Data reconciliation and fusion methods: a survey

Abdelghani Bakhtouchi
Ecole Militaire Polytechniques (EMP), Algiers, Algeria and
Ecole Nationale Supérieure d'Informatique (ESI), Algiers, Algeria

Received 26 September 2018
Revised 3 July 2019
Accepted 3 July 2019

Abstract

With the progress of new technologies of information and communication, more and more producers of data exist. On the other hand, the web forms a huge support of all these kinds of data. Unfortunately, existing data is not proper due to the existence of the same information in different sources, as well as erroneous and incomplete data. The aim of data integration systems is to offer to a user a unique interface to query a number of sources. A key challenge of such systems is to deal with conflicting information from the same source or from different sources. We present, in this paper, the resolution of conflict at the instance level into two stages: references reconciliation and data fusion. The reference reconciliation methods seek to decide if two data descriptions are references to the same entity in reality. We define the principles of reconciliation method then we distinguish the methods of reference reconciliation, first on how to use the descriptions of references, then the way to acquire knowledge. We finish this section by discussing some current data reconciliation issues that are the subject of current research. Data fusion in turn, has the objective to merge duplicates into a single representation while resolving conflicts between the data. We define first the conflicts classification, the strategies for dealing with conflicts and the implementing conflict management strategies. We present then, the relational operators and data fusion techniques. Likewise, we finish this section by discussing some current data fusion issues that are the subject of current research.

Keywords Data reconciliation, Records linkage, Data fusion, Conflicts resolution, Integration
Paper type Original Article

1. Introduction

Resolving schema-level conflicts or even using the same schema can not, however, avoid conflicts between the values of instances themselves. Indeed, when data from different sources is integrated, different values can be references to the same entity in reality. These variations are due to: *Different conventions and different vocabularies*, *Incomplete information*, *The presence of the erroneous data* or *Data freshness*.

The three main goals of data integration systems is to increase [1]: *completeness*, *concision* and *accuracy* of the data. *Completeness* indicates the quantity of data measured by the number of entities as well as the number of attributes. *Concision* measures the uniqueness of object representations in integrated data, both in terms of unique entities number and the number of single entity attributes. Finally, *accuracy* indicates the data correctness, which means, how data are consistent with reality.

Publishers note: The publisher wishes to inform readers that the article "Data reconciliation and fusion methods: A survey" was originally published by the previous publisher of Applied Computing and Informatics and the pagination of this article has been subsequently changed. There has been no change to the content of the article. This change was necessary for the journal to transition from the previous publisher to the new one. The publisher sincerely apologises for any inconvenience caused. To access and cite this article, please use Bakhtouchi, A. (2020), "Data reconciliation and fusion methods: A survey", New England Journal of Entrepreneurship. Vol. 18 No. 3/4, pp. 182-194. https://10.1016/j.aci.2019.07.001. The original publication date for this paper was 06/06/2019.

Declaration of Competing Interest: None.

Applied Computing and Informatics
Vol. 18 No. 3/4, 2022
pp. 182-194
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1016/j.aci.2019.07.001

High completeness can be achieved by integrating more sources, whereas so as to be concise and accurate, two levels of tasks are performed by a data integration system[2]: schema-level matching and instance-level matching.

The schema-level matching is intended to establish semantic links between the content of different data sources. This task is accomplished during the generation of the integration system global schema. While the purpose of instance-level matching is to produce the correct data for each entity after determining reconciliation between instances that represent the same entity in reality. Some integration frameworks suppose the existence of a common identifier in all sources referring to the same concept. If such unique identifier exists, query results over sources can be reconciled thanks to relational operations. Nevertheless, taking into account that sources are autonomous, a common unique identifier doesn't exist in most of cases; How can a data integration system judge that two descriptions refer to the same entity or not? For this purpose, integration systems use entity reconciliation methods. Instance-level matching is processed through two phases: reference reconciliation and data fusion.

Even if schema conflicts are resolved, instances are not necessarily homogeneous; conflicts can occur when the same data is copied to multiple sources and the results of querying these sources contain conflicting data values for the same entities. Instance conflicts are often grouped into two classes: (i) reference conflicts and (ii) attribute value conflicts (see Figure 1).

The reference reconciliation step (key conflict resolution) aims to figure out the conflicts at the instance level by determining entities that refer to the same entity in reality. The purpose of the data fusion step (attribute conflict resolution) is to merge records referring to the same entity in reality by combining them in one representation and resolving eventual conflicts.

The remaining of this paper is organized as follows: Section 2 present the reference reconciliation through the definition of its main bases and techniques. In Section 3, we discuss the data fusion by classifying the various conflicts that may appear as well as the different techniques to resolve such conflicts. both Section 2 and Section 3 end by a discussion on issues related to step. Finally, the paper is concluded in Section 4.

2. Reference reconciliation (resolving key conflicts)

The problem of reconciling data is one of major problems for data integration systems. It consists in deciding whether two data descriptions are references to the same entity in reality or not. Data reconciliation is also denoted by the `reference reconciliation`, *record matching* [3,4], *record linkage* [5], *entity resolution* [6,7], *object identification*, *duplicate detection* [8] or *data cleaning* [9,10]. We consider that a data is defined by an identifier

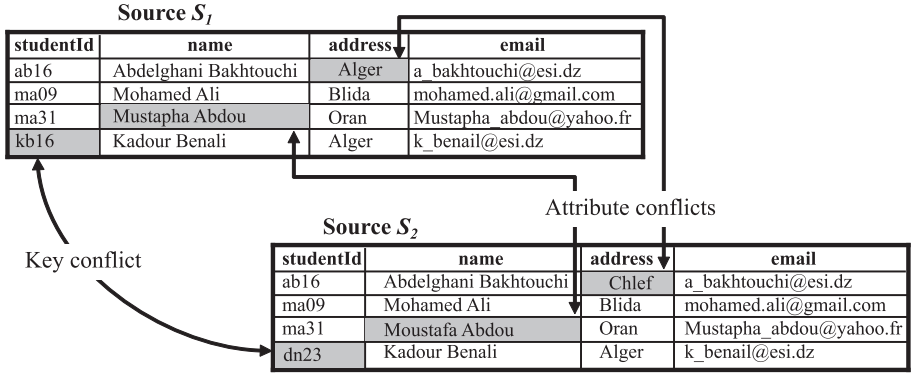

Figure 1.
Example of keys conflict and attribute values conflict.

(reference) and by a description. It is therefore a question of reconciling different identifiers described in relation to the same schema.

Historically, reference reconciliation has been mentioned for the first time in the terminology of "*Record Linkage*" by Halbert Dunn [11]. In fact, in the 1950s, data was represented by records in files, which justifies the use of the term "*Record Linkage*" to name the task of reference reconciliation in a data integration system. The problem of reference reconciliation performed by computer was introduced by Newcombe et al. in 1959 [12]; It was formalized ten years later by Felligi and Sunter [5]. After that, the problem of reference reconciliation has been considered in different terminologies by different communities. In the field of *databases*, data matching and duplicate detection are used when merging or cleaning multiple databases. In the *automatic natural language* processing, they make a co-reference and resolution of the anaphors when they want to find which are the nominal groups that refer to the same entity. In order to automate and make effective the reference reconciliation, a large number of methods have been proposed. Summaries of these methods can be found in [13–20].

The different reconciliation approaches available in the literature can be classified according to two criteria [21]: (i) the exploitation of the relations between data, and (ii) the exploitation of the knowledge for the reference reconciliation.

2.1 Evaluation of reference reconciliation methods

The quality of the obtained results of a reference reconciliation method is evaluated using measures of *Information Retrieval* field: *Recall*, *Precision* and *F-Measure*.

Recall: Proportion, among all possible pairs, of those for which the method has produced a correct result.

Precision: Proportion, among the couples for which the method has produced a result (of reconciliation or non-reconciliation), of those for which the result is correct.

F-Measure: Finding a good compromise between recall and precision is a more difficult goal to achieve, so results can be evaluated by calculating the combination of the two measures.

$$F - Measure = (2 \times Recall \times Precision)/(Recall + Precision)$$

2.2 Similarity measures

One of the most important challenges that reconciliation methods face is the difficulty of the syntactical variation of the vocabulary used to describe the data. These variations may be due to typographical variations, abbreviations, the presence of acronyms, coding differences and synonyms. In such a context, most reference reconciliation methods rely on similarity measurement techniques between values.

Similarity measure: A function: Sim: $E \times E \rightarrow [0, 1]$ applied to a pair of elements $(e_1, e_2) \in E \times E$ and which returns a real number in the interval $[0, 1]$ expressing the similarity between these two elements.

Similarity Score: The similarity score is the actual value calculated by a similarity measure for a pair of elements.

A multitude of similarity measures between character strings have been developed, each one is effective for particular types of syntactic variations. These similarity measures are categorized into three classes [21]: *Character-based measures*, *Atomic chains-based measures* and *Hybrid measures*.

The comparison of the different measures shows that no non–hybrid measure is appropriate for all datasets. The best measures for some datasets may be the worst for other datasets [22]. As a result, more flexible measures that combine several similarity measures

are needed. More generally, the selection of these measures must be customizable in a reference reconciliation system to be able to adapt to data.

2.3 Exploitation of relationships between data

Reference reconciliation methods can be distinguished by their ability to exploit the relationships between data. Two types of possible approaches are distinguished [17]: *local* approaches and *global* approaches.

2.3.1 Local approaches. Some of the local approaches exploit references descriptions not structured in attributes. The works [23–26], adopting this vision, calculate similarity score using textual data only, in the form of a single string. Such approach is interesting (1) when one requires to carry out a quick similarity calculation [24,23,25], (2) when one requires to have a set of candidate pairs for the reconciliation to apply then a more sophisticated similarity calculation [26] or even (3) when the association (*attribute*, *value*) is not sure.

A second part of local approaches consider structured data (attributes). A multitude of methods dealing with structured data proposed probabilistic models [5,27,3], or calculate a score of similarity for reference pairs [28–30]. However, the methods, in the first case, are very dependent to the probabilistic model parameters estimation, contrary to the last case method where it is a question of calculating the similarity scores of the attribute values and then combining them using an aggregation function to capture possibly influence degrees of the different attributes on the reference pairs similarity.

2.3.2 Global approaches. To improve the results quality of reconciliation methods, global approaches hexploit the relationships between data. These relationships are expressed explicitly like foreign keys in relational databases or semantic relations in other data models [7,31,32] or, are detected and used in the reference reconciliation algorithm [33]. Relationships, when they exist, allow to take into account more information when comparing reference descriptions. Some of these approaches, such as [31], perform simultaneous reconciliations of references that are of several types (people, articles, conferences) in order to capture the dependencies between the reference pairs induced by the relationships linking these references. Thus, global methods improve the obtained results quality by reducing the number of false reconciliations (false positives) or the number of false non-reconciliations (false negatives).

2.3.3 Comparison between the two approaches. Some local reference reconciliation approaches such as Fellegi and Sunter [5] and Winkler [27] use a probabilistic model. However, they are very dependent on the estimation of the different probabilistic model parameters. This dependency can be an obstacle for some applications such as news websites or blogs. In fact, these parameters can hardly be estimated because we cannot have labelled data and their estimation that would makes the method execution ineffective. Unlike local approaches based on probabilistic models, local approaches based on similarity measures do not confront the dependency problem.

Local approaches consider each reference as an entity that is completely independent of other entities and ignore that data is often represented in schemas expressing relationships between data. While global approaches exploit the references description in terms of attributes, and relationships linking different references in order to make a decision of reconciliation or non-reconciliation.

2.4 Exploitation of knowledge for reconciliation

Different types of knowledge can be used to improve the effectiveness of reference reconciliation approaches in term of choice and parameters of similarity measures as well as the importance of different attributes in the reconciliation decision. As example of such knowledge, we can cite the knowledge about references and attributes, and knowledge about

values such as synonymy. Reconciliation approaches that exploit knowledge in a fixed way (by coding it directly in the approach) are very sensitive to changes of data sources and domain (medical, commercial, etc.). Whereas, dynamic exploitation of domain knowledge allows adaptation of changes in data and application characteristics. Two types of approaches can be distinguished [34,17] according to the manner of acquiring knowledge namely *learning-based* approaches (supervised) and *rule-based* approaches (unsupervised).

2.4.1 Supervised approaches. Supervised approaches [22,8] use learning algorithms to learn different knowledge from expert-labelled data. For these approaches, if data changes, a new sample must be created on which the learning method will be reapplied.

2.4.2 Unsupervised approaches. Unsupervised approaches [35–37] use declarative language to allow an expert to specify the knowledge needed for reconciliation, in rules form for example. For these approaches, it is necessary to re-specify, taking into account the changes of data, the necessary knowledge such as, transformation or reconciliation rules or the concept profile descriptions.

The knowledge exploitation allows the approaches to be adaptable to the data evolution and the changes of domains (trade, trip, medicine, Web) in addition to the results quality improvement by reducing of false decisions. However, these approaches are dependent on human effort to label the learning data or to manually declare the rules. This makes these approaches difficult to apply to large volumes of data or evolving data.

2.4.3 Comparison between the two approaches. Reference reconciliation methods that exploit knowledge in a fixed way by implementing it directly into the source code of the method are highly vulnerable to changes of data sources and domain. For example, population census data in Algeria or China have different characteristics with regard to the names of persons and their date of birth; non-explicit (segmented) information can be presented in different orders according to the population culture. Country, for example, in China the last name precedes the first name and the date of birth is represented beginning with the year followed by the month ending with the day whereas in Algeria the format (First Name Last Name, Day Month Year) is the most used. More generally [17] the quality of the data sources may be different. Unsupervised or supervised exploitation of domain knowledge makes methods capable of adapting to changes in data and application characteristics. For methods that acquire domain knowledge through supervised learning, when the data changes, a new sample is created and the learning method is reapplied.

For methods using declarative language to specify knowledge [17] such as Hernandez and Stolfo [35], Wai Lup Low et al. [36] and Doan et al. [37], it is needed to perform a re-specification of the necessary knowledge such as transformation or reconciliation rules or the concept profiles descriptions, taking into account changes in the data.

The methods exploiting domain knowledge are generic, and provide less false positives and false negatives. contrariwise, the cost in terms of human effort required to either label the training data or to manually declare the rules makes these methods difficult to apply to large volumes of data or to frequently changing data.

2.5 Discussion on issues related to data reconciliation

Based on our reading of the literature and our work in this area, we summarize, in the following, some open issues and research directions for improving data reconciliation and data integration in general.

2.5.1 Compromise between effectiveness, efficiency, genericity and reduction of human intervention. A key challenge in developing an effective reconciliation solution is that some of the requirements are in conflict with one another. To reduce the search space and so improves effectiveness, blocking methods are used. Nevertheless, this can reduce the efficiency by the elimination of some relevant entity pairs. otherwise, the efficiency can be improved by

combining the use of various reconciliation algorithms, but increase computing time and thus reduce the effectiveness. Reconciling entities in a single domain easy than for different domains with a generic reconciliation solution. A specific reconciliation solution requires less human intervention to provide the training data. More effort are needed to ensure such compromises [19,38].

2.5.2 Reconciliation in big data integration. Big data community is giving today more interest to data reconciliation from big and heterogeneous data sources [39]. Big data reconciliation is particularly difficult because big data sources contains unstructured data in addition to their heterogeneous structure and are in dynamic evolution [40]. In practice, existing reconciliation methods can't keep their ineffectiveness when the volume of data become very large and in a huge number of sources. New methods are proposed offering parallel and scalable reconciliation to deal with the volume dimension [41,39,42]. These include blocking techniques and techniques that dispatch the charge between different nodes. To deal with the execution time dimension, incremental clustering methods are suggested [43].

2.5.3 Unstructured data reconciliation. The explosion of the number of data sources and the emergence of unstructured data led information to be unclean, incomplete and redundant. [44]. For years, the worlds of structured systems (transactions, databases, Oracle, DB2, Teradata, and others) have grown side by side with the world of unstructured systems (email, phone conversation transcripts, spreadsheets, reports, documents, and others). And for years these worlds existed as if they were isolated. After these two environments matured and grown to a large size, it was recognized that it is necessary to integrate the two worlds. Of course, data reconciliation is a very crucial step in this integration. Hence, the unstructured data reconciliation research field still relevant today. Highly heterogeneous information spaces require also new data reconciling techniques. To overcome the variety aspect, proposed techniques generate structured data by tagging and matching text. Furthermore, available data are usually inaccurate and contains noise (http://tdan.com/matching-unstructured-data-and-structured-data/5009). To deal with this aspect of truthfulness, proposed clustering and assembly techniques can treat noise and values changing.

2.5.4 Real-time reconciliation. Real-time reconciliation (also known as dynamic reconciliation, online reconciliation, just in time reconciliation) assumes that data is dynamically matched, and therefore changes regularly (for example, customer data in a widely used system). For this reason, some systems do not copy the work data. Instead, the data is indexed to the source system using the same clustering rules that are used in the reconciliation process [45]. The usefulness of this type of reconciliation requires the proposition of even more efficient methods and algorithms. Many recent efforts address this issue [46–49].

3. Data fusion (resolving attribute conflicts)

The last phase of a data integration system is to merge records referring to the same entity in reality by combining them in one representation and resolving eventual conflicts.

The problem of conflicting values of attributes in data integrating field was evoked for the first time by *Umeshwar Dayal* [50]. Since then, even if the problem got less attention, but some techniques have been proposed [51]. We present in the following, data fusion [52] also known as *data consolidation* or *entities resolution*.

3.1 Conflicts classification

Data conflicts can be classified into two classes: (a) Attribute value uncertainty when information are missing, and (b) contradictions when the attribute have different values [1,51].

Uncertainties: The existence of the attribute value in one source and a null value in one or more sources.

Contradictions: The attribute value in one source is different than the attribute value in one or more sources.

3.2 Strategies for dealing with conflicts

Conflict-handling strategies are operations that interpret the way in which incoherent data are handled. To provide a unique and coherent representation, sometimes one vale is selected, other times values are combined or a new value is created.

There are several strategies for dealing with incoherence, some of which are mentioned repeatedly in the literature [53–57]. They are divided according to how they manage (or do not manage) conflicting data into three main classes [58]: *ignorance*, *avoidance* and *resolution* of conflicts.

3.2.1 Conflicts ignorance. Ignoring conflicts strategies don't make decisions about conflicting values, they are not sometimes aware of the conflict, and can therefore produce incoherent results.

"*Skip conflicts*" strategy and "*Consider all possibilities*" strategy are examples of ignoring conflicts strategies.

3.2.2 Conflicts avoidance. These strategies are aware of conflicts, although they don't resolve individually each conflict. On the contrary, only one decision is made, for example, the preference of a particular source. Since the decision is often made before or without looking at the data values, conflicts are not always handled. They are more effective in terms of processing time than conflict resolution strategies to the detriment of the precision lose due to non-consideration of all information that may be useful for conflict resolution.

Two classes of conflicts avoidance techniques are identified, a first class of strategies that take into account meta-data when making decisions (based on metadata) and a second class of strategies that don't take them into account (based on instances).

"*Confident source*" strategy is an example of metadata-based strategies, "*Take information*" strategies and "*Only coherent data*" are examples of instance-based strategies.

3.2.3 Conflicts resolution. Conflict resolution strategies, unlike previous classes, take into account all values and metadata in the decision of conflict resolution. These approaches are less efficient in processing time, but provide conflict resolution capabilities that are as flexible as possible.

Likewise the previous class, decisions taken by conflict resolution strategies can be *instance-based*, or *metadata-based*, using for instance, the freshness of data or the reliability of a source. According to the produced result they can be also classified into: *decision* strategies that choose one of existing values, or *mediation* strategies that create a new value, such as the average of existing numbers.

The *most common value* strategy and the *draw* strategy are examples of conflict resolution strategies based on instances that produce decisions. An example of mediation strategies is the *middle value* strategy. The *most recent value* strategy is a representative strategy of metadata strategies that produce mediated values.

3.3 Relational operators and data fusion techniques

Data fusion can be performed by standard (`union` and `join`) or advanced relational operators. The join and union (and their variants) merge data of all kinds. Approaches based on union generate a common schema, according to which records from sources tables are added. Other approaches define new operators and combine them to existing ones [51].

3.3.1 Joining approaches. To do the join operation on two tables, their schemas are expanded, to allow adding unknown values to tuples. *Outer join* operation is a variant that avoid losing tuples. The variant *complete disjunction* merges tuples of two or more relations into the same tuple. [59].

3.3.2 Union approaches. The union of two relations merges the data of the same tuples, that is to say, the same values. The minimal union operation is an improved variant that eliminates tuples having null-values but share the same non-null values with other tuples.

3.3.3 Other techniques. In addition to resolving uncertainties, it exist relational operators for contradictions elimination. The *matchjoin* operator starts by generating all possible tuples, it reduce them then following a way defined by the user [60]. The *prioritized merge* operator [61] follow the same procedure except the values selection from preferred sources.

We mention at the end that data fusion can be achieved also by SQL-based techniques through functions defined by user, aggregation functions or others.

3.4 Discussion on data fusion issues

Similar to what we did with data reconciliation, we also summarize, in the following, some open issues and search directions for improving data fusion and data integration in general.

Advanced conflict resolution techniques

Trusting the most accurate sources is not always the best solution because even these sources can contains errors. The work in [62–64] proposes to examine the accuracy of sources when deciding real values through probabilistic models that calculate the iterative accuracy of sources.

It is difficult to distinguish between incorrect values and outdated ones. Thus, the most common value may be an outdated value, while the most recent value may be a wrong value. To find the correct values, a probabilistic model integrating the notion of sources freshness is proposed in [65].

Sources can integrate instances from other sources. Therefore, errors can propagate quickly and slant the conflict resolution decision. Works in [66–68] propose to take dependencies between the sources into account during the discovery of the correct values. They use algorithms that iteratively detect dependency between sources.

Fusion in big data integration

To meet the veracity challenges in big data, approaches extend existing ones to deal with the big data volume, the response time, and the data variety. For instance [69], proposed a framework offering three levels of transparency: data resource integration level, data fusion level and data service provision level.

From data fusion to knowledge fusion

The knowledge fusion identifies the true subject-predicate-object triple extracted by several information extractors from several information sources [70]. Traditionally, they process to mapping schemas then data reconciliation focusing on conflicts. Therefore, knowledge sources are different that involve more efforts to deal with [71]. While adapting and improving data fusion methods can solve some of the knowledge fusion problems, there still a place for improvement. However, quality improvement involves more fundamental modifications to the first assumptions that data fusion techniques suppose.

Multi-sensor data fusion

To provide a description of an environment or a process, observations from many sensors are combined. This combination from disparate sources is somehow better than if these sources were used individually [72]. Multi-sensor data fusion is a difficult task due to the fact that sensor technologies are imperfect and diverse, furthermore their application environment are also of different natures. Even if some of these problems have been addressed, there is no existing technique can overcame all multi-sensor data fusion challenges [73,74].

4. Conclusion

Conflicts resolution is accomplished following two phases: reference reconciliation first then data fusion. Reference reconciliation methods aim to answer one question, two given data descriptions refer to the same real world entity or to two different entities. This problem occurs when more than one representation are used to describe one entity in real world, when the data contains errors, and when the information is incomplete. We defined the principles of a reconciliation method and then distinguished reference reconciliation methods, first on how to use reference descriptions, then how to acquire knowledge.

Data fusion is about the fusion of duplicates in the same representation and at the same time the resolution of possible conflicts between different values of the same attribute. This problem was not the preoccupation of researchers until the last two decades when several works and methods have emerged. In this paper, we have also discussed some current data reconciliation and fusion issues that are the subject of current research at the end of each section.

References

[1] X.L. Dong, F. Naumann, Data fusion – resolving data conflicts for integration, PVLDB 2 (2) (2009) 1654–1655.

[2] F. Naumann, A. Bilke, J. Bleiholder, M. Weis, Data fusion in three steps: resolving inconsistencies at schema-, tuple-, and value-level, in: Bulletin of The Technical Committee On Data Engineering, 2006, pp. 21–31.

[3] V.S. Verykios, A.K. Elmagarmid, E.N. Houstis, Automating the approximate record-matching process, Inf. Sci. Inf. Comput. Sci. 126 (1–4) (2000) 83–98, https://doi.org/10.1016/S0020-0255(00)00013-X.

[4] R. Baxter, P. Christen, A comparison of fast blocking methods for record linkage, 2003, pp. 25–27.

[5] I.P. Fellegi, A.B. Sunter, A theory for record linkage, J. Am. Stat. Assoc. 64 (1969) 1183–1210.

[6] O. Benjelloun, H. Garcia-Molina, H. Kawai, T.E. Larson, D. Menestrina, Q. Su, S. Thavisomboon, J. Widom, Generic entity resolution in the serf project, IEEE Data Eng. Bull. 29 (2) (2006) 13–20.

[7] I. Bhattacharya, L. Getoor, Entity Resolution in Graphs, chapter Mining Graph Data, Wiley, 2006.

[8] S. Tejada, C.A. Knoblock, S. Minton, Learning object identification rules for information integration, Inf. Syst. 26 (8) (2001) 607–633, https://doi.org/10.1016/S0306-4379(01)00042-4.

[9] E. Rahm, H.H. Do, Data cleaning: problems and current approaches, IEEE Data Eng. Bull. 23 (2000) 2000.

[10] H. Galhardas, D. Florescu, D. Shasha, E. Simon, Ajax: an extensible data cleaning tool, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM Press, Dallas, Texas, United States, 2000, p. 590.

[11] H.L. Dunn, Record linkage, Am. J. Public Health 36 (1946) 1412–1416.

[12] H. Newcombe, J. Kennedy, S. Axford, A. James, Automatic linkage of vital records, Science 130 (1959) 954–959.

[13] L. Gu, R. Baxter, D. Vickers, C. Rainsford, Record linkage: current practice and future directions, CSIROMathematical and Information Sciences, 2003, Tech. rep.

[14] W.E. Winkler, Overview of record linkage and current research directions, Bureau of the Census, 2006, Tech. rep.

[15] N. Koudas, S. Sarawagi, D. Srivastava, Record linkage: similarity measures and algorithms, in: Proceedings of the 2006 ACM SIGMOD international conference on Management of data, SIGMOD '06, ACM, New York, NY, USA, 2006, pp. 802– 803. doi: https://doi.org/10.1145/1142473.1142599.

[16] A.K. Elmagarmid, P.G. Ipeirotis, V.S. Verykios, Duplicate record detection: a survey, IEEE Trans. Knowl. Data Eng. 19 (2007) 1–16, https://doi.org/10.1109/TKDE.2007.9.

[17] F. Sans, N. Pernelle, M.C. Rousset, RTconciliation de rTfTrences: une approche adaptTe aux grands volumes de donnTes, Colloque sur l'Optimisation et les SystFmes d'Information (COSI) (2007) 521–532.

[18] C. Batini, M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques, first ed., Springer Publishing Company, Incorporated, 2010.

[19] H. Koepcke, E. Rahm, Frameworks for entity matching: a comparison, Data Knowledge Eng. J. 69 (2) (2010) 197–210.

[20] L. Getoor, A. Machanavajjhala, Entity resolution: Theory, practice & open challenges, in: International Conference on Very Large Data Bases, 2012.

[21] F. Sa-S, IntTgration sTmantique de donnTes guidTe par une ontologie, UniversitT Paris-Sud (2007) (Ph.D. thesis).

[22] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, S. Fienberg, Adaptive name matching in information integration, IEEE Intell. Syst. 18 (5) (2003) 16–23, https://doi.org/10.1109/MIS.2003.1234765.

[23] W.W. Cohen, Data integration using similarity joins and a word-based information representation language, ACM Trans. Inf. Syst. 18 (3) (2000) 288–321, URL: http://doi.acm.org/10.1145/352595.352598.

[24] A. Monge, C. Elkan, An efficient domain-independent algorithm for detecting approximately duplicate database records, in: proceedings of the 2nd ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97), 1997, pp. 23–29.

[25] L. Gravano, P.G. Ipeirotis, N. Koudas, D. Srivastava, Text joins in an rdbms for web data integration, in: Proceedings of the 12th international conference on World Wide Web,WWW'03, ACM, New York, NY, USA, 2003, pp. 90–101. doi: https://doi.org/10.1145/775152.775166.

[26] A. Bilke, F. Naumann, Schema matching using duplicates, in: Proceedings of the 21st International Conference on Data Engineering, ICDE '05, IEEE Computer Society, Washington, DC, USA, 2005, pp. 69–80, https://doi.org/10.1109/ICDE.2005.126.

[27] W.E. Winkler, Methods for record linkage and bayesian networks, Tech. Rep. Statistical Research Report Series RRS2002/05, U.S. Bureau of the Census, Washington, D.C., 2002.

[28] W.W. Cohen, Integration of heterogeneous databases without common domains using queries based on textual similarity, SIGMOD Record 27 (2) (1998) 201–212, URL: http://doi.acm.org/10.1145/276305.276323.

[29] D. Dey, S. Sarkar, P. De, Entity matching in heterogeneous databases: A distance based decision model, in: Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences-Volume 7 – Volume 7, HICSS '98, IEEE Computer Society, Washington, DC, USA, 1998, p. 305, https:// doi.org/10.1109/HICSS.1998.649225.

[30] S. Guha, N. Koudas, A. Marathe, D. Srivastava, Merging the results of approximate match operations, in: Proceedings of the Thirtieth international conference on Very large data bases – Volume 30, VLDB '04, VLDB Endowment, 2004, pp. 636–647, URL: http://dl.acm.org/citation.cfm?id=1316689.1316745.

[31] X. Dong, A. Halevy, J. Madhavan, Reference reconciliation in complex information spaces, in: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, SIGMOD '05, ACM, New York, NY, USA, 2005, pp. 85–96. URL: http://doi.acm.org/10.1145/1066157.1066168.

[32] R. Ananthakrishna, S. Chaudhuri, V. Ganti, Eliminating fuzzy duplicates in data warehouses, in: Proceedings of the 28th international conference on Very Large Data Bases, VLDB '02, VLDB Endowment, 2002, pp. 586–597, URL: http:// dl.acm.org/citation.cfm?id=1287369.1287420.

[33] D.V. Kalashnikov, S. Mehrotra, Domain-independent data cleaning via analysis of entity-relationship graph, ACM Trans. Database Syst. 31 (2) (2006) 716–767, URL: http://doi.acm.org/10.1145/1138394.1138401.

[34] H. Zhao, S. Ram, Entity matching across heterogeneous data sources: An approach based on constrained cascade generalization, Data Knowledge Eng. J. 66 (3) (2008) 368–381.

[35] M.A. Hernández, S.J. Stolfo, Real-world data is dirty: data cleansing and the merge/purge problem, Data Min. Knowl. Discov. 2 (1) (1998) 9–37, https://doi. org/10.1023/A:1009761603038.

[36] W.L. Low, M.L. Lee, T.W. Ling, A knowledge-based approach for duplicate elimination in data cleaning, Inf. Syst. 26 (8) (2001) 585–606.

[37] A. Doan, Y. Lu, Y. Lee, J. Han, Object matching for information integration: A profiler-based approach, in: IIWeb, 2003, pp. 53–58.

[38] J. Li, C. Huang, X. Wang, S. Wu, Balancing efficiency and effectiveness for fusion-based search engines in the big data environment, Inf. Res.: Int. Electron. J. 21 (2) (2016) n2.

[39] J.G. Enríquez, F. Domínguez-Mayo, M. Escalona, M. Ross, G. Staples, Entity reconciliation in big data sources: a systematic mapping study, Expert Syst. Appl. 80 (2017) 14–27.

[40] X.L. Dong, D. Srivastava, Big Data Integration, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2015.

[41] S. Bergamaschi, D. Beneventano, F. Mandreoli, R. Martoglia, F. Guerra, M. Orsini, L. Po, M. Vincini, G. Simonini, S. Zhu, et al., From data integration to big data integration, in: A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, Springer, 2018, pp. 43–59.

[42] C.Le Fèvre, L.Poty, G.Noël, Les big data, généralités et intégration en radiothérapie, Cancer/ Radiothérapie.

[43] A. Gruenheid, X.L. Dong, D. Srivastava, Incremental record linkage, PVLDB 7 (9) (2014) 697–708.

[44] R.Blanco, J.G.Enríquez, F.J.Domínguez-Mayo, M.Escalona, J.Tuya, Early integration testing for entity reconciliation in the context of heterogeneous data sources, IEEE Transactions on Reliability.

[45] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Springer Publishing Company, Incorporated, 2012.

[46] S. Xie, C. Yang, X. Wang, Y. Xie, Data reconciliation strategy with time registration for the evaporation process in alumina production, Can. J. Chem. Eng. 96 (1) (2018) 189–204.

[47] D.S. Almeida, C.S. Hara, R.R. Ciferri, C.D. Aguiar Ciferri, , An asynchronous collaborative reconciliation model based on data provenance, Software: Pract. Experience 48 (1) (2018) 197–232.

[48] M. Salame, Predictive and adaptive queue flushing for real-time data reconciliation between local and remote databases, uS Patent App. 15/ 243,960 (Aug. 17 2017).

[49] L. Sun, S.M. Zoldi, Method and apparatus for reconciliation of multiple sets of data, uS Patent 9,535,959 (Jan. 3 2017).

[50] U. Dayal, in: M. Schkolnick, C. Thanos (Eds.), Processing queries over generalization hierarchies in a multidatabase system Proceedings of the 9th International Conference on Very Large, Data Bases, Morgan Kaufmann, 1983, pp. 342–353.

[51] J. Bleiholder, F. Naumann, Data fusion, ACM Comput. Surv. 411 (1) (2008) 1– 41.

[52] X. Li, X.L. Dong, K. Lyons, W. Meng, D. Srivastava, Truth finding on the deep web: is the problem solved?, PVLDB 6 (2) (2012) 97–108.

[53] A. Fuxman, E. Fazli, R.J. Miller, Conquer: efficient management of inconsistent databases, in: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, SIGMOD '05, ACM, New York, NY, USA, 2005, pp. 155– 166, http://doi.acm.org/10.1145/1066157.1066176.

[54] A. Motro, P. Anokhin, Utility-based resolution of data inconsistencies, in: Proceedings of the 2004 international workshop on Information quality in information systems, IQIS '04, ACM, New York, NY, USA, 2004, pp. 35–43. http://doi.acm.org/10.1145/1012453.1012460.

[55] Y. Papakonstantinou, S. Abiteboul, H. Garcia-Molina, Object fusion in mediator systems, in: Proceedings of the 22th International Conference on Very Large Data Bases, VLDB '96, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996, pp. 413–424, URL: http://dl.acm.org/citation.cfm?id=645922. 673481.

[56] E. Schallehn, K.-U. Sattler, G. Saake, Efficient similarity-based operations for data integration, Data Knowl. Eng. 48 (3) (2004) 361–387, https://doi.org/10.1016/j.datak.2003.08.004.

[57] V.S.Subrahmanian, S.Adali, A.Brink, J.J.Lu, A.Rajput, T.J.Rogers, R.Ross, C.Ward, Hermes a heterogeneous reasoning and mediator system.

[58] J. Bleiholder, F. Naumann, Conflict handling strategies in an integrated information system, in: Proceedings of the International Workshop on Information Integration on the Web (IIWeb), no. 197 in Informatik-Berichte, Institut fnr Informatik, Edinburgh, UK, 2006, URL: http://edoc.hu-berlin. de/docviews/abstract.php?id=27030.

[59] C.A. Galindo-Legaria, Outerjoins as disjunctions, SIGMOD Record 23 (2) (1994) 348–358, URL: http://doi.acm.org/10.1145/191843.191908.

[60] S. Greco, L. Pontieri, E. Zumpano, Integrating and managing conflicting data, in: Revised Papers from the 4th International Andrei Ershov Memorial Conference on Perspectives of System Informatics: Akademgorodok, Novosibirsk, Russia, PSI '02, Springer-Verlag, London, UK, UK, 2001, pp. 349–362, URL: http://dl. acm.org/citation.cfm?id=646802.705967.

[61] L.L. Yan, M.T. Özsu, Conflict tolerant queries in aurora, in: Proceedings of the Fourth IECIS International Conference on Cooperative Information Systems, COOPIS '99, IEEE Computer Society, Washington, DC, USA, pp. 279–290, URL: http://dl.acm.org/citation.cfm?id=520790. 793790.

[62] A.D. Sarma, X.L. Dong, A.Y. Halevy, Data integration with dependent sources, in: EDBT 2011, 14th International Conference on Extending Database Technology, ACM, Uppsala, Sweden, 2011, pp. 401–412.

[63] M. Wu, A. Marian, Corroborating answers from multiple web sources, in: Proceeding of WebDB, 2007.

[64] X. Yin, J. Han, P.S. Yu, Truth discovery with multiple conflicting information providers on the web, IEEE Trans. Knowl. Data Eng. 20 (6) (2008) 796–808.

[65] X. Dong, L. Berti-Equille, D. Srivastava, Truth discovery and copying detection from source update history, Tech. rep., Technical report, AT&T Labs-Research, Florham Park, NJ, 2009.

[66] J. Slaney, B.W. Paleo, Conflict resolution: a first-order resolution calculus with decision literals and conflict-driven clause learning, J. Autom. Reasoning 60 (2) (2018) 133–156.

[67] C. Marsh, J. Thomas, W. Webb, D. Bell, P. Nguyen, Apparatus and method for conflict resolution in remote control of digital video recorders and the like, uS Patent 9,706,160 (Jul. 11 2017).

[68] M.N. Maunder, K.R. Piner, Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets, Fish. Res. 192 (2017) 16–27.

[69] Z. Xie, W. Lv, L. Qin, B. Du, R. Huang, An evolvable and transparent data as a service framework for multisource data integration and fusion, Peer-to-Peer Networking Appl. 11 (4) (2018) 697–710.

[70] X.L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, W. Zhang, From data fusion to knowledge fusion, PVLDB 7 (10) (2014) 881–892.

[71] M. Ringsquandl, S. Lamparter, R. Lepratti, P. Kröger, Knowledge fusion of manufacturing operations data using representation learning, in: IFIP International Conference on Advances Production Management Systems, Springer, 2017, pp. 302–310.

[72] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, Inf. Fusion 14 (1) (2013) 28–44, https://doi.org/10.1016/j.inffus.2011.08.001, URL: https://doi.org/10.1016/j.inffus.2011.08. 001.

[73] W. Jiang, W. Hu, C. Xie, A new engine fault diagnosis method based on multisensor data fusion, Appl. Sci. 7 (3) (2017) 280.

[74] Y. Tang, D. Zhou, Z. He, S. Xu, An improved belief entropy–based uncertainty management approach for sensor data fusion, Int. J. Distrib. Sens. Netw. 13 (7) (2017), 1550147717718497.

Corresponding author
Abdelghani Bakhtouchi can be contacted at: a_bakhtouchi@esi.dz