MEDIATION FRAMEWORK FOR ENTERPRISE INFORMATION SYSTEM INFRASTRUCTURES: APPLICATION-DRIVEN APPROACH

Leonid Kalinichenko, Dmitry Briukhov, Dmitry Martynov, Nikolay Skvortsov, Sergey Stupnikov
Institute of Informatics Problems, Russian Academy of Science, Vavilov Street,44-2, Moscow, Russian Fed.
leonidk@synth.ipi.ac.ru, brd@ipi.ac.ru, domartynov@gmail.com, nskv@ipi.ac.ru, ssa@ipi.ac.ru

Keywords: mediator, canonical models, application-driven approach, ontology, refinement, information sources, registration, query rewriting, semantic identification and integration of sources.

Abstract: This position paper provides a short summary of results obtained so far on an application-driven approach for mediation-based EIS development. This approach has significant advantages over the conventional, information source driven approach. Basic methods for the application-driven approach are discussed including synthesis methods of canonical information models, unifying languages of various kinds of heterogeneous information sources in one extensible model, methods of identification of sources relevant to an application and their registration at the mediator applying GLAV techniques as well as ontological contexts reconciliation methods. Methodology of EIS application development according to the approach is briefly discussed emphasizing importance of a mediator consolidation phase by the respective community, application problem formulations in canonical model and their rewriting into the requests to the registered information sources. The technique presented is planned to be used in various EIS and information systems.

1 INTRODUCTION

One of the widely spread views of Enterprise Information Systems (EIS) is considering them as infrastructures provided for collecting information and services from across the entire enterprise thus giving all its workers complete and transparent access to information. Diverse forms of enterprise inter- and intra-organizational models were developed (such as, virtual corporation, extended enterprise). A basic issue for modeling of such distributed business activities, their collaboration is that there are many types of elements to be modeled in an enterprise, and many perspectives and contexts in which those definitions would be "viewed." Often such elements can be implemented as legacy applications. Advanced enterprise modeling approaches share the fundamental strategy of integrating at the model level - taking fragments of information within the enterprise and placing them in a larger context. 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.

The following concerns will be touched during these discussions. Various applications are to be supported by EIS, each of them working in its own, specific context (probably, overlapping with contexts of other applications) that should be semantically supported. Heterogeneous information sources of various kinds (data sources, service sources, process sources, ontological sources) relevant to EIS are to be used in a specific context of an application. Many of such sources are autonomous and evolve with time. A set of sources relevant to a specific EIS may be changed quite rapidly. The technologies applied for relevant sources are also rapidly evolving. Therefore justifiable identification of relevant to EIS sources, reaching semantic integration of various kinds of them in contexts of appropriate applications, making EIS stable in the rapidly evolving world constitute serious challenges. New methods and tools for EIS application development over multiple distributed collections of data and services are required.

We emphasize two principally different approaches to the problem of integrated representation of multiple information sources for an EIS: 1) moving from sources to problems (an integrated schema of multiple sources is created independently of specific applications) and 2) moving from an application to sources (a description

of an application subject domain (in terms of concepts, data structures, functions, processes) is created, in which sources relevant to the application are mapped). The first approach *driven by information sources* is not scalable with respect to the number of sources, does not make semantic integration of sources in a context of specific application possible, does not lead to justifiable identification of relevant to EIS sources, does not provide for enhancing of EIS stability w.r.t. evolution of the relevant to EIS sources. These deficiencies are inherent to the Global as View (GAV) approach studied elsewhere (Ullman, 1997), (Halevy, 2001). GAV might be used as a basic technique for the information sources driven approach.

The second approach (*application-driven*) assumes creation of subject mediator that supports an interaction between an application and sources on the basis of the application domain definition (description of the mediator). Subject mediator approach (new technology) has obvious advantages over the approach driven by specific information sources. Some of the results of our research and experimental work oriented on application-driven, subject mediation development for various kinds of information systems are summarized in this paper.

2 APPLICATION-DRIVEN APPROACH METHODS

The method for synthesis of the mediator canonical information model as the common language for semantic mapping of various source models into it as well as methods for identification of information sources relevant to a mediator and their registration at the mediator are discussed in this section.

2.1 Synthesis of Canonical Information Models

The present period of IT development is characterized by the process of explosive growth of various information representation models. This development takes place in frame of specific distributed infrastructures (such as OMG architectures (in particular, the model driven architecture (MDA)), SemanticWeb and Web service architectures, digital library architectures as collective memories of information in various subject domains, architectures of the information Grid), as well as in the standards of languages and data models (such as, for example, ODMG, SQL, UML, XML and RDF stacks of data models), process models and workflow models, semantic models (including ontological models and models of metadata), models of digital repositories of data and knowledge in particular domains.

This process is accompanied by another trend — intensive development of based on such models information components and services. This growth causes the accelerating need for integration in various applications of components and services represented in heterogeneous models, as well as their reuse and composition implementing new information systems. The more variety of applied models we meet in various components and services, the more complex become problems of their integration and composition. Research and development of adequate methods for manipulation of various information models are required.

The basis of these methods is constituted by the concept of a *canonical information model* serving as the common language, ”Esperanto”, for adequate uniform expression of semantics of various information models surrounding us. To prove that a definition in one language can be substituted with a definition in another one, formal specification facilities and commutative model mapping methods are provided. Initially ideas of mapping *structured data models* and canonical model construction for them were developed. The basic definitions of equivalence of database states, database schemas and data models were introduced to preserve operations and information while constructing of mappings of various structured data models into the canonical one (Kalinichenko, 1990). According to this approach, each data model was defined by syntax and semantics of two languages – data definition language (DDL) and data manipulation language (DML). The main principle of mapping of an arbitrary source data model into the target one (the canonical model) constituted the principle of *commutative data model mapping*. According to it, preserving of operations and information of a source data model while mapping it into the canonical one could be reached under the condition, that the diagram of DDL (schemas) mapping and the diagram of DML (operators) mapping are commutative. At that time in the process of data model mappings construction the denotational semantics were used as a formalism (metamodel), allowing to prove a commutativity of the diagrams mentioned (Kalinichenko, 1990).

Later, for the *object data models*, the method of data model mapping and canonical models constructions was modified as follows. As a formalism (metamodel) of the method the Abstract Machine Notation (AMN) (Abrial, 1996) was used instead of the denotational semantics. It allowed to define the model-theoretic specifications in the first

order logics and to prove the fact of *specification refinement*. Instead of equivalence of respective specifications, it became possible to reason on their refinement (Kalinichenko, 1997). It is said that specification *A refines* specification *D*, if it is possible to use *A* instead of *D* so that the user of *D* does not notice this substitution. Existence of B-technology for AMN (Abrial, 1992) provides for conducting proofs of commutativity of model mappings semi-automatically.

The main principle of *canonical model synthesis* is that its *extensibility* is required for semantic integration and information interoperability in heterogeneous environment, including various 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 models *M* of the environment. The source schema refines the canonical model schema. The refinement of the schema mapping is formally checked. Another existing methods (e.g., used in Clio project (Haas, 2005)) supporting generation of mappings (queries) between source and target schemas are applicable to limited class of structured data models and do not allow to create an application specifications providing for registration of various kinds of relevant sources (including objects, services, processes). On the contrary, our canonical data model synthesis method provides a seminal role for synthesis of canonical models for various kinds of source information models including process models (Kalinichenko, 2005), service models (Stupnikov, 2006), ontological models (Kalinichenko, 2002).

Canonical model synthesis method we applied recently to the *process models*, required for describing application activities of enterprises (Kalinichenko, 2005). While mapping processes for synthesis of their canonical model, it is required to preserve the semantics of concurrency. A possibility of interpretation of concurrent process events in logics, and specifically, in the AMN has been discovered recently. Algorithms of process specifications mappings into AMN were constructed (Butler, 2000) (Stupnikov, 2002). This approach allows to construct provable refinements of process specifications, applying the B-technology. Simultaneously classification of the diversity of workflow models by means of workflow patterns has been obtained (Aalst, 2003). Due to these findings, the possibility of creation of a canonical process model kernel and construction of its extensions, refined by various workflow patterns, became possible. Finally we showed how to synthesize the canonical process model (Kalinichenko, 2005).

Similarly, to work with ontologies of different sources a *unified representation of ontologies* is required. Again, it is provided by a *canonical ontological model* that includes a kernel and its extensions for every specific ontological model used in the information sources. Extensions are developed so that the ontological model of a source should serve as a refinement of an extended canonical ontological model of the mediator.

After analysis of various ontological models the kernel of canonical ontological model has been identified as a subset of the object model with logical capabilities (Kalinichenko, 1995) (Kalinichenko, 2002). This kernel has been defined as a subset of the general mediator canonical model kernel mentioned above. This subset includes:

- concepts defined as abstract data types (ADT);
- type invariants expressing concept constraints;
- generic ontological metaclasses, instances of which are ADTs defining concepts;
- verbal definitions of concepts applying metaframes annotating types expressing concepts.

One of the popular ontological models, OWL DL has been mapped into the ontological canonical model and respective definition of the canonical model kernel extension has been defined (Kalinichenko, 2002). Correctness proof of such extensions is provided by AMN tools to justify refinement of the extension by the source ontological model.

During practical synthesis of canonical information models it is advisable to produce reversible language mappings that will be suitable for the process of heterogeneous sources registration in the mediator as well as for the run time support when wrappers (adapters) should interpret operations of the canonical level in terms of the source model language. Taking into account the labor intensive character of development of the required mappings (compilers) it is reasonable to apply specific *Meta Environment* facilities providing for declarative specification of compilers and generating them according to such specifications (Brand, 2002). It is such meta approach that is used in frame of our project.

Using of extensible canonical models with formal semantics and possibilities of proof of the fact of their refinement by source models provides for systematic consideration of EIS application development issues in the environment of multiple heterogeneous distributed information sources and

for development of respective basic methods and facilities. Some of them are considered further.

2.2 Identification of Relevant Information Sources and their Registration in Mediator

Definition of a subject mediator and registration of information sources in mediator is based on our experience of *compositional development* of information systems (Briukhov, 2001). Registration of sources is a process of purposeful specification transformation including decomposition of mediator specifications into consistent fragments, search among specifications of relevant sources of data types treating as candidates for refining by them of the mediator specification types, construction of expressions defining source classes as composition of the mediator classes. For such manipulation a specification composition calculus has been developed (Kalinichenko, 1999). A principle of type specification decomposition into a set of specification *reducts* serving as the basic units of reuse and composition has been declared. An operation of identification of the *most common reduct* of source type and mediator type specifications has been introduced. Type lattice and *type algebra* have also been defined. Important point in this scheme consists in implementation of the type refinement proof applying logical model of the mediator and source type specifications in AMN.

Our work on compositional calculus emphasizes complete type specifications and expressiveness sacrificing tractability in complex cases. Such decision is motivated by orientation of the modeling facilities on type refinement and composition. The benefits we get include rigorous way of identification of common fragments of source and mediator type specifications leading to justifiable mapping of source type into mediator type. Investigation oriented on simpler data types considered functions as components modeled with relations and refinement ordering of functions (Mili, 1997). But this is too restrictive for our purposes.

A process of registration of heterogeneous information sources in a subject mediator is based on GLAV that combines two approaches - Local As View (LAV) and Global As View (GAV). According to LAV the schemas of sources being registered are considered as materialized views over virtual classes of a mediator. GAV views provide for reconciliation of various conflicts between source and mediator specifications and provide rules for transformation of a query results from source into mediator representation. Such registration technique provides for stability of EIS application specification during any modifications of specific information sources and of their actual presence (removing, addition new ones, etc.) as well as for scalability of mediators w.r.t. the number of sources registered in them.

Identification of sources relevant to a mediator (that precedes the registration) is based on three models: metadata model, characterizing source capabilities represented in external registries, canonical ontological model, providing for definition of application domain concepts, and canonical model providing for definition of structure and behavior of application and information source objects. Reasoning in canonical models is based on the semantics of the canonical model and facilities for proof of refinement. Reasoning in the metadata model is a heuristic one based on nonfunctional requirements to the sources needed in application (e.g., indexes of data quality in sources). For the design, the mediator and sources specifications are given uniformly in canonical model, though in process of design a transformation into such model from another specification language (e.g., from UML) might be required.

Complicated problem of registration consists in reconciliation of contexts of an EIS application and specific sources. Ontological definitions annotate elements of application specification (a mediator) and of source specifications given in a form of types, classes, processes. A similarity of concepts is established in two steps: first, by means of verbal ontologies, and then by establishing refinement relationship of concepts as abstract data types. Thus the conformity of the concepts is verified. Depending on complexity of specifications such verification can be provided automatically or interactively. In the simple cases the verification can be reduced to justifying of concept subsumption in description logics applying practically such systems as FaCT or Pellet.

The techniques listed are used as a basis for the tool prototype for identification and registration of information sources in mediator. In this process ontological specifications are used for identification of mediator classes semantically relevant to a source class. The maximal subset of a mediator class specification semantically relevant to a source class is identified as the most common fragment (reduct) of specifications of respective instance types given for these classes (Kalinichenko,1999). Concretizing types reconciling the conflicts (of values, structures, behaviors) are defined so that an instance type of the mediator class would be refined by an instance type of the source class. The main registration result is a *GLAV expression* defining how a source class is determined as a composition of the mediator classes. In process of sources evolution a specification of

mediator remains stable, only such expressions need to be modified. The tool supporting techniques of registration of sources in mediator includes:

- facilities identifying relevant information sources by metadata;
- facilities for reaching consensus of ontological contexts of information sources being registered and of the mediator;
- facilities for automation of the heterogeneous information sources registration in the mediator based on the GLAV approach;
- metainformation repository storing specifications of mediators, information sources and the results of registration.

It is important to note that in the infrastructure considered the EIS *application-driven mediators* are treated as information sources in their turn. They are provided with meta-definitions and become elements of registries making possible to construct applications as composition of other applications developed also as mediators.

3 APPLICATION DEVELOPMENT METHODOLOGY

3.1 Application-driven Mediator Specification

Application specification for EIS includes definition of terminology and concepts of the subject domain of the application. They are expressed by the respective dictionaries (thesauri) and ontological definitions in the canonical model. Application definitions include also specifications of object classes corresponding to the application subject domain, specification of instance types of these classes and of their methods defining behavior of the objects, specification of processes characteristic for the application. On the early stages of design the specifications mentioned can be expressed by means of various specification languages (e.g., by UML). However, it is assumed that finally such specifications (including ontological ones) are mapped into the canonical model specifications having formal semantics. It is worth of remark that application-driven specifications are formulated independently of specific pre-existing information sources. The result of this specification activity fulfilled by a community interested in the specific EIS application constitutes the mediator specification created as the result of reaching a consensus in such community. The mediator specification activity is called the mediator *consolidation* phase.

3.2 Methods and Facilities for Application Problems Formulation and Solution

General approach consists in problem formulation in terms of subject mediator specifications and transformation of this formulation into set of tasks (queries) to the real information sources registered at the mediator. Such transformation in the database theory is known as view based query rewriting (information sources are treated as materialized views over virtual classes of the mediator) (Halevy, 2001). This is complicated problem having various (usually limited) solutions applicable to specific data models. In context of the current work rewriting techniques applying inverse rules are used (Halevy, 2001)(Wang, 2002). The respective algorithms are characterized by conceptual clarity, modularity, capability of producing maximally contained queries in time polynomial w.r.t. the size of query and view specifications. Our contribution consists in a method of the mediator programs rewriting applying the inverse rule technique *in a typed object environment* (Kalinichenko, 2004). The method is based on the use of refinement relationship between mediator data types and source data types helping to get containment of the rewritten queries in the original mediator level queries expressed in canonical model. It is important to note that a combination of LAV and GAV approaches is supported during rewriting queries as GLAV.

4 CONCLUSION

Thus, the basic principles of application-driven EIS development over multiple heterogeneous information sources are the following:

1. independence of application (mediator) specification of the existing information sources;
2. definition of an application mediator as a result of consolidation effort of the respective community;
3. emphasizing semantic canonical definitions for the mediator specification;
4. independence of user interfaces of the sources registered at the mediator: application users should be only conscious of definition of the application domain (definition of mediator);
5. independence of publication of the newly developed information sources of the

mediators (as a result of such publication, actions can be initiated on registration of such sources in respective mediators);

6. three stage identification of information sources relevant to mediator providing 1) source relevance to nonfunctional requirements expressed by metadata, 2) source ontological relevance, 3) source structural and behavioral relevance;
7. semantic integration of relevant heterogeneous information sources in canonical mediator specification;
8. integrated access to the information sources registered at mediator applying the canonical model and query rewriting system;
9. recursive structure of mediators: each mediator can be registered as a new information source. Such mediator capability makes possible integration of different applications that is specifically important for virtual organizations development.

Principles numbered as 1 – 8 define advantages of application-driven approach over information source-driven approach for mediation-based EIS development.

The approach presented is being implemented and applied to systems in e-science (e.g., virtual observatories) and is planned for administrative EIS for the city of Moscow and for life-long learning. The work reported was partially supported by the RFBR grants 06-07-08072 and 06-07-89188.

REFERENCES

van der Aalst W.M.P., et al, 2003. *Workflow Patterns.* Distributed and Parallel Databases, 14(3)

Abrial J.-R. 1992. *B-Technology. Technical overview.* BP International Ltd.

Abrial J.-R., 1996. *The B-Book.* Cambridge University Press.

van den Brand M.G.J., et al, 2002. *Compiling Language Definitions: The ASF+SDF Compiler.* ACM TOPLAS, Vol. 24, No. 4.

Briukhov D. et al, 2001. *Information sources registration at a subject mediator as compositional development. In* Proceedings of ADBIS'01 Conference, Springer, LNCS, vol. 2151.

Butler M., 2000. *csp2B: A Practical Approach to Combining CSP and B.* Formal Aspects of Computing, Vol. 12.

Laura M. Haas, et al Clio, 2005. *Grows Up: From Research Prototype to Industrial Tool.* In Proc. of the ACM SIGMOD Conference, June 14-16, 2005, Baltimore, Maryland, USA.

Halevy A., 2001. *Answering queries using views: a survey.* VLDB Journal, 10(4).

Kalinichenko L., 1990. *Methods and tools for equivalent data model mapping construction.* In Proceedings of the EDBT'90 Conference. Springer, LNCS, vol. 416

Kalinichenko L., 1995. SYNTHESIS: the language for description, design and programming of the heterogeneous interoperable information *resource environment.* IPI RAS, Moscow.

Kalinichenko L., 1997. *Method for Data Models Integration in the Common Paradigm.* In Proc. of the ADBIS'97 Conference, St.Petersburg.

Kalinichenko L., 1999. *Compositional Specification Calculus for Information Systems Development. In* Proceedings ADBIS'99 Conference, Maribor, Slovenia, September, Springer, LNCS, vol.1691.

Kalinichenko L., Skvortsov N., 2002. *Extensible ontological modeling framework for subject mediation.* In Proceedings of the 4-th RCDL Conference, October, Dubna.

Kalinichenko L., Martynov D., Stupnikov S., 2004. *Query rewriting using views in a typed mediator environment* In Proceedings of the ADBIS'04 Conference, Hungary, Budapest, Springer, LNCS, Vol. 3255, September.

Kalinichenko L., Stupnikov S., Zemtsov N., 2005. *Extensible Canonical Process Model Synthesis Applying Formal Interpretation.* In Proceedings of the ADBIS'05 Conference, Estonia, Tallinn, Springer, LNCS, Vol. 3631, September.

Mili R., Mili A., Mittermeir R., 1997.S*toring and retrieving software components: a refinement based systems.* IEEE Transactions on Software Engineering, v. 23, N 7, July 1997

Stupnikov S., Kalinichenko L., DONG Jin Song, 2002. *Applying CSP-like Workflow Process Specifications for their Refinement in AMN by Pre- existing Workflows.* In Proceedings of the ADBIS'2002 Conference. Slovak University of Technology.

Stupnikov S.A., Kalinichenko L.A., Bressan S., 2006. *Interactive discovery and composition of complex Web services.* In Proceedings of the East-European Conference on "Advances in Databases and Information Systems" (ADBIS'06), Springer, 2006, p. 216 – 231

Ullman J.D. 1997. *Information Integration Using Logical Views.* In Proc. of the 6th Int. Conf. on Database Theory (ICDT'97)

Wang J., Maher M., Topor R., 2002. *Rewriting Unions of General Conjunctive Queries Using Views.* In Proc. of the EDBT'02 Conference, Springer, LNCS, 2287, Prague, March.