Semantic-based Process Analysis

Mauro Dragoni
Fondazione Bruno Kessler (FBK), Shape and Evolve Living Knowledge Unit (SHELL)

https://shell.fbk.eu/index.php/Mauro_Dragoni - dragoni@fbk.eu

work done in collaboration with
Piergiorgio Bertoli\textsuperscript{2}, Francesco Corcoglioniti\textsuperscript{1}, Chiara Di Francescomarino\textsuperscript{1}, Chiara Ghidini\textsuperscript{1}, Michele Nori\textsuperscript{2}, and Marco Pistore\textsuperscript{2}, and Roberto Tiella\textsuperscript{1}

\textsuperscript{1}Fondazione Bruno Kessler, Trento
\textsuperscript{2}SAYService, Trento

ISWC 2014 – Riva del Garda, Trento
October, 23\textsuperscript{rd} 2014
Process Analysis

Extracts analytical knowledge about the performances of a business process starting from collected process execution data

Three Challenges

- **Challenge 1**: Combining three different dimensions.
  - D1: the procedural dimension (P)
  - D2: the domain of interest (K)
  - D3: the execution dimension (T)

- **Challenge 2**: Semantic Reasoning

- **Challenge 3**: Scalability
Semantic Process Analysis

Employs (SW) techniques that leverage the explicit formalization of the semantics of a business process and the data it manipulates.

Our approach / contributions:

- Integrated OWL 2 / RDF model of P + K + T queried with SPARQL
  
  → address Challenge 1

- OWL 2 reasoning for making explicit inferrable knowledge
  
  → address Challenge 2

- Implementation based on SW triplestores
  
  → address Challenge 3
Outline

1. Our Use Case
2. The Proposed Model
3. The Architectural Solution
4. Evaluation
Use case: Birth Management Process
The Proposed Model: an Integrated View

- Reconciliation of knowledge and information related to different dimensions:

[Diagram showing the proposed model with different dimensions and their reconciliation process.]
The Proposed Model: an Integrated View

- Reconciliation of knowledge and information related to different dimensions:
The Integrated Ontological Model
The Integrated Ontological Model

- BPMN Ontology


https://shell-static.fbk.eu/resources/ontologies/bpmn2_ontology.owl
The Integrated Ontological Model

- BPMN Ontology
- Domain Ontology
The Integrated Ontological Model

- BPMN Ontology
- Domain Ontology
- Trace Ontology
The Integrated Ontological Model
The Architectural Solution

- Challenges to cope:
  - collect trace data at fast rate
  - answer to complex queries
- Investigated solution: architecture based on triplestores
How data are organized?

- Process model and traces are stored in separated graphs
- Both explicit and implicit (inferred) data are stored
How data are populated?

- Process model (defined at design time) is stored once per all offline.
- Trace update operation occurs every time a new piece of information is available.
- Separate triplestores are used in order to allow different optimizations based on their purpose.
How data are retrieved?

- Queries performed by using SPARQL 1.1
- SPARQL aggregates turned out to be useful for analytical queries
- We introduced SPARQL extension mechanism
Evaluation - 1

- **Process** $P$: 4 pools, 19 activities, 11 domain objects, 19 events, 14 gateways, 54 sequence flows, 6 message flows.

- **Domain ontology** $K$:
  - 379 classes covering 28 activities and 12 data objects;
  - average of ~25 fields for each data object;
  - max field level-depth 4;
  - 5 properties.

- **Set of execution traces** $T$:
  - average of ~10 events for each trace;
  - average of 2040 triples for each trace;
  - further 1260 triples can be inferred.
## Evaluation - 2

<table>
<thead>
<tr>
<th>Query</th>
<th>Description</th>
<th>P</th>
<th>K</th>
<th>T</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q.1</td>
<td>Average time per process execution spent by a specific municipality.</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Q.2</td>
<td>Total number of Registration Request documents filled from January, 1st, 2014.</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Q.3</td>
<td>Percentage of times in which the flow followed is the one which passes first through the APSS pool and then through the municipality one.</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Q.4</td>
<td>Number of cases and average time spent by each public office involved in the birth management procedure for executing optional activities.</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Q.5</td>
<td>Number of times in which the municipality sends to SAIA a request without FiscalCode.</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Q.6</td>
<td>Last event of trace TraceID.</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Q.7</td>
<td>Average time spent by trace TraceID.</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Q.8</td>
<td>Does the trace TraceID pass through the activity labeled with “PresentAtTheHospital”?</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Traces</td>
<td>Stored triples</td>
<td>Storing</td>
<td>Querying</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>---------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asserted</td>
<td>Inferred</td>
<td>Total</td>
<td>Throughput</td>
<td>Total time</td>
</tr>
<tr>
<td>1500</td>
<td>3062349</td>
<td>1895471</td>
<td>4957820</td>
<td>37.89 trace/min</td>
<td>2426.88 s</td>
</tr>
<tr>
<td>10500</td>
<td>21910269</td>
<td>13057464</td>
<td>34967773</td>
<td>37.41 trace/min</td>
<td>16851.21 s</td>
</tr>
<tr>
<td>42000</td>
<td>87503538</td>
<td>52045200</td>
<td>139548738</td>
<td>37.34 trace/min</td>
<td>67537.95 s</td>
</tr>
</tbody>
</table>
### Evaluation - 3

Daily, weekly, and monthly load.

<table>
<thead>
<tr>
<th>Traces</th>
<th>Stored triples</th>
<th>Throughput</th>
<th>Total time</th>
<th>Avg. Time Q.4</th>
<th>Avg. Time Q.8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asserted</td>
<td>Inferred</td>
<td>Total</td>
<td>Throughput</td>
<td>Total time</td>
</tr>
<tr>
<td>1500</td>
<td>3062349</td>
<td>1895471</td>
<td>4957820</td>
<td>37.89 trace/min</td>
<td>2426.88 s</td>
</tr>
<tr>
<td>10500</td>
<td>21910269</td>
<td>13057464</td>
<td>34967773</td>
<td>37.41 trace/min</td>
<td>16851.21 s</td>
</tr>
<tr>
<td>42000</td>
<td>87503538</td>
<td>52045200</td>
<td>139548738</td>
<td>37.34 trace/min</td>
<td>67537.95 s</td>
</tr>
</tbody>
</table>
### Evaluation - 3

<table>
<thead>
<tr>
<th>Traces</th>
<th>Stored triples</th>
<th>Storing</th>
<th>Querying</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asserted</td>
<td>Inferred</td>
<td>Total</td>
</tr>
<tr>
<td>1500</td>
<td>3062349</td>
<td>1895471</td>
<td>4957820</td>
</tr>
<tr>
<td>10500</td>
<td>21910269</td>
<td>13057464</td>
<td>34967773</td>
</tr>
<tr>
<td>42000</td>
<td>87503538</td>
<td>52045200</td>
<td>139548738</td>
</tr>
</tbody>
</table>

Throughput independent of the load.

Daily, weekly, and monthly load.
## Evaluation - 3

<table>
<thead>
<tr>
<th>Traces</th>
<th>Stored triples</th>
<th>Throughput</th>
<th>Storing</th>
<th>Querying</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asserted</td>
<td>Inferred</td>
<td>Total</td>
<td>Throughput</td>
</tr>
<tr>
<td>1500</td>
<td>3062349</td>
<td>1895471</td>
<td>4957820</td>
<td>37.89 trace/min</td>
</tr>
<tr>
<td>10500</td>
<td>21910269</td>
<td>13057464</td>
<td>34967773</td>
<td>37.41 trace/min</td>
</tr>
<tr>
<td>42000</td>
<td>87503538</td>
<td>52045200</td>
<td>139548738</td>
<td>37.34 trace/min</td>
</tr>
</tbody>
</table>

Throughput independent of the load

Time required for queries is acceptable for real-time usage

Daily, weekly, and monthly load.
Lessons Learned

- Divide-et-impera approach for inference to scale
- Usability feedbacks and expertise requirements
Thank You!

Mauro Dragoni
https://shell.fbk.eu/index.php/Mauro_Dragoni
dragoni@fbk.eu