A BEST-EFFORT INTEGRATION FRAMEWORK FOR IMPERFECT INFORMATION SPACES

Ahmed Abu Halimeh, Aziz Deraman, Fadi Safieddine.

College of Engineering and Technology. American University of the Middle East (AUM), Egaila, Kuwait.
School of Informatics & Applied Mathematics, University Malaysia Terengganu, Kuala Terengganu, Terengganu, Malaysia.
College Of Business Administration. American University of the Middle East (AUM), Egaila, Kuwait.

Email: Ahmed.AbuHalimeh@aum.edu.kw
Email: a.d@umt.edu.my
Email: Fadi.Safiedinne@aum.edu.kw

Ashraf Jaradat*

College Of Business Administration. American University of the Middle East (AUM), Egaila, Kuwait.
Email: Ashraf.Jaradat@aum.edu.kw
*Corresponding author.

Abstract: Entity resolution (ER) with imperfection management has been accepted as a major aspect while integrating heterogeneous information sources that exhibit entities with varied identifiers, abbreviated names, and multi-valued attributes. Many of novel integration applications such as personal information management and web-scale information management require the ability to represent and manipulate imperfect data. This requirement signifies the issues of starting with imperfect data to the production of probabilistic database. However, classical data integration (CDI) framework fails to cope with such requirement of explicit imperfect information management. This paper introduces an alternative integration framework based on the best-effort perspective to support instance integration automation. The new framework explicitly incorporates probabilistic management to the ER tasks. The probabilistic management includes a new probabilistic global entity, a new pair-wise-source-to-target ER process, and probabilistic decision model logic as alternatives. Together, the paper presents how these processes operate to support the current heterogeneous sources integration challenges.

Keywords: data integration; information integration; uncertainty management; best-effort integration framework; probabilistic instance integration; data quality

Reference to this paper should be made as follows: Jaradat, A., Abu Halimeh, A., Deraman, A., & Safieddine, F., (2018) 'A best-effort integration framework for imperfect information spaces.' Int. J. of Intelligent Information and Database Systems, Vol.11, No.4, pp.296-314.

Biographical notes:

Ashraf Jaradat (ashraf.jaradat@aum.edu.kw) currently is an assistant professor in the department of management information system at American University of the Middle East. He received his PhD in information system from the national university of Malaysia. After obtaining his doctoral degree, Dr Ashraf was an assistant professor in the department of management information system at Yarmouk University. He also worked in SpringHill groups as advisor business development. His research interest includes data integration, data fusion, uncertainty management, information system modelling, e-learning, and knowledge management.

1 Introduction

Information integration (II) in its various appearances is one of the most relevant and critical problems studied in the database management and artificial intelligence (AI) fields, as well as in related areas such as the semantic web. The common definition of II is being the general process of obtaining a single source out of some heterogeneous information sources (Ziegler and Dittrich, 2004, p. 4). Challenges in II originate from its sources and data quality.

The participated sources are developed and maintained independently with different interfaces, data models, schemas, data representation, and may contain a collection of semi-structured or unstructured objects that have the potential to be integrated while providing useful services to users. The information space term is used to refer to such collaboration (Magnani and Montesi, 2007, p.18; Ioannou, Niederee, and Nejdl, 2008; Dong and Halevy, 2005, p.26; Motro and Anokhin P, 2006, p.177).

There are many definitions of data quality. Data can be considered high quality if it is fit for its intended uses in operations, decision making and planning (Strong, Lee, and Wang, 1997, p.103). Furthermore, as data volume increases, the question of internal data consistency becomes significant, regardless of fitness for use for any particular purpose. People's views on data quality can often be in disagreement, even when discussing the same set of data used for the same purpose. Therefore, Dong and Naumann (2009) have suggested approaches in data fusion to ensure data quality.

Best-effort or schema-based is a dataspace approach proposed toward reaching the DataSpace Support Platform (DSSP) vision of overcoming the CDI limitations and supporting complex information space applications (Franklin, Halevy, and Maire, 2005; Halevy, Franklin, and Maire, 2006; Kuicheu et al, 2013; Sarma, Dong, and Halevy, 2009, p. 123). This approach recognises that imperfection inherently unavoidable in the integration process, yet imperfect information is valuable and even more useful than losing it, and hence, it is an important result to be managed and viewed to users (Zhang et al., 2008; Hedeler et al., 2010, p.114). An approach is presented by Dong and Halevy (2005) as a best-effort integration solution exhibits various CDI's benefits at lower cost, on-demand and more automated integration, however, their approach has been criticised for producing lower integration and merged content quality (Dittrich, Salles, and Blunschi, 2009; Sarma, Dong, Halevy, 2011). The approach takes a CDI framework, as mediator-based, and attempts to relax it by reducing or removing the manual intervention in its integration process while providing useful service to both technical and non-technical users. Users can relax the CDI framework to simplify, remove or make several integration

components less precise while maintaining an automated process (Blanco et al., 2010; Sarma, Dong, and Halevy, 2011).

The term Entity Resolution (ER) is used in this paper to refer to the process of handlings the actual data entity and the data conflicts. ER also refers to the process of handling conflicts and imperfections raised at the instance level due to the actual nature of data, the probabilistic matching heterogeneity at the instance integration level (Ioannou, Niederee, and Nejdl, 2008; Ioannou et al., 2012; Li et al., 2015). ER process composes of two sub-tasks to respond to these conflicts: entity linkage that handles the entity conflict, and data fusion that handles the data inconsistency (Dong and Naumann, 2009; Jiang, 2008; Haase and Volker, 2008; Panse et al., 2010; Ioannou and Staworko. 2013). Therefore, ER with imperfection management is recognised by several researchers as one of the crucial requirements and challenges in the age of information spaces integration (Magnani and Montesi, 2010; Hedeler et al., 2010; Agrawal and Yu, 2009).

On the other hand, probability theory is presented in the literature as an appropriate modelling strategy for imprecise integration approaches through the possible-worlds manipulation and probabilistic database generation (Ioannou, Niederee, and Nejdl, 2008; Magnani and Montesi, 2010; Hedeler et al., 2010). Probability theory provides explicit representation and numeric quantifications to imperfect data. The overall process presents the issues of starting with imperfect data to the construction of probabilistic database(s) (Magnani and Montesi, 2010; Dong, Halevy, and Yu, 2009; Sarma, Dong, and Halevy, 2011).

When considering the new requirements and challenges of explicit imperfect information management, the CDI framework fails to cope and support this complex scenario of integration as stated by the recent literature studies (Cooper and Devenny, 2009; Magnani and Montesi, 2010; Ioannou et al., 2011). The CDI framework's components are formulated based on precise integration handling and outcome, which is a major undertaking that requires significant upfront human intervention. Moreover, the CDI framework treats instance integration as a secondary issue by assuming the existence of universal key identifier and a fair number of attributes' values (Jaradat, 2015). The key to these challenges is the traditional Decision Model (DM) in the CDI process. The DM encapsulates its logic with the manual matching process due to the precise and manual review enforcement and the iterative process at different entity orders' needs. Thus, DM presents a major obstacle to achieving automated and efficient linkage results. As a consequence, the CDI framework cannot automatically and efficiently address the information integration over heterogeneous information spaces using the traditional pre-selected thresholds and decision model logic; and hence, it needs to be extended or replaced (Jaradat, 2015).

The main contributions of this paper are introducing a new integration framework that takes imperfect information management and instance integration as a primary component in its structure. This work supports the best-effort perspective by considering the ER as a non-trivial problem and uses the traditional source-to-target framework as its base in adding a new element to its original structure. In this paper, the authors relate the CDI framework's notation and its ER associated approaches. In particular, the paper presents the implemented resolution process, DM logic approach, and the CDI

framework's limitations to address the ER problem over heterogeneous information sources. This paper's main contribution is to present a framework including concept, formulation and process. Finally, the paper concludes by identifying corresponding functionality resolutions challenges that require further probabilistic management modelling attention.

2 Related Work

From the traditional perspective, the CDI framework is formulated based on precise integration components of (Θ, S, M), where Θ can be a target (T) or global schema (G), S is a local schema, and M is a mapping from S to Θ. Throughout this formulation, the DM logic is performed using a comparison function, i.e. $sim\ (r_1, r_2)$, predefined threshold value(s), i.e. $\{\delta_m, \delta_p\}$, and a precise DM of $(Match\ [M], Possible - Match\ [P], Non - Match\ [U])$ classes or (M, U) classes, to determine to which class the compared pair will be assigned (Subramaniyaswamy and Pandian 2012; Elmagarmid, Ipeirotis, and Verykios, 2007). Furthermore, classes' assignment process is preceded iteratively in independent logic, or dependent logic at recent ER works by jointly identifying and resolving related items (Kalashnikov et al., 2008; Bhattacharya, Getoor, and Licamele, 2007; Ioannou and Staworko 2013; Ioannou and Velegrakis 2016). Figure 1 shows the dependent integration process between three sources, where ER are collectively done in iterative and propagated process.

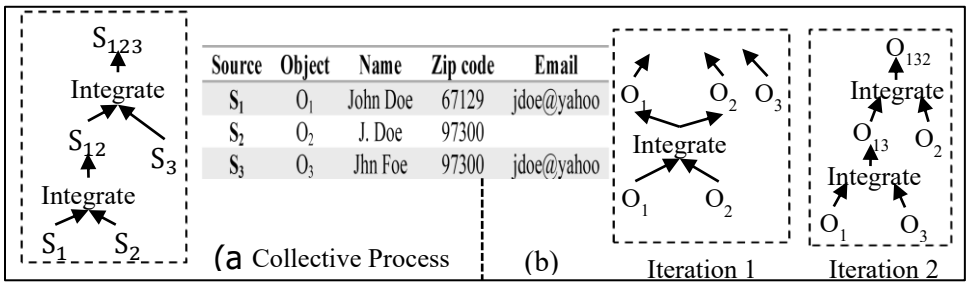

Fig. 1 The traditional pair-wise iteration for integrating three sources (Jaradat, 2015)

Recently the works of Cooper and Devenny (2009), Magnani and Montesi (2009), and Sarma, Dong, and Halevy (2009) have demonstrated how to implement or extend CDI framework based on its traditional processes and DM using varied collective ER approaches by varied probabilistic schema generation proposals.

Xu and Embley (2003) defined the framework I as triple $(T, \{S_i\}, \{M_i\})$, where T is a target schema, $\{S_i\}$ is a set of n source schemas, and $\{M_i\}$ is a set of n source-to-target mappings, such that for each S_i schema there is a mapping M_i from S_i to T, $1 \leq i \leq n$ (Xu and Embley, 2003). Both target and source schemas in I are presented in rooted graphs, where the graph includes a set of *objects sets* O and a set of *relationship sets* R. For a schema H, which is either a source or a target schema, the union of O and R is denoted by $\sum_H$ and V_H depicts the extension of $\sum_H$ with derived O and R sets. A source-to-target mapping M_i for S_i schema on T the schema is a function of $f_i(V_{S_i}) \longrightarrow \sum_T$. The mapping shows inter-schema correspondences between a S_i schema and a T schema.

Pankowski (2008) extended the above framework by adding probability to its mapping element. He defined a probabilistic XML integration framework as $(S, T, M_{ST}, Prob)$, where (S, T, M_{ST}) is an ordinary source-to-target framework and $Prob$ is a probability function over M_{ST}: $\forall\, m \in M_{ST}$, $Prob(m) \in [0,1], \sum_{m \in M_{ST}} Prob(m) = 1$. Moreover, Sarma (2009) provided a framework for DIS with the uncertainty that extended the source-to-target framework by providing probabilistic schema mappings (pM), and probabilistic mediated schemas (M) (Sarma, Dong, and Halevy, 2011). He refers to a *source schema* as $\bar{S}$, and to a relation in $\bar{S}$ as $S = \langle s_1, \dots, s_m \rangle$. Similarly, he refers to the *target schema as* $\bar{T}$, and to a relation in $\bar{T}$ as $T = \langle t_1, \dots, t_n \rangle$. Schema mappings are considered in a limited form. It was in sort of attribute correspondences as $c_{ij} = (s_i, t_j)$, where s_i is a source attribute in the schema S and t_j is a target attribute in the schema T. Based on that, a pM is a triple (S, T, m), where $S \in \bar{S}, T \in \bar{T}$, and $\boldsymbol{m}$ is a set of $\{(m_1, Pr(m_1)), \dots, (m_l, Pr(m_l))\}$, such that for $i \in [1, l], m_i$ is a one-to-one mapping between S and T, for every $i, j \in [1, l], i \neq j \Longrightarrow m_i \neq m_j$, and $\sum_{i=1}^{l} Pr(m_i) = 1$. A probabilistic mediated schema for $\{S_1, \dots, S_n\}$ sources is a set of $= \{(M_1, Pr(M_1)), \dots, (M_l, Pr(M_l))\}$, where for each $i \in [1, l]$, M_i is a mediated schema for $(S_1, \dots, S_n)$, and for each $i, j \in [1, l], i \neq j, M_i$ and M_j correspond to different clustering of the source attributes, and $\sum_{i=1}^{l} Pr(M_i) = 1$. Moreover, each M_i for the set $\{S_1, \dots, S_n\}$ sources is denoted by $M_i = \{A_1, \dots, A_m\}$, where each of the A_i is called a mediated attribute. The mediated attributes are *sets* of attributes from the data sources, i.e. $A_i \subseteq A$; for each $i, j \in [1, m], i \neq j \Rightarrow A_i \cap A_j = \emptyset$, such that A is the set of all source attributes, i.e. $A = (attr(S_1) \cup \dots \cup attr(S_n))$, and $attr(S_i)$ denotes the attributes in the schema $S_i, i \in [1, n]$.

As stated by Magnani and Montesi (2010) that the goal of uncertain or imperfect data integration is the usage of the imperfection presented in the data sources and/ or generated during the matching process to build an uncertain integrated view of the data that is represented in a sort of several alternative results. They also sated in Magnani and Montesi (2007) that models for data integration with uncertainty management can represent integrated data sources resulted from uncertain data integration processes.

In fact, providing an explicit uncertainty management integration approach requires us to deal with uncertain or imperfect data at different levels or results during the matching and integration processes and phases, while obtaining a compact representation of alternative results associated with confidence values that indicates the level of correctness. Therefore, the uncertainty or imperfection management modelling are required to represent and handle imperfect and heterogeneous data sources, the generated results from the matching processes and its Decision Model (DM) logic, and the generated results from the merging processes, i.e. the mediated schema, the global instance, and/ or the merged data values. (Magnani and Montesi, 2010; Magnani and Montesi, 2007, Dong, Halevy, and Yu, 2009, Sarma, Dong, and Halevy, 2009).

In table 1, we compare the main papers that are related to our work with regards to the data integration framework with explicit uncertainty management. In particular, we focus on the main data integration tasks and elements that require explicit uncertainty management handling and how the related works addressed them.

Table 1 Comparison between different data integration approaches.

		Matching Process	**Merging Process**

Criteria / Related works	The type data sources	Similarity Technique (Collective / Source to target)	DM Logic (Precise/ Most likely Probabilistic)	Level of Integration (Schema, Instance, Data fusion)	The Merging Result (Precise/ Probabilistic)
Cooper and Devenny (2009)	Text data represented in sort of entities. Data sources are associated with a reliability measure	Collective process	*K* Most likely: using a predefined threshold value	Data fusion level	Precise outcomes based on predefined threshold values.
Magnani and Montesi (2009)	Text data represented in sort of sets of schema objects.	Source to target process	Probabilistic	Schema level	Probabilistic
Sarma, Dong, and Halevy (2011)	Text data represented in sort of sets of schema objects.	Source to target process	Probabilistic	Schema level	Probabilistic
Xu and Embley (2003)	Text data represented as sets of schema objects.	Source to target process	Precise DM logic	Schema level	Precise outcomes
Pankowski (2008)	Text data represented in XML Format	Source to the target process	Probabilistic	Schema level	Probabilistic outcomes
Whang and Garcia-Molina (2012)	Text data represented in sort of records	Collective and iterative process	Precise and collective DM logic	Instance level	Precise outcomes based on the predefined threshold value.
Nikolov et al. (2008)	Text data represented in OWL knowledge bases	Iterative and collective process	Precise DM logic based on Threshold values	Instance and data fusion levels	Precise outcomes based on predefined threshold values
Panse Approach (Panse & Ritter 2010; Panse et al., 2010; Panse 2015)	digital data represented in sort of relational probabilistic data	Source to the target process	Precise DM logic based on Threshold values	Instance and data fusion levels	Probabilistic outcomes based on predefined rules with threshold values
Ioannou Approach (Ioannou et al., 2013; 2012; 2011)	Text data represented in sort of entities	Source to target and collective processes	Probabilistic DM Logic	Instance and data fusion levels	Probabilistic outcomes
Bhattacharya, Getoor, and Licamele (2007)	Text data represented in sort of returned query's records	Collective process	Precise DM logic	Instance level	Precise outcomes based on collective resolution
Christen (2008)	Text data represented in sort of records	Source to the target process	Precise DM based on Threshold values	Instance level	Precise outcomes based on predefined threshold values

Based on the summarised discussion above, we can outline several lessons learned regarding the need of a proper uncertainty management modelling that can explicitly

address and represent the imperfection at the entity resolution level, i.e. instance and data fusion phases. These remarks have directly motivated researcher to extends the CDI framework, the data matching process and its DM logic, and the ER process to explicitly cope, represent and manage uncertain data. The lessons learned are stated as follows:

- The above CDI framework definitions maintain instance integration as either a precise or in a static manner when processed based on the traditional collective and iterative resolution. Therefore, existing ER methodologies based on the CDI framework suffer from one of the main drawbacks of the database systems in supporting heterogeneous, imperfect and volatile data (Magnani and Montesi, 2010, Blanco et al., 2010, Ioannou et al., 2011, Ioannou et al., 2012). Our review shows a large number of collective ER efforts on extending the CDI framework at the schema level are failing to completely or closely produce accurate linkage results without manual reviews (Cooper and Devenny, 2009; Whang and Garcia-Molina 2012; Nikolov et al., 2008; Bhattacharya, Getoor, and Licamele, 2007). The failure occurs due to the usage of the preselected threshold value(s) that requires users to be extra careful. In fact, tuning threshold values are a difficult and very much a domain-dependent process.
- The works Magnani and Montesi (2010), Ioannou (2011), and Panse & Ritter (2010) show that ER over heterogeneous information sources is not a secondary issue in the integration framework. Furthermore, the process can handle the schema integration by assuming the existence of universal key identifiers and a fair number of attribute values.
- ER with explicit uncertainty management up to the end of the resolution process cannot be efficiently addressed using the traditional collective and iterative pair-wise resolution process, which modifies the actual data based on the DM's resulted data to be collectively used in the followed iterations. Several iterative processes at different orders are needed since collective ER's results are highly depended on the order of the examined items (Bhattacharya, Getoor, and Licamele, 2007; Christen, 2008; Ioannou et al., 2012). Figure 1(b) shows the integration of three objects, as originated from three sources, due to the use of the traditional iteration. Note here that iterations one and two results in different objects' order and may lead to different resolution results.
- ER with uncertainty management cannot be automatically handled using the traditional DM logic due to the volatile data nature, and the encapsulation process of the matching outputs with the DM logic, which has been treated as a part of the matching classification and leads to a major limitation concerning its impracticability for probabilistic ER. If the new attribute value is added or updated, the user needs to make new decisions. DM cannot be adjusted individually, or during clustering, hence, it needs to be continuously repeated. Therefore, it becomes a discouraging, difficult and inefficient process. Moreover, adding a new source might cause the process to reach a stage that does not reflect the reality of the participated sources since the process is limited to two sources, i.e. the new source and the merged source with the modified or removed data. Previous merging decisions cannot be modified even if new negative evidence might appear in the newly participated source (Blanco et al., 2010; Ioannou et al., 2012).
- The ER with uncertainty management cannot be explicitly incorporated using traditional probabilistic approaches. Traditional probabilistic approaches failed to maintain matching outputs in a sort of probabilistic data sets, or the probabilistic matching outputs had been replaced by predefined rules to generate precise mapping decisions, such as the works presented in Panse et al. (2010), and Magnani and Montesi (2007).

- The work of Panse (2015) shows CDI's inefficient handling of data fusion with explicit uncertainty management by using probabilistic data fusion strategies that do not consider the probabilistic linkage and merge outcomes.
- Majority of the works, except for Panse (2015), represent the integration of text/ string data format. The integration of data sources containing both string and digital data such as images is not taken into consideration so far.
- Finally, the promising efforts presented by Sarma, Dong, and Halevy (2011), Magnani and Montesi (2010), and Ioannou et al. (2012) show encouraging results toward data integration with explicit uncertainty management.

Consequently, this paper proposes an alternative integration framework that considers volatile data, instance integration, and uncertainty management as primary elements in the framework structure.

The proposed framework in this paper can complement the works done by Sarma, Dong, and Halevy (2011) and by Magnani, and Montesi (2010) toward a complete and formal information integration framework with explicit uncertainty management incorporation. It also goes in parallel with recent ER efforts done by Ioannou et al. (2013; 2012; 2011) and Panse (2015) toward a comprehensive probabilistic ER solution. It extends the probabilistic database definition to capture the probability at the linkage level besides type-I and type-II levels. Furthermore, our proposed framework in regards to its ER features and process comes with additional features to suit the problem space of imperfect and heterogeneous data, which can be exited in multiple formats such as text and images. The framework generates a probabilistic global entity (nDO) that exhibits the benefits of the iDO concept representation and its categorisation rules. Utilizing the iDO concept in the proposed framework makes it more practical and effective; as it allows the generation of domain-independent resolution rules based on the attributes (sub-) categories, and their mapping correlations with the *EoI*. Thus, more-general dependency rules can be created that suit varied information space domains without the need for further modification. These rules can reduce the uncertainty in the possible-worlds generation, and hence, enhance the ER results. Besides that, the proposed framework avoids pitfalls that may result from the one-time prior and collectively merging decisions. It also can support volatile data more efficiently as the linkage decision is separated from the merging decision, computing and maintain decision and merger separately, and merge on the fly upon a user's request. Therefore, no entities' merging is performed in advance, where any addition of new entities requires only the computation of the linkages or, in some cases, the re-computation of the probability of existing linkage. Moreover, the clear separation of matching, linkage, and merging processes and results in the implemented ER process, i.e. chain of subsequent phases of the resolution would make the process more efficient as any update can be accepted without the need for re-executing the whole process.

3 The Proposed Framework

This paper introduces a framework to support the automation of instance integration by explicitly incorporating probabilistic management to the ER tasks. It also aims to extend the pair-wise-source-to-target CDI framework based on the best-effort perception and to consider entities integration over multiple heterogeneous and volatile information sources. The proposed new framework presents a new alternative to pair-wise-source-to-target ER process that effectively copes with the uncertainty management manipulation and avoids the traditional iterative and collective cost. The framework implements a process by

maintaining the actual data of the participated sources, separating the matching process from the ER decision logic, assigning a probabilistic value of similarity between the matched data pairs, as well as the probabilistic linkage information. Moreover, the new framework operates over a collection of entities that describe Real-World Objects (RWOs) originated from heterogeneous and volatile information sources. These entities are represented based on the iDO concept to be the unified data model for the participated sources (Deraman et al., 2009, Deraman et al., 2005). The participated entities are described using iDO conceptualisation and attributes categorisations to enable rich entity representation, distinguish between strong and weak attribute evidence, and provide domain independent ER rules. Finally, the framework introduces an alternative probabilistic DM to replace the precise one, which assures the production of a probabilistic global entity (nDO_v) and represents multi-valued attributes over pairs of probabilistic entity linkages. Figure 2 shows the dataflow processes and outputs in our proposed framework.

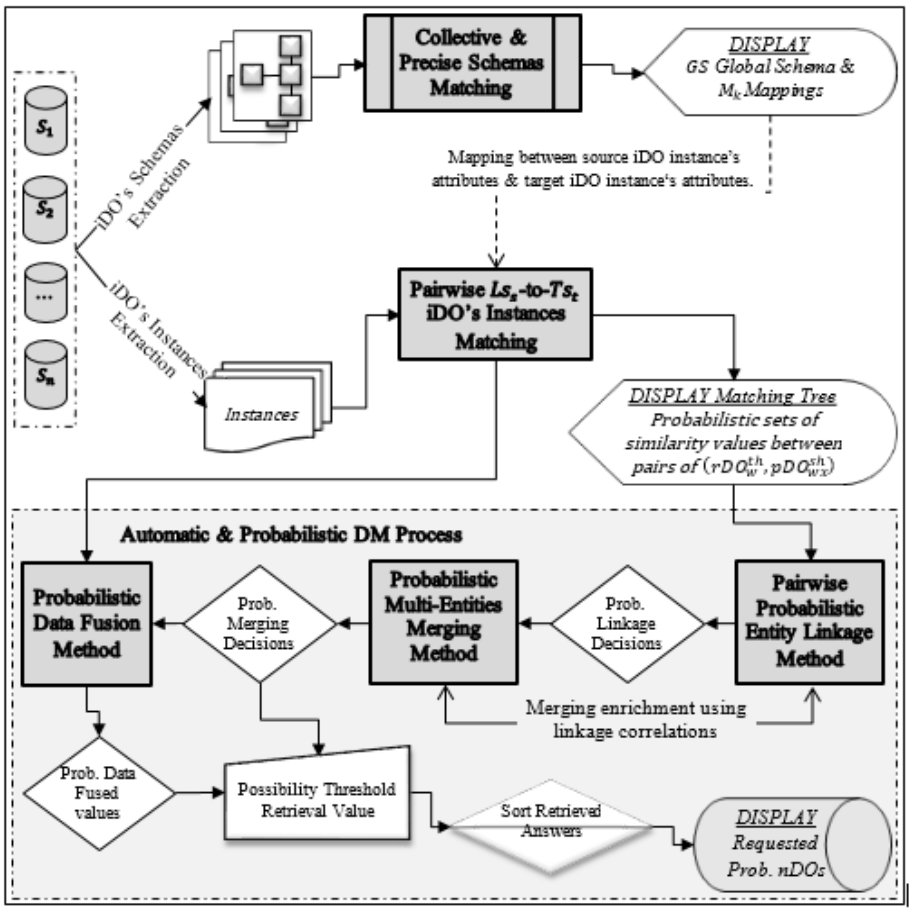

Fig. 2 The data flow process and outcomes for the proposed best-effort framework

In the following sections, the authors discuss related concepts to the proposed best-effort integration framework; the unified data model, the possible world generation

rules, the framework formulation, the implemented resolution process and its matching stage, and the probabilistic ER DM.

3.1 The Unified iDO Model

The participated information sources are sets of n sources, i.e. $S = \{S_1, S_2, \dots, S_n\}$, that are roughly assumed to refer to the same domain. Each participated source $S_i.tp \in S$ is a type of semi-structured (tp_1), or unstructured (tp_2), i.e. $tp \in Tp = \{tp_1, tp_2\}$. This helps to distinguish the matching process among these types of sources. The $S_i.tp_1$ sources are assumed to contain unique instances; hence internal comparisons aren't required since no duplicated objects can exist. Yet, $S_i.tp_2$ sources may contain duplicated instances; hence, matching process will proceed internally. A source schema is denoted by $\bar{S_i}$ that consists of a set of C attributes, i.e. $\bar{S_i} = \{A_1, A_2, \dots, A_C\}$. Each S_i contains a set of m $iDOs$, where the ranges of m in these sources are different. iDO_{ih} is a triple of a comprise set of attribute names, values and types, i.e. $(A_{ik}, a_{k.g}^{ih}, ty): i \in [1,n], h \in [1,m], k \in [1,C], and\ g \in [1,q]$. An attribute may have a single value, i.e. $A_k.a_k^{ih}$, or multi-valued, i.e. $A_k.a_{k.g}^{ih}: 1 < g \leq q$.

To facilitate distinguishing between strong and weak attribute evidence, the authors classify relationship by mapping into four categories based on different relationship-cardinalities between attributes and their iDOs. Table 2 shows four types of relationships as identified by the iDO model.

Table 2 The mapping types that influence the ER decision

MAPPING TYPE	DEFINITION
$M_1: (iDO\begin{smallmatrix}1 \to 1\\ 1 \leftarrow 1\end{smallmatrix}A_k)$	One iDO can only have one data value for an attribute, and one attribute value can only correspond to one iDO, such as the mapping between a Person with its IC.
$M_2: (iDO\begin{smallmatrix}1 \to 1\\ M \leftarrow 1\end{smallmatrix}A_k)$	One iDO can only have one data value for an attribute, but one attribute value may refer to many iDOs, such as the mapping between a Person and its Age attribute.
$M_3: (iDO\begin{smallmatrix}1 \to M\\ 1 \leftarrow 1\end{smallmatrix}A_k)$	One iDO can have many data values for an attribute, but one attribute value may refer to one iDOs, such as the mapping between a Person and its Phone-No.
$M_4: (iDO\begin{smallmatrix}1 \to M\\ M \leftarrow 1\end{smallmatrix}A_k)$	One iDO can have many data values for an attribute, and one attribute value may refer to many iDOs, such as the mapping between object Person and its Address.

Additionally, to provide domain independent ER rules, we classify attributes into three major categories, based on their content.

- *Identification category*: this category represents the *main parameter* attribute, i.e. $(\mathrm{A}_k.ty = \mathrm{Idn} \to \ddot{\mathrm{A}}_k)$ as well as other fundamental features of an iDO that indicate the nature of an object and provide clear identity information to a specific one. Based on the mapping types presented in Table 2, It is sub-divided into five subcategory's types to distinguish the *main parameter*, which represents the entity of interest for resolution

(*EoI)*, from other identification attributes that may have different mappings to their *EoI*, i.e. $ty \in Iden = \{Idn, Idn\text{-}I, Idn\text{-}II, Idn\text{-}III, Idn\text{-}IV\}$, such as a person's *Name*, *IC-No*, and *Passport No.*

- *Descriptive category:* this category represents the associated features that provide additional and related information about a specific iDO. Those features might be changed, such as person's age, facial features, address, occupation and salary. By using the mapping types above, the authors can subdivide this category into four subcategories, i.e., $ty \in Desc = \{Desc\text{-}I, Desc\text{-}II, Desc\text{-}III, Desc\text{-}IV\}$, such as phone-No, age, and address.
- *Supportive category:* It contains the external information that an iDO may have. This information is the features from associated objects to provide more details about an iDO. The associated objects' information can be identifiable or descriptive properties for the external objects that link it to the selected iDO. This category is also sub-divided into four subcategories types, i.e.$ty \in Supp = \{Supp\text{-}I, Supp\text{-}II, Supp\text{-}III, Supp\text{-}IV\}$, such as Venue, Article Name, and Co-author Name.

3.2 Possible world generation rules

By classifying the iDO attributes based on the above sub-categories, the new framework can generate general possible-worlds rules by assigning objects to their corresponding subcategories, and hence, an independent domain's rules can be created to determine the impossible worlds for each possible linkage case or alternative, i.e. $lc_l = (lc_1 \parallel lc_2)$. Based on that, the possible worlds set (Psw_{Lc_l}) that are associated with each linkage case can be generated accordingly. This generation is achieved based on the sample space production of the probabilistic similarity value for a pair of iDOs attribute's values, i.e., $\Omega_{a_{k.g}}^{int(wx)}$={$(a_{k.g}^{wx}, Pr(a_{k.g}^{wx}))$, $((a_{k.g}^{w}, a_{k.g}^{x}), Pr(a_{k.g}^{w}, a_{k.g}^{x}))$}: w and x refers to the compared iDOs. Possible world generation rules for each linkage case are stated in Table 3.

Table 3 The mapping rules to obtain the possible worlds sets for (lc_1, lc_2) cases

MAPPING TYPE	**Possible Worlds Generation Rule**
M_1	$If\ Mapping = M_1, then\ (a_k^{wx}, Pr(a_k^{wx})) \in Psw_{Lc_1}, \&\& \left((a_k^w, a_k^x), Pr(a_k^w, a_k^x)\right) \in Psw_{Lc_2}$
M_2	$If\ Mapping = M_2, then\ (a_k^{wx}, Pr(a_k^{wx})) \in Psw_{Lc_1}, \&\& \left((a_k^{wx}, Pr(a_k^{wx})) \vee ((a_k^w, a_k^x), Pr(a_k^w, a_k^x))\right) \in Psw_{Lc_2}$
M_3	$If\ Mapping = M_3, then\ \left((a_k^{wx}, Pr(a_k^{wx})) \vee \left((a_k^w, a_k^x), Pr(a_k^w, a_k^x)\right)\right) \in Psw_{Lc_1}, \&\& \left((a_k^w, a_k^x), Pr(a_k^w, a_k^x)\right) \in Psw_{Lc_2}$
M_4	$If\ Mapping = M_4, then \left((a_k^{wx}, Pr(a_k^{wx})) \vee \left((a_k^w, a_k^x), Pr(a_k^w, a_k^x)\right)\right) \in Psw_{Lc_1}, \&\& \left((a_k^{wx}, Pr(a_k^{wx})) \vee \left((a_k^w, a_k^x), Pr(a_k^w, a_k^x)\right)\right) \in Psw_{Lc_2}$

3.3 Framework Formulation

The schema integration is beyond the research scope of this paper. Therefore, we assumed a global schema and mapping production is prior performed by creating the global schema (GS). This process is performed to initiate the ER process by selecting all attributes included in the GS, or a subset from GS.

The proposed best-effort framework takes instance integration as a non-trivial process that requires probability management capabilities. Hence, a probabilistic global entity named *digital network object* (nDO_v) is added to the framework formulation. This framework also aims to remove the manual interventions by obtaining less precise but automatic ER answers. In correspondence, the proposed framework is formulated as given below:

Definition. The proposed best-effort information integration framework is four components of $(Ls_s, Ts_t, M_{s,t}, nDO_v)$; such that $(Ls, Ts_t, M_{s,t})$ are a precise source-to-target framework and nDO_v is the added probabilistic global entity, where;

- Ls_s Is a local source or relation that belongs to a Ls set of n local sources or relations, i.e., $Ls = (Ls_1, Ls_2, \dots, Ls_n): s \in [1,n], Ls_s \in S$. Each Ls_s is a type of $(tp_1\ or\ tp_2)$, in which its local instances (*reference entities)* are denoted as $pDO_x^{sh} = \{a_{1.1}^{ty}, \dots, a_{c.q}^{ty}\}: x \in [1,y], h \in [1,m], k \in [1,c], g \in [1,q], \exists\ a_{k.g}^{ty} = a_{k.g}^{idn}, \left|a_{k.g}^{idn}\right| = 1$.

- Ts_t is a target source that belongs to a Ts set of n target sources or relations, i.e. $Ts = (Ts_1, Ts_2, \dots, Ts_n): t \in [1,n], Ts_t \subseteq TS_t \in S$. Each Ts_t can be in type of $(tp_1 or\ tp_2)$, in which its target object instances (*underlying entities)* are denoted as $rDO_w^{th} = \{a_{1.1}^{ty}, \dots, a_{c.q}^{ty}\}: w \in [1,z], h \in [1,m], k \in [1,c], g \in [1,q], \exists\ a_{k.g}^{ty} = a_{k.g}^{idn}, \left|a_{k.g}^{idn}\right| = 1$.

- $M_{s,t}$ is a triple of $\left(Ts_t.rDO_w^{th}.\left(\ddot{a}_1, \dots, a_{c.q}^{ty}\right), Ls_s.pDO_x^{sh}.\left(\ddot{a}_1, \dots, a_{c.q}^{ty}\right), m_{t,s.k}\right)$. $M_{s,t}$ Mapping is a set of one-to-one probabilistic matching for each target attribute value $A_{tk}.a_{k.g}^{th}.ty \in rDO_w^{th}$ against a local attribute value $A_{sk}.a_{k.g}^{sh}.ty \in pDO_x^{sh}$, if initially there is $M_{s,t.k}\left(\ddot{a}_k^{sh} \sim \ddot{a}_k^{th}\right) \geq \delta: th \neq sh, and\ A_{tk} = A_{sk}, \exists\ tk, sk \in [11, nc]$. Thus, for each instance pairs from $iDO_{ih}^{Ins} = \cup_{i=1}^{n} \cup_{h=1}^{m} \cup_{k=1}^{c} \cup_{g=1}^{q} (iDO_{ih}^{tp}.a_{k.g}^{ty})$ there exists $Ls_s.iDO_{sh}$ against $Ts_i.iDO_{ih}$ source-to-target entities matching in the form of $pDO_x^{sh} \sim rDO_w^{th}$ between their shared attributes values $\ddot{A}_1^{sh}.\ddot{a}_1 \sim \ddot{A}_1^{th}.\ddot{a}_1, \dots, A_c^{sh}.a_{c.q} \sim A_c^{th}.a_{c.q}: \ddot{a}_1^{th} \sim \ddot{a}_1^{sh} \geq \delta$, $(\sim)$ denotes the pair-wise matching operation, and (δ) is the similarity threshold value for considering the matching between the pairs of main parameter's data values.

- nDO_v is a set of z mutual probabilistic global entities alternatives (nDOs) that are generated from merging their possible corresponding iDOs, which have pair-wise linkage in sort of an underlying entity to possible reference linkages, i.e. $\left(rDO_w^{th}: pDO_{w1}^{sh}[Pr_{w1}], \dots, pDO_{wy}^{sh}[Pr_{wy}]\right): 1 \leq w \leq z, 1 \leq x \leq y$. A probabilistic global entity contains a set of possible ordinary entities merged from the underlying entity with its possible references, i.e. $\left(nDO_v = nDO_{v.1}, \dots, nDO_{v.f}\right): 1 \leq j \leq f$. Each possible merge has an assigned probability distribution value generated from multiplying the probability linkages of its probable linked references, i.e. $nDO_{v.j} = \{(rDO_w^{th}: pDO_{w1}^{sh}, pDO_{w2}^{sh}, \dots, pDO_{wy}^{sh}), Pr(M_{v.j})\}: \sum_{j=1}^{f} Pr(M_{v.j}) = 1$. Furthermore, for each requested attribute and within each possible merge, there might be a multi-valued

attribute in which each possible attribute value alternative is assigned with a probabilistic fusion score obtained from updating and conditionally combining the reliability scores of its attribute values, i.e., $nDO_{v.j}.A_k = \{(a(\Omega_{Tv.1}), \mu(a(\Omega_{Tv.1}))), \dots, (a(\Omega_{Tv.L}), \mu(a(\Omega_{Tv.L})))\}: 1 \leq l \leq L, \sum_{l=1}^{L} \mu(a(\Omega_{Tv.l})) = 1$. An $a(\Omega_{Tv.1})$ world (alternative) may contain single or multi-possible true values, i.e. $a(\Omega_{Tv.l}) = \{a_{k.1}^{v.j}, \dots, a_{k.g}^{v.j}\}$.

3.4 The Implemented Resolution Process

The ER is performed in a pair wise-source-to-target process with correspondence to the extracted *local and target instances*. In this process, pair-wise probabilistic linkage results between an underlying entity and a set of possible local instances are produced based on the pair-wise similarity matching between the correlated data items, i.e. $rDO_w = \{rDO_w^{th}: pDO_{w1}^{sh}[Pr(L_{w1})], \dots, pDO_{wy}^{sh}[Pr(L_{wy})]\}$. Then, the probabilistic entity merging can be computed to produce a global entity, i.e. $nDO_v^w = \{(nDO_{v.1}^w, Pr(nDO_{v.1}^w)), (nDO_{v.2}^w, Pr(nDO_{v.2}^w)), \dots, (nDO_{v.f}^w, Pr(nDO_{v.f}^w))\}$. Figure 3(a) shows the ER process within three unstructured information sources, while figure 3(b) shows the ER process between their contained objects.

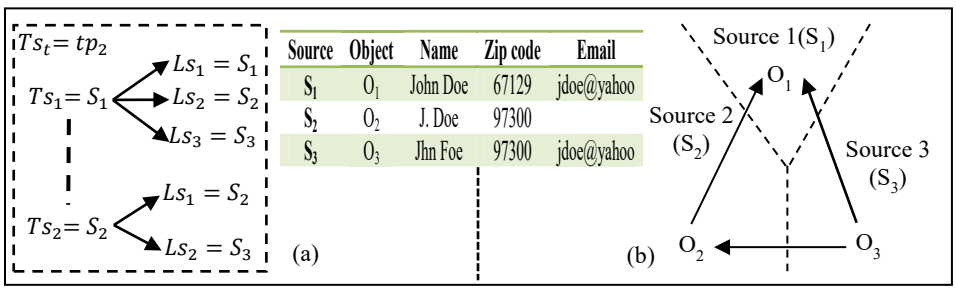

Fig. 3 The utilised pair-wise-source-to-target process for three information sources

From figure 3, we see that the iterations at different sources'/objects' order will not lead to different outputs. This process allows the maintenance of the original data, their lineage, and the separation of the matching, linkage, merging, and fusing outputs. It also allows the production, the storage, and the manipulation of probabilistic results. Thus, the proposed process overcomes the limitations of the traditional process and provides good practicability for probabilistic ER.

Despite the above illustration of the general resolution process, the instances comparison, which denotes the initial stage in this process, will be cover in further details next including the setting up of data inputs for resolution challenges and solutions.

3.5 The Matching Comparison Process

From the extracted data values and in correspondence to their referred instances, a pair-wise matching mechanism is carried pair-wisely by comparing a *local instance* to a *target instance* in the form of $|Ts_t.iDO_{th}|.|Ls_s.iDO_{sh}|$. Accordingly, the matching function is processed as shown in equation 3, where the abbreviated form of $|Ts_{th}|.|Ls_{sh}|$ is used.

$$|Ts_{th}|.|Ls_{sh}| = \begin{cases} \bigcup_{t=1}^{n}\bigcup_{h=1}^{m}|Ts_{th}|.\bigcup_{s=1}^{n}\bigcup_{h+1}^{m}|Ls_{sh}| : Ts_{th}.tp = tp_2, pDO_{sh} \neq rDO_{th} \\ \bigcup_{t=i=1}^{n-1}\bigcup_{h=1}^{m}|Ts_{th}|.\bigcup_{s=i+1}^{n}\bigcup_{h=1}^{m}|Ls_{sh}| : Ts_{th}.tp = tp_1; Ts_{th} \neq Ls_{sh}; s,t,i \in [S_1, S_n] \end{cases} \quad \text{(eq. 1)}$$

The example below is given to demonstrate the implemented matching process.

Example 1: Given multi-sources as $S_1(\mathrm{iDO}_{11}, \ldots, \mathrm{iDO}_{1m}), \ldots, S_n(\mathrm{iDO}_{n1}, \ldots, \mathrm{iDO}_{nm})$, such that $(S_i = \{1,2,\ldots,n\}, iDO_{ih} = \{i1, i2, \ldots, im\}: 1 \leq i \leq n, 1 \leq h \leq m)$. Then, the matching process is implemented as figure 4 shows.

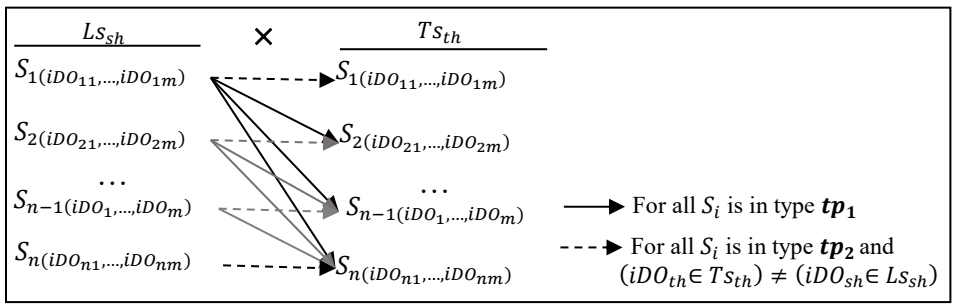

Fig. 4 The matching process according to the source type

Given the above matching process and the n-gram matching function in equation 2, the framework can generate probabilistic matching outputs in the form of a tree that views both target iDOs and their potential candidates of local iDOs with their matched attributes values pairs, i.e. $(pDO_x^{sh.Ins}.\ddot{a}_1, \ldots, a_{c.q}^{ty}) \sim (rDO_w^{th.Ins}.\ddot{a}_1, \ldots, a_{c.q}^{ty}) \rightarrow Pr_{w \sim x}(\ddot{a}_1, \ldots, a_{c.q}^{ty})$: $sh \neq th, w \in [1,z], pDO_x^{sh} \in Ls_s$, and $pDO_x^{sh.Ins}, rDO_w^{th.Ins} \in iDO_{ih}^{Ins}$, where iDO_{ih}^{Ins} is the union set of the extracted iDO instances from the participated Ls_s and Ts_t information sources due to a selected Gs schema, i.e. $iDO_{ih}^{Ins} = \cup_{i=1}^{n} \cup_{h=1}^{m} \cup_{k=1}^{c} \cup_{g=1}^{q} (iDO_{ih}^{tp}.a_{k.g}^{ty})$.

$$\mathrm{F}\ (\mathrm{Sim}\ \langle \mathrm{a}, \mathrm{b} \rangle) = \frac{\left|2 \times \sum_{t \epsilon\ n-grams\ (a) \cap n-grams\ (b)} \log \mathrm{P(t)}\right|}{\left|\sum_{t \epsilon\ n-grams\ (a)} \log \mathrm{P(t)}\right| + \left|\sum_{t \epsilon\ n-grams\ (b)} \log \mathrm{P(t)}\right|} \quad \text{eq. 2}$$

Where, *n-grams(a)* and *n-grams(b)* are the set of n-grams in *a* and *b* respectively, and *P(t)* is the probability of n-grams occurring in a word (Chen, Promparmote, and Maire, 2006). The n-gram function is used to find the probability of a string matching between pairs of attributes values by their objects construct. The process would decompose text strings into a set of tokens (n-grams), which are the contiguous N characters of the text string.

In this process, the similarity scores between *rDO* instance and its set of *pDOs*, are kept and stored alongside the original data to be used for the decision model stage. Accordingly, the process would separate original data items from the data that represent

the linkage decision. The separation enables suitable reusability when new data become available and helps to achieve automated and efficient probabilistic DM process for linkage.

3.6 The Probabilistic DM

In this proposed framework, a probabilistic DM is introduced to replace the traditional, precise DM. The proposed probabilistic DM considers a new resolution stage of multiple entities is merging instead of the collective pair-wise merging that is practically impossible and ineffective to be able to manage imperfection until the end. It also extends the probabilistic database definition to capture the probability management at the linkage level, besides type-I and type-II. The implemented DM takes the matching outputs for a pair of iDOs in sort of probabilistic similarity events and possible-worlds to produce probabilistic pair-wise linkage results without using preselected threshold values. Also, in this DM the probabilistic linkage values are stored alongside the linkage decisions to be used by the entities merging operation. Consequently, the process generates new probabilistic global entities (nDOs), and their multiple data values inconsistencies managed and fused. This outcome is a probabilistic database containing uncertainty management not only on the attributes level but also on their linkage and merging results.

The probabilistic DM is in charge of conceptually representing the probabilistic pair-wise entity linkage decisions and determining their posterior linkage scores, i.e. $L(rDO_{\mathrm{w}}^{th}: pDO_{\mathrm{wx}}^{sh}) \rightarrow \{(Lc_1, Pr(Lc_1)), (Lc_2, Pr(Lc_2))\}$. It is in charge of representing the multi-entities merging's alternatives and computing their probability distributions, i.e. $\{(nDO_{v.1}^{w}, \grave{P}r(nDO_{v.1}^{w})), \dots, (nDO_{v.f}^{w}, \grave{P}r(nDO_{v.f}^{w}))\}$. It is also in charge of representing the data fused values' alternatives and computing their updated reliability scores, i.e. $\{(a_{k.Tv}^{v.j} = a(\Omega_{tv.1_p}), \mu(a_{k.Tv}^{v.j} = (\Omega_{tv.1_p})|Pws_{Tv})), \dots, (a_{k.Tv}^{v.j} = a(\Omega_{tv.L_p}), \mu a_{k.Tv}^{v.j} = a((\Omega_{tv.L_p})|Pws_{Tv}))\}$.

4 Conclusion and further research

This paper presents a new best-effort information integration framework that copes with the current information integration requirements and challenges, where imperfect data management and the integration over heterogeneous information spaces are the major aspects. The proposed framework realises the dataspace vision by supporting instance integration automation with explicitly incorporating probabilistic management into the ER tasks. Moreover, the proposed framework relaxes the traditional source-to-target CDI framework by making the instance integration less precise but automatic, while continuing to provide useful services to both technical and non-technical users.

The proposed framework identifies, manages, and probabilistically models three functionality resolution challenges. These are (i) Probabilistic pair-wise-source-to-target entity linkage, (ii) Probabilistic entity merging over multiple possible linked references to an underlying entity, and (iii) Probabilistic data fusion values over multi-valued attributes. These challenges deploy three subsequent phases of resolution process based on the proposed framework and its probabilistic matching outputs. Each one of these sub-problems may comprise several cases and types of imperfection and conflict that need to

be specified and handled. Therefore, there is a great potential for future investigations and research that can be carried on in regards to the proposed framework and its additional challenges.

The proposed framework has been theoretically validated, however in our future publication there will be a practical demonstration using real-world datasets and a developed prototype named Impiana-I that will show the effectiveness of the framework to cope with the challenges outlined above. (Jaradat, 2015).

References

Aggarwal, C.C. and Philip, S.Y. (2009) 'A survey of uncertain data algorithms and applications.' *IEEE Transactions on Knowledge and Data Engineering*, *21*(5), pp.609-623.

Bhattacharya, I. and Getoor, L. (2007) 'Query-time entity resolution.' *Journal of Artificial Intelligence Research*, *30*, pp.621-657.

Blanco, L., Crescenzi, V., Merialdo, P. and Papotti, P. (2010) 'Probabilistic models to reconcile complex data from inaccurate data sources.' In *International Conference on Advanced Information Systems Engineering* (pp. 83-97). Springer Berlin Heidelberg.

Ceri, S., Braga, D., Corcoglioniti, F., Grossniklaus, M. and Vadacca, S. (2010) 'Search computing challenges and directions.' In *International Conference on Object and Databases* (pp. 1-5). Springer Berlin Heidelberg.

Chen, Y.P.P., Promparmote, S. and Maire, F. (2006) 'MDSM: Microarray database schema matching using the Hungarian method.' *Information Sciences*, *176*(19), pp.2771-2790.

Christen, P. (2008) 'Automatic record linkage using seeded nearest neighbour and support vector machine classification.' In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 151-159). ACM.

Cooper, R. and Devenny, L. (2009) 'A Database System for Absorbing Conflicting and Uncertain Information from Multiple Correspondents.' In *British National Conference on Databases* (pp. 199-202). Springer Berlin Heidelberg.

Deraman A, Jaradat A, Mokhtar N, Din L, and Said N. (2005) 'Digital object modeling: An approach to the development of danum valley biodiversity repository in danum valley.' *Conservation Area: Physical, Biological & Social Science Environments*, pp 384-392.

Deraman A, Yahaya J, Salim J, Idris S, Jaradat A, Jambari D, Komoo I, Leman M, Unjah T, Sarman M, and Sian L. (2009) 'The development of myGeo-RS: A knowledge management system of geodiversity data for tourism industries'. *Communications of the IBIMA*, 8(19), pp. 142-146.

Dittrich J, Salles Dittrich, J., Salles, M.A.V. and Blunschi, L. (2009) 'iMeMex: From Search to Information Integration and Back.' *IEEE Data Eng. Bull.*, *32*(2), pp.28-35.

Dong, X.L. and Halevy, A. (2005) 'August. A platform for personal information management and integration.' In *Proceedings of VLDB 2005 PhD Workshop* (p. 26).

Dong, X.L. and Naumann, F. (2009) 'Data fusion: resolving data conflicts for integration.' *Proceedings of the VLDB Endowment*, *2*(2), pp.1654-1655.

Dong, X.L., Halevy, A. and Yu, C. (2009) 'Data integration with uncertainty.' *The VLDB Journal—The International Journal on Very Large Data Bases*, *18*(2), pp.469-500.

Elmagarmid, A.K., Ipeirotis, P.G. and Verykios, V.S. (2007) 'Duplicate record detection:

A survey.' *IEEE Transactions on knowledge and data engineering*, *19*(1).

Franklin, M., Halevy, A. and Maier, D. (2005) 'From databases to dataspaces: a new abstraction for information management.' *ACM Sigmod Record*, *34*(4), pp.27-33.

Haase, P. and Völker, J. (2008) 'Ontology learning and reasoning—dealing with uncertainty and inconsistency.' In *Uncertainty Reasoning for the Semantic Web I* (pp. 366-384). Springer Berlin Heidelberg.

Halevy, A., Franklin, M. and Maier, D. (2006) 'Principles of dataspace systems.' In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 1-9). ACM.

Ioannou, E. and Velegrakis, Y. (2016) 'Searching web 2.0 data through entity-based aggregation.' In *Transactions on Computational Collective Intelligence XXI* (pp. 159-174). Springer Berlin Heidelberg.

Ioannou, E. and Staworko, S. (2013) 'Management of inconsistencies in data integration.' In *Dagstuhl Follow-Ups* (Vol. 5). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Ioannou, E., Nejdl, W., Niederée, C. and Velegrakis, Y. (2012) 'Embracing Uncertainty in Entity Linking.' In *Semantic Search over the Web* (pp. 225-253). Springer Berlin Heidelberg.

Ioannou, E., Nejdl, W., Niederée, C. and Velegrakis, Y. (2011) 'LinkDB: a probabilistic linkage database system.' In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (pp. 1307-1310). ACM.

Ioannou, E., Nejdl, W., Niederée, C. and Velegrakis, Y. (2010) 'On-the-fly entity-aware query processing in the presence of linkage.' *Proceedings of the VLDB Endowment*, *3*(1-2), pp.429-438.

Ioannou, E., Niederée, C. and Nejdl, W. (2008) 'Probabilistic entity linkage for heterogeneous information spaces.' In *Advanced Information Systems Engineering* (pp. 556-570). Springer Berlin/Heidelberg.

Jaradat, A. (2015) *Best effort resolution model for imperfect instance management using probabilistic network digital object approach*. Ph.D. Thesis. The National university of Malaysia, Malaysia.

Jiang, Z. (2008) 'Reconciling Continuous Attribute Values from Multiple Data Sources.' *PACIS 2008 Proceedings*, p.264.

Kalashnikov, D.V., Chen, Z., Mehrotra, S. and Nuray-Turan, R. (2008) 'Web people search via connection analysis.' *IEEE Transactions on Knowledge and Data Engineering*, *20*(11), pp.1550-1565.

Kuicheu, N.C., Wang, N., Tchuissang, G.N.F., Xu, D., Dai, G. and Siewe, F. (2013) 'Managing Uncertain Mediated Schema and Semantic Mappings Automatically in Dataspace Support Platforms.' *Computing and Informatics*, *32*(1), pp.175-202.

Li, P., Dong, X.L., Guo, S., Maurino, A. and Srivastava, D. (2015) 'Robust group linkage.' In *Proceedings of the 24th International Conference on World Wide Web* (pp. 647-657). ACM.

Magnani, M. and Montesi, D. (2007) 'Uncertainty in data integration: current approaches and open problems.' *In Proceedings of the first International VLDB Workshop on Management of Uncertain Data (MUD) in conjunction with VLDB* (pp. 18-32).

Magnani, M. and Montesi, D. (2010) 'A survey on uncertainty management in data integration.' *Journal of Data and Information Quality (JDIQ)*, *2*(1), p.5.

Magnani, M. and Montesi, D. (2009) *Probabilistic data integration*. Tech. rep. UBLCS-2009-10, University of Bologna.

Motro, A. and Anokhin, P. (2006) 'Fusionplex: resolution of data inconsistencies in the integration of heterogeneous information sources.' *Information fusion*, *7*(2), pp.176-196.

Nikolov, A., Uren, V., Motta, E. and De Roeck, A. (2008) 'Integration of semantically annotated data by the KnoFuss architecture.' In *International Conference on Knowledge Engineering and Knowledge Management* (pp. 265-274). Springer Berlin Heidelberg.

Pankowski, T. (2008) 'Reconciling inconsistent data in probabilistic XML data integration.' In *British National Conference on Databases* (pp. 75-86). Springer Berlin Heidelberg.

Panse, F. (2015). Duplicate Detection in Probabilistic Relational Databases. PhD thesis. University of Hamburg, Germany.

Panse, F. and Ritter, N. (2010) 'Tuple Merging in Probabilistic Databases.' In *MUD* (pp. 113-127).

Panse, F., Van Keulen, M., De Keijzer, A. and Ritter, N. (2010) 'Duplicate detection in probabilistic data.' In *Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on* (pp. 179-182). IEEE.

Papadakis, G., Ioannou, E., Palpanas, T., Niederee, C. and Nejdl, W. (2013) 'A blocking framework for entity resolution in highly heterogeneous information spaces.' *IEEE Transactions on Knowledge and Data Engineering*, *25*(12), pp.2665-2682.

Sarma, A.D., Dong, X.L. and Halevy, A.Y. (2009) 'Data modeling in dataspace support platforms.' In *Conceptual Modeling: Foundations and Applications* (pp. 122-138). Springer Berlin Heidelberg.

Sarma, A.D., Dong, X.L. and Halevy, A.Y. (2011) 'Uncertainty in data integration and dataspace support platforms.' In *Schema Matching and Mapping* (pp. 75-108). Springer Berlin Heidelberg.

Strong, D.M., Lee, Y.W. and Wang, R.Y. (1997) 'Data quality in context.' Communications of the ACM, 40(5), pp.103-110.

Subramaniyaswamy, V. and Pandian, S.C. (2012) 'A complete survey of duplicate record detection using data mining techniques.' *Information Technology Journal*, *11*(8), p.941.

Whang, S.E. and Garcia-Molina, H. (2012) 'Joint entity resolution.' In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on* (pp. 294-305). IEEE.

Xu, L. and Embley, D.W. (2003) 'Using schema mapping to facilitate data integration'. *ER03*.

Zhang, W., Lin, X., Pei, J. and Zhang, Y. (2008) 'Managing uncertain data: Probabilistic approaches.' In *Web-Age Information Management, 2008. WAIM'08. The Ninth International Conference on* (pp. 405-412). IEEE.

Ziegler, P. and Dittrich, K.R. (2004) 'Three Decades of Data Integration—all Problems Solved?' In *Building the Information Society* (pp. 3-12). Springer US.