Populating the FLE Financial Knowledge Graph

Manuel Peña-Muñoz[1] and Alejandro Llaves[1] and Terunobu Kume[1]

[1] Fujitsu Laboratories of Europe, Pozuelo de Alarcon, 28224 Madrid, Spain
manuel.p.munoz|alejandro.llaves|terunobu.kume@uk.fujitsu.com

1. Integration of Financial Data for Knowledge Processing

In Fujitsu Laboratories of Europe (FLE), we are developing a platform to get insights in the financial sector from the analysis of multiple heterogeneous data sources, such as stock quotes, corporate financial reports, news, and social network data, among others. At the core of the platform, there is a knowledge graph that is populated with new nodes and relationships as new data are ingested.

Most public data sources in the financial domain are provided by country regulators, e.g. by enforcing companies to publish their annual reports online. However, we still find very important financial datasets, such as corporate announcements published daily as non-structured texts in PDF, which require an enormous adaptation effort and ad-hoc ETL tools to be automatically ingested and later analysed. In the case of private sources, we realized that financial datasets tend to be siloed and fragmented. Additionally, entity identifiers and formats for the same type of data sometimes differ between organizations. Another common problem is that the utilised data properties are not always properly documented, and they are rarely self-explanatory.

The benefit of using a knowledge graph in our platform is threefold. First, it reduces the time spent on simple and repetitive data integration tasks. Second, the acquired knowledge is used by a virtual assistant to facilitate the data reconciliation process, thus users do not need to be experts to merge a new dataset. And third, it adds extra value to the customer information by linking knowledge to open and private data sources.

2. FLE Knowledge Graph Population

We have developed a method called **Dynamic Data Loading** (DDL – related patents [1,2], other patents awaiting publication**)** that makes use of semantic technologies and machine learning techniques to add new data to the knowledge graph during runtime. Entities in the graph can be of various types, such as companies, people, locations or events. We store ontologies and vocabularies in Apache Jena.[1] Among others, Jena contains instances of data properties annotated with language variations, its usage as entity identifiers, and the domain of the data property in some cases. When new data are ingested, these annotations together with NLP tools are used to recognize the data properties and the potential identifiers. Entity types and the dataset domain are in-

[1] https://jena.apache.org/

ferred using probabilistic methods with respect to the percentage of data properties successfully recognized.

The DDL workflow comprises four steps: (1) Data Property Reconciliation and Potential Identifiers Finder, (2) Entity Type and Domain Recognition, (3) Entity Disambiguation, and (4) Knowledge Base Storage. During workflow execution, when the system reaches a point where there is not enough evidence to perform an action, it offers the user various options to continue. The system learns from user decisions and applies the acquired knowledge in future executions of the workflow.

Our technology has been tested on the reconciliation of corporate data for a financial regulator. The datasets include companies from Japan, US and Spain from different sources. In total, 286,440 entities were added to the knowledge graph – from which 19,185 were merged automatically – and 473,375 relationships were created. In average, our approach speeds up the full process in the order of 10 times and also reduces disambiguation tasks compared with state of the art technologies. This is possible thanks to advanced features, such as context inference and disambiguation suggestions. Therefore, the use of our platform for the ingestion of heterogeneous datasets in the financial domain provides a significant competitive advantage.

3. Lessons Learned

A topic of discussion at the beginning of the project was the use of one or more databases to store the knowledge graph. By depending only on a graph database, we were missing semantic relationships and meaningful entity types. Document databases, even using a semantically enriched format as JSON-LD, lack the built-in link analysis capabilities, e.g. shortest path. The triplestores we considered were not fast enough to scale up to huge amounts of data. At the end, we decided to incorporate a smart storage system that uses a combination of these three types of databases (graph, document and triplestore) with an API layer on top.

The design and development of the platform capability to learn from user actions was challenging as well. Our goal was to improve the global reconciliation knowledge with individual usage. For instance, when a user suggests that certain data property can be used as an identifier, this is learned by the platform and added to the global property metadata. As a consequence, the platform can recommend in forthcoming reconciliation processes a new potential identifier.

We believe that other areas for future applications may involve any industry handling massive data volumes, such as healthcare, retail, transport and manufacturing.

References

1. Peña-Muñoz, M., Llaves, A. and de la Torre, V., Fujitsu Ltd, 2018. *Apparatus program & method for data property recognition.* U.S. Patent Application 15/679,406.
2. Llaves, A., Peña-Muñoz, M. and de la Torre V., Fujitsu Ltd, 2018. *Apparatus program & method for data property recognition.* U.S. Patent Application 15/679,296.