Exploring the Application of Ontologies in Organizations for Data Harmonization

Authors:

- MSc. Carlos Tubbax
- PhD. Jan Verelst

Presenter:

- Carlos Tubbax, carlos.tubbax@uantwerpen.be

Universiteit Antwerpen

Presenter

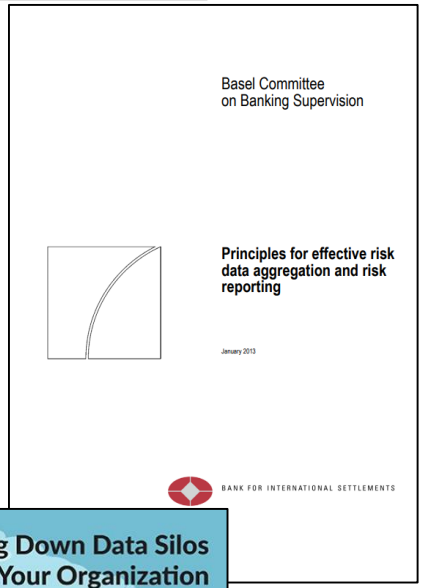
- Carlos Tubbax
- MSc. Business Engineering and Management Information Systems
- Currently a PhD candidate at the UAntwerpen.

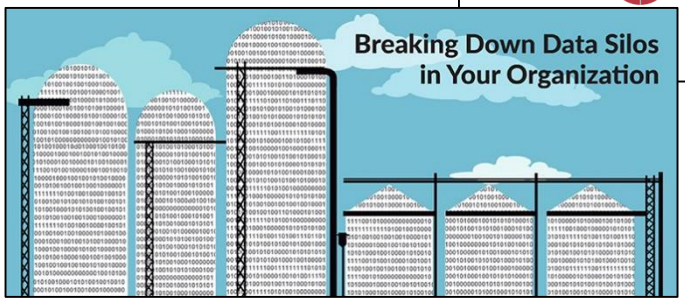
Purpose of this work

- Purpose: illustrate how ontologies can be used to harmonize (i.e. integrate) data from heterogenous data sources.

Problem statement

- Financial crisis and the role of data management problems.
- Compliance with:
  - Self-regulation (e.g. BCBS239).
  - Regulation from authorities.
- Heterogeneous data sources in terms of technology, providers, stakeholders, etc.
- Limitations of relational databases and data-warehouses.





Ontologies

- An ontology consist of:
  - a taxonomy.
  - a set of inference rules.

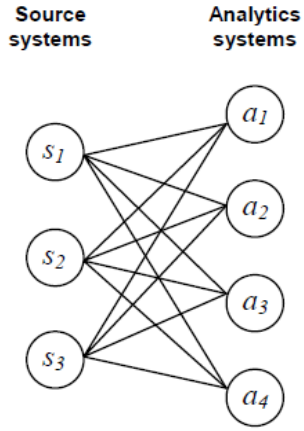
- The Financial Industry Business Ontology (FIBO):
  - Aimed at representing the business logic of financial organizations in a standardized and unambiguous way that is readable by computers and humans.

Data harmonization

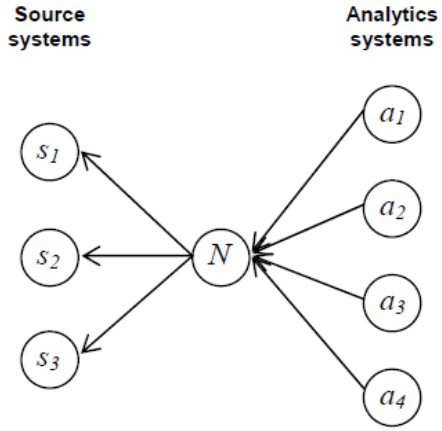
- Data harmonization consists of 3 dimensions, which are:
  - **Technical harmonization**
  - **Semantic harmonization (based on NS theory)**
  - **Data quality harmonization**

Data harmonization architecture

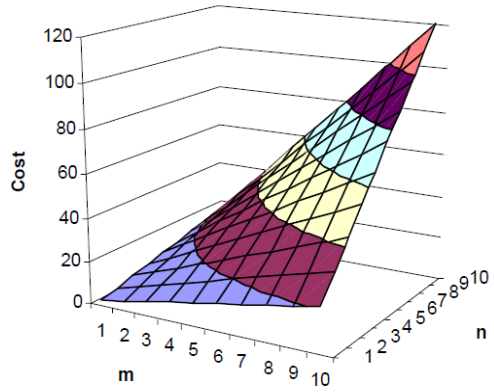
- Two architectures (Flood, 2006):
  - Stovepipe (left)
  - Numeraire (right)

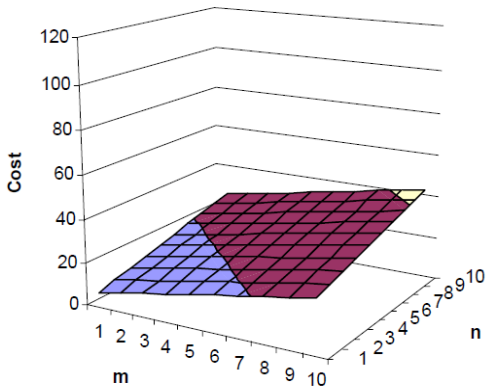


$$Cost = (m+n)k_{spec} + (mn)k_{map}$$



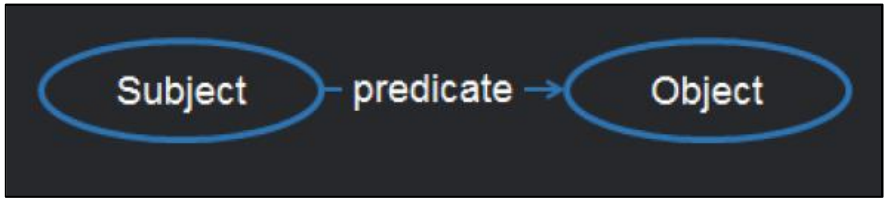
$$Cost = k_N + (m+n)k_{spec} + (m+n)k_{map}$$





Triples

- Data is represented as triples.
- A triple consists of:
  - A subject.
  - A predicate.
  - An object.
- Triples can be merged into a knowledge graph if they share nodes.
- Inferencing can be conducted on triples.



Research methodology

- Systematic top-down approach consisting of 3 steps:

  1. Formulate business questions that need to be answered.
  2. Define a data harmonization architecture and its components according to Flood (2006).
  3. Define requirements for each of the three data harmonization dimensions based on the business questions previously formulated.

- Research questions:
  - **Main research question:** How can business organizations use ontologies for data harmonization?
  - **Subquestions:**

    Work Package 1. How can different sources of financial data be harmonized and stored in different knowledge graphs?

    Work Package 2. How can these different knowledge graphs be merged into federated knowledge graphs?

    Work Package 3. What are the results obtained from such federated knowledge graphs and how can such results be utilized by business users?

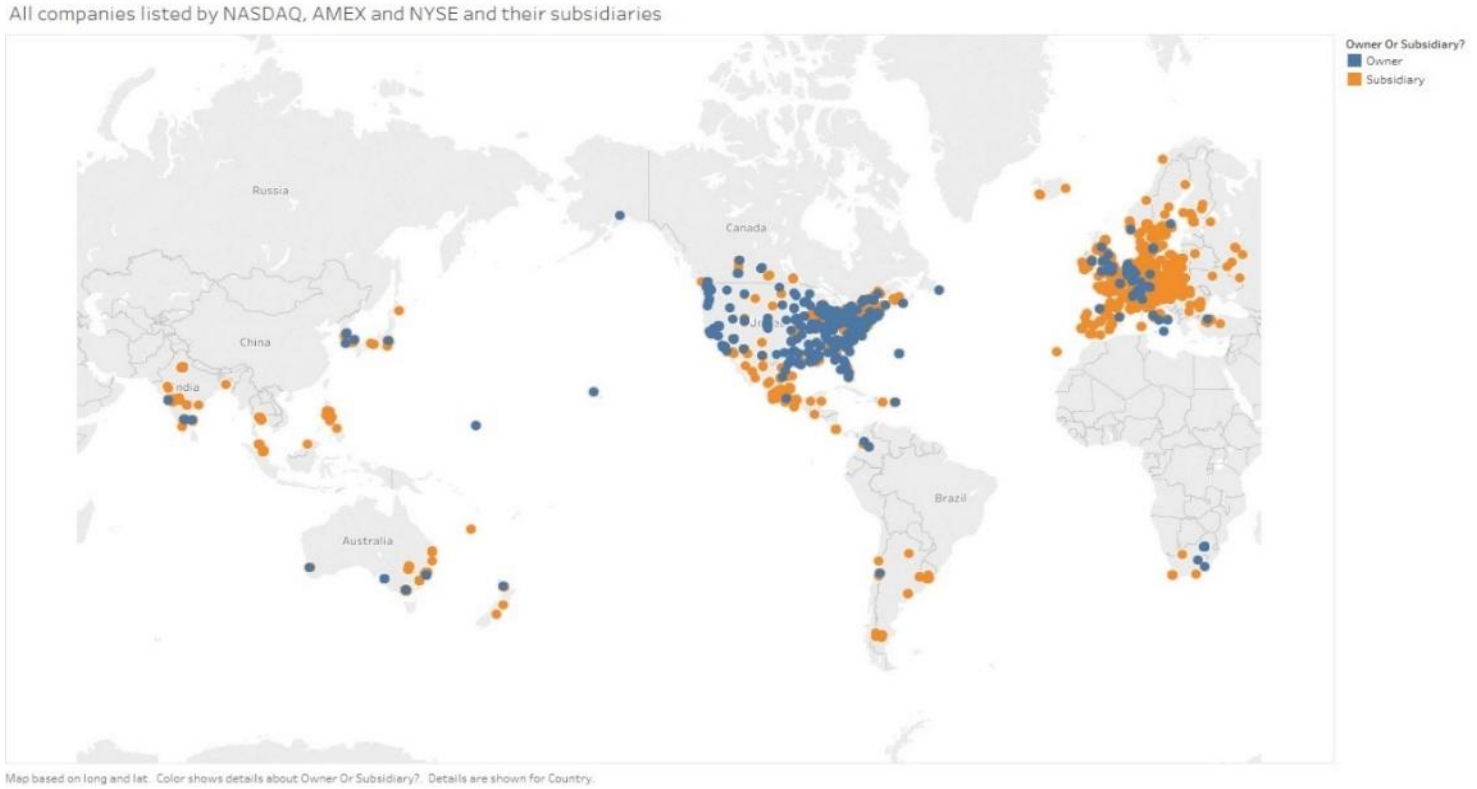
RESEARCH RESULTS

Powered by:

data.world

tableau®

Where are all listed companies and their subsidiaries located?



What Belgian companies are owned by Johnson & Johnson and where are their headquarters?

Subsidiary Name	Headquarters Address	Postal Code	City
AMO Belgium BVBA	Culliganlaan 2B	1831	Machelen
GMED Healthcare BVBA	Leonardo da Vincilaan 15	1831	Machelen
J.C. General Services CVBA/SCRL	Turnhoutseweg 30	2340	Beerse
Janssen Infectious Diseases-Diagnostics BVBA	Turnhoutseweg 30	2340	Beerse
Janssen Pharmaceutica NV	Turnhoutseweg 30	2340	Beerse
Janssen-Cilag NV	Antwerpseweg 15-17	2340	Beerse
Johnson & Johnson Belgium Finance Company CVBA	Turnhoutseweg 30	2340	Beerse
Johnson & Johnson Medical NV	Leonardo da Vincilaan 15	1831	Machelen
Omrix Biopharmaceuticals NV	Leonardo da Vincilaan 15	1831	Machelen

Conclusion

- Dependencies (URIs) critical for reuse and integration of knowledge graphs!

- Timeliness metric for compliance with *BCBS239.*

- Systematic approach for data harmonization with ontologies and semantic technologies.

Purpose of this work

- Purpose: illustrate how ontologies can be used to harmonize (i.e. integrate) data from heterogenous data sources.

Questions


OBRIGADO
gracias
どうも
ARIGATO
hvala
DANK U
takk
謝謝
THANK YOU
OBRIGADO
MERSI
merci
danke schön
köszi
PALDIES
ありがとう
TEŞEKKÜR EDERIM
muchas gracias
grazas
GRAZZI
GRACIAS
благодаря
OBRIGADO
どうも
THANKS
OBRIGADO
quan
TAK
vielen dank
TAK
PALDIES
OBRIGADO
danke
DANK U
mesi
감사합니다
köszi
Благодаря
grazie
TACK
Gràcies
TEŞEKKÜR EDERIM
muchas gracias
obrigada
спасибо
MULTUMESC
多謝
NA GODE

The Semantic Web

- An extension of the current internet by Tim Berners-Lee.
- A web consisting of individual pieces of data concerning *resources.*
  - *A resource* = a thing from the real world such as a person.
- A web in which data receives meaning and context that is understandable by machines thanks to computer inferencing.
- A web in which the content of data is referred to rather than its presentation.
  - URLs refer to web sites, excel-sheets, .pdf-files spread over the internet (at presentation level).
  - URIs (Uniform Resource Identifiers) refer to individual pieces of data concerning *resources (at content level).*

Semantic Modeling

- Semantic modeling languages are categorized by the inference rules that they allow:
  - **Inference rule:** "An enunciate by which the computer derives new information from other given information".
  - It allows users to give context to data also understandable by computers.
- There are 4 semantic modeling languages:
  - Resource Description Framework (RDF).
  - RDF Schema (RDFS).
  - RDFS-Plus.
  - The Ontology Web Language (OWL).

Timeliness ratios

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