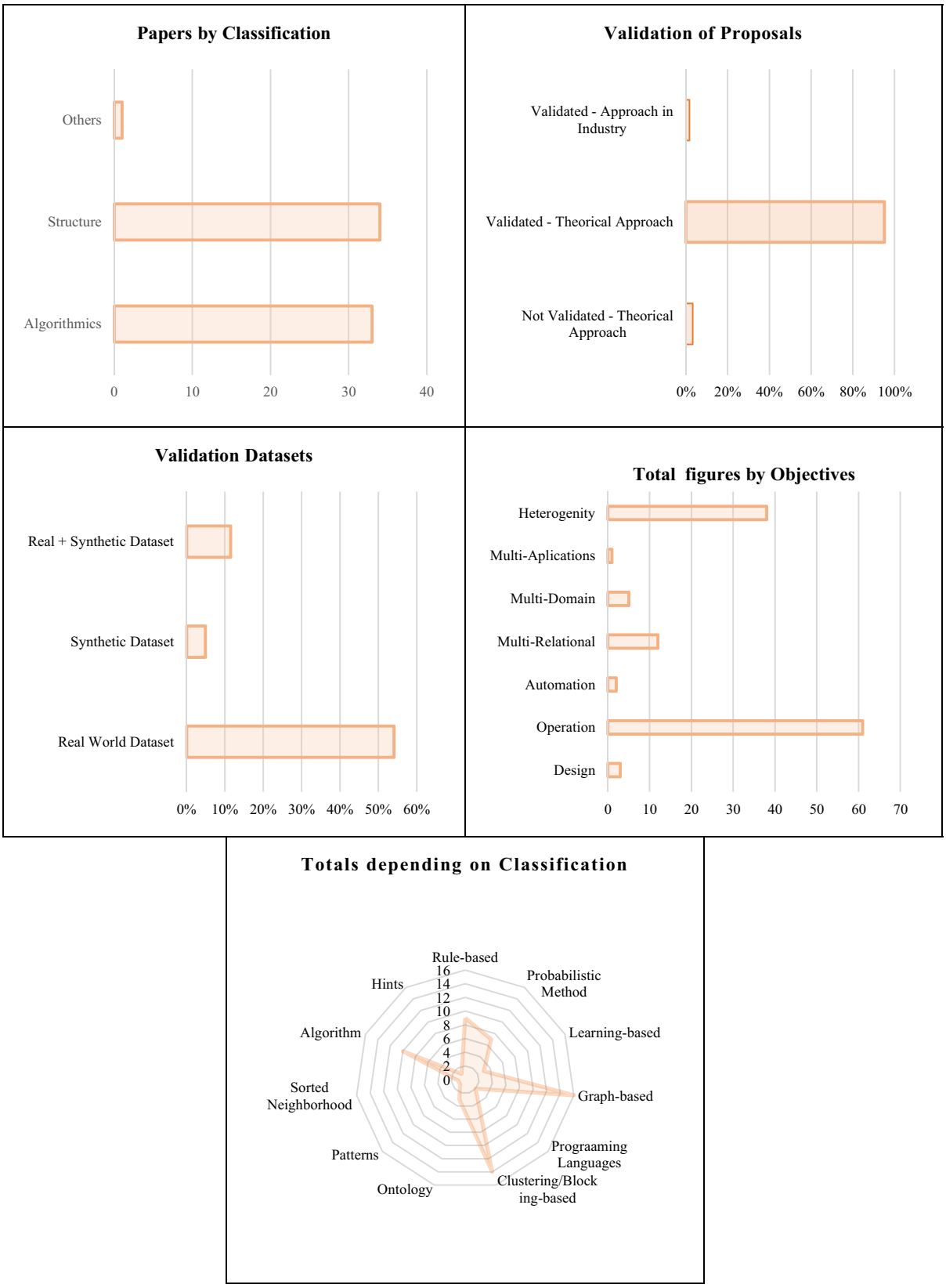

Fig. 3. Data Synthesis Graphics.

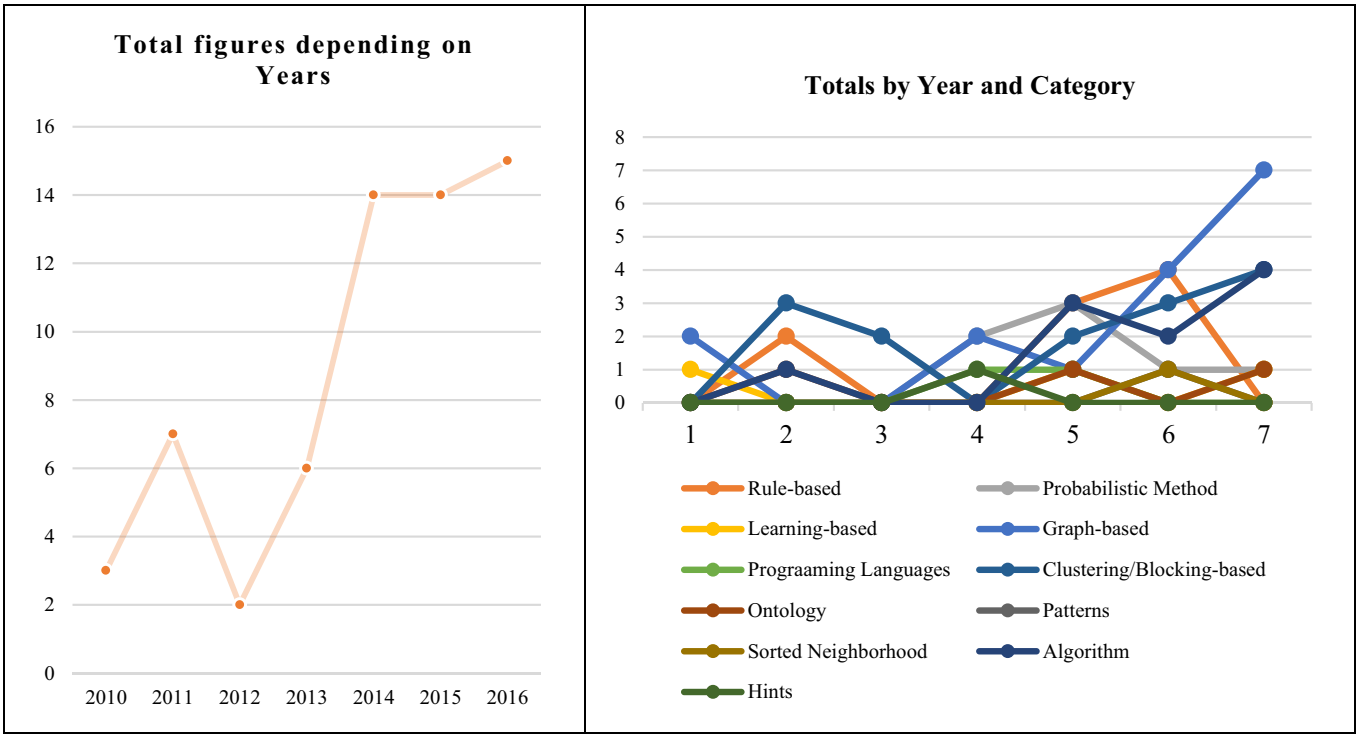

Fig. 4. Total of papers by Year and Category.

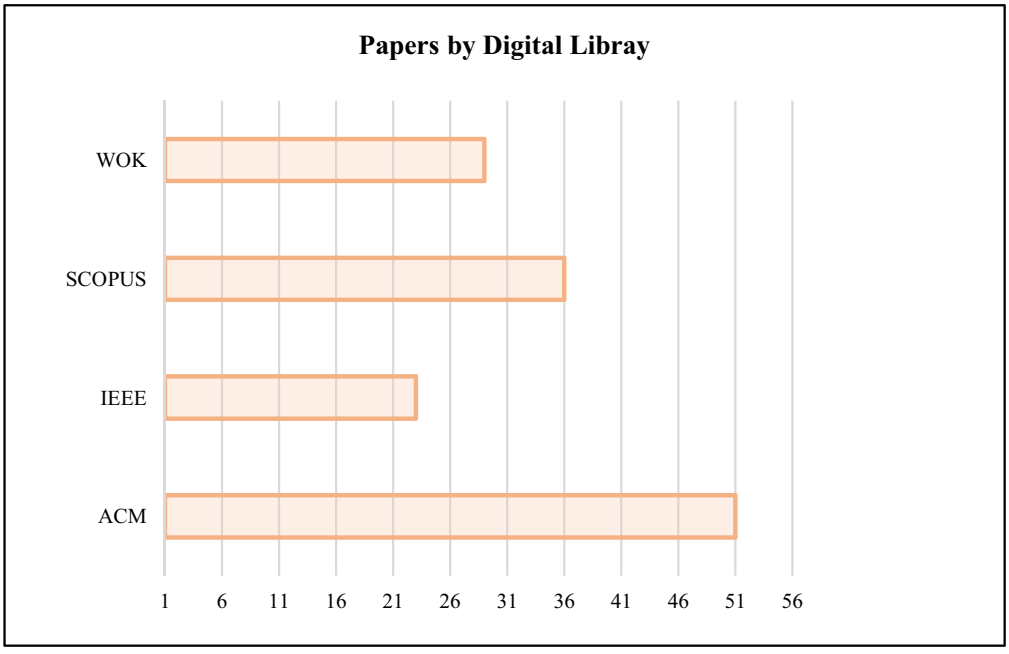

Fig. 5. Total of papers by Digital Library.

noticed that most of the datasets used for testing the proposals are real-world datasets. However, the solutions are not applied to real-world problems in industry. It is necessary to invest more research efforts in the automation of the proposed solutions because we have found just a few alternatives in this regard. Moreover, a new search is required in order to increase the domains where this study has been applied.

It is also necessary to apply these solutions to multi-applications problems, where different applications with different requirements need to be served with the results of the reconciliation process. To conclude, it is very important to point out that this type of studies should continue to keep it updated and do not let it become obsolete.

Acknowledgments

This research has been supported by the MeGUS project (TIN2013-46928-C3-3-R), Pololas project (TIN2016-76956-C3-2-R), by the SoftPLM Network (TIN2015-71938-REDT) of the Spanish the Ministry of Economy and Competitiveness and Fujitsu Laboratories of Europe (FLE).

References

Ali, O., & Cristianini, N. (2010). Information fusion for entity matching in unstructured data. *IFIP Advances in Information and Communication Technology, 339 AICT*, 162–169. http://doi.org/10.1007/978-3-642-16239-8_23.

Ayat, N., Akbarinia, R., Afsarmanesh, H., & Valduriez, P. (2013). Entity resolution for distributed probabilistic data. *Distributed and Parallel Databases, 31*(4), 509–542.

Ayat, N., Akbarinia, R., Afsarmanesh, H., & Valduriez, P. (2014). Entity resolution for probabilistic data. *Information Sciences, 277*, 492–511. http://doi.org/10.1016/j.ins.2014.02.135.

Balaji, J., Javed, F., Kejriwal, M., Min, C., Sander, S., & Ozturk, O. (2016). An ensemble blocking scheme for entity resolution of large and sparse datasets.

Beheshti, S.-M.-R., Benatallah, B., Venugopal, S., Ryu, S. H., Motahari-Nezhad, H. R., & Wang, W. (2016). A systematic review and comparative analysis of cross-document coreference resolution methods and tools. *Computing, 1*, 1–37. http://doi.org/10.1007/s00607-016-0490-0.

Bortoli, S., Bouquet, P., & Bazzanella, B. (2014). An identification ontology for entity matching. *On the Move to Meaningful Internet Systems: Otm 2014 Workshops, 8842*, 587–596.

Brando, C., Frontini, F. b, & Ganascia, J.-G. (2015). Disambiguation of named entities in cultural heritage texts using linked data sets. *Communications in Computer and Information Science, 539*, 505–514. http://doi.org/10.1007/978-3-319-23201-0_51.

Bratus, S., Rumshisky, A., Khrabrov, A., Magar, R., & Thompson, P. (2011). Domain-specific entity extraction from noisy, unstructured data using ontology-guided search. *International Journal on Document Analysis and Recognition, 14*(2), 201–211. http://doi.org/10.1007/s10032-011-0149-5.

Brizan, D. G., & Tansel, A. U. (2006). A survey of entity resolution and record linkage methodologies. *Communications of the IIMA, 6*(3), 41–50.

Calvanese, D., Keet, C. M., Nutt, W., Rodríguez-Muro, M., & Stefanoni, G. (2010). Web-based graphical querying of databases through an ontology: The Wonder system. In *Proceedings of the 2010 ACM symposium on applied computing* (pp. 1388–1395).

Carvalho, L. F. M., Laender, A. H. F., & Meira, W., Jr (2015). Entity matching: A case study in the medical domain. In *Alberto Mendelzon International Workshop on Foundations of Data Management* (p. 57).

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. In *Mobile networks and applications: Vol. 19* (pp. 171–209). Springer Berlin Heidelberg.

Costa, G., Cuzzocrea, A., Manco, G., & Ortale, R. (2011). Data de-duplication: A review. *Studies in Computational Intelligence, 375*, 385–412. http://doi.org/10.1007/978-3-642-22913-8_18.

Costa, G. D. A. (2016). Large-scale entity resolution for semantic web data integration LARGE-SCALE ENTITY RESOLUTION FOR SEMANTIC, (October 2014).

Da Silva, F. Q. B., Santos, A. L. M., Soares, S., Frana, A. C. C., Monteiro, C. V. F., & Maclel, F. F. (2011). Six years of systematic literature reviews in software engineering: An updated tertiary study. *Information and Software Technology, 53*(9), 899–913. http://doi.org/10.1016/j.infsof.2011.04.004.

de Assis Costa, G., & de Oliveira, J. M. P. (2016). A blocking scheme for entity resolution in the semantic web. In *Advanced Information networking and applications (AINA), 2016 IEEE 30th international conference on* (pp. 1138–1145).

Dharavath, R., & Kumar, C. (2015). The journal of systems and software entity resolution based EM for integrating heterogeneous distributed probabilistic data. *The Journal of Systems & Software, 107*, 93–109. http://doi.org/10.1016/j.jss.2015.05.035.

Dharavath, R., & Singh, A. K. (2016). Entity resolution-based jaccard similarity coefficient for heterogeneous distributed databases. In *Proceedings of the second international conference on computer and communication technologies* (pp. 497–507).

Dorneles, C. F., Gonçalves, R., & dos Santos Mello, R. (2011). Approximate data instance matching: A survey. *Knowledge and Information Systems, 27*(1), 1–21. http://doi.org/10.1007/s10115-010-0285-0.

Efthymiou, V., Efthymiou, V., Papadakis, G., Papastefanatos, G., Stefanidis, K., & Palpanas, T. (2016). Parallel meta-blocking for scaling entity resolution over big heterogeneous data. *Information Systems, 65*(December 2016), 137–157. http://doi.org/10.1016/j.is.2016.12.001.

Efthymiou, V., Stefanidis, K., & Vassilis, C. (2016). Minoan ER : Progressive entity resolution in the web of data. In *19th international conference on extending database tech- nology, EDBT 2016* (pp. 670–671).

Fan, W., Geerts, F., Tang, N., & Yu, W. (2014). Conflict resolution with data currency and consistency. *Journal of Data Information Quality, 5*(1–2), 6:1–6:37. http://doi.org/10.1145/2631923.

Fan, W., Li, J., Ma, S., Tang, N., & Yu, W. (2011). Interaction between record matching and data repairing. In *Proceedings of the 2011 international conference on management of data - SIGMOD '11: Vol. 1* http://doi.org/10.1145/1989323.1989373.

Firmani, D., Saha, B., & Srivastava, D. (2016). Online entity resolution using an oracle. In *Proceedings of the VLDB endowment: 9* (pp. 384–395).

Fisher, J., Christen, P., Wang, Q., & Rahm, E. (2015). A clustering-based framework to control block sizes for entity resolution. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '15* (pp. 279–288).

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How "big data" can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165*, 234–246. http://doi.org/10.1016/j.ijpe.2014.12.031.

Frontini, F., Brando, C., Ganascia, J.-G., Domain-adapted named-entity linker using Linked Data. Workshop on NLP Applications: Completing the Puzzle co-located with the 20th International Conference on Applications of Natural Language to Information Systems (NLDB 2015), Jun 2015, Passau, Germany. Proceedings of the Workshop on NLP Applications: Completing the Puzzle co-located with the 20th International Conference on Applications of Natural Language to Information Systems (NLDB 2015), 2015, http://ceur-ws.org/Vol-1386/. hal-01203356

Gaikwad, S., & Bogiri, N. (2015, January). A survey analysis on duplicate detection in hierarchical data. In Pervasive Computing (ICPC), 2015 International Conference on (pp. 1–6). IEEE.

Genero, M., Cruz-Lemus, J. A., & Piattini, M. (2014). *Métodos de Investigación en Ingeniería del Software. RA-MA Editorial.*

Getoor, L., & Machanavajjhala, A. (2013). Entity resolution for big data. In *KDD '13: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining: 4503* http://doi.org/10.1145/2487575.2506179.

Group, O. M. (2017). OMG unified modeling language TM (OMG UML), version 2.5. *InformatikSpektrum, 21*(May), 1–758. http://doi.org/10.1007/s002870050092.

Gu, Q., Zhang, Y., Cao, J., Xu, G., & Cuzzocrea, A. (2014). A confidence-based entity resolution approach with incomplete information. In *Data science and advanced analytics (DSAA), 2014 international conference on* (pp. 97–103).

Hermansson, L., Kerola, T., Johansson, F., Jethava, V., & Dubhashi, D. (2013). Entity disambiguation in anonymized graphs using graph kernels. In *Proceedings of the 22nd ACM international conference on conference on information & knowledge management - CIKM '13* (pp. 1037–1046).

Hernández, M., & Koutrika, G. (2013). HIL: A high-level scripting language for entity integration. In *Proceedings of the 16th international conference on extending database technology* (pp. 549–560).

Hsueh, S. C., Lin, M. Y., & Chiu, Y. C. (2014). *A load-balanced mapreduce algorithm for blocking-based entity-resolution with multiple keys: 152* (pp. 3–9).

Irmak, U., & Kraft, R. (2010). A scalable machine-learning approach for semi-structured named entity recognition. In *Proceedings of the 19th international conference on world wide web WWW 10* (p. 461).

Kejriwal, M. (2014). Populating entity name systems for big data. In *Proceedings of the 13th international semantic web conference* (pp. 521–528).

Kitchenham, B., & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information and Software Technology, 55*(12), 2049–2075. http://doi.org/10.1016/j.infsof.2013.07.010.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *Engineering, 2*, 1–65. http://doi.org/10.1145/1134285.1134500.

Kitchenham, B., Pretorius, R., Budgen, D., Pearl Brereton, O., Turner, M., & Niazi, M. (2010). Systematic literature reviews in software engineering – A tertiary study. *Information and Software Technology, 52*(8), 792–805. http://doi.org/10.1016/j.infsof.2010.03.006.

Kolb, L., Köpcke, H., Thor, A., Databases, H. S. D., & Systems, H. (2011). Learning-based entity resolution with MapReduce. *Cikm, 1*, 1–6. http://doi.org/http://doi.acm.org/10.1145/2064085.2064087.

Kong, C., Gao, M., Xu, C., Quian, W., & Zhou, A. (2016). Entity matching across multiple heterogeneous data sources. In *Proceedings of the 21st International conference on database systems for advanced applications (DASFAA 2016): Vol. 9642* (pp. 133–146).

Leitaõ, L., Calado, P., & Herschel, M. (2013). Efficient and effective duplicate detection in hierarchical data. *IEEE Transactions on Knowledge and Data Engineering, 25*(5), 1028–1041. http://doi.org/10.1109/TKDE.2012.60.

Li, L., Li, J., & Gao, H. (2015). Rule-based method for entity resolution. *IEEE Transactions on Knowledge and Data Engineering, 27*(1), 250–263. http://doi.org/10.1109/TKDE.2014.2320713.

Li, L., Li, J., Wang, H., & Gao, H. (2011). Context-based entity description rule for entity resolution. *Cikm, 1*, 1725–1730. http://doi.org/10.1145/2063576.2063825.

Lin, Y., Wang, H., Li, J., & Gao, H. (2016). Efficient entity resolution on heterogeneous records. http://doi.org/1610.09500.

Liu, H., Kumar, T. K. A., & Thomas, J. P. (2015). Cleaning framework for big data – object identification and linkage. In *2015 IEEE international congress on big data* (pp. 215–221).

Liu, Y., Wang, H., & Gao, H. (2011). A fast entity resolution method based on wave of records. In *2011 international conference on consumer electronics, communications and networks, CECNet 2011 – proceedings: 60933001* (pp. 4642–4645). http://doi.org/10.1109/CECNET.2011.5768200 .

Maddodi, S., Attigeri, G. V., & Karunakar, A. K. (2010). Data Deduplication techniques and analysis. In *Proceedings - 3rd international conference on emerging trends in engineering and technology, ICETET 2010* (pp. 664–668).

(2011). *Big data: the next frontier for innovation, competition, and productivity* (p. 156). McKinsey Global Institute. (June).

Mendeley Support Team. (2011). Mendeley guide. *Mendeley Desktop*, 1–16. http://doi.org/10.1063/1.476396.

Ngai, E. W. T., & Gunasekaran, A. (2007). A review for mobile commerce research and applications. *Decision Support Systems, 43*(1), 3–15. http://doi.org/10.1016/j.dss.2005.05.003.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011a). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569. http://doi.org/10.1016/j.dss.2010.08.006.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011b). The application of data mining techniques in financial fraud detection: A classification framework

and an academic review of literature. *Decision Support Systems, 50*(3), 559–569. http://doi.org/10.1016/j.dss.2010.08.006.

Ngai, E. W. T., Moon, K. K. L., Riggins, F. J., & Yi, C. Y. (2008). RFID research: An academic literature review (1995–2005) and future research directions. *International Journal of Production Economics, 112*(2), 510–520. http://doi.org/10.1016/j.ijpe.2007.05.004.

Nguyen, K., & Ichise, R. (2016). Linked data entity resolution system enhanced by configuration learning algorithm. *IEICE TRANSACTIONS on Information and Systems, 99*(6), 1521–1530.

Nuray-turan, R., Kalashnikov, D. V., & Mehrotra, S. (2013). Adaptive connection strength models for relationship-based entity resolution. *Jdiq, 4*(2), 1–22. http://doi.org/10.1145/2435221.2435224.

Otero-Cerdeira, L., Rodríguez-Martínez, F. J., & Gómez-Rodríguez, A. (2015). Ontology matching: A literature review. *Expert Systems with Applications, 42*(2), 949–971. http://doi.org/10.1016/j.eswa.2014.08.032.

Papadakis, G., Ioannou, E., Niederée, C., & Fankhauser, P. (2011a). Efficient entity resolution for large heterogeneous information spaces. In *Proceedings of the fourth ACM international conference on web search and data mining - WSDM '11: 535* http://doi.org/10.1145/1935826.1935903.

Papadakis, G., Ioannou, E., Niederée, C., Palpanas, T., & Nejdl, W. (2011b). To compare or not to compare: Making entity resolution more efficient. In *Proceedings of the international workshop on semantic web information management* (pp. 3:1–3:7).

Papadakis, G., Ioannou, E., Niederée, C., Palpanas, T., & Nejdl, W. (2012). Beyond 100 million entities: Large-scale blocking-based resolution for heterogeneous data. In *Proceedings of the fifth ACM international conference on web search and data mining* (pp. 53–62).

Papadakis, G., Svirsky, J., Gal, A., & Palpanas, T. (2016). Comparative analysis of approximate blocking techniques for entity resolution. *Pvldb, 9*(9), 684–695. http://doi.org/10.14778/2947618.2947624.

Ponce, J., Escalona, M. J., Gómez, A., Luque, M., & Molina, A. (2010). Definición de una política de pruebas en la gestión cultural : Aplicación al desarrollo del proyecto Mosaico, 6, 25–43.

Priya, P. A., Prabhakar, S., & Vasavi, S. (2015). Entity resolution for high velocity streams using semantic measures. In *Advance computing conference (IACC), 2015 IEEE international* (pp. 35–40).

Rahmani, H., Ranjbar-Sahraei, B., Weiss, G., & Tuyls, K. (2016). Entity resolution in disjoint graphs: An application on genealogical data. *Intelligent Data Analysis, 20*(2), 455–475.

Ramadan, B., Christen, P., & Liang, H. (2014). Dynamic sorted neighborhood indexing for real-time entity resolution. *Databases Theory and Applications, Adc 2014, 8506*, 1–12. http://doi.org/10.1145/2816821.

Selenium. (2017). Selenium website documentation, last access February. *RA-MA Editorial*.

Shen, W., Han, J., & Wang, J. (2014). A probabilistic model for linking named entities in web text with heterogeneous information networks. In *Proceedings of the 2014 ACM SIGMOD international conference on management of data* (pp. 1199–1210).

Shin, K., Jung, J., Lee, S., & Kang, U. (2015). BEAR: Block elimination approach for random walk with restart on large graphs. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data* (pp. 1571–1585). ACM.

Shu, L., Chen, A., Xiong, M., & Meng, W. (2011). Efficient Spectral neighborhood blocking for entity resolution. In *Proceedings - International conference on data engineering* (pp. 1067–1078).

Shu, L., Lin, C., Meng, W., Han, Y., Yu, C. T., & Smalheiser, N. R. (2012). A framework for entity resolution with efficient blocking. In *Information reuse and integration (IRI), 2012 IEEE 13th international conference on* (pp. 431–440) .

Shull, F., Singer, J., & Sj??berg, D. I. K. (2008). *Guide to advanced empirical software engineering. Guide to advanced empirical software engineering*. http://doi.org/10.1007/978-1-84800-044-5.

Sleeman, J., & Finin, T. (2013). Entity type recognition for heterogeneous semantic graphs. In *2013 AAAI fall symposium series* (pp. 63–67).

Song, D., Luo, Y., & Heflin, J. (2016). Linking heterogeneous data in the semantic web using scalable and domain-independent candidate selection. *IEEE Transactions on Knowledge and Data Engineering*, 143–156 *PP*(99) http://doi.org/10.1109/TKDE.2016.2606399.

Sun, C., Shen, D., Kou, Y., Nie, T., & Yu, G. (2014a). ERGP: A combined entity resolution approach with genetic programming. In *2014 11th web information system and application conference* (pp. 215–220).

Sun, Z., Xiao, N., Liu, F., & Fu, Y. (2014b). DS-Dedupe: A scalable, low network overhead data routing algorithm for inline cluster deduplication system. In *2014 International conference on computing, networking and communications, ICNC 2014* (pp. 895–899).

Wang, H., Li, J., & Gao, H. (2015). Efficient entity resolution based on subgraph cohesion. *Knowledge and Information Systems, 46*(2), 285–314. http://doi.org/10.1007/s10115-015-0818-7.

Wang, J., Oyama, S., Kurihara, M., & Kashima, H. (2014). Learning an accurate entity resolution model from crowdsourced labels. In *Proceedings of the 8th international conference on ubiquitous information management and communication – ICUIMC '14* (pp. 1–8).

Wang, Q., Cui, M., & Liang, H. (2016). Semantic-aware blocking for entity resolution. *IEEE Transactions on Knowledge and Data Engineering, 28*(1), 166–180.

Whang, S. E., & Garcia-Molina, H. (2014). Incremental entity resolution on rules and data. *VLDB Journal, 23*(1), 77–102. http://doi.org/10.1007/s00778-013-0315-0.

Whang, S. E., Marmaros, D., & Garcia-Molina, H. (2013). Pay-as-you-go entity resolution. *IEEE Transactions on Knowledge and Data Engineering, 25*(5), 1111–1124. http://doi.org/10.1109/TKDE.2012.43.

Wohlin, C., & Prikladniki, R. (2013). Systematic literature reviews in software engineering. *Information and Software Technology, 55*(6), 919–920. http://doi.org/10.1016/j.infsof.2013.02.002.

Xue, X., & Wang, Y. (2016). Using Memetic Algorithm for instance coreference resolution. In *2016 IEEE 32nd international conference on data engineering, ICDE 2016: 28* (pp. 1450–1451).

Yang, Y., Sun, Y., Tang, J., Ma, B., & Li, J. (2015). Entity matching across heterogeneous sources. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1395–1404).

Yu, M. (2014). Entity linking on graph data. *WWW, 1*, 21–25. http://doi.org/10.1145/2567948.2567954.

Yumusak, S., Dogdu, E., & Kodaz, H. (2014). A Short Survey of Linked Data Ranking, 14–17. http://doi.org/10.1145/2638404.2638523.

Zhang, H., Babar, M. A., & Ali Babar, M. (2013). Systematic reviews in software engineering: An empirical investigation. *Information and Software Technology, 55*(7), 1341–1354. http://doi.org/10.1016/j.infsof.2012.09.008.

Zhang, H., Babar, M. A., & Tell, P. (2011). Identifying relevant studies in software engineering. *Information and Software Technology, 53*(6), 625–637. http://doi.org/10.1016/j.infsof.2010.12.010.

Zhu, L., Ghasemi-gol, M., Szekely, P., Galstyan, A., & Knoblock, C. A. (2016). Unsupervised entity resolution on multi-type graphs. *International Semantic Web Conference, 9981*, 649–667. http://doi.org/10.1007/978-3-319-46523-4.