# Process-Based Integration of Heterogeneous Information Sources

Sudhir Agarwal and Peter Haase

Institute of Applied Informatics and Formal Description Methods (AIFB),
University of Karlsruhe, Germany.
{agarwal,haase}@aifb.uni-karlsruhe.de

**Abstract.** The number of information sources in the Web is growing day by day with a tremendous speed. At the same time, users tend to manage more and more tasks electronically, thereby using publicly available information sources. Since the Web is an open, distributed and dynamic environment, in which information sources are offered and controlled autonomously and independently, the interoperability of available information is still a big challenge. In this paper, we present a process-based approach for integrating various information sources and illustrate our approach by an example. Our approach also allows dynamic and flexible integration of information sources provided by the users.

## 1 Introduction

The ever growing number of information sources in the Web and the ever increasing tendency of users to accomplish more and more tasks electonically give rise to the need of dynamic and flexible methods for knowledge integration.

An information source consists of information and operations that can be performed to access or modify the information. Current approaches for schema integration consider the integration on the data level and can be used for generating integrated view of the schemas [1]. Hence, they are suitable for integrating information but not for information sources and offer solutions for relatively static situations.

To enable dynamic and flexible integration of information sources, the operations of an information source must also be taken into account. In this paper, we propose a semantic Web services based approach for flexible and dynamic integration of heterogeneous information sources. Our approach also allows integration of information sources provided by the users. The basic idea lies in abstract specification of information as well the operations of an information source. We specify the schema of an information source with ontologies and operations (use cases) with a process description language SWPDL, that we introduce in section 2. Once the information sources are described in such a way, the integration of the sources can be specified by composing various relevant processes, again with SWPDL.

### 1.1 Sample Scenario

In our sample scenario, we consider the integration of multiple bibliographic information sources in a semantic portal. These information sources are heterogeneous both with respect to their schema as well as the operations they provide. Suppose, one of the information sources is the DBLP bibliographic database, organized according to the DBLP metadata scheme. The DBLP database can be queried using a simple operation with the query as the input and the query result as the output. Suppose, another data source in the scenario is the ACM Digital Library, which uses its own metadata scheme for organizing its data. Also, the process for querying the ACM Digital Library is more complex: The user has the choice to either logon to the library and use more advanced search capabilities (e.g. on a more extensive dataset), or he may perform query operations without providing account information. When integrating these information sources to provide transparent querying in a semantic portal, we need to provide an integration on both the schema and the process level. We will use this sample scenario throughout the following sections to explain the model of process-based integration of information sources.

## 2 Semantic Web Process Description Language

In this section, we give a short introduction to SWPDL, an ontology for describing processes semantically [2]. An SWPDL process may perform three types of activities: (1) local operation activity, (2) communication activity and (3) Web process activity [3, 4]. The order of execution of various activities is controlled by the control constructs like *sequence*, *parallel* and *choice*.

A *local operation activity* $o \in O$ has a set of input parameters, a set of output parameters and a grounding. A grounding connects an operation with its concrete implementation for example a Java method or another Web service. To describe a communication activity the notion of a *message* is necessary. A *message* is an object that carries a set of objects [5]. With a *communication activity* an actor (sender) sends a message to another actor (receiver). Each communication activity is either of type *send* or of type *receive*, depending on whether the communication activity sends or receives a message from or to an actor. Communication activities also have a grounding, that specifies which communication protocol, e.g. HTTP, SOAP etc. and which message format, e.g. HTML, SOAP-message etc. is to be used for the communication. A *process activity* is an activity to invoke a process from inside a process. The intuitive semantics of the control constructs *sequence*, *parallel* and *choice* is straightforward. For technical details of their execution semantics we refer to [2].

## 3 Integration of Heterogeneous Information Sources

In section 3.1, we show how the schemas of information sources can be specified with ontologies, giving them formal and hence machine understandble semantics.

In section 3.2, we show how the operations offered by an information source can be specified homogeneously with our process decription language SWPDL introduced in section 2. In section 3.3, we show how various information sources can be integrated dynamically and flexibly, once their schemas and operations have been specified formally and abstractly, simply by composing the desired processes, again with SWPDL (cf. figure 1).

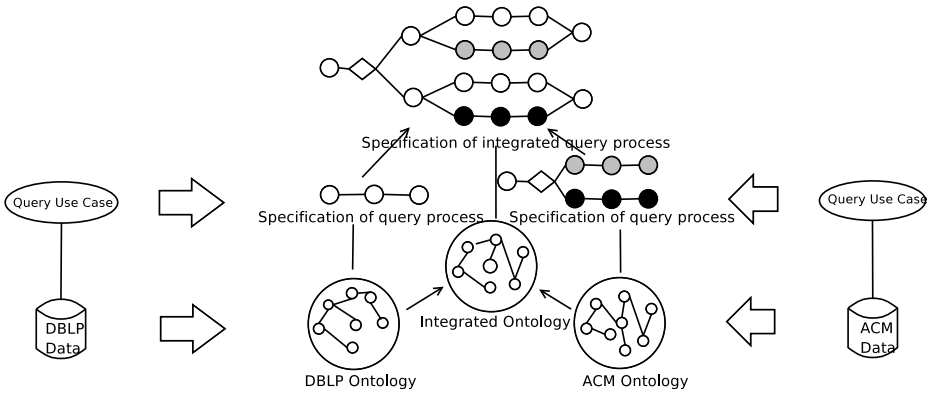


**Fig. 1.** Specification of Information Sources

### 3.1 Specifying Schemas with Ontologies

The integration of the information residing in heterogenous information sources requires to consider the intended meaning of the information to achieve an interoperability on a semantic level.

A database schema typically only describes the structural representation of the information. Ontologies feature richer modelling primitives, such as concept and relation hierarchies as well as axioms. Ontologies thus provide the means to formally specify the conceptualization of a domain, allowing for a semantic interoperability. Using ontologies to specify the meaning of the schemas therefore enables to semantically integrate the underlying information sources [6].

In many cases, a description of the structural schema of the information sources is already given. Several approaches exist to extract the conceptualization inherent in a schema, such as relational schemas and XML schema as an ontology [7]. The relationship between the ontology and the information sources they describe is established via mappings that resemble and enrich the structure of the information source and are used to define and annotate terms and resources from the information source and its schema.

To integrate the schemas of multiple heterogeneous sources, either a global ontology is derived which integrates local schemas, or each information source has its own ontology and the different ontologies are linked directly. Hybrid

solutions combine the two approaches by building ontologies of single information sources using elements from a shared vocabulary. Again, mappings are used to express the relationship between the ontologies. The mappings can either be defined manually or automatically by using lexical relations, top-level groundings and semantic correspondences.

For the sample scenario, the content of the ACM and DBLP information sources would be described using an ontology reflecting their respective metadata schemes. However, the user would be presented with an integrated ontology (obtained by one of the means described above), allowing transparent querying against the various information sources.

### 3.2 Specifying Operations as Processes

In general, an information source has many use cases. For each use case, it offers operations to users. In simple cases, such operations are like remote procedures, but in general, especially in case of Web portals, they can be more complex and need multiple user interactions. Further, an information source performs local operations as partial tasks of a user case, e.g. calculations or database query.

In this step, we model each use case of an information source as a process with our process description language SWPDL introduced in section 2. The process description uses the ontology described in the previous section in order to refer to any objects that take part in the process. The local operations are specified as local operation activities and can be mapped to a concrete implementation, e.g. a method in a Java bean. Any interaction with the user is specified as communication activities. In general, an information source can itself be an integrated information source. that is, it may use operations of other information sources. In such a case the owner of the information may or may not want to disclose this. In the first case, he specifies the operation as process activity and in the second case as a local activity.

For our sample scenario, the querying of the DBLP database can be specified as a sequence of a message, a local operation and another message. The querying of the ACM Digital Library consists of a sequence of a message and and choice. The message is for asking the user about his choice. The choice consists of two sequences, which are similar to the sequence for DBLP.

### 3.3 Integration of Sources by Composing Processes

We assume the situation that there are many information sources that have been specified as described in sections 3.1 and 3.2 and their specifications (ontologies and process descriptions) are publicly available. Process descriptions serve as semantical descriptions (at least execution semantics) of the corresponding operations. In order to have an integrated system spanning over some information sources, an actor identifies for each use case how its functionality can be implemented by using the operations provided by other information sources or locally available [8–10]. The operations of other information sources can be embedded as process activities or local activities in the description of the use case. Since

the basic activities of a process have concrete groundings, such a process specification is directly executable. Mappings between different ontologies can be integrated as local operations to support interoperability at the data level.

For our sample scenario, the process for querying the integrated system consists of a sequence of a message and a choice. As in ACM, the message is for asking the user about his choice. The choice contains for each choice value two sequences (one for querying DBLP and the other for querying ACM) that are executed in parallel.

## 4 Conclusion

In this paper, we presented a novel approach for dynamic and flexible integration of information sources. We showed by an example, how ontologies and semantic Web services can be used to specify data, schema and use cases an information source in a homogeneous way and how processes can be composed to achieve desired integration of heterogeneous information systems.

**Acknowledgement** This work is supported by and Research reported in this paper has been partially financed by the German Ministry for Education and Research (bmb+f) in the SemIPort project and the EU in the IST project SEKT (IST FP6-506826 `http://sekt.semanticweb.org/`).

## References

1. Lenzerini, M.: Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM Press (2002) 233–246
2. Agarwal, S.: Specification of invocable semantic web resources. In: 1st International Conference on Web Services. (2004)
3. Hoare, C.A.R.: Communcating Sequential Processes. Prentice Hall (1985)
4. Milner, R.: A Calculus for Communicating Processes. Volume 92 of Lecture Notes in Computer Science. Springer (1980)
5. Milner, R.: A polyadic pi-calculus: a tutorial. Technical report (1991)
6. Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Huebner, S.: Ontology-based integration of information — a survey of existing approaches. In: IJCAI-01 Workshop: Ontologies and Information Sharing. (2001) 108–117
7. Volz, R., Oberle, D., Staab, S., Studer, R.: Ontolift prototype. WonderWeb Deliverable D11 (2003) http://wonderweb.semanticweb.org.
8. Agarwal, S., Handschuh, S., Staab, S.: Surfing the service web. In: ISWC2003: 2nd International Semantic Web Conference. Lecture Notes in Computer Science, Springer (2003)
9. Agarwal, S., Handschuh, S., Staab, S.: Annotation, composition and invocation of semantic web services. Journal of Web Semantics (To appear) (2004)
10. Sycara, K., Paolucci, M., Ankolekar, A., Srinivasan, N.: Automated discovery, interaction and composition of semantic web services. Journal of Web Semantics **1** (2003) 27–46