Integrating Data using Graphs and Semantics

Juan F. Sequeda
juan@capsenta.com
What do you mean by ...

How many orders were placed in May 2016?

- E-Commerce: 317,595
- Shipping: 317,124
- Billing: 316,899
What do you mean by ...

**What is an Order?**

- **Data resides in different sources**
  - When a user clicks “Order” on the website

- **Ambiguity**
  - When it comes out of the billing system and the CC has been charged

- **No Shared Understanding**
  - When the customer has received the product

- **E-Commerce**

- **Billing**

- **Shipping**

- **Lack of Semantics**
Status Quo 1

Did the Biz User communicate the correct message to IT?

Did IT understand correctly what the Biz User wanted?

Did IT deliver the correct/precise results?

Total net sales of all Orders today

Data Architect

SELECT ...

FROM ...

IT

MS Access

XLS

XLS

XLS

Reports

T=1

T=2

T=3

CSV

CSV

CSV
Status Quo 2

Data Architect

Enterprise Data Warehouse

IT

Biz

Total net sales of all Orders today

Total net sales of all Orders today with FX

Reports

Time and $
Cross Organizational Data Integration
GRAPHS ARE COOL!
“Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted.”
“Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted.”

“A prison official’s ‘deliberate indifference’ to a substantial risk of a serious harm to an inmate violates the Eighth Amendment”
“Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted.”
Common denominator

**XML**

```xml
<constitution id="US_Constitution_1992">
    <section id="US_Constitution_1992/section/123">
        <text>Excessive bail shall ...</text>
    </section>
    <topic>Cruelty</topic>
</constitution>
```

**Text**

“Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted.”

**Tabular**

<table>
<thead>
<tr>
<th>id</th>
<th>text</th>
<th>topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>Excessive bail shall</td>
<td>Cruelty</td>
</tr>
</tbody>
</table>
“Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted.”

“A prison official’s ‘deliberate indifference’ to a substantial risk of a serious harm to an inmate violates the Eighth Amendment.”
“Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted.”
(Summary) Why are Graphs Cool?

- Flexible
- Integration
- Data and Metadata are one

- Common Denominator
- Traversal, Navigation, Reachability
- Semantics

Survey of Graph Database Models

RENZO ANGLES and CLAUDIO GUTIERREZ

Universidad de Chile

Graph database models can be defined as those in which data structures for the schema and instances are modeled as graphs or generalizations of them, and data manipulation is expressed by graph-oriented operations and type constructors. These models took off in the eighties and early nineties alongside object-oriented models. Their influence gradually died out with the emergence of other database models, in particular geographical, spatial, semistructured, and XML. Recently, the need to manage information with graph-like nature has reestablished the relevance of this area. The main objective of this survey is to present the work that has been conducted in the area of graph database modeling, concentrating on data structures, query languages, and integrity constraints.

ACM Computing Surveys 2008
Integrating Data using Graphs and Semantics
MAPPING RELATIONAL DATABASES TO GRAPHS
Relational Database to RDF (RDB2RDF)

**Person**

<table>
<thead>
<tr>
<th>ID</th>
<th>NAME</th>
<th>AGE</th>
<th>CID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alice</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>Bob</td>
<td>NULL</td>
<td>100</td>
</tr>
</tbody>
</table>

**City**

<table>
<thead>
<tr>
<th>CID</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Austin</td>
</tr>
<tr>
<td>200</td>
<td>Madrid</td>
</tr>
</tbody>
</table>

Mapping:
- Alice: foaf:name
  - foaf:age: 25
- Bob: foaf:name
- <Person/1> foaf:based_near <City/100>
- <Person/2> foaf:based_near <City/200>

**City**
- <City/100> rdfs:label: Austin
- <City/200> rdfs:label: Madrid
W3C RDB2RDF Standards

• Standards to map Relational Data to RDF

• A Direct Mapping of Relational Data to RDF
  – Default automatic mapping of relational data to RDF

• R2RML: RDB to RDF Mapping Language
  – Customizable language to map relational data to RDF
W3C Direct Mapping

Input:
- Database (Schema and Data)
- Primary Keys
- Foreign Keys

Output:
- RDF graph

Diagram:
- Relational Database
- Direct Mapping Engine
- RDF

A Direct Mapping of Relational Data to RDF
W3C Recommendation 27 September 2012

This version: http://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/
Latest version: http://www.w3.org/TR/rdb-direct-mapping/
Previous version: http://www.w3.org/TR/2012/PR-rdb-direct-mapping-20120814/
Editors:
- Marcelo Arenas, Pontificia Universidad Católica de Chile
- Alexandre Bertela, W3C <bertela@w3.org>
- Eric Prudhommeaux, W3C <eric@w3.org>
- Juan Sequeda, University of Texas at Austin <sequeda@cs.utexas.edu>
W3C Direct Mapping Result

<table>
<thead>
<tr>
<th>ID</th>
<th>NAME</th>
<th>AGE</th>
<th>CID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alice</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>Bob</td>
<td>NULL</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CID</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Austin</td>
</tr>
<tr>
<td>200</td>
<td>Madrid</td>
</tr>
</tbody>
</table>

Person

City

Direct Mapping
R2RML: RDB to RDF Mapping Language

W3C Recommendation 27 September 2012

This version:
http://www.w3.org/TR/r2rml/

W3C Recommendation 27 September 2012

This version:
http://www.w3.org/TR/r2rml/

Previous version:
http://www.w3.org/TR/r2rml/

Editors:
Sanghoon Dae, Oracle
Sara Eroz, Oracle
Richard Cyganiak, National University of Ireland, Galway

Capsenta
The Smart Data Company™
Example R2RML

```xml
<TriplesMap1>
  a rr:TriplesMap;
  rr:logicalTable [ rr:tableName "Person" ];

  "http://www.ex.com/Person/{ID}";
  foaf:Person

  rr:subjectMap [ rr:template "http://ex.com/City/{CID}";
                  rr:class ex:City ];

  rr:predicateObjectMap [ 
    rr:predicate foaf:name;
    rr:objectMap [ rr:column "TITLE" ]
  ]
</TriplesMap1>
```
Graph Data Virtualization
NoETL Architecture
Hybrid NoETL and ETL Architecture
Scalability

• Seconds vs Months
• Reuse existing relational infrastructure
  – 30+ years of optimizations
  – Semantic Query Optimizations
• Result: SPARQL as fast as SQL under mappings

The Tipping Point Problem

- Flexible
- Integration
- Data and Metadata are One
- Common Denominator
- Traversal, Navigation, Reachability
- Semantics

An overarching theme is the need to create systematic and real-world benchmarks in order to evaluate different solutions for these features.

Sequeda (2015) Integrating Relational Databases with the Semantic Web