Entity Identity Reconciliation based Big Data Federation - A MDE approach

J.G. Enríquez *jose.gonzalez@iwt2.org*
Department of Computer and Language Systems,
University of Seville, Seville, Spain

F.J. Domínguez-Mayo *fjdominguez@us.es*
Department of Computer and Language Systems,
University of Seville, Seville, Spain

M.J. Escalona *mjescalona@us.com*
Department of Computer and Language Systems,
University of Seville, Seville, Spain

J.A. García-García *julian.garcia@iwt.org*
Department of Computer and Language Systems,
University of Seville, Seville, Spain

Vivian Lee *vivian.lee@uk.fujitsu.com*
Fujitsu Laboratories of Europe
Hayes, Middlesex, U.K.

Masatomo Goto *masatomo.goto@uk.fujitsu.com*
Fujitsu Laboratories of Europe
Hayes, Middlesex, U.K.

Abstract

"*Information is power*" is a sentence attributed to Francis Bacon that acquired a high important in the current era of the information. However, too much information can be a negative aspect. The term of "*Infoxication*" refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information. With the increasing of relevance of open data and big database, the application of mechanisms and solutions to manage information is critical. This paper introduces the problem of unique identification and data reconciliation and offers a discussion about how to solve this problem in big and open data environment. The problem of data reconciliation in multiple databases and the unique identification of entities is not a new problem, but, how effective are classical mechanisms in the new internet environment? In this paper a solution based on model-driven engineering and virtual graph is presented in order to improve the processing of information in big open repositories. The paper illustrates the idea with a real example for the right exploitation of heritage information in the south of Spain.
Keywords: Entity Reconciliation, Deduplication, Model-Driven Engineering, Graphs, Big Data, Open Data.

1. Introduction

The era of information that we are living provokes that the management of information is critical in practically full aspects of our life. However, with the incorporation of information and communication technologies (ICT) in our life, society, companies and people are

suffering the effect of "Infoxication". It refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information.

In the first ear of ICT, the main problem was how to store information, how to search and how to manage it in an efficient way. Currently, with Big Data and Cloud Computing present in our lives, the main problem is how I could use this information to extract the knowledge that it can offer me.

This paper is focused in this sense. When we have a high number of data or heterogeneous information in a concrete field, it is necessary to explore techniques that help me to manage them and to extract all the knowledge. In our concrete study, we are want to explore techniques for improving the unique identification of objects in heterogeneous, disperse and big data bases.

The necessity of unique identification is no new. Since database starting to be used, the necessity of this identification appears. When we have to integrate or extract data from a same identity from different databases, not always is a simple task to determinate which the same entities are. However, if we work in the context of big and open databases, which are disseminate via internet, the unique identification and data reconciliation is more critical for two main reasons: 1) databases in this area are continually changing so we cannot make a static reconciliation, we have to produce a dynamic unique identification and reconciliation. 2) Frequently, we do not know the internal structure of our source of data.

Big Data is the latest buzzword that gains the popularity in recent years. In addition to the original "3Vs" characteristics defined by Gartner [1] – volume (large amount of data), velocity (high speed of data in and out), and variety (heterogamous data types and sources), recently, many organizations have further added two new features: veracity (the quality of the data be captured), and complexity (data management can become a very complex process). These two latest features proved that the more experiences organizations gain from dealing with Big Data, the more they see the importance of the quality of the data source, and how to link, connect and correlate data from multiple sources, in comparison to only the speed and quantity. The view is also confirmed by Gregory Piatetsky [2], in his interview with Michael Berthold, that in many Big Data projects, there is no large data analysis happening, but the challenge is in the extract, transform, load part of data preprocessing.

The benefits of connecting data from heterogeneous sources are in two fold, firstly, it is very unlikely that multiple sources all contain the same information on a concerned entity; therefore, combining different information provides users with a more comprehensive view, hence helps with a better decision making. Secondly, when there is information overlapping among different sources, connected information provides an effective way to cross validate the quality of each information source.

In this paper we present a model-driven approach for improving the task of unique identification and data reconciliation. Our solution is based in the use of a dynamic data structure.

For this aim, the paper is structured as follow. In Section 2 we present the background analyzing related works and solutions related with reconciliation of entities in bid data. In Section 3 we present a global view of our approach and we illustrate it with real example. The Paper finishes with Conclusions and Future Works.

2. Background

In most of the Big Data processing scenarios, physical data migration or storage is not practical because of the data size and the speed that data are generated. Thus, in order to be able to analyzing Big Data at runtime, data federation technology is the most feasible way to create a view on interested data sources at certain time. However, building such kind of virtual database can be difficult for Big Data. This is not only because the Big Data can be very large in volume, but also because of its variety – data can be structured, semi-structured, or unstructured.

Currently, most of the existing data federation technologies and products are developed and tested for structured data environment, even the well-known Big Data product Hadoop is mainly focusing on large data set processing after the data are stored inside its system. It is authors' believe that there are no known technologies are dealing with Big Data federation at the time of writing.

In order to look for previous works that can help us to define our environment, in this Section we present a global view of the problem of entity resolution problem, which is no new but it offering new challenges in Big Data. Presented the current situation for this problem, we start looking for solutions in the context of structured databases, which has widely worked with this problem and later we present why new solutions in the big data context have to be offered.

2.1. Entity Resolution Problem

Entity resolution (ER) problem is a fundamental problem in data integration dealing with the combination of data from different sources to a unified view of the data. Entity resolution is inherently an uncertain process because the decision to map a set of records to the same entity cannot be made with certainty unless these are identical in all of their attributes or have a common key.

ER problem consist on extracting, matching and resolving entity mentions structured and unstructured data. This, is a long-standing challenge in database management. In other words, entity resolution, aims at "cleaning" a database by identifying tuples that represent the same entity. However, this problem is called by many different names like: record linkage, deduplication, co-reference resolution, reference reconciliation, object consolidation, identity uncertainty, uncertainty management, merge-purge and database hardening.

The need for data integration stems from the heterogeneity of data the lack of sufficient semantics to fully understand the meaning of data, and errors that may stem from incorrect data inser-tion and modifications. Solving the ER problem, we can get solutions for a lot of different domains like commercial interests, scientific studies and veracity of data.

Despite the long history of work on ER there is still a surprising diversity of approaches like: rule based methods, pair-wise classification, clustering approaches, and richer forms of probabilistic inference and a lack of guiding theory. Meanwhile, in the age of big data, the need for high quality entity resolution is only growing. We are inundated with more and more data that needs to be integrated, aligned and matched before further utility can be extracted. [5], [6], [7].

With a body of research that spans over multiple decades, data integration has a wealth of formal models of integration [13], [14], [15], [16], algorithmic solutions for efficient and effective integration [17], [18], [19], [3], and a body of systems, benchmarks and competitions that allow comparative empirical analysis of integration solutions [20], [21], [7].

For instance, Ioannou et Al. [3] describe a novel framework for entity linkage with uncertainty. In this paper, instead of using the linkage information to merge structures a-priori, possible linkages are stored alongside the data with their belief value proposing a new probabilistic query answering technique that is used to take the probabilistic linkage into consideration. Ioannou et Al. formally define the semantics, describe an efficient implementation and report on the findings of their experimental evaluation:

- Entity Linkage. Most of the existing entity linkage techniques focus on off-line identification and linkage of the data describing the same real world objects.
- Probabilistic Data. Few existing data integration proposals focus on dealing with uncertainty that appears in the data through the applied entity linkage algorithms.

2.2. A structured database perspective

Entity reconciliation in a structured database, e.g. a relational database, is a well-explored issue. The technology name can be further divided into two overlapping categories:

1. Physical integrity - deals with the correctness of data storing and fetching.

2. Logical integrity - makes sure the piece of data is correct or reasonable given a particular context.

The integrity in such kind of database system is reinforced by integrity constrains, for example, entity integrity through primary key, or referential integrity through foreign key.

Some techniques founded on the literature for solving this king of problem are:

1. ARTEMIS [22]: this tool environment is developed to support the analyst in the process of analyzing and reconciling sets of heterogeneous data schemas. Schema analysis is performed according to the concept of affinity, while schema reconciliation is performed interactively on clusters of elements with affinity, based on unification rules.
2. Freebase [23]: is a practical, scalable, graph-shaped database of structured general human knowledge, inspired by Semantic Web research and collaborative data communities such as the Wikipedia.
3. MFIBlocks [18]: is based on iteratively applying an algorithm for mining Maximal Frequent Itemsets. It waives the need to manually design a blocking key, the value of one or more of a tuple's attributes. In the other hand, it localizes the search for similar tuples and is able to uncover blocks of tuples that are similar in multiple, possibly overlapping sets of attributes.
4. Febrl [20]: allows even inexperienced users to learn and experiment with both traditional and new record linkage techniques. It is written in Python and its source code is available, it is fairly easy to integrate new record linkage techniques into it. Febrl can be seen as a tool that allows researchers to com-pare various existing record linkage techniques with their own ones, enabling the record linkage research community to conduct their work more efficiently.
5. Swoosh: Benjelloun et All formalize the generic ER problem, treating the functions for comparing and merging records as black-boxes, which permits expressive and extensible ER solutions. Benjelloun et All identify four important properties that, if satisfied by the match and merge functions, enable much more efficient ER algorithms. In this paper, it is shown the development of three efficient ER algorithms: G-Swoosh for the case where the four properties do not hold, and R-Swoosh and F-Swoosh that exploit the four properties. F-Swoosh in addition assumes knowledge of the "features" (e.g., attributes) used by the match function.

However, these kinds of integrity rules are normally predefined with rigid structure, which, if not completely impossible, is very difficult to reinforce, in a Big Data environment.

2.3. Virtual Graphs approach for Big Data perspective

In the age of big data, the need for high quality entity resolution is growing, as we are inundated with more and more data, all of which needs to be integrated, aligned and matched, before further utility can be extracted. [1]

Entity identity reconciliation in the Big Data context means two things: firstly, data come in free form, structured, semi-structure, or unstructured, and secondly, data are from open source, e.g. most of them are not rigidly curated. In comparison to a structured database reconciliation, these characterises require the underlying solution to be flexible, and also able to produce accurate results. In this sense, the solutions proposed on the previous section, are not compatible with this envirоinment that is the reason why it is needed to look for new solutions.

In this envirоinment we propose a solution based on virtual graphs. Graph technology is a nature solution to handle Big Data, especially for modelling relationships between entities. The variety of graph algorithms, for example, Dijkstra, A*, and Kruskal offers great flexibility in different situations. Theoretically, graphs can be represented in two ways: explicitly and implicitly. An explicit graph is a collection of elements that can be completely stored in memory, which means each vertex and edge of the graph is fixed at the time it is

stored. On the other hand, an implicit graph is a graph that cannot be stored in memory because of its large size (Mondal and Deshpande, 2012).

At the University of Seville, we have been conducting research into implicit graph for many years; we have formally named our concept Virtual Graphs. The ability of building the graph at runtime, allows us to build different solutions to tackle many business scenarios, where the fixed predefined data model cannot cope with the extensibility or the unpredictable availability of the data sources.

Thus, this solution helps to solve the problems that big data presents in respect to the rigid structure of the databases.

3. A MDE approach to Entity Identity Reconciliation

Model-Driven Engineering (MDE) [8] emerged to address the complexity of software systems in order to express the concepts of the problem domain in an effective way. In this line, the basic principle of MDE is "Everything is a model" [9].

The main idea of the MDE is use a set of models for decreasing the level of abstraction. Thus, on the early stages of the development, the models are more abstract than the final stages where the models are closer to implementation. However, working on this way, there are some necessities that must be satisfied:

- Have a set of common elements for developing all the models by the same way.
- Define mechanisms to make possible to new models from other ones. [10]

For the first one, comes up the concept of metamodel. The objective of a metamodel is defining the relationships between concepts of a problem domain and defining the semantic associated to them.

For second one, is used the mechanism of transformations. A transformation between two models, represent a relationship between two abstract syntaxes and it is defined by a set of relationships between the elements of the metamodel. [11]

MDE can be applied by different ways. One of them is defined by the MOF (Meta-Object Facility) standard. MOF is considered a metamodel, it means, a tool for building metamodels or even transformation languages that use the metamodels for specify transformations. [12]

In the following picture, it is shown how these concepts are combined to generate new models from other existing models.

MDE appears as a possible guide to pose a solution because:

- MDE paradigm works with models for representing the information of a domain. In this context, with the use of models, we are going to be able to structure the information.
- MDE paradigm also works with transformations. In this sense, we are going to be able to extrapolate our problem to different domains.

3.1. Using MDE to Entity Identity Reconciliation

In this section, we propose an approach based on MDE and virtual graphs methodology for solving the problem of the entity reconciliation.

This approach is based on four main sections which will provide us the mechanisms to generate solutions for different environments. These principal sections are: "virtual graph metamodel", "connection metamodel", "entity reconciliation metamodel" and "view metamodel".

1. Virtual graph metamodel: As it is mentioned before, our solution is based on MDE and Virtual Graphs. For this reason, the first section the metamodel of a virtual graph. Basically, it is an extended version of a graph metamodel.
2. Entity Reconciliation metamodel: this metamodel will let us reconcile different data sources. Here, we will define the problem inheriting from the previous metamodel and adding different data sources.

3. Data Source metamodel: with this metamodel, we are going to be able to connect the different data sources mentioned before for getting the information that is wanted to be reconciled through specific connectors defined for each data source.
4. View metamodel: with this metamodel, we will represent the information that we have reconciled before.

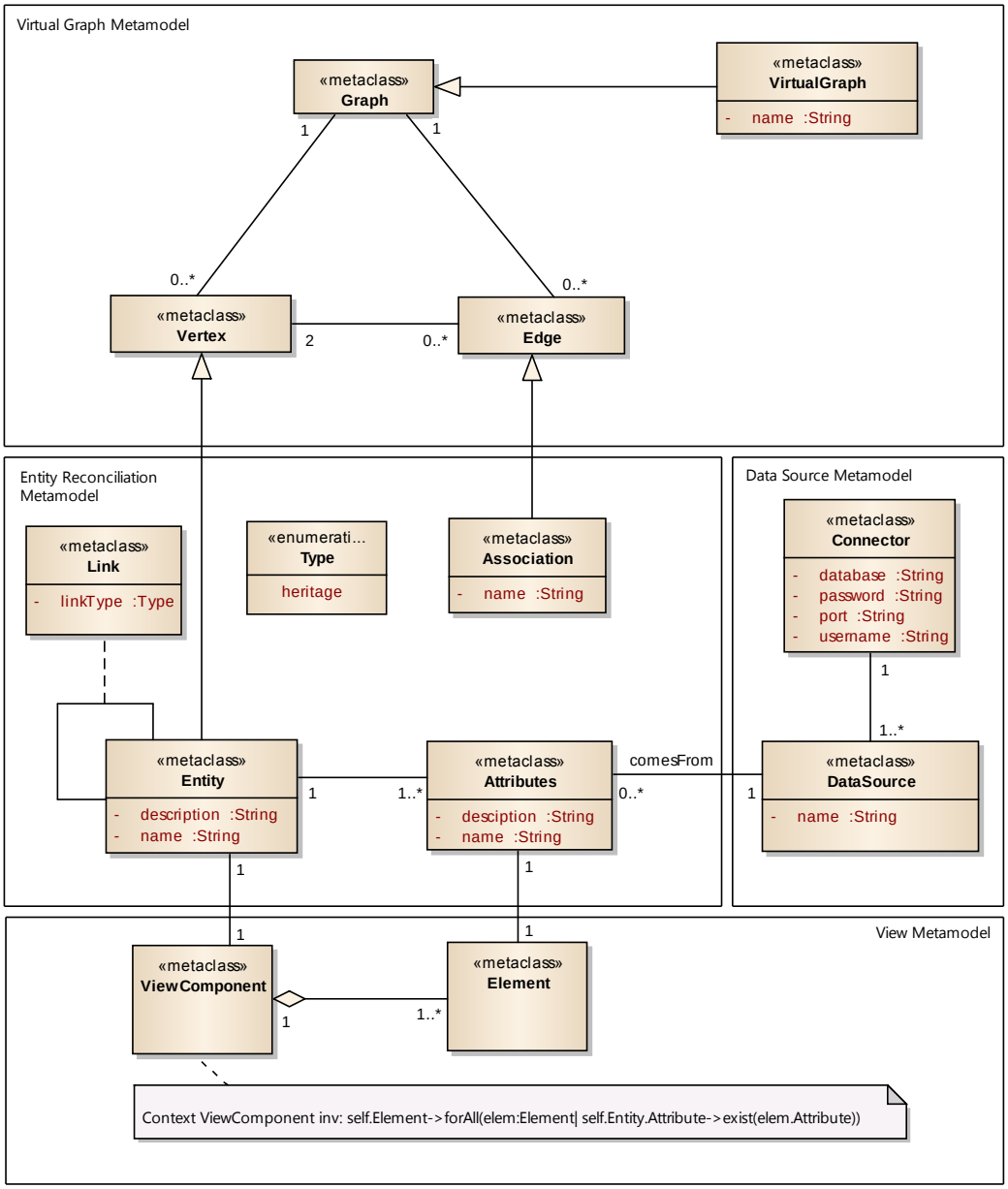

Fig 1. Metamodel for solving the ER problem

Explaining in depth the previous model:

1. Virtual Graph Metamodel:

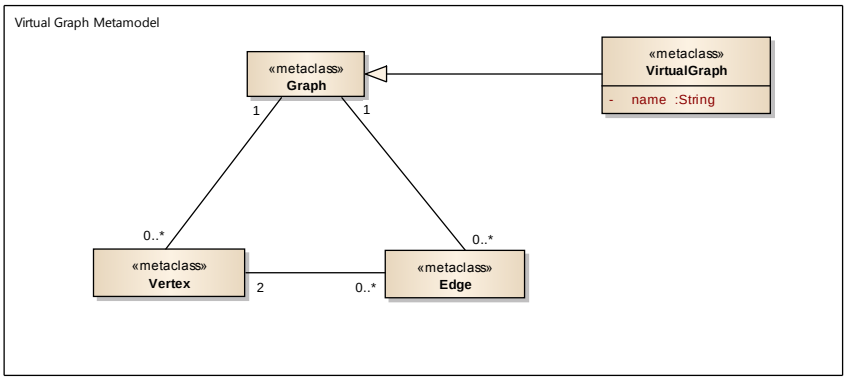

Fig 2. Virtual Graph Metamodel Section

- Graph: it is the main metaclass. It represents information about an explicit graph.
- Vertex: it represents information about the set vertexes that compose a graph. It has an association with the metaclass "Graph" where one graph can contain zero or more vertexes and one vertex, is a part of one graph.
- Edge: it represents information about the set of edges that compose a graph. It has two associations: one with the class "Graph" where one graph can contain zero or more edges and an edge is a part of one graph and another one with the metaclass "Vertex", where one edge consists of two vertex and one vertex, can contain zero or more edges.
- VirtualGraph: it represents information about an implicit graph and it inherits from the metaclass "Graph".

2. Entity Reconciliation Metamodel:

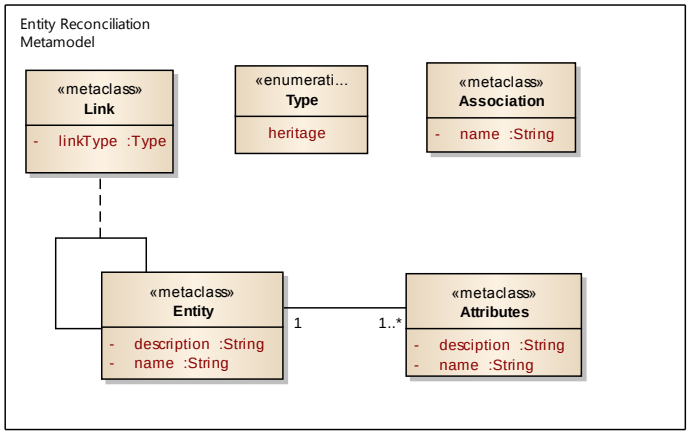

Fig 3. Entity Reconciliation Metamodel Section

- Entity: it is one of the most important metaclasses of the metamodel. It represents information about the entities that we want to reconcile. It inherits from the metaclass "Vertex", it means, that all the entities that we want to reconcile are going to be transformed in vertexes of our graph. Also, it has a related metaclass called "Link".
- Link: it is a metaclass that save information about the entity. This metaclass has an enumeration type called "Type" that contains an attribute called heritage that is going to be the relation between entities.
- Type: it is an enumeration type for defining the types of links that the entities can be composed of.

- Association: this metaclass represents information about the relations between data sources and inherits of the metaclass “Edges”.
- Attributes: the attributes metaclass represents information about the attributes of the entities. It has a an association with the metaclass “Entity” where an specific entity has one or more attributes and one attribute is related just with one entity.

3. Data Source Metamodel:

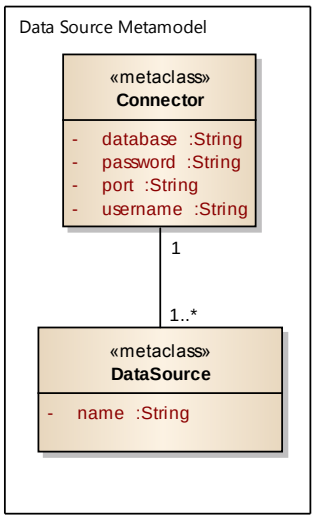

Fig 4. Data Source Metamodel Section

- Connector: it is a metaclass that store information about the way of connecting with a data source.
- Data Source: it represents information about the data sources that we want to reconcile. It has relation with the metaclass “Connector” where one data source have one connector an a connector is a part of one or more data sources.

4. View Metamodel:

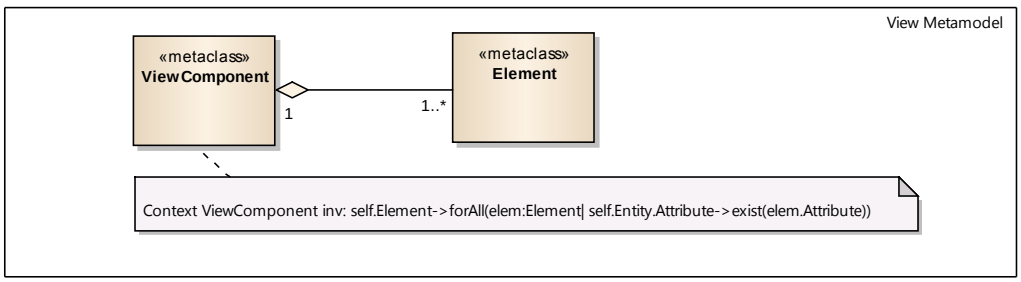

Fig 5. View Component Metamodel Section

- ViewComponent: the instantiation of this metaclass will let create the component necessary for representing the consolidated information. It has a composition relation with the “Element” metaclass where one view component is composed of one or more elements and one element is a part of one view component. This metaclass has a constraint comment also.
- Comment: this comment is a constraint for defining that one element is a part of a view component only if the attributes that represents, are contained in the entity that we want to reconcile.
- Element: this metaclass represent information about the elements that are going to be showed on the view components.

4. Conclusions

The era of ICT (Information and Communication Technologies) has covered our life with a high number of data and available information that, currently, it is difficult to manage and explore. The necessity of produce mechanisms, techniques and suitable tools to support this management is being every day a more critical aspect for companies and society in general.

In this paper we analyze a mechanism to explore a very concrete problem. In context of big data and open data, the unique identification and its right unification is critical. This concept is no new because expert in database have worked in algorithms and solutions for that, however, these classical approaches are not always useful in the new open context of Internet.

The paper presents a solution to solve identity reconciliation and unification under a model-driven perspective. The paper has presented a global view of the solution and a first draft of metamodels used for this aims. Besides, virtual graphs as data structure were also introduced.

5. Future Works

An intensive systematic literature review is being performed in the context of Big Data and Data Bases environments to know what the current state of existing ER solutions is and to compare all existing solutions in the literature with our approach. The main objective is to know the strengths and weaknesses of our solution with existing approaches but, up to now, no ER solutions based on dynamic structures in Big Data environments has been found.

In addition, it has been presented a set of metamodels that will let us instantiate them and design different solutions depending on the scenario where we are working with. So, it is necessary to apply our approach to different scenarios in order to see how well our solution works and to be able to validate it. Then, there are two real scenarios in which our approach can be applied:

- The first one is for the management of cultural heritage information in the Andalusian region (Spain), this is a big issue in which there are lots of monuments and several data sources where the information is stored. In addition, the size and complexity of these data sources make complicated the management of these systems due to the large amount of information stored on them. An illustrative example is "MOSAICO" (the official database of cultural heritage information of government of the Andalusian region) that contains many terabytes of information. Then, it is necessary to uniquely identify the existing information about monuments from all data sources. In this environment, it is being developed the application “DIPHDA” (Dynamic Integration for Patrimonial Heritage Data in Andalucía) with the collaboration of the Fujitsu Laboratories of Europe (FLE). The objective of DHIPDA is to achieve significantly improved accuracy and data management efficiency, based on reconciliation logic applied to open data information, as opposed to simple string matching reconciliation. This solution will be capable of integrating management systems, such as “MOSAICO”, diffusion systems such as “Europeana”, as well as open data information provided by Wikipedia and Yelp, as examples.
- The second one is in the domain of e-Health, more specifically to be able to accurately identify patients. This is a big challenge since it requires the advanced solutions to allow different clinics to exchange healthcare information in a reliable and secure way. Moreover, for those organizations that exchange healthcare information without using a common unique identifier or reconciled identity information, the successfulness of the information linkage is relying entirely on the accuracy and completeness of the key demographic data. So, it is very important to help these organizations to retrieve patient information from several data sources.

Other important future work is the definition of a development process lifecycle for the application of our approach to different scenarios. Then, it is needed to define a set of phases

which will let us analyze and design models, deploy them in an executable environment and finally, execute the solution.

As regards the deployment and executable environment is concerned, it is also necessary to develop a platform to support the derived code from designed models. This platform must support the previous lifecycle process and will let us deploy ad execute any solution designed for our approach to different domains.

Acknowledgements

This research has been supported by the MeGUS project (TIN2013-46928-C3-3-R) of the Spanish Ministry of Science and Innovation.

References

[1] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011. Retrieved 13 July2011.

[2] Gregory Piatetsky (2014-08-12). "Interview: Michael Berthold, KNIME Founder, on Research, Creativity, Big Data, and Privacy, Part 2". KDnuggets. Retrieved 2014-08-13.

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

[4] Mondal, J., Deshpande. 2012. Managing large dynamic graphs efficiently. A. pp. 145–156.

[5] Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). Swoosh: a generic approach to entity resolution. The VLDB Journal—The International Journal on Very Large Data Bases, 18(1), 255-276.

[6] Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: theory, practice & open challenges. Proceedings of the VLDB Endowment, 5(12), 2018-2019.

[7] Gal, A. (2014). Tutorial: Uncertain Entity Resolution. Proceedings of the VLDB Endowment, 7(13).

[8] Schmidt, D. C. (2006). Guest editor's introduction: Model-driven engineering. Computer, 39(2), 0025-31.

[9] Bézivin, J. (2005). On the unification power of models. Software & Systems Modeling, 4(2), 171-188.

[10] Fondement, F., & Silaghi, R. (2004, October). Defining model driven engineering processes. In Third International Workshop in Software Model Engineering (WiSME), held at the 7th International Conference on the Unified Modeling Language (UML).

[11] Thiry, L., & Thirion, B. (2009). Functional metamodels for systems and software. Journal of Systems and Software, 82(7), 1125-1136.

[12] OMG. *Meta Object Facility (MOF™) Core*. Object Management Group, http://www.omg.org/spec/MOF/. 2011b.

[13] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C. Saita. Declarative Data Cleaning: Language, Model and Algorithms. In Proc. of Int'l Conf. on Very Large Databases (VLDB), 2001.

[14] I. P. Fellegi and A. B. Sunter. A Theory for Record Linkage. Journal of the American Statistical Association, 64(328):1183–1210, 1969.

[15] A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi. A framework for modeling and evaluating automatic semantic reconciliation. The VLDB Journal, 14(1):50–67, 2005.

[16] Z. Bellahsene, A. Bonifati, and E. Rahm, editors. Schema Matching and Mapping. Springer, 2011.

[17] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02, pages 269–278, New York, NY, USA, 2002. ACM.

[18] Kenig, B., & Gal, A. (2013). MFIBlocks: An effective blocking algorithm for entity resolution. Information Systems, 38(6), 908-926.

[19] Gal, A., & Sagi, T. (2010). Tuning the ensemble selection process of schema matchers. Information Systems, 35(8), 845-859.

[20] Christen, P. (2008, August). Febrl-: an open source data cleaning, deduplication and record linkage system with a graphical user interface. InProceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1065-1068). ACM.

[21] Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. Knowledge and Data Engineering, IEEE Transactions on,24(9), 1537-1555.

[22] Castano, S., & De Antonellis, V. (1999, August). A schema analysis and reconciliation tool environment for heterogeneous databases. In Database Engineering and Applications, 1999. IDEAS'99. International Symposium Proceedings (pp. 53-62). IEEE.

[23] Bollacker, K., Cook, R., & Tufts, P. (2007, July). Freebase: A shared database of structured general human knowledge. In AAAI (Vol. 7, pp. 1962-1963).

[24] Getoor, L., & Machanavajjhala, A. (2013, August). Entity resolution for big data. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1527-1527). ACM.