An Ontology-based Framework for Enriching Event-log Data

Thanh Tran Thi Kim, Hannes Werthner
e-commerce group
Institute of Software Technology and Interactive Systems, Vienna, Austria
Email: kimthanh@ec.tuwien.ac.at, werthner@ec.tuwien.ac.at

***Abstract*—Using process-aware information systems in enterprises is becoming popular in the business environment. The systems have the capability to generate event log data that capture information about what is practically happening within enterprises. Event log data is used for process mining to extract the hidden knowledge which can assist the manager in business process management. However, the knowledge hidden in event logs would be more useful if the event logs are enriched by relevant external data sources. In this paper, we propose an approach to enrich event logs with external data sources by using ontology based data integration. We use database-to-ontology mapping techniques to integrate data sources and use semantic reasoning techniques for inferring the knowledge hidden in the data sources. A framework for the approach, illustrating examples for the implementation and expected results are presented in this paper.**

Keywords-process-aware information systems; data integration; process mining.

I. INTRODUCTION

Process-aware Information Systems (PAISs) are increasingly used by many enterprises in the modern business environment. A PAIS is defined as a software system that manages and executes operational processes involving people, applications, and/or information sources on the basis of process models [1]. Moreover, the system has the capability to generate event log files, which record the information of real executions within enterprises. The knowledge hidden in the event logs is extracted by process mining techniques and used for model construction and analysis [2]. In particular, process mining application includes features of three categories: model construction, statistical performance analysis and knowledge discovery. Model construction refers to the dynamic building of business process based on the information contained in event logs. Statistical performance analysis aims to extract predefined statistical measures. Knowledge discovery is the incorporation of event log data with other data sources to search for hidden patterns and relationships [3]. Several studies have been carried out to show the potential of this incorporation. Most of them use data warehouse techniques for integrating data sources and extracting knowledge from the data sources [3], [4]. However, complexity problems are raised as challenges for this approach [4], [5]. Workflow executions may generate different kinds of facts about workflow activities, resources, and instances. Because of the multiple, related types of facts, the approach may be faced with semantic problems. Particularly, the presence of these kinds of facts needs to ensure semantic correctness to avoid information loss [5].

To avoid the problems of the data warehouse approach, we propose the framework for integrating event logs with other data sources based on the TOVE ontology [6], [7]. TOVE (TOronto Virtual Enterprise) is an integrated ontology for supporting enterprise modeling which contains concepts related to business models, such as activity, organization agent, cost, resources, etc. Event logs are exported by PAISs to record the operations of business processes in companies, such as the information about who performs which activities at what time. The approach is raised by the question how to enrich event log data and what knowledge could be gained from the enrichment. Merging data in event logs with other data sources are mentioned in [3] as a potential approach for knowledge discovery in process mining. The benefit of the approach could be seen in the enriched event logs which is extended with relevant information by linking to ontologies. Therefore, the knowledge extracted from event logs is collected not only from the event logs but also from others company related data sources, which are related and linked to them. For instance, cost data is not included in event logs but can be inferred by reasoning from the cost ontology in TOVE. Therefore, the results of process mining in can be opened to new perspectives, e.g., cost perspective.

In general, our approach contains two main parts: ontology based data integration and knowledge discovery. Ontologies are very useful in knowledge sharing and integration as well as knowledge research and extraction [8]. In our study, we use TOVE ontology as a conceptual framework for integrating data sources. In particular, event log data and organizational data are migrated to TOVE ontology as instances. Hence, TOVE becomes a knowledge base and can be used for knowledge discovery. As a result, competency questions related to business process management can be answered by querying the axioms constructed in TOVE.

The remainder of the paper is structured as follows: Section 2 introduces the various data sources and the TOVE ontology which are the main objects of the integration. Section 3 presents the framework for mapping event logs and other data sources to the TOVE ontology. Section 4 illustrates the querying axioms for answering questions related to business process management and the expected results.

 ISBN: 978-1-61208-175-5

Section 5 presents the related work, including knowledge discovery in process mining, semantic process mining and TOVE ontology. Finally, Section 6 concludes the paper.

II. THE TOVE ONTOLOGY AND VARIOUS DATA SOURCES

A. The TOVE ontology

TOVE is an integrated set of ontologies for supporting enterprise modeling [9]. The development of the TOVE ontology is driven by the specification of tasks that arise from enterprise engineering within the TOVE project [7]. The goal of enterprise engineering is to formalize the knowledge required for business process reengineering and create an environment that facilitates the application of this knowledge to a particular company. The ontology consists of a set of generic core ontologies, including an activity ontology, resource ontology, organization ontology, product ontology. It also includes a set of extensions to these generic ontologies to cover concepts such as cost and quality.

The primary component of the ontology is its terminology for classes of processes and relations of processes and resources, along with definitions of these classes and relations. Within TOVE, the activity ontology plays an important role and relates to most of axioms [9], [10]. In TOVE, activities are defined as the basic entities that specify a transformation in the world. An activity in TOVE is accompanied with its corresponding states which defines what has to be true in the world in order for the activity to be performed. Moreover, an activity is performed by an organization agent with a particular amount of resources. Based on the relations between activity, organization, resource ontologies, most of questions related to enterprise management are satisfied by querying the axioms built in the ontology. Another prominent part of TOVE is the cost ontology. Costs are related to consuming resources and time when performing activities. Figure 1 shows a set of generic core ontologies in TOVE.

The TOVE ontology presents a mature framework whereas event log data have a simple data structure. Event logs contain information about activites, originators who perform the activities, the process instances which the activites belong to, and the timestamp when the activities occur. Opposite with the simplicity of event log data, TOVE contains many concepts, as shown in Figure 1 and most of the concepts of TOVE are not related directly to event log data elements. Therefore, we use a part of TOVE which are simplified to be suitable with the event log data. For example, the activity ontology in TOVE has relations with the product requirement constraints concept. However, we bypass the product requirement constraints concept because the data of the product requirement constraints do not exist in event log data.

In our approach, we select the activity ontology, organization ontology, resource ontology and cost ontology. In addition, we add a new concept to TOVE (i.e., process concept) and modify some properties of concepts in TOVE to correspond with the properties of event log data and organization database. The knowledge derived from TOVE will be used to enhance process models as results of the process mining.

B. Various data sources

The different data sources in our project are event log data and organization databases. We assume that in companies which are using information systems to support business management, event log data can be received from a PAIS and organization databases obviously exist in a particular database system. The details of event log data and organization database are described as follows.

PAISs produce event log files to record the operation of business processes. Depending on the particular PAISs in use, event log data may contain various types of information in different formats. Generally, an event log data record is consisting of an activity (task name), originator, timestamp, event type and case identification elements [2]. The activity element indicates the name of the activity or the task which is operated. Originator implies entities who initiate or perform the activity. Timestamp is the point of time when the activity happened. Event type denotes the state of the activity (e.g., the start or completion or postpone of the activity). And case identification is a unique number that identifies a specific process instance to which the activity belongs. Although the contents of a log data record may vary, event logs need to contain at least activity and case identification elements.

As the example of Table I shows, *activity A* was performed by Mark at the time *17-05-2008:16:09*; the activity was in the *start* state and belongs to the *case 1*. *Case 1* includes a number of activities, such as *activity A* and *activity B*. All the activities are ordered by their respective timestamp.

Table I
EXAMPLE OF AN EVENT LOG

case id	activity id	originator	timestamp	event type
case 1	activity A	Mark	17-05-2008:16:09	start
case 2	activity B	Chris	18-05-2008:09:12	start
case 1	activity C	Tom	18-05-2008:10:06	complete
case 3	activity B	Mary	18-05-2008:15:02	start
...	...	...	...	...

In terms of semantics, a log file refers to a set of process instances (i.e., cases). Each process instance includes a number of events happened within the process. An event occurs when an activity is operated by an originator at a certain point of time (i.e., timestamp). Each event has an event type representing the status of the event when it is performed, e.g., start or complete. Hence, one can observe that TOVE ontological concepts for enterprise operation are considerably similar to the concepts appearing in event logs.

Considering the data fields in event logs, there is the originator element which contains information of employees

 ISBN: 978-1-61208-175-5

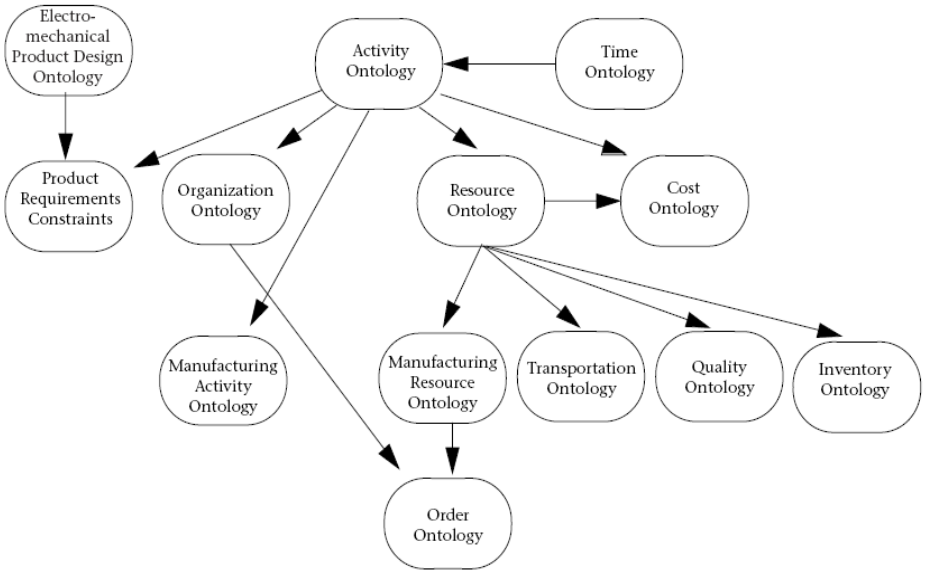

Figure 1. TOVE Ontology [11]

who perform activities. For business management, it is obvious to store the data of employees in a database, i.e., an organizational database with data schema as presented in Figure 2.

The important tables in the database are *employee* and *activity* which are related directly to *originator* and *activity* respectively in event log data. Based on the properties in these table, the information of *originator* and *activity* in event log could be extended. For instance, an originator has information about address, experience year or the labour cost, etc.

III. Framework for integrating event logs and other data sources based on the TOVE ontology

There are two main functionality blocks in the framework: mapping and knowledge discovery. In this context, mapping refers to the adding of instances into the TOVE ontology from data sources. The derived result of the mapping is the TOVE ontology with instances which is regarded as a knowledge base. Knowledge discovery is performed by querying axioms in the knowledge base. Figure 3 represents briefly the framework of the ontology based integration in our approach.

We have two types of data sources, event log data and organizational database. As mentioned in Section 2, we suppose event log data contains information about activities, originators, timestamps, and cases identifications. Organizational database contains the information support for enterprise management dealing with cost accounting, human resources management or resources. The mapping

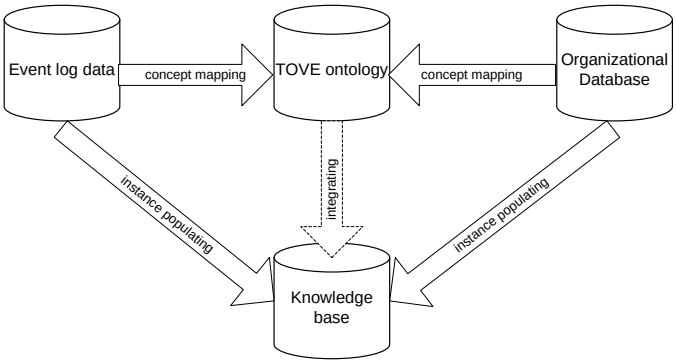

Figure 3. Mapping event log data and extra data sources to the TOVE ontology.

from event log data to TOVE ontology is considered as the migration instances from data fields (i.e., activity, originator and case) to concepts (i.e., activity, organization-agent and process) respectively. Particularly, the values of the name properties in TOVE ontology is filled by the values of the data fields in the event logs. The values of the rest of the properties in TOVE are filled by the values of data fields in the organizational database. The mapping is referred to database-to-ontology mapping whereby a database and an ontology are semantically related at a conceptual level [12], [13]. In our approach, we assume the concept of originator in event logs is similar to an organization agent in TOVE. Likewise, event and timestamp correspond to activity and timestamp, respectively. Therefore, the integration based on

 ISBN: 978-1-61208-175-5

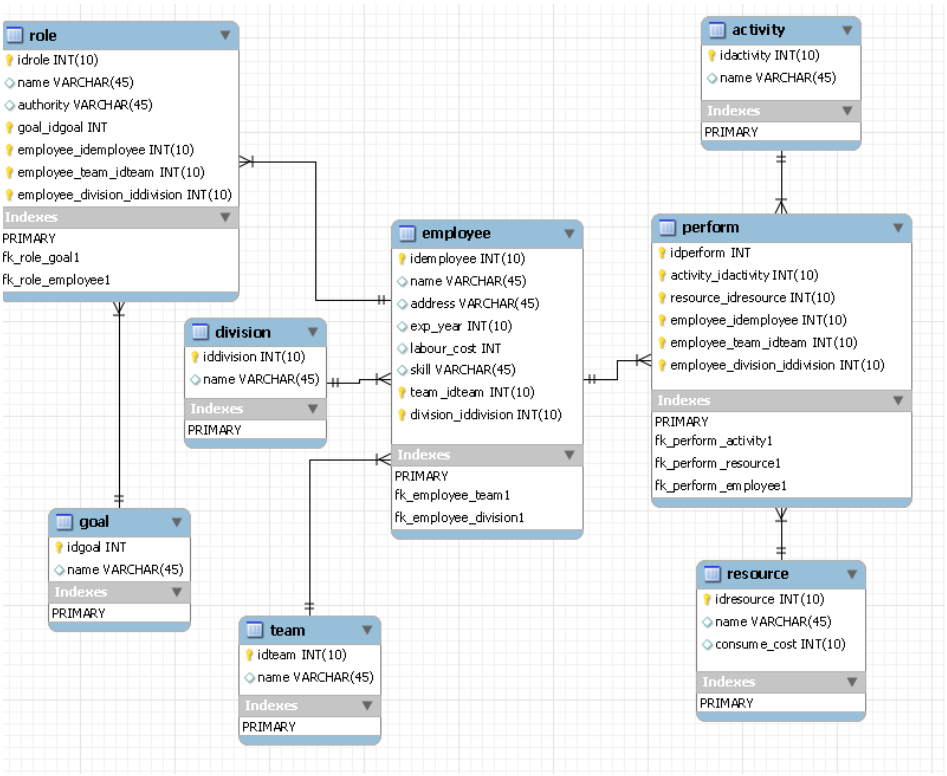

Figure 2. Organizational Database

the TOVE ontology is feasible.

Using reasoning techniques over the ontologies can discover knowledge hidden in the data sources. The reasoning is done by querying the axioms in the TOVE ontology. Note that there are a huge number of axioms in TOVE which support for answering the question related to enterprise management and modeling [9], [10], [14]. Thus, the reasoning may be valuable for knowledge discovery. As a result, combining the semantic reasoning and process mining techniques for discovering knowledge in the enriched event log data represents a sound approach for semantic process mining.

To implement the framework, we use Java [15] as a foundation to combine several techniques. In particular, the event log files are stored in XML format and the organizational database is managed by MySQL [16] . The TOVE ontology and the knowledge base are encoded and stored in WSML format [17]. Besides, several java packages are utilized for data integration (e.g., javax.xml.xpath, java.sql, etc.) and knowledge extraction (e.g., wsmo4j). Within this paper, we introduce a part of knowledge base and the expected results of the knowledge extraction in Section 4.

IV. QUERYING AXIOMS FOR ANSWERING QUESTIONS RELATED TO BUSINESS PROCESS MANAGEMENT

As a result of the ontology-based data integration process, we obtain an knowledge base containing event log data and organizational data. In this section, we illustrate an example about querying axioms for answering questions related to costs of business processes. Figure 4 shows a part of the knowledge base as a diagram of concepts with their properties.

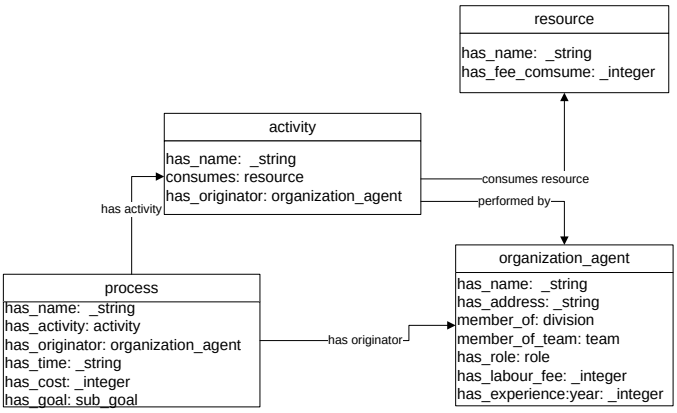

Figure 4. A part of the knowledge base

There are four concepts *resource, activity, organization-agent and process*. They are related by the relationships *consumes resource, performed by, has activity, has originator*. Considering the concept *process*, it is an additional concept which is added to TOVE to use information about process

instances in event log data. Based on this concept, questions related to processes can be answered.

Deriving costs of business processes is currently not possible with process mining. In our approach, an interesting question that can be answered is "How much does a process cost"?. We use the WSML toolkit for building the ontology and testing axioms as shown in Figure 5.

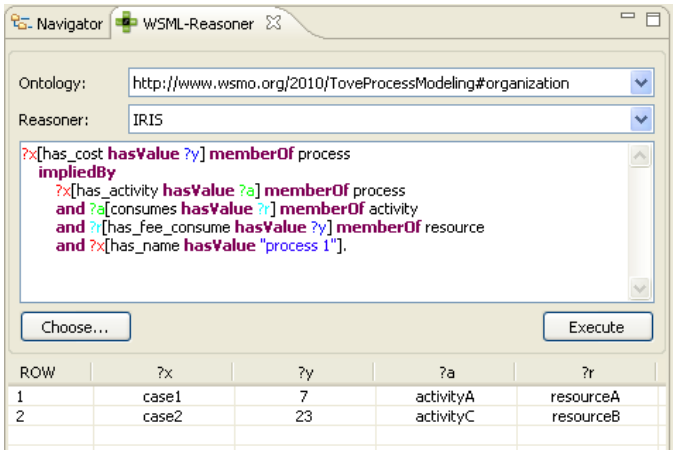

Figure 5. Reasoning with WSML toolkit

Figure 5 displays an axiom of the knowledge base in WSML format for costing a process. WSML utilizes logical expression syntax for the specification of axioms, in other words, rules are defined as logical expression in WSML. In the example, the rule "how much does a process (e.g., *process 1*) cost?" is demonstrated. In detail, the *process 1* is defined by two instances of the concept *process* (i.e., *case1* and *case2*). The *process 1* has two activities, *activityA* and *activityC*. Each activity consumes resources which have particular costs associated. In this case, the *resourceA* has cost 7 and *resourceB* has cost 23 which are values of the property *has_fee_consume* . Therefore, the cost of the *process 1* is inferred from the cost of *resourceA* and *resourceB* which are used by *activityA* and *activityC* respectively.

Moreover, based on the constraints between the concepts shown in Figure 4, various kinds of questions can be answered, such as:

- How much does the consumption of resources cost for performing *activity A* in *process 1*?
- Which resources are consumed in *process 1*?
- How much does it cost for performing *process 1*?

V. Related Work

Knowledge discovery in process mining by incorporating event logs with other data sources is mentioned in [3], [4], [5]. Most of the authors use a data warehouse approach for integrating and extracting knowledge from the data sources. It provides a platform for mining unknown and valuable patterns and relationships. Some of the significant techniques in this area, such as OLAP (online analytical processing), traditional database queries, data mining, and etc., are used in this field. OLAP technology enables data warehouses to be used effectively for online analysis, providing rapid responses to interactive complex analytical queries [3]. On the other hand, traditional database queries can answer simple questions. In contrast, data mining with specific algorithms can identify discernible patterns and trends in data, and it can support prediction and decision making. The merging of data from event logs with other data sources are carried out within several studies [3], [4], [18].

Process mining aims to discover what really happened in the enterprise systems based on event logs recorded by PAISs. Depending on the kind of information contained in event logs, the process mining is separated into three perspectives, i.e., process perspective, organizational perspective and case perspective which respectively answers the question "How?", "Who?" and "What?" [2]. The results delivered from process mining might be process models, analysis diagrams, or answers for questions involved to business process management. Although some process mining algorithms are borrowed from data mining or others fields, all of them are developed and adapted for the goals of process mining as mentioned above. The significant capability of process mining is to reveal the hidden knowledge in event logs to aid the enterprises to know what is really going on in their systems [2]. To practice process mining, more than 280 plug-ins have been implemented in ProM [19], [20]. Some of process mining techniques have been implemented as tools and applied in the real systems such as health care systems in hospitals or invoice processing systems, and brought out benefits for the enterprises in the domains [2], [21].

To keep improve the achievements gained in process mining, a new approach has been researched which is called semantic process mining and carried out within the SUPER project [22]. Basically, the methodology is to connect elements in event logs with adequate concepts in ontologies and cooperate the process mining and semantic techniques to deliver on expected results. With this approach, process mining has been raised from the syntactic level to the concept level in which it is more effective and useful for business analysts as well as normal users [23]. Compared with our approach, in the SUPER approach event logs are also enriched by connecting with concepts in ontologies. However, the difference is that the knowledge discovery in the SUPER approach is done by enriching event logs, whereas in our approach it is performed in the TOVE ontology (i.e., the knowledge base). Moreover, with the ontology based integration, the enrichment can be done with different data sources.

VI. Conclusion

This paper proposed a framework for integrating event logs with other data sources and mapping them to on-

 ISBN: 978-1-61208-175-5

tologies and afterwards using these results in semantic process mining. The mapping is termed as database-to-ontology mapping and supported by several existing tools. For this purpose, we use the TOVE ontology, which in our case is populated with instances extracted from different data sources. The integration enriches the event logs with extra information from the other data sources. It serves for answering questions (by reasoning) relating event logs with organizational data. This framework is already implemented and currently evaluated.

REFERENCES

[1] M. Dumas, W. M. van der Aalst, and A. H. ter Hofstede, *Process-aware information systems: bridging people and software through process technology*. New York, NY, USA: John Wiley & Sons, Inc., 2005.

[2] W. M. P. van der Aalst, H. A. Reijers, A. J. M. M. Weijters, B. F. van Dongen, A. K. A. de Medeiros, M. Song, and H. M. W. E. Verbeek, "Business process mining: An industrial application," *Inf. Syst.*, vol. 32, no. 5, pp. 713–732, 2007.

[3] J. E. Ingvaldsen and J. A. Gulla, "Model-based business process mining," *IS Management*, vol. 23, no. 1, pp. 19–31, 2006.

[4] M. zur Mühlen, "Process-driven management information systems - combining data warehouses and workflow technology," in *Fourth International Conference on Electronic Commerce Research*, B. Gavish, Ed., 2001, pp. 550–566, dallas.

[5] A. Bonifati, F. Casati, U. Dayal, and M.-C. Shan, "Warehousing workflow data: Challenges and opportunities," in *VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 649–652.

[6] M. S. Fox, M. Barbuceanu, and M. Grüningerr, *An Organisation Ontology for Enterprise Modelling: Preliminary Concepts for Linking Structure and Behaviour*, 1995, pp. 123–134.

[7] E. I. Laboratory, "Tove ontology project," http://www.eil.utoronto.ca/enterprise-modelling/tove/, retrieved: August, 2011.

[8] M. Hepp, "Ontologies: State of the art, business potential, and grand challenges," in *Ontology Management, Semantic Web, Semantic Web Services, and Business Applications*. Springer, 2008, pp. 3–22.

[9] M. Grüninger, K. Atefi, and M. S. Fox, "Ontologies to support process integration in enterprise engineering," *Comput. Math. Organ. Theory*, vol. 6, no. 4, pp. 381–394, 2000.

[10] M. Grüninger and M. S. Fox, "An activity ontology for enterprise modelling," in *Submitted to: Workshop on Enabling Technologies - Infrastructures for Collaborative Enterprises*, West Virginia University, 1994.

[11] M. S. Fox and M. Grüninger, "Enterprise modeling," *AI Magazine*, vol. 19, no. 3, pp. 109–121, 1998.

[12] N. C. Raji Ghawi, "Database-to-ontology mapping generation for semantic interoperability," in *Third International Workshop on Database Interoperability*, 2007.

[13] J. Barrasa, s. Corcho, and A. Gómez-pérez, "R2o, an extensible and semantically based database-to-ontology mapping language," in *in In Proceedings of the 2nd Workshop on Semantic Web and Databases(SWDB2004*. Springer, 2004, pp. 1069–1070.

[14] D. Tham, M. S. Fox, and M. Grüninger, "A cost ontology for enterprise modelling," in *Proceedings of third Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, Morgantown,WV*, 1994, pp. 111–117.

[15] Oracle, "Oracle technology network for java developers," http://www.oracle.com/technetwork/java/index.html, retrieved: August, 2011.

[16] "Mysql:: The world's most popular open source database," http://www.mysql.com, retrieved: August, 2011.

[17] S. T. I. S. I. ESSI WSML working group, "Web service modeling language wsml," http://www.wsmo.org/wsml/wsml-syntax#l, retrieved: August, 2011.

[18] D. Grigori, F. Casati, M. Castellanos, U. Dayal, M. Sayal, and M.-C. Shan, "Business process intelligence," *Comput. Ind.*, vol. 53, pp. 321–343, April 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=982250.982256

[19] B. F. van Dongen, A. K. A. de Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters, and W. M. P. van der Aalst, "The proM framework: A new era in process mining tool support." in *Lecture Notes in Computer Science: Applications and Theory of Petri Nets 2005: 26th International Conference, ICATPN 2005, Miami, USA, June 20-25, 2005. / Gianfranco Ciardo, Philippe Darondeau (Eds.)*, vol. 3536. Springer Verlag, Jun. 2005, pp. 444–454.

[20] E. T. U. Process Mining Group, "Prom - the leading process mining toolkit," http://prom.win.tue.nl/tools/prom/, 2009, retrieved: August, 2011.

[21] R. S. Mans, H. Schonenberg, M. Song, W. M. P. van der Aalst, and P. J. M. Bakker, "Application of process mining in healthcare - a case study in a dutch hospital," in *BIOSTEC (Selected Papers)*, 2008, pp. 425–438.

[22] SUPER, "Super integrated project," http://www.ip-super.org/content/view/71/73/, retrieved: August, 2011.

[23] A. K. Alves De Medeiros, W. Van Der Aalst, and C. Pedrinaci, "Semantic process mining tools: Core building blocks," in *16th European Conference on Information Systems*, W. Golden, T. Acton, K. Conboy, H. van der Heijden, and V. K. Tuunainen, Eds., Galway, Ireland, 2008, pp. 1953–1964.

 ISBN: 978-1-61208-175-5