Entity Identification Problem in Big and Open Data

J. G. Enríquez[1], Vivian Lee[2], Masatomo Goto [2], F. J. Domínguez-Mayo [1] and M. J. Escalona [1]
[1]*Department of Computer and Language Systems, University of Seville, Seville, Spain*
[2]*Fujitsu Laboratories of Europe, Hayes, Middlesex, U.K.*

Keywords: Software Engineering, Big Data, Open Data, Entity Identification, Intelligent Reconciliation, Virtual Graphs.

Abstract: Big and Open Data provide great opportunities to businesses to enhance their competitive advantages if utilized properly. However, during past few years' research in Big and Open Data process, we have encountered big challenge in entity identification reconciliation, when trying to establish accurate relationships between entities from different data sources. In this paper, we present our innovative Intelligent Reconciliation Platform and Virtual Graphs solution that addresses this issue. With this solution, we are able to efficiently extract Big and Open Data from heterogeneous source, and integrate them into a common analysable format. Further enhanced with the Virtual Graphs technology, entity identification reconciliation is processed dynamically to produce more accurate result at system runtime. Moreover, we believe that our technology can be applied to a wide diversity of entity identification problems in several domains, e.g., e-Health, cultural heritage, and company identities in financial world.

1 INTRODUCTION

In the current Big Data era, quintillions bytes of data are produced everyday. Better utilization of Big Data has shown great benefits to organizations for accurate and faster decision-making, thus enhancing business performance and competitive advantage. However, the 3Vs' nature of the Big Data, e.g. velocity, variety, and volume, have also presented grand challenges to most of the traditional information systems, in terms of data processing, integration, and analysis (Manyika et al, 2011).

In addition to the Big Data, Open Data has also emerged as another hot topic recently. The idea is that certain data should be freely available to everyone, without any restrictions. Open data has gained more popularity in the recent years, especially with the new initiatives of open-data government such as Data.gov and Data.gov.uk (Official websites of government: data.gov and data.gov.uk, 2014). The power of Big and Open Data is enormous, if one can properly unlock and discover the insight. However, one of the major hurdles during the chained process - from receiving the Big and Open Data to applying proper data analytic tool is that, data are always siloed. Further more, even data are integrated from different sources, reconcile the information to refer to the same entity, proved to be a big challenge, the reasons being different data sources issue different local identity to the entity that may already has other identity in other systems, in the mean time, different business domain has proprietary way of defining local identity without a standard mechanism.

In this paper, we present our work to tackles this issue. We introduce our technical background in section II, and give out scenarios where entity identification is the key issues in section III. Section IV further explains our technology in details, and we draw conclusions to our work in section V. Finally, we present our future road maps in section VI.

2 BACKGROUND

2.1 Big Data Platform

Fujitsu Laboratories of Europe Ltd. has been conducting Big Data and Open Data related research since the emergence of the phenomena. Over the years, we have matured our Big Data platform - BigGraph for data integration, storage, and processing. Empowered by Linked Data (Berners-Lee, 2006) technology, the BigGraph platform is able to efficiently extract, and integrate Big Data from heterogeneous sources in variety of types, into a common analysable format. Figure 1 is a simple illustration of the platform:

G. Enríquez J., Lee V., Goto M., J. Domínguez-Mayo F. and J. Escalona M..
Entity Identification Problem in Big and Open Data.
DOI: 10.5220/0005470704040408
In *Proceedings of the 17th International Conference on Enterprise Information Systems* (ICEIS-2015), pages 404-408
ISBN: 978-989-758-096-3

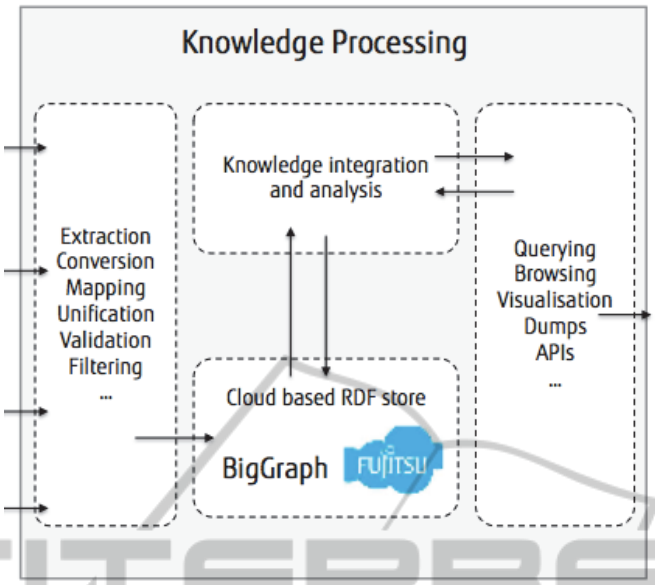

Figure 1: Fujitsu Big Data Platform.

At the time of writing, the platform is able to handle data sources that contain structured, semi-structured, and un-structured data types.

It is also worth explaining that Linked Data relies on two technologies that are fundamental to the Web: Uniform Resource Identifiers (URIs) (Berners-Lee, 2005) and the HyperText Transfer Protocol (HTTP) (Fielding, 1999). While Uniform Resource Locators (URLs) have become familiar as addresses for documents and other entities that can be located on the Web, Uniform Resource Identifiers provide a more generic means to identify any entity that exists in the world.

2.2 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.

3 INSPIRATION

In a recently published white paper, James and Nigel stated “managing identifiers is easier in a closed system. The web has many advantages, but it presents challenges for identifiers because of its vast scale and their ad hoc usage” (Powell and Shadbolt, 2014). In real life, we clearly see problems in different domains that require entity identifications to be reconciled. Some detailed examples are explained as follows:

Management of cultural heritage information is a big issue in the Andalusian region (Spain), because there are lots of monuments that need to be uniquely identified. In addition, there are different kinds of relationships between monuments. For instance, a monument can be associated with another one, or it can be grouped within a set of monuments. Furthermore, a monument can contain lots of artworks, and these artworks can be associated to other existing monuments. Assigning/reconciling identifications and build relationships between monuments requires a very comprehensive system that current does not exist yet.

In e-Health, to be able to accurately identify patients 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.

Another typical example of the identification reconciliation issue is the identifiers of the companies worldwide. In financial world, there are many different regulators who produce variety of financial reports for companies all over the world. Each of these regulators issues unique identifiers within their own system to the companies. However, currently there is no any good system that can help users to integrate all the identifiers issued by different regulators, and relate them to the same company entity. Thus making integration of reports from different regulator almost impossible.

4 OUR IDEA

4.1 EIDER Model

Motivated by the problems and challenges described in the previous section, we feel the urge to develop a platform that is able to solve the entity identification problems, which allows organizations to gain even more benefits when utilizing Big and Open Data – hence the birth of EIDER platform.

Based on the mature BigGraph platform previously mentioned, EIDER platform is further enhanced to tackle the entity identification problem. The system architecture of EIDER is illustrated in the figure 2.

The system consists of three main functionalities:

• Entity Extraction
• Entity Reconciliation
• Data Integration

Each of the functionalities is further explained as follows.

• Entity Extraction

This function is responsible for extracting entity identity information from heterogeneous sources. We assume that each entity, for example, an organization, or a person, has a name, which the name can be formally registered name or a known/nick name. With this information, external data sources are scanned in order to detect and extract identification information related to this entity name.

• Entity Reconciliation

This is the key function that forms the heart of the EIDER platform. It reconciles variety of entity identities, which are extracted from different sources, and applies the intelligent algorithms for reconciliation. The algorithms are comprehensive methods that consist of string similarity metrics, string distance function, natural language processing, text mining, and graph traversal (Maali et Al, 2011).

• Data Integration

After entity identities are reconciled, more information can be extracted and integrated by using the reconciled identities. A very important usage of this function is to build reports, which the user might want to extract, for example, balance sheet, or publicly available finance performance data and integrate them into a single report. This allows users to gather more comprehensive knowledge from the Big and Open Data to enhance their decision-making (Insights on Data Integration Methodologies, 2008),

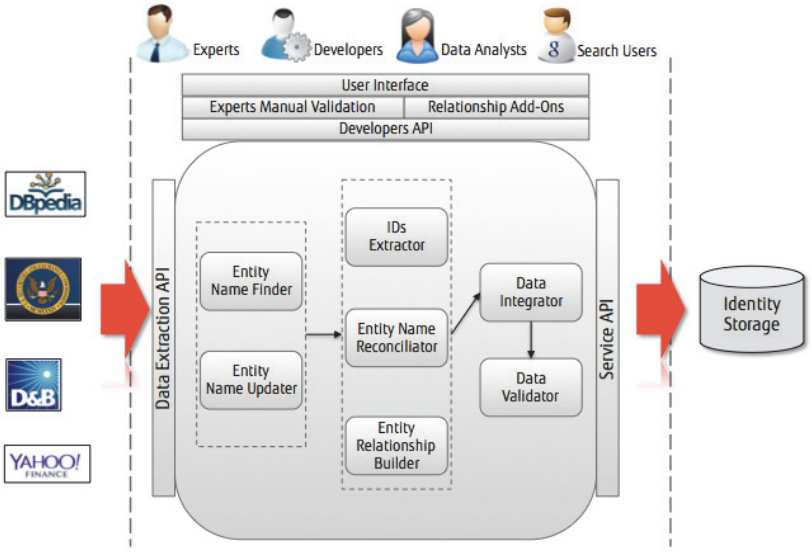

Figure 2: EIDER System Architecture.

although we do not restrict the application on private data as well.

Apart from the key functionalities that the EIDER system provides, another important aspect that worth mentioning here is the data freshness. Since information extracted by the EIDER system are Open Data, freshness is crucial in order to maintain the quality of data. EIDER employs patented method (EP14176955.4) to retrieve data at system runtime thus guarantees information in the EIDER system is as fresh as possible.

4.2 Virtual Graphs and EIDER Model

We see two layers of data reconciliations in the underlying system:

- EIDER System initial entity reconciliation
- Virtual Graphs enhanced entity reconciliation

The initial entity reconciliation is mainly text mining technology focused, which the main purpose is to allow accurate external entity identification extraction from multiple Open Data sources. It is a pre-reconciliation that makes the Big and Open Data ready for integration.

At the end of EIDER system processing chain, Big and Open Data are transformed into a common analyzable format that is Linked Data. These data are then stored into storage as graphs. To further strengthen the entity reconciliation function of the EIDER system, here we employ the Virtual Graphs technology as post-reconciliation that can apply graph algorithms to solve problems in the graph space.

It is worth noting that we have conducted a systematic literature review, and the results show that currently, there are no any methodologies, tools, or proposals that use Virtual Graphs for solving this kind of problems.

To further explain our solution, we illustrate our theory in the following figure:

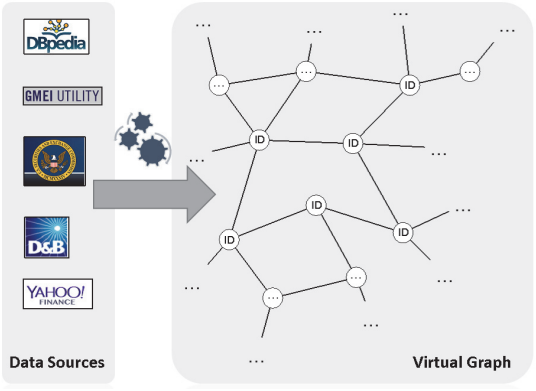

Figure 3: Virtual Graphs and EIDER Model.

In this diagram, Big and Open data from heterogeneous sources in different data formats are extracted, (for example, DBpedia (Auer et al, 2007), Yahoo Finance, GMEI Utility etc.); these data are then integrated into a graph storage, whose variety of entity identifications are managed by the Virtual Graphs. Each node of the graph represents a unique identifier from external data sources, and the edges are the relationships between the different nodes. With the dynamicity nature of the Virtual Graphs, we are able to maintain an entity reconciliation system, which is capable of adding/removing any new/old entity identities at anytime, without breaking the integrity of the whole data structure in the system.

5 CONCLUSIONS

Entity identification problem is a relatively new topic, which is emerged with the Big and Open Data phenomena. The only reference we have found so far is the white paper presented by James and Nigel (Powell and Shadbolt, 2014) in 2014. Nevertheless, the focal point of that paper, is only to present concerns to organizations and individuals, that they should be cautious whenever they inventing a new unique identifier within their system, for the entities that are already exist.

Our paper addresses the entity identification issue from a research and engineering perspective. With our solution - a system that consists of an intelligent reconciliation platform, the EIDER model, and the Virtual Graphs technology, we are able to reconcile multiple entity identification from heterogeneous Open Data sources. Furthermore, the two layers of reconciliation make sure the accuracy of the reconciled entity identifiers. This is not only based on historical data, but we are also able to incorporate any new identities at dynamically at system runtime.

From a software engineering's point of view, our system is engineered in a generic manner, so that the solution is applicable to many different domains that share the same issue, as described in section III.

6 FUTURE WORKS

This paper presents an initial investigation of the entity identification problem in the Big and Open Data era. At the time of writing, we have completed our initial system architecture design; some of the core components in the EIDER system have also been implemented, e.g. entity extraction, data integration, and some basic entity reconciliation. For the forth-

coming period, we will focus more on enhancing the intelligent reconciliation algorithms, in particular, the Virtual Graphs technology for entity identification and the integration into the EIDER system. A schedule of tasks for this research work is show in the figure 4.

The first step is to define a formal mechanism to for the Virtual Graphs in the context of the Big and Open Data. The second step is then to define a methodological environment, which allows us to be able to customize our approach to more concrete scenarios, e.g. monument identification, e-Health, and financial application.

Since the Virtual Graphs design relies completely on practical experiences for each of the scenarios, therefore, we need to work along these lines:

- The definition of the methodological process for the adaptation and application of unique identification in Big and Open Data.
- The definition of a formal procedure that will be instanced for each concrete application scenario.
- The definition of the procedure to carry out this instance in a Virtual Graphs.
- The practical evaluation of the approach in real contexts.

Figure 4: Schedule of tasks.

ACKNOWLEDGEMENTS

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

REFERENCES

Manyika, J., Chui, M., Brown, B. 2011. Big data: The next frontier for innovation, competition, and productivity.

Official websites of government: data.gov and data.gov.uk. Last check November 2014.

Berners-Lee. July 2006. Linked Data – Design Issues.

Berners-Lee. 2005. Uniform Resource Identifier (URI): Generic Syntax.

Fielding, 1999. R. Hypertext Transfer Protocol.

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

Powell, J., Shadbolt. S.N. March 2014. Creating Value with Identifiers in an Open Data World. Open Data Institute and Thomson Reuters.

Maali, F., Cyganiak, R., Peristeras. V. 2011. Entity Reconciliation Against LOD Hubs.

Insights on Data Integration Methodologies. ESSnet-ISAD workshop, Vienna, 29-30 May 2008.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. 2007. DBpedia: A Nucleus for a Web of Open Data. pp 722-735.