Intermediator framework protocol for information sources registration at heterogeneous mediators

Dmitry Briukhov, Leonid Kalinichenko, Nikolay Skvortsov, Iliya Tyurin

Institute for Problems of Informatics, Russian Academy of Science

e-mail: {brd,leonidk,scvora,turin}@synth.ipi.ac.ru

Abstract

The intermediator framework is required to support heterogeneous subject mediators interoperability in diverse world of mediation platforms that can be observed in distributed digital libraries and other areas. The intermediator framework based on the "local as view" mediation approach is introduced. The paper[1] focuses on a protocol supporting registration of a mediator information (source) at another mediator. It is proposed to use a subset of OAI protocol to support such registration exchanging metainformation uniformly represented in the canonical model of the intermediator framework.

1 Introduction

Mediator middleware positioned between heterogeneous information sources and information consumers provides modeling facilities and services for conversion of unorganized, nonsystematic population of autonomous information sources kept by different information providers into a well-structured information collection defined by the integrated uniform metainformation. Mediators provide also a uniform query interface to the multiple sources, thereby freeing the user from having to locate the relevant sources, query each one in isolation, and combine manually the information from them. Important application areas greatly benefit from the *subject mediation* approach supporting information integration in a particular subject domain. Each subject domain is defined by the experts specifying terminologies (thesauri), concepts (ontologies), data (objects) structuring, methods applicable to data, processes (workflows), characteristic for the domain. These definitions constitute a subject mediator *schema*. After consolidating the schema, an *operational* phase of the mediator starts. During operational phase information providers can disseminate their information for integration in the subject domain independently of each other and at any time. They should register their information sources at the mediator to make further information dissemination possible. During the registration, each local source class (modeled as a set of instances (objects) of the class instance type) should be defined in terms of the mediator schema. Such definition has a form of a materialized view over virtual classes of the mediator [11, 3]. This mediation approach is called 'local as view' (LAV) that is in contrast to 'global as view' (GAV) [4] where information sources should be pre-selected before a mediator formation and a global mediator's schema should be defined for them [5, 16]. LAV approach is intended to cope with a dynamic, possibly incomplete set of sources. Sources may change their exported schemas, become unavailable from time to time. Registration method for the LAV approach treated as the process of compositional information systems development is introduced in [2]. According to this method, local source definitions are treated as specifications of requirements and classes of the mediator schema with the related metainformation -- as specifications of pre-existing components. The main burden of registration effort in LAV is imposed on the information providers making mediators scalable with respect to a number of sources involved. Registrations for a certain mediator proceed in parallel and require specific protocol for their support.

Note that mediators form a recursive structure (Figure 1) so that a mediator can be registered as an information source at another mediator. Thus interdisciplinary subject mediators can be formed. Mediators themselves constitute a heterogeneous world, each mediator type introducing specific *mediator middleware platform*.

Each mediator platform is characterized by a set of inherent features, including mediator (meta)information model (defining terminological, ontological, typing and query facilities, metainformation repository structuring and access), an approach for a mediator schema consolidation, a process for information source registration at a mediator, mediator services, interfaces and protocols.

[1] This research was supported by the Russian Foundation for Basic Research, grants 01-07-90084 and 00-07-90086

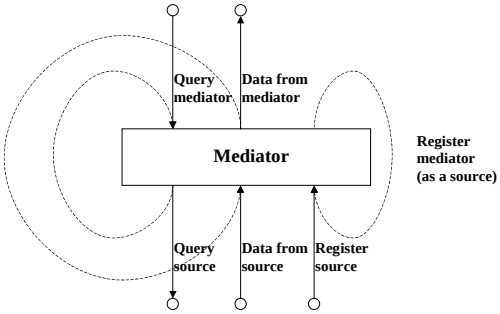

Fig. 1. Mediator's Recursion

The diversity of existing architectures exhibiting certain features of mediation platforms can be observed in distributed digital libraries (DL) area where different approaches for networked DL are in use, such as z39.50 architecture [1]; wide network of DL which enhances connectivity across document repositories and provides a quasi standardized access [10]; community-oriented DL which appear as a common need, subject gateways and broker-based architectures [13]; federated architectures and mediators [9, 14]. It is likely that for one subject domain an information can be found that has been previously registered at heterogeneous mediator platforms (Figure 2). The *intermediator framework* is required to support heterogeneous mediators interoperability.

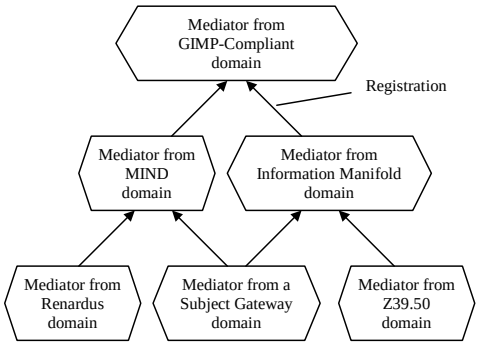

Fig. 2. Subject Domain Registration Structure

This paper introduces the LAV-based intermediator framework based on the LAV approach focusing on a protocol supporting registration of a mediator (source) *A* at another mediator *B*. The registration requires contextualizing of metainformation of *A* at the mediator *B*. Contextualizing means proper correlation of terminological, ontological, structural and behavioral metainformation of *A* w.r.t. the corresponding metainformation entities of the mediator *B*. For scalability reason the registration process should be performed at the mediator (source) *A* (using specific tools and involving personnel of *A*). The registration procedure should include selecting and transferring of respective metainformation definitions of *B* into *A*, performing contextualization of *A* in *B* leading to proper terminological and ontological correlation, defining *A* classes as materialized views over *B* classes, communicating the registration results into *B*.

2 Intermediator framework and normative registration procedure

In the intermediator framework a *domain* is a distinct mediator scope, within which certain common mediator features defined by a common mediator platform are exhibited. A mapping or *bridging* resides at the boundary between the domains providing the required transformations (Figure 3).

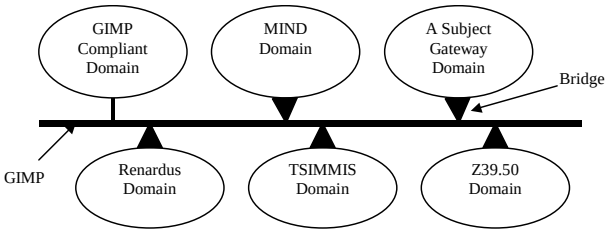

Fig. 3. Intermediator Framework

In the *intermediator framework* the open, mediator platform-neutral (canonical) information model is distinguished leading to unifying models and metainformation representation for thesauri, ontologies and hybrid

object/semistructured data and behaviors for various mediator platforms. We denote such modeling facilities in the sequel as M_c. Such modeling, its respective metainformation and representation is called *normative*. Specific bindings and tools are required for mapping a platform-specific mediator model into a canonical definition and back. The approach for such mapping that should be information and operations preserving is based on an idea of fixing a core of the canonical model and developing for each specific mediator (source) model such core extension that together with the core can be refined by this mediator (source) model [7]. Thus each new mediator (source) platform can be abstracted in the common middleware concepts. Mediator services and facilities can be specified as normative, platform independent in frame of the canonical model. A registration procedure we consider here is such normative procedure independent on the models of participating mediators. Only one such registration procedure is required for the intermediator framework. It can be used for any pair of mediators *A, B* after proper mapping of their models into M_c. We denote such transformation as a *bridging* between two different mediator domains.

The *Generalized Intermediator Protocol* (GIMP) is required for supporting of the intermediator framework and identifying main compliance points. Basically, GIMP is investigated as a combination of two protocols – slightly extended Simple Digital Library Interoperability Protocol **(**SDLIP) [15] providing facilities for data access in the intermediator environment and a subset of Open Archive Initiative Protocol (OAIP) [12] providing facilities for mediator metainformation exchange during the registration process. Only part of GIMP, supporting process of registration, is considered here.

3 Representation of metainformation in GIMP

A *mediator metainformation repository* is a network accessible server to which GIMP protocol requests, embedded in HTTP, can be submitted. The GIMP protocol provides access to metainformation from GIMP-compliant repositories. Such metainformation is output in the form of a *record*. A record is the result of a protocol request issued to the repository to disseminate metainformation from a particular subject mediator (source) that is considered a constituent of the repository (OAI *item*). For mediators interoperation their metainformation repositories are assumed to be GIMP-compliant (conformant to the canonical model). A record is an XML-encoded byte stream that is returned by a repository in response to a GIMP protocol request for metainformation from a mediator schema in that repository. GIMP records are structured exactly as the records of the OAIP [12].

The canonical model in GIMP is syntactically represented in the XML Schema language. An XML Schema is viewed as a collection (vocabulary) of type definitions and element declarations whose names belong to a particular namespace called a target namespace. Namespace is a concept used for modularization of schemas in the Web environment. A collection of schema entity specifications related to a specific context is contained in a namespace. To make interoperation framework completely determined, the SYNTHESIS information model has been chosen as the canonical one. SYNTHESIS is a hybrid object/semistructured model [8]. XML Schema definitions providing namespaces for the SYNTHESIS model are contained in *http://www.ipi.ac.ru/synthesis/projects/ XMLBIS/synxmls/*. According to GIMP, a canonical schema may be organized in modules, each of them containing all subject mediator specifications of one of the following kinds: structural specifications (definitions of subject mediator types and classes), ontological specifications, thesaurus specifications, classifier specifications, etc.. For each kind of a metainformation format an XML Schema definition and a respective namespace are provided (Figure 4). In terms of OAI, a subject mediator can disseminate metainformation in multiple formats mentioned. All disseminated records of this group of possible metadata formats share the same unique identifier – the identifier of a subject mediator. Each record disseminated by a *GetRecord* protocol request is identified by the combination of this unique identifier and a metadata prefix, which identifies the metadata format.

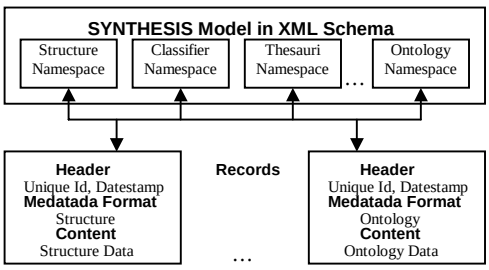

Fig. 4 GIMP Items and Metadata Formats

4 Heterogeneous information sources registration protocol

The approach is demonstrated on an example taking cultural heritage as a subject domain of the mediator. A protocol of registration of the Uffizi museum Web site at the subject mediator is shown. This information source has been chosen to make example shorter comparing to registration of possible mediators (e.g., CIMI museum profile for z39.50). The protocol of registration of any mediator (source) remains to be the same. Detailed description of the example chosen showing a method and process of an information source contextualization at a mediator is contained in [2]. Here we focus on the protocol issues.

A mediator (source) provider starts the registration by locating a mediator belonging to a subject domain to be registered in. For a known mediator and a known subject domain, a provider applies an OAI request:

ListMetadataFormats <identifier of a subject mediator>

The metadata formats are required for a choice of further strategy (e.g., if no thesaurus or ontological definitions are available in the schema then respective steps of the registration process will be omitted) as well as for using as obligatory arguments for further *GetRecord* requests.

For our example we get:

ListMetadataFormats Request

```
http://culturalheritage/gimp/OAI-script?verb=ListMetadataFormats
```

Response

```
<ListMetadataFormats
    xmlns="http://www.openarchives.org/OAI/1.0/OAI_ListMetadataFormats"
    xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
xsi:schemaLocation=
"http://www.openarchives.org/OAI/1.0/OAI_ListMetadataFormats
http://www.openarchives.org/OAI/1.0/OAI_ListMetadataFormats.xsd">
  <responseDate>2001-01-01T19:20:30-04:00</responseDate>
  <requestURL>http://culturalheritage/gimp/OAI-
script?verb=ListMetadataFormats
  </requestURL>
  <metadataFormat>
     <metadataPrefix>syn</metadataPrefix>
     <schema>http://www.ipi.ac.ru/synthesis/synthesis.xsd</schema>
  </metadataFormat>
  <metadataFormat>
    <metadataPrefix>syn_structure</metadataPrefix>
    <schema>http://www.ipi.ac.ru/synthesis/syn_structure.xsd</schema>
  </metadataFormat>
  <metadataFormat>
    <metadataPrefix>syn_ontology</metadataPrefix>
    <schema>http://www.ipi.ac.ru/synthesis/ontology.xsd</schema>
  </metadataFormat>
</ListMetadataFormats>
```

GetRecord <identifier of a subject mediator> <metadataPrefix>

for a given subject domain and for a specific metainformation format results in getting an information record containing all metainformation specifications of the given kind. For our example we get (for brevity, a specification of only one data type *Painting* and one class *painting* obtained from a mediator schema is shown):

GetRecord Request

```
http://www.culturalheritage.org/gimp/OAI-script?verb=
GetRecord&identifier=gimp:CulturalHeritageMediator&metadataPrefix=structure
```

Response

```
<GetRecord ...>
```

```
...
<record>
  <header>
    <identifier>CulturalHeritageMediator</identifier>
    <datestamp>1999-01-01</datestamp>
  </header>
  <metadata>
    <structure xmlns="http://www.ipi.ac.ru/synthesis/syn_structure.xsd"
               xmlns:syn="http://www.ipi.ac.ru/synthesis/synthesis.xsd">
      <ADT id="Painting">
        <superType id = "Heritage_Entity" />
        <attribute id="dimensions">
          <attributeType id="sequence">
            <type_of_element id="integer"/>
          </attributeType>
        </attribute>
      </ADT>
      <Class id="painting">
        <instance_section id = "Painting" />
      </Class>
…
    </structure>
  </metadata>
</record>
</GetRecord>
```

After obtaining the required specifications from the mediator, the provider contextualizes its source metainformation in the mediator's context, maps local source structural definitions into the canonical model and constructs representation of local classes in terms of the mediator's classes. More on these methods can be found in [2].

The results of the registration are formed as a local source schema expressed in terms of the canonical model with the appropriate terminological, structural and ontological links to the respective components of the mediator schema. This local source schema is made known to the mediator. The schema uses metainformation formats similar to that of the mediator. The mediator in its turn applies

ListMetadataFormats <identifier of a local source>

Similarly to the provider, the metadata formats are required for choice of a strategy of processing the local source metainformation at the mediator to complete the registration as well as for using formats as obligatory arguments for further requests:

GetRecord <identifier of a local source> <metadataPrefix>

For our example for the structured data obtained we get (only a definition of the Uffizi class *canvas* defined as a view over the mediator classes *painting* and *creator* is shown here):

Request

```
http://uffizi/gimp/OAI-script?
verb=GetRecord&identifier=gimp:UffiziRegistration&metadataPrefix=structure
```

Response

```
<GetRecord ...>
...
<record>
  <header>
    <identifier>UffiziRegistration</identifier>
    <datestamp>1999-01-01</datestamp>
  </header>
```

```
  <metadata>
    <structure xmlns="http://www.ipi.ac.ru/synthesis/syn_structure.xsd"
               xmlns:syn="http://www.ipi.ac.ru/synthesis/synthesis.xsd" >
      <Class id="v_canvas">
        <class_section>
          <attribute id = "key">
            <attributeType id="invariant" />
            <attributeType>
              <frame id = "unique">
                <slot><frame>title</frame></slot>
              </frame>
            </attributeType>
          </attribute>
          <attribute id = "lav">
            <attributeType id="invariant" />
            <attributeType>
              <formula>
               <![CDATA[ subseteq(v_canvas, painting(p/R_Painting_Canvas) &
                p.in_collection.in_repository = 'Uffizi' &
                creator(c/R_Creator_Cavas) & ex w/Painting (in(w, c.works)
&
                w.in_collection.in_repository = 'Uffizi'))
returns=c.painter)))]]>
              </formula>
            </attributeType>
          </attribute>
        </class_section>
        <instance_secton id = "CR_Painting_Creator_Canvas" />
      </Class>
...
    </structure>
  </metadata>
</record>
</GetRecord>
```

Mediator takes registration information from the provider, processes and checks the results to complete the registration. The success or failures are reported to the provider.

Providers can register several information sources at a time. In such cases they apply OAIP set concept for grouping resulting items (sources defined by related local resource registration metainformation). **ListSets** and **ListRecords** are additional OAIP requests that can be applied by the mediator in this case.

5 Conclusion

An approach for the intermediator framework based on usage of common, mediator platforms neutral canonical information model is sketched. Instead of inventing a new protocol for the framework support, it is proposed to build the Generalized Intermediator Protocol on the existing protocols developed for distributed DL environments, such as SDLIP and OAIP.

One of the outcome of this investigation is a proposal to use OAIP in course of a mediator (source) registration at another mediator to disseminate metadata positioned on a higher level than those for which OAIP has been originally intended for: the mediator (source) schema definitions are exchanged conformant to canonical information model represented by the respective namespaces in XML Schema.

References

1. ANSI/NISO Z3950-1995. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. Bethesda, MD: NISO Press.

2. D.O. Briukhov, L.A. Kalinichenko, N.A. Skvortsov. Information sources registration at a subject mediator as compositional development. To be published in the Proceedings of the Conference on Advances in Databases

and Information Systems, Vilnius, September 2001 (preliminary version of the paper can be found at http://www.ipi.ac.ru/synthesis/publications/registration/)

3. O. Duschka and M. Genesereth. Answering Queries Using Recursive Views. In Principles Of Database Systems (PODS), 1997

4. M. Friedman, A. Levy, and T. Millstein. Navigational Plans for Data Integration, 1999

5. H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The Tsimmis Approach to Mediation: Data Models and Languages. Journal of Intelligent Information System, 1997

6. The CIMI Profile: Z39.50 Application Profile for Cultural Heritage Information, Release 1.0H. <http://www.cimi.org/products/cimi_products.html#THREE>.

7. L. A. Kalinichenko. Method for data models integration in the common paradigm. In Proc. of the First East European Workshop 'Advances in Databases and Information Systems', St. Petersburg, September 1997

8. L.A. Kalinichenko. Integration of heterogeneous semistructured data models in the canonical one. In Proc. of First Russian National Conference on Digital Libraries, Saint-Petersburg, October 1999

9. L.A. Kalinichenko, D.O. Briukhov, D.V. Kravchenko, V.N. Zakharov. Infrastructure of the subject mediating level aiming at semantic interoperability of heterogeneous digital library collections. In Proc. of Second Russian National Conference on Digital Libraries, Protvino, September 2000

10. B.M. Lainer. The NCSTRL approach to Open Architecture for the Confederated Digital Library. D-Lib Magazine, December 1998

11. A. Levy, A. Rajaraman, and J. Ordille. Querying Heterogeneous Information Sources using Source Descriptions. In Proc. of the 22nd Conf. on Very Large Databases, pages 251-262, 1996

12. The Open Archives Initiative Protocol for Metadata Harvesting Protocol Version 1.0 of 2001-01-21, Document Version 2001-04-24, http://www.openarchives.org/OAI/openarchivesprotocol.htm

13. Renardus evaluation report of existing broker models. http://www.renardus.org/deliverables/D1_1_final.doc

14. Resource Selection and Data Fusion for Multimedia International Digital Libraries, MIND IST R&D PROJECT, http://www.iei.pi.cnr.it/DELOS/delos2/International/Presentations/mind-overv.ppt

15. The Simple Digital Library Interoperability Protocol.
http://www-diglib.stanford.edu/testbed/doc2/SDLIP/, 1999

16. V. S. Subrahmanian. Hermes: a Heterogeneous Reasoning and Mediator System.
http://www.cs.umd.edu//projects/hermes/publications/postscripts/tois.ps