Using MDE for the Reconciliation of Entities in Large Data Sources

José González Enríquez, Francisco José Domínguez Mayo, María José Escalona Cuaresma
Grupo de Ingeniería Web y Testing Temprano (IWT2). Computer Languages and Systems Department.
University of Seville, Seville.
jose.gonzalez@iwt2.org, {fjdominguez, mjescalona}@us.es

Abstract

With the birth of the Big and Open Data new business opportunities have been created in which the handling of a great quantity of information and its quality, positions any company with clear advantage with respect to its competitors. However, when working with large volumes of data, we find a very important problem called entity reconciliation. This problem is based on the difficulty of identifying the same real-world entity in different data sources. In this article, a proposal based on the Model-Driven Engineering (MDE) paradigm of and virtual graphs technology is presented to solve this problem. MDE is chosen for the ease that it provides to make our solution scalable to any domain and virtual (or implicit) graphs, for the advantages of using any type of algorithm based on graphs to find a solution to a specific problem and the necessity of not storing all the information we have in the same structure because of its size. Finally, a real case study is presented in which this solution is applied, concretely, for the management of large volumes of data of the historical heritage of the Andalusia region of Spain.

1 Introduction

Actualmente, la gestión de la información es un aspecto crítico en muchas aspectos de nuestra vida cotidiana. La incorporación de las Tecnologías de la información y la Comunicación (TIC) en ella, hace que la las personas experimente un exceso de información, también conocido con el término de "infoxicación". Este término, se refiriere a la dificultad que alguien tiene a la hora de entender un problema y tomar decisiones sobre él debido a la presencia de un exceso de información (Yang et al., 2003).

En la primera era de las TIC, el principal problema que los investigadores tenían era cómo encontrar información y cómo guardarla de una forma eficiente. Actualmente, solventado ese problema, con la presencia del Big Data y el Cloud Computing, el mayor problema para los investigadores es cómo extraer el conocimiento de toda esa gran cantidad de datos existente, dependiendo de las necesidades de cada usuario (Enríquez et al., 2015).

La complejidad de la información se ha ido incrementando, no solo por la alta capacidad de producción de información, sino por estas enormes cantidades datos, que en muchas ocasiones, no representan la información de un determinado tema en su totalidad o bien, pueden contener errores. Otro de los factores críticos que la información sobre algo en concreto, puede estar en múltiples bases de datos, no solo en una, y estas bases de datos, pueden ser diferentes, pueden no tener la misma estructura, en algunas hay información repetida y no siempre es sencillo realizar una comparación entre los registros que guardan para comprobar si mencionan a un mismo elemento. De esta forma, la integración de las bases de datos, toma un papel muy importante. Esto, no significa crear una nueva base de datos unificado todas las demás existentes, significa que es necesario proporcionar una consolidación de la información proveniente de todas las bases de datos consultadas y otra cosa muy importante, es hacerlo rápido. En este contexto, el problema de la reconciliación de entidades toma un valor muy importante.

La reconciliación de entidades, (también conocida como los términos en inglés: "Entity Reconciliation", "Entity Resolution" o "ER", ver Figura 1), es un problema fundamental en la integración de datos. Este problema, se refiere a la combinación de diferentes fuentes de datos en una visión unificada, en otras palabras, identificar entidades (registros de una base de datos) del mundo real, que se refieran a la misma entidad del mundo digital. Éste, es un proceso incierto ya que la decisión de asignar varios registros con una misma entidad, no puede ser tomado con certeza a menos que esos registros sean idénticos en todos sus atributos o tengan una clave común que los identifique (Getoor and Machanavajjhala, 2012; Wang et al., 2013). Este problema, puede ser aplicado a una gran cantidad de escenarios tales como: la gestión de información bibliográfica, sanitaria o la gestionada por sistema de geolocalización entre otras.

 In: Zoe Falomir, Juan A. Ortega (eds.): Proceedings of JARCA 2016, Almería, Spain, June 2016, published at http://ceur-ws.org

id	FN	LN	phone	address
r_1	J.	Zhang	012-6006	Harbin
r_2	Jian	Zhang		2 Dazhi St
r_3	W.	Wang	236-6308	2 Dazhi St
r_4	Mei	Wang	236-6308	
r_5	Wei	Wang	124-2635	5 Dazhi St

Figura 1. Ejemplo de Reconciliación de Entidades [6]

Aunque este problema no es nuevo, la gestión de grandes volúmenes de datos presenta nuevos retos y la necesidad de realizar una reconciliación de entidades de buena calidad es cada vez mayor (Enríquez et al., 2015; Gal, 2014). En el artículo escrito por Getoor y Machanavajjhala (Getoor and Machanavajjhala, 2013), los autores exponen algunos de los retos más importantes de la reconciliación de entidades en el entorno del Big Data, éstos son:

1. Más fuentes de datos y más grandes: con la necesidad de técnicas paralelas eficientes que las soporten.
2. Heterogeneidad: las fuentes de datos, cada vez se están volviendo más heterogéneas, algunas son no estructuradas, otras, contienen datos con cierto nivel de ruido o incompletos. A todo esto, también hay que añadir los diferentes tipos de fuentes de datos estructuradas existentes, lo que provoca que no haya un consenso a la hora de guardar la información.
3. Entidades más relacionadas: con la necesidad de inferir relaciones más allá de comprobar solo la "igualdad".
4. Datos multirelacionales: tratando la estructura de entidades de forma que el contexto de cada una de las entidades, también pueda determinar si son la misma o no.
5. Sistemas multidominio: métodos personalizables que se puedan extender a través de cualquier dominio sin importar la temática tratada.
6. Sistemas multiaplicación: de forma que los datos que se generen, los puedan consumir aplicaciones con diferentes requisitos u objetivos.

Este trabajo presenta una propuesta para resolver el problema de la reconciliación de entidades basada en ingeniería guiada por modelos (ModelDriven Engineering o MDE), grafos virtuales.

El presente trabajo se estructura de la siguiente forma: la sección 2, expone una serie de trabajos relacionados a la hora de resolver este tipo de problema. La sección 3, presenta la propuesta definda para esta investigación. Seguidamente, la sección 4 explica caso de estudio real en el que la propuesta presentada se está aplicando. La sección 5 muestra un ejemplo real de instancia de la propuesta y, finalmente, la sección 6 enumuera una serie de conclusiones y trabajos futuros.

2 Trabajos Relacionados

La reconciliación de entidades es un problema bien conocdico, que ha sido investigado desde que surgieron las bases de datos relacionales. Con el nacimiento del Big Data, ha tomado una especial atención para los invesgitadores debido a los nuevos retos que fueron mencionados anteriormente.Las técnicas para resolver este tipo de problema pueden clasificarse en términos generales en: métodos basados en reglas deterministas, métodos probabilísticos, técnicas basadas en el aprendizaje y técnicas basadas en grafos.

En cuanto a métodos deterministas basados en reglas encontramos varias propuetas. Los autores del artículo (Lee et al., 2013), propusieron un enfoque para la reconciliación de la coreferencia que combina la información global y las características precisas de los modelos de aprendizaje automático con sistemas deterministas basados en reglas. En el artículo (Galhardas et al., 2001), una vez presentado el problema de la limpieza de datos, los autores presentan un lenguaje, un modelo de ejecución y algoritmos que permiten a los usuarios describir expresamente las especificaciones de limpieza de datos utilizando para su demostración un ejemplo de un conjunto de referencias bibliográficas. En el artículo (Bhattacharya and Getoor, 2005), los autores proponen un modelo probabilístico para la resolución de entidades colectivas para dominios relacionales donde las referencias están conectadas entre sí. Además, presentan un algoritmo para la reconciliación de entidades colectivas que no está supervisado y también tiene en cuenta las relaciones entre entidades, finalemnete, aplican el enfoque en dos conjuntos de datos bibliográficos reales.

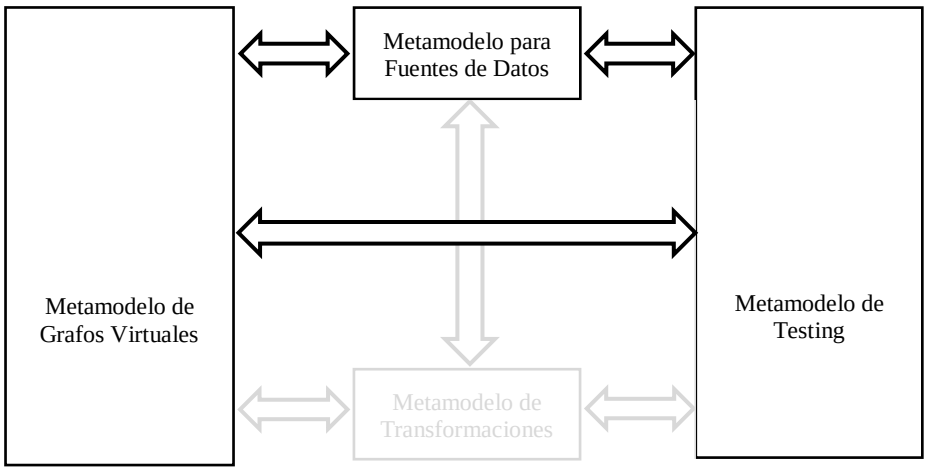

Figura 2. Arquitectura de la Propuesta

Otras propuestas relacionadas con métodos probabilísticos son: artículo (Winkler, 2002) donde el autor presentó una serie de métodos para el enlace de registros y redes bayesianas. Los autores del artículo (Verykios et al., 2003), presentaron un modelo de decisión bayesiano para la coincidencia de los registros de coste óptimo, utilizando la relación de las probabilidades anteriores de una coincidencia junto con valores apropiados de umbrales para dividir el espacio de decisión en tres áreas de decisión. Siendo éste, una versión mejorada del modelo propuesto por Fellegi y Sunter (Fellegi and Sunter, 1969).

Poniendo el foto en técnicas basadas en el aprendizaje podemos encontrar: la propuesta de Sarawagi y Bhamidipaty (Sarawagi and Bhamidipaty, 2002). En este trabajo, los augores presentaron un sistema que utiliza un método para descubrir interactivamente pares de entrenamiento desafiantes usando el aprendizaje activo. Los autores Cohen y Richman en (Cohen and Richman, 2002), describieron técnicas para agrupar y emparejar nombres de identificadores que son escalables y adaptables, en el sentido de que pueden ser entrenados para obtener un mejor rendimiento en un dominio en particular.

Finalemnte, los trabajos basados en grafo son los que han tenido más repercusión en los últimos años. Algunos de las propuestas más relevantes son: la propuesta presentada por Ioannou en el artículo (Ioannou et al., 2010). Aquí los autores describieron un marco para la vinculación de entidades con incertidumbre donde los posibles vínculos se almacenan junto con los datos con su valor de creencia. Utilizaron una técnica de respuesta de consultas probabilísticas para tener en cuenta la vinculación probabilística. En el artículo (Wang et al., 2013), los autores se centraron en la construcción de una tabla de referencia efectiva, basándose en la relación de coocurrencia entre fichas para identificar nombres de entidad adecuados. El primer conjunto de datos del modelo lo representan como un grafo, y luego agrupan los vértices del propio grafo. Los autores del artículo (Wang et al., 2016), modelaron el problema de reconciliación de entidades como la partición de los vértices de un grafo ponderado en subgrafos cohesivos. Proponen un algoritmo aproximado con la relación de aproximación y un algoritmo heurístico para realizar la reconciliación de entidades en un gran conjunto de datos de manera eficiente.

3 Propuesta

Nuestra propuesta, como ya se ha mencionado anteriormente, pretende asegurar la calidad de la reconciliación de entidades mediante la incorporación de testing temprano a dicho proceso, y para ello se basa en MDE y en el concepto de grafos virtuales.

El uso de lenguajes de modelado ayuda a especificar cierto nivel de abstracción a la hora de definer un problema. Además, el uso de los modelos definidos se puede utilizar para apoyar el desarrollo de aplicaciones de software (Rodrigues Da Silva, 2015). Considerando que el principio de MDE es: “Todo es un modelo” (Bézivin, 2005), utiliza un conjunto de modelos para disminuir el nivel de abstracción. Por lo tanto, los modelos son más abstractos en las primeras etapas que en las etapas finales del proceso de desarrollo de software. Una de las ventajas de MDE es su apoyo a la automatización, ya que los modelos pueden transformarse automáticamente desde las primeras etapas de desarrollo hasta las etapas finales (Enríquez et al., 2016).

La tecnología de grafos, es una de las soluciones naturales para tratar problemas relacionados con Big Data y especialmente, para las relaciones que existen entre las entidades. La variedad de algoritmos, por ejemplo: Dijkstra, A*, Kruskal, etc. ofrecen una gran flexibilidad en diferentes situaciones. Teóricamente, los grafos se pueden representar de dos formas: explícita e implícita. Un grafo explícito es una colección de elementos que pueden ser almacenados en memoria, lo que quiere decir que cada vértice y cada arista del grafo puede ser completamente almacenado en memoria. Por otra parte, un grafo implícito es un grafo que no puede

ser almacenado en memoria completamente por diversas razones como pueden ser: su tamaño o limitaciones de hardware (Mondal and Deshpande, 2012).

En nuestra propuesta, el grafo virtual se utiliza para representar la estructura de la solución diseñada para reconciliar las entidades procedentes de diversas fuentes de información y para almacenar los datos de dicha solución una vez realizada la reconciliación. Con esta tecnología, se tiene la posibilidad de construir los grafos en tiempo de ejecución, lo que permite construir diferentes soluciones para hacer frente a muchos escenarios dentro de una lógica de negocio, donde el modelo de datos predefinido no puede hacer frente a la extensibilidad o la disponibilidad que se requiere de las fuentes de datos. De esta forma, se ayuda a solucionar problemas que se presentan en Big Data, como puede ser la rigidez de las estructuras de las bases de datos. Por otra parte, gracias a MDE, se puede obtener una solución fácilmente escalable.

En el trabajo previo (Enríquez et al., 2015), se propuso un enfoque para abordar la problemática de la reconciliación de entidades utilizando grafos virtuales, dentro del contexto de MDE. La propuesta presentada en este artículo extiende el trabajo previo y añade un nuevo pilar fundamental en la reconciliación de entidades: el testing.

En la Figura 2 se muestra el metamodelo propuesto, el cual permite al usario crear modelos que definan tanto los objetivos de la reconciliación de entidades a realizar, como los objetivos de testing que indiquen los aspectos importantes a probar del comportamiento de dicha reconciliación. Los cuatro pilares fundamentales de este metamodelos son los grafos virtuales, las fuentes de datos, las transformaciones y el testing.

- Metamodelo de grafos virtuales: permite al usuario crear la estructura que representará la solución obtenida de la reconciliación de entidades, mediante un grafo virtual. Este metamodelo es una versión extendida del metamodelo de un grafo.
- Metamodelo para fuentes de datos: permite representar las fuentes de informaciónque contienen los datos que se quieren reconciliar y cómo se puede acceder a las mismas, teniendo en cuenta dichas fuentes pueden ser una base de datos estructurada o no estructurada, un servicio web o cualquier otro tipo de almacén o generador de información.
- Metamodelo de transformaciones: permite representar las diferentes transformaciones que los datos deben sufrir para estar en consonancia con el modelo de datos que representa la reconciliación de entidades definida por el usuario.
- Metamodelo de Testing: permite representar de forma temprana los objetivos de testing correspondientes a la reconciliación de entidades a realizar, de modo que se pueda determinar cuanto antes si dicha reconciliación es la que realmente se desea llevar a cabo y si los resultados que se están obteniendo son también los realmente deseados.

4 Caso de Estudio

En este caso de estudio se está llevado a cabo en la gestión de la información en el ámbito de las bases de datos de patrimonio histórico de la comunidad de Andalucía, concretamente, poniendo el foco en la gestión de los monumentos inmuebles.

El sistema que el Instituto Andaluz de Patrimonio Histórico (IAPH) utiliza para la gestión de este tipo de información se llama “MOSAICO” (“MOSAICO: Sistema de Información para la Gestión del Patrimonio Cultural en Andalucía,” 2016). Éste, es un sistema horizontal y global cuyos objetivos son:

1. Ofrecer los recursos tecnológicos y herramientas para la gestión del patrimonio histórico.
2. Ofrecer un sistema de información global que almacene información sobre todo el patrimonio cultural.
3. Acercar al público en general y al gobierno la información relacionada con patrimonio

Este sistema fue desarrollado por el IAPH para cumplir con sus propios objetivos tales como:

1. La gestión de la información sobre el patrimonio cultural.
2. La protección de la información patrimonial cultural de Andalucía.
3. La preservación del patrimonio cultural de Andalucía, la difusión de los valores de los bienes culturales.
4. El acercamiento del gobierno al ciudadano.

Existen una gran cantidad de fuentes de datos que almacenan información relacionada con los monumentos de las cuidades por lo que para el IAPH, mantener en control toda la información publicada sobre el patrimonio en el mundo suponen una tarea muy difícil.

Además, el tamaño y la complejidad de estas fuentes de datos complican la gestión de estos sistemas debido a la gran cantidad de información almacenada en ellos (por ejemplo, sólo MOISAICO, almacena Terabytes de información). Por tanto, es necesario reconciliar la información existente sobre los monumentos de todas las fuentes de datos.

Considerando este problema, se está desarrollando en colaboración con los Laboratiorios Fujitsu de Europa (FLE), la aplicación “DIPHDA” (“Dynamic Integration for Patrimonial Heritage Data in Andalucía”, Integración Dinámica para el Patrimonio Histórico de Andalucía)

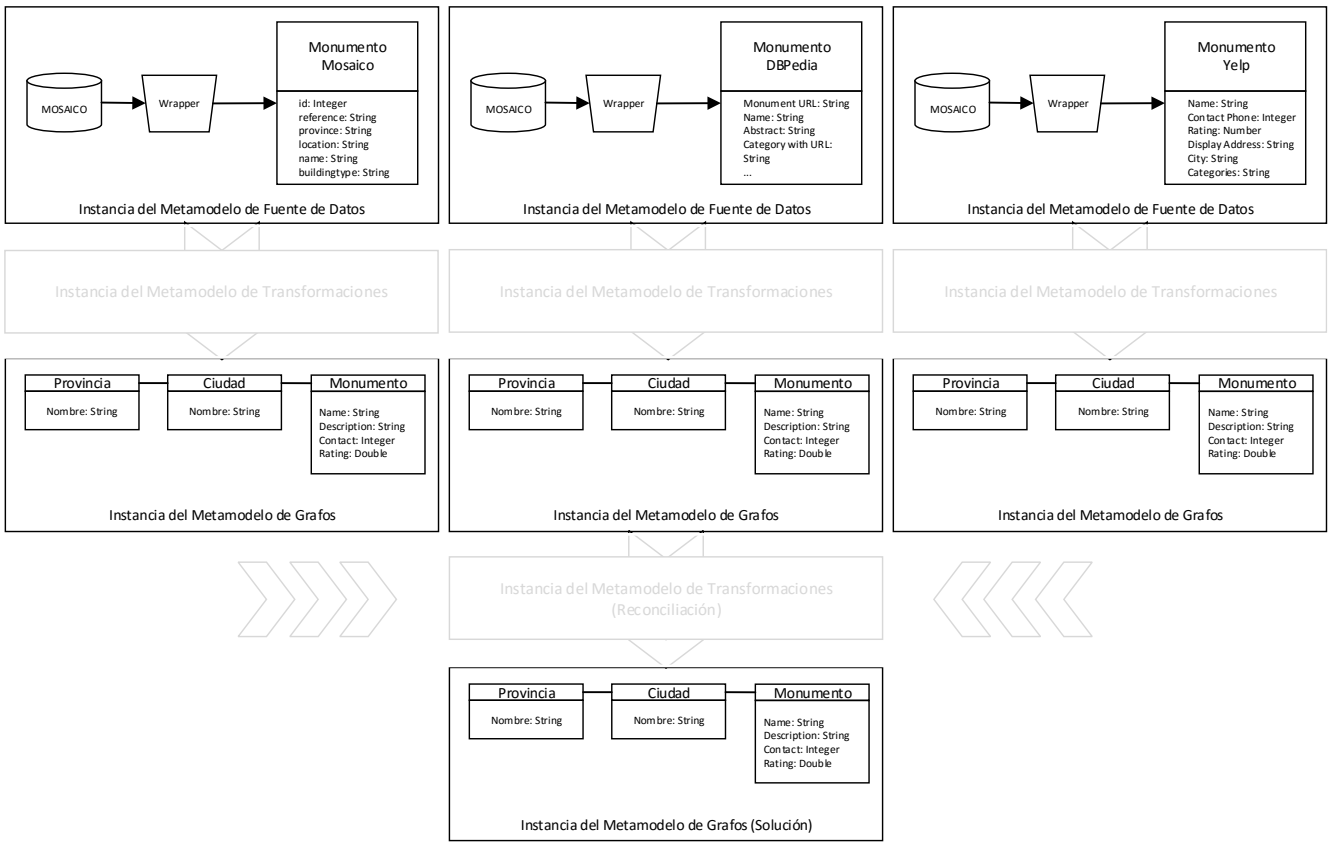

Figura 3. Arquitectura de la Propuesta

El objetivo de DHIPDA es lograr una mejora significativa de la precisión y la eficiencia de la gestión de datos, basada en la reconciliación de entidades lógica, aplicada a la información abierta de datos, en oposición a la simple reconciliación de concordancia de cadenas que se está utilizando en estos momentos.. Esta solución será capaz de integrar diferentes sistemas de gestión. Para este caso en particular se utilizaron los sistemas "MOSAICO", Wikipedia y Yelp.

La información que maneja DIPHDA se recupera del proceso de reconciliación realizado en una de sus funcionalidades donde el usuario tiene que definir la estructura de datos donde se almacenarán los resultados del proceso de reconciliación. Esta funcionalidad se basa en un Lenguaje Específico de Dominio (DSL o Domain Specific Language).

Van Deursen y el resto de autores definieron en el artículo (Deursen et al., 2000), un DSL, como un lenguaje de programación o un lenguaje de especificación ejecutable que ofrece, a través de anotaciones y abstracciones apropiadas, un poder expresivo centrado en, y generalmente restringido, a un dominio de problema particular. En este contexto, DIPHDA proporciona un DSL para diseñar el problema concreto a tratar. Para este ejemplo, el usuario diseñó la estructura de datos con todos los atributos y operaciones necesarios para llevar a cabo el proceso de reconciliaciíon de entidades.

5 Ejemplo de Instanciación del Metamodelo

En esta sección, se muestra un ejemplo real de instanciación para un dominio específico. El ejemplo está basado en la reconciliación de entidades para la gestión del patrimonio histórico, concretamente, para la sección de monumentos.

Para el modelado de la solución, lo primero que el usuario debe tener en cuenta es el dominio del problema, es decir, conocer el entorno en el que se está trabajando y qué solución es la que se quiere obtener. Una vez conocido el contexto, el usuario tendrá que instanciar el metamodelo de forma que la solución que se modela de respuesta al problema que se plantea.

En este caso, se cuenta con tres fuentes de datos que almacenan información de patrimonio histórico correspondiente a monumentos localizados en las diferentes ciudades y pueblos de Andalucía: MOSAICO, DBPedia y Yelp. MOSAICO es la fuente de datos oficial a través de la que el Instituto Andaluz de Patrimonio Histórico (IAPH) gestiona toda la información relativa a monumentos en Andalucía, mientras que DBPedia y Yelp son fuentes de datos que se consultan a través de Internet.

El problema que se plantea consiste en reconciliar toda esta información, teniendo en cuenta que, como se ha comentado en la introducción, la que se puede encontrar por Internet no siempre es de buena calidad al no estar contrastada por las

personas encargadas de protegerla, mantenerla y difundirla (en este caso concreto, el IAPH). Por otra parte, también se puede observar en la figura 3 que no todas las fuentes de datos contienen la misma información o la misma estructura. Por todo ello, surge claramente la necesidad de diseñar una solución para llevar a cabo la reconciliación de entidades de estas fuentes.

Para cada una de estas fuentes de datos, será necesaria una instanciación del modelo de fuente de datos que obtenga la información de cada una de ellas y las convierta en las entidades. Seguidamente, e instanciando el modelo de tranformaciones y el de grafos virtuales, se pasará la información desde las diferentes fuentes de datos a una estructura de grafo que el usuario habrá definido para representar la solución de la reconciliación. Como se puede observar, el grafo está compuesto por nodos Provincia, nodos Ciudad y nodos Monumento. Por tanto, una vez realizado este paso, se obtendrá un grafo para cada una de las fuentes de datos con la información estructurada tras haber hecho las transformaciones correspondientes.

A continuación, se llevará a cabo la reconciliación, aplicando una serie de transformaciones sobre los grafos virtuales correspondientes a las fuentes de datos. El resultado de estas transformaciones quedará reflejado en una nueva instancia del metamodelo de grafo virtual que representa la solución final de la reconciliación. Por último, la instanciación del metamodelo de testing permitirá probar las transformaciones que se realizan tanto en el primer nivel (de fuentes de datos a estructura de grafo virtual) como en el segundo nivel (reconciliación entre diferentes grafos virtuales), así como realizar pruebas sobre la integración de todo el proceso de reconciliación (desde las fuentes de datos a la solución final).

6 Conclusiones y Trabajos Futuros

En este trabajo se presenta una propuesta para la problemática de la reconciliación de entidades en el ámbito del Big Data que integra técnicas de testing para mejorar la calidad de dicho proceso. Con el objeto de facilitar la automatización de todo el proceso, la propuesta se basa en la ingeniería guiada por modelos y en los grafos virtuales.

Como trabajo futuro se propone ampliar la definición del metamodelo propuesto, centrándose especialmente en las partes de transformaciones y testing. Asimismo, se pretende identificar casos de uso en los que se pueda aplicar el enfoque propuesto para, posteriormente, llevarlo a cabo con el fin de verificar su validez. Actualmente se está trabajando en un caso de uso real en colaboración con el IAPH y Fujistu Laboratories Europe (FLE). Este proyecto denominado "DIPHDA" se desarrolla para resolver las dificultades que tiene el mantenimiento del gran número de datos del patrimonio histórico y monumental de una forma exhaustiva, precisa y rentable. El objetivo del proyecto es conseguir una mejora significativa en la precisión y eficiencia en la gestión de datos, basada en la reconciliación lógica aplicada a la información de Open Data, a diferencia de la simple búsqueda de correspondencia.

Acknowledgments

Este trabajo ha sido cofinanciado por el proyecto MeGUS (TIN201346928C33R), el proyecto Pololas (TIN201676956C32R), la red de investigación SoftPLM (TIN201571938REDT) del ministerio de Economía y Competitividad de España y por los Fujitsu Laboratirios de Europa (FLE).

References

Bézivin, J., 2005. On the unification power of models. Software and Systems Modeling 4, 171–188. doi:10.1007/s10270-005-0079-0

Bhattacharya, I., Getoor, L., 2005. A latent dirichlet allocation model for entity resolution. Proceedings of the 2005 SIAM International Conference on Data Mining 47–58.

Cohen, W.W., Richman, J., 2002. Learning to match and cluster large high-dimensional data sets for data integration. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 475–480. doi:10.1145/775107.775116

Deursen, A. Van, Klint, P., Visser, J., 2000. Domain-specific languages: an annotated bibliography. ACM Sigplan Notices 35, 26–36. doi:10.1145/352029.352035

Enríquez, J.G., Blanco, R., Domínguez-Mayo, F.J., Tuya, J., Escalona, M.J., 2016. Towards an MDE-based Approach to Test Entity Reconciliation Applications, in: ACM (Ed.), Proceedings of the 7th International Workshop on Automating Test Case Design, Selection, and Evaluation. ACM, New York, NY, USA, pp. 74–77. doi:10.1145/2994291.2994303

Enríquez, J.G., Domínguez-Mayo, F.J., Escalona, M.J., García-García, J.A., Lee, V., Masatomo, G., 2015. Entity Identity Reconciliation based Big Data Federation-A MDE approach, in: International Conference on Information Systems Development (ISD2015).

Fellegi, I.P., Sunter, A.B., 1969. A Theory for Record Linkage. Source Journal of the American Statistical Association 64, 1183–1210. doi:10.1080/01621459.1969.10501049

Gal, A., 2014. Uncertain Entity Resolution: Re-evaluating Entity Resolution in the Big Data Era: Tutorial. Proc. VLDB Endow. 7, 1711–1712. doi:10.14778/2733004.2733068

Galhardas, H., Florescu, D., Shasha, D., Simon, E., Saita, C.-A., 2001. Declarative Data Cleaning: Language, Model, and Algorithms. Proceedings of 27th International Conference on Very Large Data Bases

371–380.

Getoor, L., Machanavajjhala, A., 2013. Entity resolution for big data. KDD ’13: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 4503. doi:10.1145/2487575.2506179

Getoor, L., Machanavajjhala, A., 2012. Entity resolution: Theory, practice & open challenges. Proceedings of the VLDB Endowment 5, 2018–2019. doi:10.14778/2367502.2367564

Ioannou, E., Nejdl, W., Niedere’e, C., Velegrakis, Y., Nieder, C., 2010. On-the-Fly Entity-Aware Query Processing in the Presence of Linkage. Proceedings of the VLDB Endowment 3, 429–438. doi:10.14778/1920841.1920898

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D., 2013. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. Computational Linguistics 39, 885–916. doi:10.1162/COLI

Mondal, J., Deshpande, A., 2012. Managing large dynamic graphs efficiently. Proceedings of the 2012 international conference on Management of Data - SIGMOD ’12 145. doi:10.1145/2213836.2213854

MOSAICO: Sistema de Información para la Gestión del Patrimonio Cultural en Andalucía, 2016.

Rodrigues Da Silva, A., 2015. Model-driven engineering: A survey supported by the unified conceptual model. Computer Languages, Systems and Structures 43, 139–155. doi:10.1016/j.cl.2015.06.001

Sarawagi, S., Bhamidipaty, A., 2002. Interactive Deduplication using Active Learning. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 269–278. doi:10.1145/775047.775087

Verykios, V.S., Moustakides, G. V., Elfeky, M.G., 2003. A Bayesian decision model for cost optimal record matching. VLDB Journal 12, 28–40. doi:10.1007/s00778-002-0072-y

Wang, F., Wang, H., Li, J., Gao, H., 2013. Graph-based reference table construction to facilitate entity matching. Journal of Systems and Software 86, 1679–1688. doi:10.1016/j.jss.2013.02.026

Wang, H., Li, J., Gao, H., 2016. Efficient Entity Resolution Based on Subgraph Cohesion. Knowl. Inf. Syst. 46, 285–314. doi:10.1007/s10115-015-0818-7

Winkler, W.E., 2002. Methods for Record Linkage and Bayesian Networks. Research Report 29.

Yang, C.C., Chen, H., Hong, K., 2003. Visualization of large category map for Internet browsing. Decision Support Systems 35, 89–102. doi:10.1016/S0167-9236(02)00101-X