Towards Semantic Integration of XML-based Business Process Models

Jan Mendling[1], Cristian Pérez de Laborda[2], and Uwe Zdun[1]

[1] Dept. of Information Systems and New Media, WU Vienna, Austria
jan.mendling@wu-wien.ac.at, uwe.zdun@wu-wien.ac.at
[2] Dept. of Computer Science, Databases and Information Systems,
University of Düsseldorf, Germany
perezdel@cs.uni-duesseldorf.de

Abstract This paper discusses the applicability of schema integration methodology for the integration of XML Schemas for business process modelling. This methodology builds upon the assumption that the integrated schema has to support queries and updates on all underlying local schemas. The heterogeneous schemas of Business Process Execution Language for Web Services (BPEL4WS) and Petri Net Markup Language (PNML) are used as an example to illustrate potential problems of integrating semantically related models. We identify schema integration and domain modelling as two areas of research that need to be balanced in order to specify integrated schemas.

1 Introduction

Models play an important role for business process management. They serve both as documentation of complex procedures and interactions and as a blueprint for information systems. Recently, it has become common practice in the area of business process modelling (BPM) to express metamodels using *XML Schema* [1,2] in order to facilitate XML-based interchange of models. Although there have been standardization efforts in the area of BPM for more than ten years, the lack of a commonly accepted schema is still a major hindrance for business process management [3]. Competing standardization bodies have proposed numerous specifications and competing XML Schemas that capture only parts of the business process life cycle (see e.g. [4]). There is a need for an integration methodology helping to merge the heterogeneous proposals into a reference model for BPM that is likely to be accepted in the industry.

In the database community there has been research into integration of heterogenous schemas for almost 30 years. In general, schema integration methodology is also suitable for integrating XML Schemas in the area of BPM. This paper aims to identify the peculiarities of semantic integration of XML-encoded models by analyzing the case of BPM. The rest of the paper is structured as follows. Section 2 gives a brief overview of schema integration research and technologies. Section 3 discusses problems of integrating XML Schemas for BPM, while Section 4 identifies two areas of research related to these problems. Section 5 gives some concluding remarks and a prospect on future research.

2 Schema Integration

Schema integration refers to the construction of a global schema from a set of local schemas. In general, the local schemas are heterogeneous, i.e. semantically related concepts are captured by different local schemas in a different way, e.g. using different names or different structure (cf. e.g. [5]). The global schema is expected to be *complete* in capturing all concepts of the local schemas, *minimal* by including semantically related concepts only once, and still *understandable* [6]. Discovering semantic relationships like equivalence, subsumption, intersection, disjointness, and incompatibility between concepts of local schemas plays a central role for schema integration. Basically three approaches can be distinguished in this context: manual, semi-automatic, and automatic schema integration.

Manual schema integration builds on semi-formal instructions to schema designers. A survey reported in [6] uses the four steps of preintegration, comparison, conformation, and merging and restructuring to compare different integration methodologies. Manual integration leverages the knowledge of a domain expert. *Semi-automatic schema integration* relies on assertions to state semantic relationships between concepts of different schemas. These assertions represent integration rules that are used by a so-called integrator to generate the global schema [7]. Although this approach is less time-comsuming, it also depends on a domain expert to state assertions. *Automatic schema integration* uses techniques from information retrieval and artificial intelligence to detect semantic relationships. An overview available in [8] describes different research prototypes that mainly discover equivalence relationships. Recently, an approach has been presented to automatically discover equivalence, subsumption, intersection, disjointness, and incompatibility [9]. In general, a certain trade-off between human effort and quality of the integrated schema can be expected. In practice, a fully automated approach still requires validation by the domain expert.

Completeness implies that the global schema has to support queries and updates on all underlying databases [7]. Semantics can get lost on the way towards the global schema, e.g. when the attributes *firstname* and *lastname* are merged into a *name* attribute. When integrating models this might not be desired.

3 Integration of XML Schemas for BPM

The example of Figure 1 illustrates a major problem when integrating heterogeneous BPM schemas. There are different formalisms available to represent control flow. These formalisms are quite different from a syntactical perspective, although they represent similar semantics. Figure 1 gives an example of an AND split where one flow of control branches into two parallel threads of execution. The first grey column provides the XML code for this process in Petri Net Markup Language (PNML) [10], which uses a graph-based representation with places and transitions as special nodes linked via control flow arcs. The second grey column follows the representation of the Business Process Execution Language for Web Services (BPEL4WS) [11] to represent this process semantics

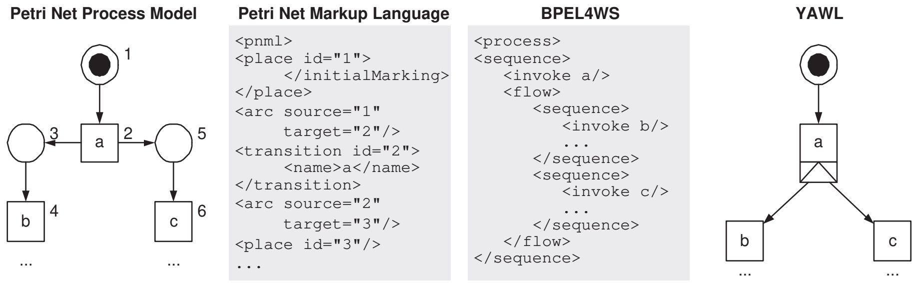

Figure 1. A sample process model, its PNML, BPEL4WS, and YAWL representation

in a block-oriented algebraic syntax. The `<flow>` structured activity is used to specify parallel execution of all its child elements. The order of syntax elements in a BPEL4WS process is crucial, but the order is irrelevant in PNML. Analyzing XML Schemas for BPM reveals three integration challenges: canonical representation, complex semantic relationships, and guidelines for XML design.

Canonical Representation: BPM specifications use different XML schema and ontology languages including XML Schema, RELAX NG, or OWL. Accordingly, a canonical representation is needed before formal integration methods can be applied. The AutoMed system uses e.g. a hypergraph model as such a format [12]. In [13] integration of RDF and XML sources is reported. However, both of these approaches treat XML data as a tree. This is not correct when key references are defined in the schema. Therefore, a canonical representation is required, capable to describe all possible types of relationship among schema elements. The loss of information during the integration process can be minimized with such an expressive BPM representation. Accordingly, a global schema for BPM can best be modelled using a semantically rich language like e.g. OWL.

Complex Semantic Relationships: Consider the different formalisms of PNML and BPEL4WS to express control flows. Although the two code fragments given in Figure 1 are semantically equivalent, their syntax elements cannot be directly related in terms of equivalence, subsumption, etc. The semantic relationship would rather be that the flow element in BPEL4WS is mappable to PNML. Yet, Petri Nets do not offer all control flow constructs used for BPM. The YAWL workflow language [14] has been developed to express the whole set of control flow patterns reported in [15]. The diagram on the right hand side of Figure 1 shows an AND split in YAWL syntax. An integrated BPM schema would have to express control flow in terms of YAWL to grant that various BPM schemas can be mapped to it.

Guidelines for XML Design: Most XML representations of BPM do not follow any naming or structural design guideline as described e.g. in [16]. As a consequence, semantically equivalent XML representations may have drastic differences within their syntax. While most of the naming conflicts, i.e. different names for the same property can easily be resolved using lexical databases (e.g.

WordNet), more complex techniques are needed for structural conflicts. For a classification of schema and data conflicts refer to [5].

4 Towards Building Integrated BPM Schemas

The case of BPEL4WS and PNML illustrates that different control flow representations can hardly be handled by schema integration technology alone. It is desirable to map both control flow representations to a more general representation like graph-based YAWL. Consequently, the integrated schema might no longer support arbitrary updates on the local schemas. We identify two areas of research related to this problem that need to be balanced for building integrated schemas: bottom-up schema integration and top-down domain modelling.

Classical schema integration can be employed to build a global schema following a *bottom-up* strategy. Any of manual, semi-automatic, or automatic schema integration methodology seems applicable here. It would be helpful to first transform the different local schemas into a semantically rich representation to minimize loss of information during the integration process. Yet, the global schema built with this methodology might still include complex semantic relationships like different control flow representation in BPEL4WS and PNML.

A domain expert is needed to discover those complex semantic relationships that cannot be expressed in terms of set operators. Following a *top-down* strategy, the domain expert has to innovate more general concepts that capture the semantics of both representations used in the local schemas. In our BPM example, this involves mapping the different control flow representations to the general graph-based representation of YAWL. This task yields an integrated schema that is also reflects the modelling competence of the domain expert.

The balancing of bottom-up schema integration and top-down domain modelling has the potential to provide for a more comprehensible specification of integrated schemas. Instead of choosing between a top-down and a bottom-up strategy (see e.g. [17]) an integrated approach is needed that reflects the advantages of both methodologies. Such an integrated approach might prove especially valuable for the standardization of heterogenous BPM schemas.

5 Conclusion and Future Work

In this paper, we have discussed the applicability of schema integration methodology for the integration of BPM schemas. The case of BPEL4WS and PNML illustrates that schema integration can solve only a subset of integration problems in the area of BPM. Specific integration problems are caused by differences in representation of control flow semantics. We identify a bottom-up schema integration and top-down domain modelling as two approaches that need to be balanced in order to build integrated schemas. Future research will be dedicated to the definition of an integrated schema for BPM building on YAWL control flow semantics. Furthermore, we will use this domain as a case to define in detail how schema integration and domain modelling can best be balanced.

References

1. Beech, D., Lawrence, S., Moloney, M., Mendelsohn, N., Thompson, H.S.: XML Schema Part 1: Structures. W3C Recommendation 02 May, World Wide Web Consortium (2001)
2. Biron, P.V., Malhorta, A.: XML Schema Part 2: Datatypes. W3C Recommendation 02 May, World Wide Web Consortium (2001)
3. Delphi Group: BPM 2003 – Market Milestone Report. White Paper (2003)
4. Mendling, J., Nüttgens, M., Neumann, G.: A Comparison of XML Interchange Formats for Business Process Modelling. In: Proceedings of EMISA 2004 - Information Systems in E-Business and E-Government. LNI (2004)
5. Kim, W., Seo, J.: Classifying schematic and data heterogeneity in multidatabase systems. IEEE Computer **24** (1991) 12–18
6. Batini, C., Lenzerini, M., Navathe, S.B.: A Comparative Analysis of Methodologies for Database Schema Integration. ACM Computing Surveys **18** (1986) 323–364
7. Spaccapietra, S., Parent, C., Dupont, Y.: Model Independent Assertions for Integration of Heterogeneous Schemas. VLDB Journal **1** (1992) 81–126
8. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB Journal **10** (2001) 334–350
9. Rizopoulos, N.: Automatic Discovery of Semantic Relationships between Schema Elements. In: Proceedings of the 6th International Conference on Enterprise Information Systems (ICEIS 2004). Volume Volume I - Databases and Information Systems Integration. (2004) 3–8
10. Billington, J., Christensen, S., van Hee, K.E., Kindler, E., Kummer, O., Petrucci, L., Post, R., Stehno, C., Weber, M.: The Petri Net Markup Language: Concepts, Technology, and Tools. In W. M. P. van der Aalst and E. Best, ed.: Applications and Theory of Petri Nets 2003, 24th International Conference, ICATPN 2003, Eindhoven, The Netherlands. Volume 2679 of Lecture Notes in Computer Science. (2003) 483–505
11. Andrews, T., Curbera, F., Dholakia, H., Goland, Y., Klein, J., Leymann, F., Liu, K., Roller, D., Smith, D., Thatte, S., Trickovic, I., Weerawarana, S.: Business Process Execution Language for Web Services, Version 1.1. Specification, BEA Systems, IBM Corp., Microsoft Corp., SAP AG, Siebel Systems (2003)
12. Zamboulis, L., Poulovassilis, A.: Using AutoMed for XML Data Transformation and Integration. In Bellahsne, Z., McBrien, P., eds.: DIWeb Workshop - CAiSE'04 Workshop Proceedings. (2004) 58–69
13. Xiao, H., Cruz, I.F., Hsu, F.: Semantic Mappings for the Integration of XML and RDF Sources. In: Proceedings of the Workshop on Information Integration on the Web (IIWeb 2004). (2004)
14. van der Aalst, W.M.P., ter Hofstede, A.H.M.: YAWL: Yet Another Workflow Language (Revised Version). QUT Technical report, FIT-TR-2003-04, Queensland University of Technology (2003)
15. van der Aalst, W.M.P., ter Hofstede, A.H.M., Kiepuszewski, B., Barros, A.P.: Workflow Patterns. Distributed and Parallel Databases **14** (2003) 5–51
16. Stuhec, G., Heilig, P., Pemberton, M.: XML Naming and Design Rules. Draft 1.0, 3 August, UN/CEFACT (2004)
17. Hasselbring, W.: The role of standards for interoperating information systems. In Jakobs, K., ed.: Information Technology Standards and Standardization: A Global Perspective. Idea Group Publishing, Hershey, PA (2000) 116–130