Heterogeneous Information Model Unification as a Pre-requisite to Resource Schema Mapping

L.A. Kalinichenko[1], S.A. Stupnikov[1]

Abstract An innovative technique for formation of collaborative consortia of enterprises in virtual organizations (VO) is considered. The technique is based on semantic integration of relevant enterprise information systems (EIS) in the VO specification treated as a subject mediator over the EISs involved. The paper explains how the canonical model for the VO specification is synthesized. Main part of the paper is devoted to the presentation of the process of EIS resource information models semi-automatic mapping into the canonical one assisted by the Heterogeneous Information Model Unifier. It is important to note that the Model Unifier is a universal tool that assists in development of mapping of various kinds of information. The process of mapping includes construction of a compiler from a resource model into the canonical one with the help of metacompilation tools. The mediation technique presented is applied in the astronomical Russian Virtual Observatory (RVO).

Introduction

This paper[2] is focused on forward looking contributions as a bridge between research in semantic interoperability and information integration in general and application of the results to the enterprise interoperability (EI) domain. Many organizations around the world work in accordance with the recently defined EI roadmap [1]. According to it, one of the most important next phase enabled by EI is the sharing of knowledge within a Virtual Organization (VO) to the mutual benefit of the VO partners. “Formation of collaborative consortia to exploit product opportunities, and the application of enterprise and VO knowledge in operational and strategic decision making in VOs, leading to enhanced competitiveness and profitability” [1] identifies two primary needs by enterprises in successfully forming and exploiting VOs.

[1] Institute of Informatics Problems of the Russian Academy of Sciences, Moscow, Russia, {leonidk, ssa}@ipi.ac.ru

[2] This research has been done under the support of the RFBR (projects 06-07-89188-a, 08-07-00157-a) and the Program of the Department of Nanotechnologies and Information Technologies of RAS entitled as Basic principles of information technologies and systems (project 1-10).

According to the approach evolved during our research it is assumed that to develop some virtual organization its abstract specification should be defined independently of existing enterprises and their Enterprise Information Systems (EISs). VO is specified in terms of ontology that defines concepts of the VO subject domain, in terms of data structures, services, processes that are characteristic for the VO. Such VO specification constitutes a definition of the respective subject mediator located above EISs of the relevant to VO enterprises. A set of specific facilities supports the mediation middleware [2]. Instead of middleware solutions for technical interoperability (like IBM WebSphere, Microsoft BizTalk, Oracle Fusion), semantic interoperation-oriented middleware based on mediation approach is emphasized. The mediator specification of the VO is assumed to be consolidated by the respective VO community.

Advanced enterprise modeling approaches share the fundamental strategy of integrating (interoperating) at the model level - taking fragments of information within the EISs relevant to VO and placing them in a larger context of VO. What model is to be taken and how a proper context is to be formed and implemented are the basic issues that are discussed in this paper.

Heterogeneous information resources of various kinds supported by the respective EIS work in its own, specific context. Many of such resources are autonomous and evolve with time. Justifiable identification of the resources relevant to VO in each of the EISs involved, reaching semantic integration of them in the context of VO, making VO stable in such rapidly evolving world constitute serious challenges. New innovative technologies for VO development over multiple distributed collections of resources supported by the respective EIS are required.

In this paper we start with a brief introduction into VO infrastructure based on subject mediator approach. One of the fundamental ideas of this infrastructure consists in specifying the mediator applying the VO canonical model as a set of language facilities sufficient for the VO conceptual modeling. Various resources relevant to EISs involved in VO should be semantically integrated in the VO specification. The canonical model plays a role of a unifying model, in which the resource information models can be represented without loss of information. The aim of this part of the introduction is to create a context for considering further an approach for the unification of heterogeneous information models used for specification of various EIS resources. This approach constitutes the main focus of the paper.

Definition of VO as a subject mediator and semantic integration in mediator of information resources belonging to the respective EISs is treated as a problem of *compositional development* of information systems [3]. Registration of EIS resources in mediator is a process of purposeful specification transformation including decomposition of mediator specifications into consistent fragments, search among specifications of ontologically relevant EIS resources, construction of expressions defining resource classes as a composition of the mediator classes. For such specification manipulation a specification calculus has been developed [3]. Important point in this scheme consists in treating resource data types as *refine-*

ments of the respective mediator data types [3] and the type refinement proof applying modeling of the mediator and resource type specifications in the first order logic (the Abstract Machine Notation – AMN [4] is used for that).

A process of registration of heterogeneous information resources in a subject mediator resembles GLAV that combines two approaches - Local As View (LAV) and Global As View (GAV). According to LAV the schemas of EIS resources being registered in VO are considered as materialized views over virtual classes of a mediator. GAV views provide for reconciliation of various conflicts between resource and mediator specifications and provide rules for transformation of a mediator program results from a resource into the mediator representation. Such registration technique provides for stability of EIS application specification during any modifications of specific information resources and of their actual presence as well as for scalability of mediators w.r.t. the number of EIS resources integrated.

Identification of EIS resources relevant to the VO specification (that precedes the registration) is based on three models: metadata model (characterizing resource capabilities represented in external registries), canonical ontological model (providing for definition of VO concepts), and canonical conceptual model (providing for definition of structure and behavior of VO and EIS information resource objects). Reasoning in canonical models is based on the semantics of the canonical model and facilities for proof of the refinement. Reasoning in the metadata model is a heuristic one based on nonfunctional requirements to the resources needed in application (e.g., indexes of quality of data).

The techniques listed are used as a basis for the tool prototype [2, 3] developed for identification and registration of EIS information resources in the VO mediator. The main registration result resembles a GLAV expression defining how a resource class is determined as a composition of the relevant mediator classes. In process of resources evolution a specification of mediator remains stable, only such expressions need to be modified.

General approach for VO problem solving using subject mediators consists in problem formulation in terms of the VO mediator specifications and transformation of this formulation into the set of tasks (queries) over the real EIS information resources registered at the mediator. A method of the mediator programs rewriting in a typed object environment has been developed and implemented applying the inverse rule technique [2]. The method is based on the use of the refinement relationship between mediator data types and resource data types helping to get containment of the rewritten queries in the mediator original queries expressed in the canonical model.

It is important to note that for the design, the mediator and resource specifications should be given uniformly in the canonical model. Therefore a transformation into the canonical model of the EIS resource information models (languages) is required. These transformations are needed to map resource schemas into the canonical model. So creating the transformations of the EIS resource models into the canonical one (resource models unification) is a pre-requisite of resource schema mapping. In the following sections we introduce the heterogeneous model

unification approach. How to develop the VO canonical model and how to automate mapping into it (by specific tools) of various information models (languages) used for EIS resource specifications will be presented in the rest of the paper in more details.

A Method for the Mediator Canonical Model Synthesis

The foundation of the integration and interoperation methods proposed is formed by the concept of a *canonical information model* serving as the common language, ”Esperanto”, for adequate uniform expression of semantics of various EIS information resource models that are to be used in VO. To prove that a definition in one language can be substituted with a definition in another one the formal specification facilities and commutative model mapping methods are provided. Currently the method of information model mapping and canonical models constructions looks as follows. As a formalism of the method the AMN [4] is applied. It allows to get specifications of the mediator and EIS resources in the first order logics and to prove the fact of the *specification refinement*.

The main principle of *canonical model synthesis* consists in its *extensibility* required to reach semantic integration and information interoperability in environments that include various heterogeneous models. A *kernel* of the canonical model is fixed. For each specific information model *M* of the environment an extension of the kernel is defined so that this extension together with the kernel is refined by *M*. Such refining transformation of models should be provably correct. The canonical model for the environment is synthesized as the union of extensions, constructed for various models *M* of the environment. Each EIS resource specification model should refine the canonical model. The refinement of the specification mapping is formally checked. The canonical information model synthesis method that we have developed initially for the structured data models provided a seminal role for synthesis of canonical models for various kinds of resource information models: object, service, ontological and process [5, 6].

This paper is focused on the problem of constructing a tool assisting in development of provably correct mapping of various EIS resource information models into the canonical one. With the help of the tool we get compilers transforming schemas into their canonical representation. Due to that schema mapping can be provided in frame of the common, canonical model. Besides unification of heterogeneous schema representations, development of the canonical model gives an ability to create a unified schema matching approach instead of inventing various approaches for matching schemas having heterogeneous information model semantics.

Automation of heterogeneous information model unification

A prototype of the Heterogeneous Information Model Unifier aimed at partial automation of methods of the canonical models synthesis has been constructed [7]. One of the practical purposes of this work is to support the heterogeneous information resource integration in a specific subject mediator [2]. Due to that, the subject mediator information model (the SYNTHESIS language [8]) has been chosen as the canonical model kernel. The SYNTHESIS language, as a hybrid semistructured and object-oriented information model, includes the following distinguishing features: facilities for definitions of frames, abstract data types, classes and metaclasses, functions and processes, logical formulae facilities applied for description of constraints, queries, pre- and post-conditions of functions, assertions related to processes. For extension of the canonical model kernel, metaclasses, metaframes, parameterized constructions including assertions and generic data types are applied. Comprehensive facilities of the kernel provide for an ability to construct refining mapping of various kinds of information models into the canonical model kernel chosen.

The aim of the Model Unifier is to unify a set of information models for their interoperability in some VO. An EIS resource model *R* is said to be *unified* if it is mapped into the canonical model *C*. To unify it is required to create an extension *E* of the canonical model kernel and a refining mapping *M* of a resource model into the extended canonical one. The refinement of *C* model by *R* means that for any admissible specification *r* represented in *R* its image *M(r)* in *C* under the mapping *M* is refined by the specification *r*. Process of model mapping includes a possibility of proving that arbitrary specification *r* represented in *R* refines its image *M(r)*. Verification of model refinement is realized over a set of resource model specification samples. Hence the following languages and formal methods are required to support the process of model unification:

- formal methods allowing to declare information model syntax and semantic mappings (compilers) of one model to another;
- formal methods supporting verification of the refinement reached by the mapping.

For the formal description of model syntax and compilers the metacompilation languages SDF (Syntax Definition Formalism) and ASF (Algebraic Specification Formalism) are used. For the languages a tool support - Meta-Environment [9] is provided. The AMN language [4] based on the first order predicate logic and set theory is used for model semantics formalization and refinement verification. AMN is supported by technology and tools for proving of refinement (B-technology) [10]. Mapping of the resource model *R* into *C* constructed by an expert with a support of the Model Unifier is divided into the following stages (syntax and semantics of the canonical model kernel are supposed to be defined):

- formalization of the model *R* syntax and semantics;

- definition of *reference schemas* of the model *R* and the canonical information model (if the latter has not yet been defined);
- integration of reference schemas of the model *R* and the canonical model;
- creation of a required extension *E* of the canonical model *C*;
- construction of a compiler of the model *R* into the extended canonical model;
- verification of refinement of the extended canonical model by the model *R*.

The Reference schema of an information model is an abstract description containing concepts related to constructs of the model and significant associations among these concepts. Both concepts and associations may be annotated by verbal definitions (looking like entries in an explanatory dictionary).

The Unifier is considered as a constituent part of the subject mediator middleware [2]. The Unifier consists of the following main components:

- Meta-Environment (used for the formal description of information model languages and generation of compilers) [9];
- Atelier B (supporting formal language and tools for proving of specification refinement) [10];
- metainformation repository;
- model manager.

Meta-Environment and Atelier B are third-party tools. Metainformation repository is an object-relational database and is used for the implementation of the *model registry* and as a specification storage. Model registry contains *registration cards* of models, canonical model extensions, specification samples. All the information produced during mapping of models (including information produced during interaction of expert with Meta-Environment and Atelier B) is stored in the registration cards. Model manager provides a graphical interface allowing an expert to manipulate information model cards; to call specific components for model integration, compiler template generation, etc.

Recently the Model Unifier was applied for the unification of a subset of OWL (Web Ontology Language). Mapping of OWL into the canonical model has been constructed and verified [7].

Related work

Note that no works focusing on a semantic enterprise interoperability in VO based on subject mediation are known. In this section we concentrate on heterogeneous information model unification works. Related works are concentrated mostly on mapping of a database schema expressed in one data model into respective database schema expressed in another model [11]. Some approaches rely on properties of specific data models (e.g. CLIO system [12] allows to generate mappings between relational and XML Schemas). Other approaches intend to be generic with

respect to data models. The main idea of the ModelGen approach [13] is using a metamodel – a set of metaconstructions (independent of any data model) applied for abstraction of specific data models. Supermodel is a data model containing constructions that correspond to all metaconstructions known to a system. Mapping includes the following steps:

1. Translation of source schema into supermodel.
2. Translation of the result into target schema realized in the supermodel.
3. Translation of the target schema into the target model.

The first and the third steps are assumed to be labor-intensive and straightforward because every model (source or target) is subsumed by supermodel. Transformation coded by the authors relates only to the second step.

Comparison of an approach proposed in our work (SYNTHESIS) with existing approaches can be done by brief analysis of differences with the ModelGen.

Metaconstructions of the canonical model kernel (SYNTHESIS language [8]) are not restricted by structured data models as in ModelGen. Supermodel constructs are easily included into constructs of the canonical model.

Extending of the canonical model is a semantic process of introduction into model of the new parameterized generic data types, metaframes annotating additional properties of initial constructs, parameterized closed logical formulae patterns expressing data dependencies. A process of extending the ModelGen supermodel is mainly a mechanical introduction of new metaconstructions. In SYNTHESIS the EIS resource information models are considered to be defined by respective languages with their syntax and semantics. ModelGen approach uses only data structure specifications. Detailed analysis of language semantics, integrity constraints, functions are not considered. The SYNTHESIS approach is based on the formal definition of semantics of complete schemas in source and target models. It should be clear that the approach proposed is much more general than ModelGen.

Methods developed by the authors differ from existing ones by basing them upon the extensible canonical model constructed in a modular way by systematic extension of the kernel, having strictly defined formal foundations, widely applying of the refinement during construction of the unifying model transformations.

Conclusion

An innovative technique for formation of collaborative consortia of enterprises in VO is considered. The technique is based on semantic integration of relevant EISs in VO specification treated as a subject mediator over the EISs involved. The paper explains how the canonical model for the VO specification is synthesized. Main part of the paper is devoted to the presentation of the process of EIS resource information models semi-automatic mapping applying the Heterogeneous Information Model Unifier that assists in development of a mapping of a specific

EIS resource model into the canonical one [7]. It is important to note that the Model Unifier is a universal tool that assists in development of mapping of various kinds of information models (data models, service models, process models, ontological models). The process of mapping of EIS resource models into the VO canonical model includes construction of a compiler from a resource model into the canonical one with the help of metacompilation tools. The mediation technique presented is applied in the astronomical Russian Virtual Observatory (RVO) [14].

References

1. Li, M-S., Cabral, R., Doumeingts, G., Popplewell, K. (2006) Enterprise Interoperability Research Roadmap. Information Society Technologies. ftp://ftp.cordis.europa.eu/pub/ist/docs/directorate_d/ebusiness/ei-roadmap-final_en.pdf
2. Kalinichenko, L.A., Briukhov, D.O., Martynov, D.O., Skvortsov, N.A., Stupnikov, S.A. (2007) Mediation Framework for Enterprise Information System Infrastructures. Proc. of the ICEIS 2007, V. DISI: 246-251. Funchal.
3. Briukhov, D.O., Kalinichenko, L.A., Skvortsov, N.A. (2001) Information sources registration at a subject mediator as compositional development. Proc. of the ADBIS'01, LNCS 2151: 70-83. Berlin Heidelberg, Springer-Verlag.
4. Abrial, J.-R. (1996) The B-Book: Assigning Programs to Meanings. Cambridge, Cambridge University Press.
5. Kalinichenko, L.A. (2004) Canonical model development techniques aimed at semantic interoperability in the heterogeneous world of information modelling. Proc. of the CAiSE INTEROP Workshop: 101–116. Riga, Riga Technical University.
6. Kalinichenko, L.A., Stupnikov, S.A., Zemtsov, N.A. (2005) Extensible Canonical Process Model Synthesis Applying Formal Interpretation. Proc. of the ADBIS'05, LNCS 3631: 183-198. Berlin Heidelberg, Springer-Verlag.
7. Kalinichenko, L., Stupnikov, S. (2008) Constructing of Mappings of Heterogeneous Information Models into the Canonical Models of Integrated Information Systems. Proc. of the ADBIS 2008: 106-122. Pori, Tampere University of Technology.
8. Kalinichenko, L.A., Stupnikov, S.A., Martynov, D.O. (2007) SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAS.
9. Van den Brand M.G.J. et al. (2001) The ASF+SDF meta-environment: a component-based language development environment. Proc. of the Compiler Construction 2001, LNCS 2027: 365-370. Berlin Heidelberg, Springer-Verlag.
10. Atelier B. http://www.atelierb.eu/index_en.html
11. Atzeni P. (2007) Schema and data translation: A personal perspective. Proc. of the ADBIS 2007, LNCS 4690: 14-27. Berlin Heidelberg, Springer-Verlag.
12. Haas, L., et al. (2005) Clio Grows Up: From Research Prototype to Industrial Tool. Proc. of the ACM SIGMOD: 805-810. ACM.
13. Atzeni, P., Cappellari, P., Bernstein, P. (2005) ModelGen: Model Independent Schema Translation. Proc. of the ICDE 2005: 1111-1112. IEEE Computer Society.
14. Briukhov, D., Kalinichenko, L., et al. (2008) Application driven mediation middleware of the Russian virtual observatory for scientific problem solving over multiple heterogeneous distributed information resources. Scientific Information for Society – from Today to the Future. Proc. of the 21st CODATA Conference. - To be published.