Exploring the Application of Ontologies in Organizations for Data Harmonization

Carlos Tubbax

Faculty of Business and Economics
University of Antwerp
Antwerp, Belgium
Email: carlos.tubbax@uantwerpen.be

Jan Verelst

Faculty of Business and Economics
University of Antwerp
Antwerp, Belgium
Email: jan.verelst@uantwerpen.be

***Abstract*—In this contribution, it is explained how ontologies could be used by business organizations to integrate data from heterogenous sources in a systematic process called *data harmonization*. In order to add academic rigor, Normalized Systems Theory (NST) has been used as a rationale to study the modularity concerns of this data harmonization project. Data harmonization consists in this contribution of three steps and offers certain advantages compared to other data integration approaches such as data-warehousing. The first are its simplicity and flexibility. The second is that the mapping costs of such an approach will just increase linearly and no longer exponentially. Additionally, the author illustrates how data harmonization can be conducted through a use case in which several datasets on US stock exchanges are mapped to the Financial Industry Business Ontology (FIBO) before being integrated for information retrieval through predefined SPARQL queries. The main contribution of this work is that it shows step-by-step how data harmonization can be conducted with costs that no longer increase exponentially but linearly as the number of data sources and destinations increases by means of a numeraire topology.**

Keywords–Ontology; Data Harmonization; Financial Industry Business Ontology; Numeraire Typology; Graph

I. INTRODUCTION

Data management issues might have significantly contributed to the unfolding of the 2008-2009 financial crisis [1]. As a response to this, several news regulatory efforts have appeared to tackle this issues, such as *BCBS239* [2]. However, financial institutions and banks still have problems when integrating data from different sources due to several causes. The first set of causes consists namely of the heterogeneity of data sources in terms of granularity, formats, technologies and schemas [3]. Another cause is the large amounts of data that emerge every day (i.e., Big Data) that are still troublesome for most financial organizations. The last cause is the rigidity of traditional relational and data-warehousing systems making them not scalable and flexible enough to cope with Big Data. That being said, all these problems combined hamper financial organizations to comply with new regulatory efforts, to act upon new challenges and to reap new business models [4] [5].

This work aims at illustrating how that can be done in a rather systematic and simple way in the context of business organizations. In section 2, a literature review will cover the different theories and technologies used in this work. Section 3 dives deeper into the data harmonization methodology used in this work. In section 4, a use case has been chosen to illustrate the implementation of this data harmonization methodology. Finally, section 5 describes the execution of such a use case and its results.

II. LITERATURE REVIEW

This section will give an introduction of the different technologies and theories used in this work.

A. The Semantic Web

The current World Wide Web lacks the ability to represent meaning in a way that not only humans can understand but also computers. As means to tackle this problem, the Semantic Web is equipped with languages that express inference rules that allow computers to do reasoning on data [6]. Additionally, the Semantic Web is aimed at enabling *smart* behavior across the web consisting of different applications where data in each application are kept up-to-date, synchronized and connected to changes in other applications. This is done by assigning a Unique Resource Identifier (URI) to each individual piece of data about a *resource*, such as a person, an object or a date in order to refer to them at the level of data rather than at the level of representation in the form of excel-sheets or websites as in the case of the current World Wide Web. That being said, the Semantic Web might be a web of data instead of a web of only applications [7].

B. Data storage paradigms

Relational databases may neglect the semantic of the relationships. Additionally, as the number of rows within a table increases, the number of joins and query time may increase, such as in the case of transitive queries (e.g., 'Who-are-the-friends-of-all-my-friends?'). Another problem is the lack of rigidity of relational databases making them particularly difficult to adapt to new business requirements or to scale them up. They may not be fit to integrate data from different sources due to their rigid nature either. Finally, changes in their schemas (i.e., deleting a foreign key) may have pervasive ripple effects across the entire database [5] [8] [9].

Concerning the second paradigm, the construction of a data-warehouse is a rather complex and arduous process in which several trade-offs and decisions have to be made in advance. Some of them are the up-front selection of a certain architecture, defining the right level of data granularity, the design of Extract-Transform-Load (ETL) capabilities and the design of an access layer with OLAP capabilities. Additionally, data-warehouses do not update data to changes in sources or other systems. Therefore, they are not suited as data repositories in the context of the *smart web* [7].

The last main paradigm is graph databases. Graph databases have a number of advantages compared to the previous paradigms. The first is stable performance by just

 ISBN: 978-1-61208-783-2

performing queries over a portion of the total graph. The second is that no formal model is needed upfront making it more flexible to changes in business requirements, etc. The third is no longer having the need of a schema before ingesting data, such as in the case of ETL in traditional data-warehousing. The last advantage is the ability to increase the database's capacity by adding new servers (i.e., scale out) whereas relational databases scale up by adding more memory to a monolithic server. That means that the database is divided in several servers and only the servers containing the data needed to answer a certain query are accessed rather than the whole monolithic server. There are two types of graph databases, namely Property Graphs and RDF stores [5] [9]. This work only studies the latter.

C. Semantic Modelling

As previously mentioned, the Semantic Web is equipped with languages that allow users to define models of the domain of discourse in terms of taxonomies and inference rules. These models are called *ontologies* [6]. The more detailed a model is, the more expressive it is considered to be. The Semantic Web offers different modeling languages that offer different expressivity levels and are listed below [7].

- The Resource Description Framework (RDF)
- The RDF Schema Languages (RDFS)
- RDFS-plus
- The Ontology Web Language (OWL)

D. SPARQL

The 'S' Protocol and RDF query language or SPARQL is the query language of the Semantic Web. Every SPARQL query follows the pattern of the graph that is being queried. Although sharing many characteristics with SQL, such as the SELECT and WHERE commands, it has the unique feature of retrieving a graph as query output by using the CONSTRUCT command [10].

E. Financial Industry Business Ontology

The Financial Industry Business Ontology (FIBO) is a modular ontology aimed at representing the business logical of financial organizations in a standardized and unambiguous way that is readable by computers and humans. FIBO is jointly developed by the Enterprise Data Management (EDM) council and the Object Management Group (OMG) [11].

F. Normalized Systems theory

NST offers a set principles to build modular structures in software, organizations, etc. Such principles are based on systems stability and thermodynamics theory to reduce the number of ripple effects and increase the traceability of problems within a system. Although following such principles does not imply that all ripple effects will be eliminated, not following them will certainly lead to more ripple effects than otherwise. NST also depart from the notion of *Bounded Input Bounded Output* which implies that a bounded number of changes in a systems should always lead to a bounded number of impacts (i.e., ripple effects). In order to achieve this, NST offers four principles which are listed as follows [12].

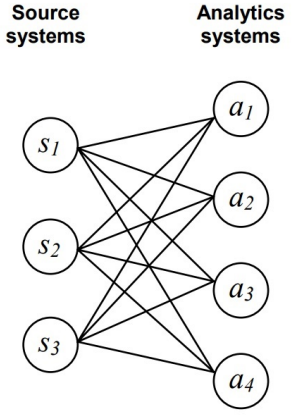
- Separation of Concerns
- Data version transparency
- Action version transparency
- Separation of states

G. Data Harmonization

In spite of the abundance of data harmonization works, no formal definition is provided. Therefore, a definition for data harmonization will be provided. Since *semantic harmonization* should be separated from *technical harmonization*, both will be considered as separate dimensions [13]. In top of that, *data quality harmonization* will be considered the third dimension as follows:

- **Technical harmonization:** it entails converting data contained in heterogenous datasets and databases to be merged into a singular format that can be stored and queried by the same technical implementation (e.g., transforming data contained as XML-files, .csv-files and in other formats into triples that can be stored in the same RDF store and queried by the same SPARQL engine). These are rather cross-cutting concerns that should be separated from other aspects of the data harmonization process as suggested by NST [13].
- **Semantic Harmonization:** this entails mapping the different concepts and data fields in the heterogeneous sources to a representation of the domain of discourse needed and agreed on by domain experts and business users [13].
- **Data quality harmonization:** heterogeneity of data sources and datasets also brings heterogeneity in terms of data quality which needs to be handled properly to create a singular view in the form of a federated database consisting of high quality data. Therefore, data quality must be brought to a level that complies with business requirements [13].

In addition to these dimensions of data harmonization, a data harmonization architecture will be needed to integrate different systems [14]. Two possible architectures are described in further detail below (see Figure 1).

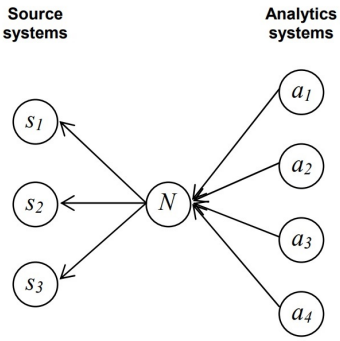


$$Cost = (m + n)k_{spec} + (mn)k_{map}$$

Figure 1. Stovepipe topology

The first architecture is the stovepipe topology as illustrated in Figure 1. Costs are considered to consists of specification

costs (k_{spec}) and mapping costs (k_{map}). Specification costs are related to the specification of the schema of each system and, in this case, they are assumed to be constant for all systems. Mapping costs (k_{map}) are the costs related to the mapping of each different source system to all the different target systems and, in this case, they are also assumed to be constant for each pair of systems. Assuming there are m source systems and n target systems. Given that the total mapping cost $(mn)k_{map}$ is proportional to the size of the graph, the total implementation cost of such a topology would increase exponentially as depicted by the equation at the bottom of Figure 1 [14]. This also implies that a bounded input (e.g., adding a new source system) may lead to an unbounded output which is highly discouraged by NST [12].



$$Cost = k_N + (m + n)k_{spec} + (m + n)k_{map}$$

Figure 2. Numeraire topology

The second architecture is the numeraire topology as illustrated in Figure 2. Costs are also considered to consist of specification costs (k_{spec}) and mapping costs (k_{map}). However, by introducing an intermediate metadata layer that decouples the target systems from the source systems, the resulting total mapping cost is no longer dependent on the size of the overall graph and only on the total number of source and target systems which might be equal to $(m+n)$. Additionally, the cost of specifying such an intermediate metadata layer is (k_N). Moreover, such an architecture uses a 'pull' strategy that departs from the inputs needed for analyses in the target systems to define the outputs that the different source systems must deliver. The total implementation cost of such a topology would increase just linearly over time as depicted by the equation at the bottom of Figure 2 [14]. Another advantage is that this architecture will remain more stable over time as a bounded input will lead to a bounded output as suggested by NST. This could only be achieved if the interfaces of the different systems in such an architecture are well insulated to comply with the *data version transparency* and *action version transparency* principles of NST. In other words, interfaces should encapsulate changes in the data and program structures within each system to avoid pervasive ripple effects on other systems [12]. Therefore, this architecture will be used to this data harmonization project [14].

III. DATA HARMONIZATION METHODOLOGY

In order to conduct any data harmonization endeavor, a sound methodology is needed to guide users. Therefore, a data harmonization methodology has been developed for this project consisting of three steps as follows. Step 1 will comprise the definition of high-level requirements in terms of business questions that need to be (graphically) answered for business users and decision makers. Finally, in Step 3, more detailed requirements will be specified for each of the three dimensions of data harmonization.

A. Defining high-level requirements (step 1)

Step 1 will comprise the definition of high-level requirements in terms of business questions that need to be (graphically) answered for business users and decision makers

B. Defining a data harmonization architecture (step 2)

Departing from the outputs of Step 1, a data harmonization architecture and its components will be designed in Step 2 following the *numeraire topology* in a '*pull fashion*'. In alignment to NST, the interfaces of the different systems within such a topology must be *data version transparent* and *action version transparent* to isolate changes in each system from other components within the overall data harmonization architecture [12].

C. Defining a low-level requirements (step 3)

Finally, in Step 3, more detailed requirements will be specified for each of the three dimensions of data harmonization as follows.

The technical harmonization requirements will cover the technical concerns of converting the data from the source systems to a format that can be stored and queried in the same storage implementation. Some of these requirements will be what serialization format will be used or what type of inferencing will be performed (*cached* or *just-in-time*). The semantic harmonization requirements will be needed to map the data fields in the different source system to their respective representations of the domain of discourse in the form of an ontology. As a matter of academic rigor and to make this methodology more generalizable, semantic harmonization principles found in the academic literature have been aligned to Normalized Systems Theory (NST) [12] [14].

Data quality harmonization is the last dimension whose requirements will be needed to bring the quality of the data in all the different source systems to a level suited to for answering the predefined business questions from Step 1. This will be done by using certain data quality metrics.

After the definition of these requirements, the data in each source will be harmonized independently from the other ones to isolate the concerns inherent to each systems and delivering loosely coupled outputs as suggested by the *Separation of Concerns* and *Separation of States* theorems from NST [12]. Accordingly, the data in each source will be converted into a RDF graph by using the CONSTRUCT command from SPARQL [10]. However, the identification and specification of dependencies between such RDF graphs are crucial since they represent the connection points between them. Therefore, an iterative approach will be followed to identify these connection points and to specify them in a way that facilitates the integration of such individual graphs into federated ones.

 ISBN: 978-1-61208-783-2

IV. DATA HARMONIZATION PLANNING

A business case, provided by D. Allemang, and A. Keen, will be used to illustrate the data harmonization methodology mentioned above. It is fictitious and has been formulated to show a realistic business scenario. Such a business scenario in this consists of unraveling the rather complex and nested ownership and control relationships between companies listed in different US stock markets and other companies. A listed company is defined as: '*a company whose shares can be traded on a country's main stock market*' [15]. AMEX, NASDAQ and NYSE are the three main stock markets in the United States that list the stocks of different companies, such as Facebook, Amazon and Apple. That being said, it would be of great value for brokers, banks, hedge funds and investors in general to have a sight of the companies that either own any, or are owned by any of these listed companies. Therefore, the planning of this data harmonization project will follow the methodology described above as follows.

A. Defining high-level requirements (step 1)

Based on the description above, the following high-level business questions have been formulated:

- What companies are listed by AMEX, NASDAQ and NYSE?
- What are the parent companies and subsidiaries of these listed companies?
- Where are all these companies located?

Additionally, the results obtained from the data harmonization process meant to answer these questions will be used to generate user-friendly visualizations for business people by means of Business Intelligence tools.

B. Defining a data harmonization architecture (step 2)

The source systems containing the data needed to answer these questions are listed as follows:

- Datasets that contain data about the companies listed on AMEX, NASDAQ and NYSE have been retrieved from NASDAQ's website [16].
- The Global Legal Entity Foundation (GLEIF) is an organization aimed at providing unique identifiers to legal entities. Additionally, they provide datasets about ownership and control relationships between these legal entities. Therefore, the dataset containing data on ownership relationships between listed companies and other companies have been imported from GLEIF's website [17].
- Additionally, a dataset containing further information (e.g., postal codes and names) of the legal entities registered by GLEIF has been imported as well [18].
- The last source system consists of datasets on postal codes and their coordinates retrieved from the GeoNames organizations [19].

As metadata layer of such an architecture, data.world is an open data platform and has been selected because of its user-friendly interface and API capabilities. More specifically, data.world will be the platform in which the different source systems will be integrated and from which outputs for the target systems will be exported through its APIs. Finally, Tableau is a Business Intelligence interface that allows users to import data and visualize them in a wide variety of forms for further analysis. Therefore, this is the target system chosen for this project.

C. Defining a low-level requirements (step 3)

The source systems containing the data needed to answer these questions are listed as follows:

- **Technical harmonization.** Because of the limited scope of this work, the technical harmonization requirements to each source will be considered to be rather simple. All datasets used in this project will be exported as .csv-files to data.world's platform in a straightforward way. However, this would not be the case if the data would need to be retrieved from a relational database through SQL queries or an API. Therefore, no requirements will be defined regarding such concerns. Moreover, the way triples will be inferred must be determined. Inferred triples can either be saved (i.e., *cached*) or inferred at the spot (i.e., *just-in-time*). However, this choice entails important change management implications because cached triples will need to be deleted if their source changes or no longer exists whereas that would not be necessary for just-in-time triples. Given that this project will have a static nature instead of a dynamic one in which sources constantly change, the data in each source system will be converted to and saved as cached triples by using CONSTRUCT queries in SPARQL and saving them in graphs. Additionally, these triples will be saved in an RDF serialization file format known as *turtle*. Such a *turtle* file can be stored and queried by any RDF store. Since URIs represent the dependencies and intersections between graphs, URIs should be standardized and properly managed across graphs. Otherwise, this would result in lots fragmented triples that are not integrable one to another.
- **Semantic harmonization.** The data fields from each data source will be mapped to their respective meanings according to FIBO. As a matter of academic rigor, semantic harmonization principles were aligned to NST and will serve as a foundation to formulate the requirements for this part of the project as follows. As first semantic harmonization requirement, technical and semantic harmonization should be done separately. Secondly, classes should be separated from inference rules. Thirdly, standardized vocabularies, such as the ones provided by FIBO should be reused. Finally, the different concepts in the datasets should be mapped to their respective meanings in FIBO via declarative CONSTRUCT queries in SPARQL [12] [14].
- **Data quality harmonization.** In order to define the requirements for data quality harmonization, data quality will be measured through different metrics provided for this work. Based on the needs of business users and by using these metrics, data will be adjusted if necessary. This will be done separately for each dataset.

 ISBN: 978-1-61208-783-2

V. DATA HARMONIZATION EXECUTION

The planning formulated above has been executed as follows. Firstly, data harmonization has been performed on the individual datasets which, in turn, were converted to individual *named* graphs. Secondly, the goal of defining them as *named* graphs is to have the ability to import them for federated queries in which several graphs are integrated and retrieved at the same time as also done in the next execution step of this work [10]. Finally, the query results were exported to Tableau for further graphical analysis. Each of these steps are described in more detail below. The entire project can be found on https://data.world/carlostubbax1/masters-thesis.

A. Data harmonization step

The four datasets mentioned above were harmonized and converted to four different graphs. For example, the dataset about ownership and control relationships has been mapped to FIBO and converted to a graph (see Figure 3). It can be noticed that the subject of the property *fibo-fnd-oac-oac:ownsAndControl* is the URI of the LEI of the parent company while the object is the one of the subsidiary. Additionally, this property could be considered as transitive given that, if A owns B and B owns C, A owns C. Therefore, *owl:TransitiveProperty* has been used to represent that as an inference rule. For example, given that General Motors Company owns Opel Bank GmbH and Opel Bank GmbH owns Opel Bank GmbH (Niederlassung Griecheland), the query engine will infer that General Motors Company also owns Opel Bank GmbH (Niederlassung Griecheland).

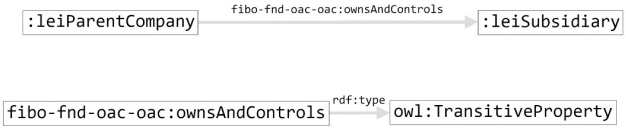


Figure 3. The *fibo-fnd-oac-oac:ownsAndControl* relationship

The namespace *https://lei.info/* has been used to generate standardized URIs of the LEIs of the companies in the datasets. Using that namespace allowed me to generate a URI for each legal entity identifier that can be recycled across several graphs or even across the semantic web. Additionally, this allows users to merge this graph to more graphs enabling its reusability.

For the purpose of data quality harmonization, the timeliness of the ownership relationships is considered to be important since relationships can change due to mergers, acquisitions, registrations problems, etc. Therefore, a timeliness metric was used to clean out ownership relationships that are not considered to be up-to-date. Such a metric is given by the equation below [20].

$$Q_{Time}(w, A) := exp(-decline(A) * age(w, A)$$

$Q_{Time}(w, A)$ denotes the probability that an data field may still be valid. $decline(A)$ depicts the marginal probability that a certain attribute value may become invalid within one period of time. $age(w, A)$ represents the duration between the last update of the data field and the current date which can be represented by variable t. In this case, the decline ratio was obtained by calculating the average percentage of ownership relationships registered by GLEIF that become inactive within just on month. The resulting ratio was 0.0193 or 1.93 percent. The value of t for each ownership relationship was determined by calculating the difference between the time each ownership relationship was updated and the current date. In turn, this was used to calculate the timeliness ratio of each consolidation relationships in the equation below [20].

$$Q_{Time}(t) := e^{-0.0193*t}, \forall t > 0$$

However, since SPARQL does not support this function, the Padé approximant was used to estimate such ratios as shown by the equation below [21].

$$e^{-0.0193*t} \approx \frac{(-0.0193 * t + 3)^2 + 3}{(-0.0193 * t - 3)^2 + 3}, \forall t > 0$$

Variable *t* represents the time period (measured in months) between the last update of each ownership relationship and 10th July 2019 which is the day the ratios were calculated. Based on this, only ratios above 0.7165 were considered to be up-to-date at a significance level of 5 percent. This means that only relationships with ratios above that threshold would be converted and saved as triples in the graph on ownership relationships. In other words, only ownership relationships with a probability higher than 71.65 percent of being up-to-date are considered for further analysis while the rest is excluded.

During the technical harmonization part, the work performed during the semantic and data quality harmonization parts has been saved in the form of triples in a turtle file called *GLEIF-Who-owns-Whom2.ttl*.

B. Data federation step

The four individual graphs made in the data harmonization step have been merged in multiple ways to build different larger graphs through federated queries. Using standardized URIs as explained earlier was crucial for this since URIs represent the intersections and dependencies between triples and graphs.

C. Results visualization step

The results of such federated queries have been exported to Tableau through data.world's APIs to answer the business questions formulated during the planning of this data harmonization project. Some of the results will be shown below. All queries can be found on https://data.world/carlostubbax1/masters-thesis.

1) How many companies are listed by AMEX, NASDAQ and NYSE?: After filtering out repeated values, 5,818 companies listed by any of these three exchanges were found.

TABLE I. BELGIAN SUBSIDIARIES OF JOHNSON & JOHNSON AND THEIR ADDRESSES.

Name	*Location*
AMO Belgium BVBA	1831 Machelen
GMED Healthcare BVBA	1831 Machelen
J.C. General Services CVBA/SCRL	2340 Beerse
Janssen Infectious Diseases-Diagnostics BVBA	2340 Beerse
Janssen Pharmaceutica NV	2340 Beerse
Janssen-Cilag NV	2340 Beerse
Johnson & Johnson Belgium Finance Company CVBA	2340 Beerse
Johnson & Johnson Medical NV	1831 Machelen
Omrix Biopharmaceuticals NV	1831 Machelen

2) What Belgian companies are owned by Johnson & Johnson and were are their headquarters?: The graph pattern in

 ISBN: 978-1-61208-783-2

the federated query necessary to answer this question has been constrained to only generate matches of companies owned by Johnson & Johnson and located in Belgium. The results of such a query are listed in Table I. There are 9 Belgian subsidiaries of Johnson & Johson and all of them are located in Machelen or Beerse.

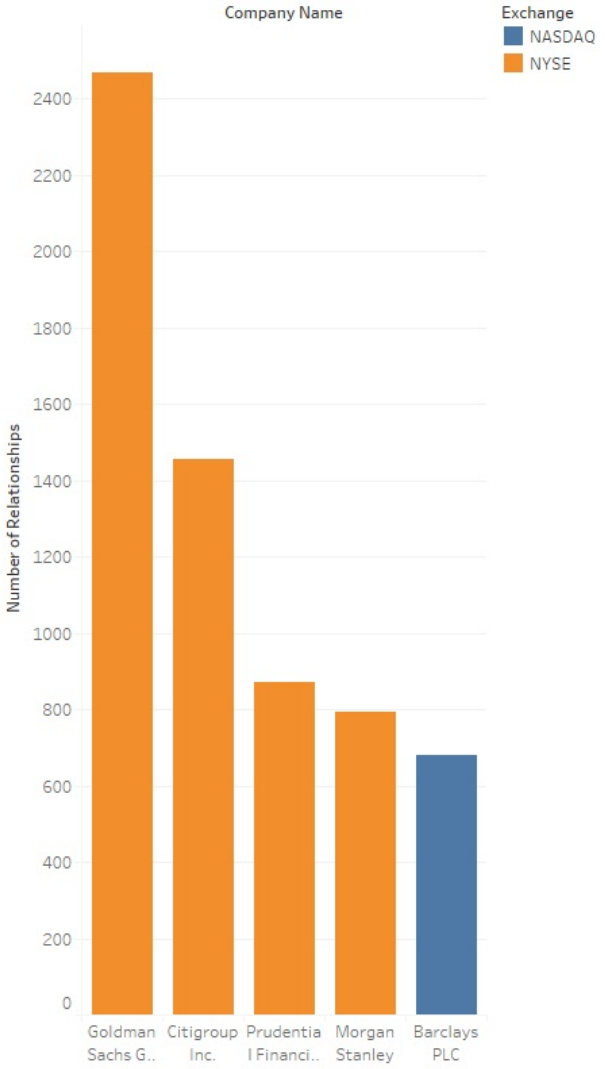


Figure 4. 5 largest companies by number of subsidiaries.

3) What are the five listed companies with the largest number of subsidiaries?: The results of a SPARQL-query meant to answer this question have been sent to Tableau to generate Figure 4. As graphically illustrated, Goldman Sachs is the company with the largest number of subsidaries followed by Citigroup and Prudentia on the second and third places respectively. It can also be seen that all companies in this top 5 come from the financial sector and 4 of them are listed on the New York Stock Exchange (NYSE).

VI. Conclusion

A. Discussion

The execution and results of this data harmonization work illustrate how data from heterogenous sources can be integrated through the use of ontologies and other Semantic Web technologies that allow computers to assign richer context to data by exploiting the power of computer inferencing. This has been shown by first converting several datasets into individual graphs before merging them into different federated graphs that served to answer various business questions about ownerships relations of companies listed in the main stock exchanges in the United States. Furthermore, these answers were exported to Tableau to generate visualizations that are more visually appealing to business users and decision makers.

The main contribution of this work is that it systematically illustrates how data harmonization can be conducted with ontologies and Semantic Web technologies by business users to integrate heterogenous datasets. Additionally, it also introduced Normalized Systems Theory to the body of knowledge on the Semantic Web and data harmonization. The main advantage of this approach is that data integration costs just grow linearly as the number of data sources and data destinations increase.

B. Recommendations

Since URIs represent the dependencies and intersections between graphs and triples, their proper design and management are crucial to ensure that graphs are integrable one to another. Therefore, standardization of URIs across graphs should be encouraged. Additionally, this may increase the reusability of graphs.

During the literature review of this work, little to none specification were found to operationalize the data quality requirements formulated in *BCBS239* [2]. One of these requirements is the timeliness principle for effective risk data aggregation and risk reporting. However, *BCBS239* provides no means nor specifications to measure the timeliness of data. Therefore, the timeliness metric used in this contribution could be used to solve that problem.

Since Semantic Web technologies are considered to be backed by a well-rooted theorical foundation provided by the World Wide Web Consortium (W3C) that may allow organizations to solve many of the interoperability problems they experience on a daily basis. Therefore, organizations should pay closer attention to these technologies.

C. Limitations

The first limitation of this work is that all datasets harmonized and integrated are .csv-files retrieved from the internet. In other words, this work is not representative of the heterogeneity in terms of data formats, sources and types that may normally found in most organizations. Therefore, this work does not capture the level of data source heterogeneity that most business and organizations deal with on a daily basis.

The second limitation of this work stems from the static nature of this data harmonization project that does not represent a more realistic dynamic business environment in which data may need to harmonized on a real-time basis. Therefore, this work does not offer an accurate representation of a dynamic data harmonization environment.

D. Further research

As a result of this research work, some hints for further research were identified. First, it would be interesting to use Normalized Systems Theory to study the modularity of ontologies. Second, it would be of value to study how the modelling languages of the Semantic Web could be used to

 ISBN: 978-1-61208-783-2

model business processes as means to exploit the capabilities provided by machine inferencing to understand data flows within them.

REFERENCES

[1] V. R. Prevosto and L. Francis, "Data and Disaster: The Role of Data in the Financial Crisis," 2010, URL: https://www.casact.org/pubs/forum/10spforum/ [accessed: 2020-04-25].

[2] B. C. on Banking Supervision, "Principles for effective risk data aggregation and risk reporting," 2013, URL: https://www.bis.org/publ/bcbs239.pdf [accessed: 2020-04-25].

[3] W. H. Inmon, *Building the Data Warehouse*. John Wiley Sons, USA, 2005, ISBN: 978-0-471-56960-2.

[4] V. Chaudhary and T. Seth, "Big Data in Finance," 2015, URL: https://www.semanticscholar.org/paper/Big-Data-in-Finance-Seth-Chaudhary/75229e144e857f15c506e87898f8aa35ac1b9852 [accessed: 2020-04-05].

[5] P. Aven and D. Burley, Eds., *Building on multi-model databases: How to manage multiple schemas using a single platform.* O'Reilly, USA, May 2017, ISBN: 978-1-491-97788-0.

[6] T. Berners-Lee and J. Hendler, "The Semantic Web," 2001, URL: https://www.scientificamerican.com/article/the-semantic-web/ [accessed: 2020-04-25].

[7] D. Allemang and J. Hendler, *Semantic Web for the Working Ontologist*. Morgan Kaufmann, USA, 2011, ISBN: 978-0-12-385965-5.

[8] J. R. Burd, S. and J. Satzinger, *Systems Analysis Design in a Changing World*. Cengage Learning, USA, 2009, ISBN: 978-0-12-385965-5.

[9] R. I. Eifrem, E. and J. Webber, *Graph Databases*. OReilly, USA, 2015, ISBN: 978-1-491-93089-2.

[10] B. DuCharme, *Learning SPARQL*. OReilly, USA, 2013, ISBN: 978-1-449-37143-2.

[11] E. D. M. council, "FIBO OWL," 2010, URL: https://spec.edmcouncil.org/fibo/doc/ [accessed: 2020-04-05].

[12] M. H. De Bruyn, P. and J. Verelst, *Normalized Systems Theory. From Foundations for Evolvable Software Toward a General Theory for Evolvable Design*. NSI, Belgium, 2016, ISBN: 978-9-07716-0 09-1.

[13] K. D. V. S. M. Cunningham, J. A. and R. Verbeeck, "Nine Principles of Semantic Harmonization," in *Proceedings of the9^{th} AMIA Annual Symposium, 2017, Somecity, USA*. AMMIA, 2017, pp. 451–459.

[14] M. Flood, "Embracing change:
nancial informatics and risk analytics In: Quantitative Finance," 2009, URL: https://doi.org/10.1080/14697680802366037 [accessed: 2020-04-25].

[15] C. Dictionary, "listed company," 2020, URL: https://dictionary.cambridge.org/fr/dictionnaire/anglais/listed-company [accessed: 2020-04-25].

[16] NASDAQ, "Listed companies dataset," 2020, URL: https://www.nasdaq.com/screening/company-list.aspx [accessed: 2020-07-02].

[17] GLEIF, "level 2 who owns whom," 2020, URL: https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom [accessed: 2020-07-02].

[18] ——, "level 1 data who is who," 2020, URL: https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom [accessed: 2020-04-25].

[19] GeoNames, "Postal Codes," 2020, URL: ttp://download.geonames.org/export/zip/ [accessed: 2020-07-02].

[20] K. M. M. Heinrich, Bernd and M. Klier, "How to measure data quality? - A metric based approach." in *Proceedings of the 28th International Conference on Information Systems, 2007*. ICIS, 2007.

[21] Stackoverflow, "Approximation of e-x," 2011, URL: https://math.stackexchange.com/questions/71357/approximation-of-e-x [accessed: 2020-07-02].

 ISBN: 978-1-61208-783-2