Subject Mediation Infrastructure and Digital Libraries

Leonid A. Kalinichenko
Institute of Informatics Problems, Russian Academy of Sciences
Vavilov street, 44 – 2, Moscow, 119333, Russian Fed.
email: leonidk@synth.ipi.ac.ru

Key Words
Subject mediator, Canonical modeling, Contextualization, Source registration at the mediator

Abstract
This paper is an attempt to attract attention to the fact that information in digital libraries for science and education (acting as collective memories) should be structured differently than in accordance with the conventional library metaphor. Papers, journals, books, textbooks, courses are not good information entities any longer. Scientists have spent centuries to reach well-defined structures, concepts and theories in the various branches of science. These definitions are more suitable as a guiding principle for information structuring and search in digital libraries as collective memories. Subject mediator technology to support collective memories in a domain of a natural science is discussed. It is assumed that specific, intermediary layer is formed by *subject mediators* providing a uniform query interface to the multiple data sources to free the user from having to locate the relevant sources, query each one in isolation, and combine manually the information from the different sources. The paper is focused on various aspects of heterogeneous sources uniform representation and contextualization at the mediator level.

1. Introduction

The notion of "Digital Library" is subject to a broad range of definitions (Kalinichenko et al 2003a). Different audiences associated with a digital library (DL) have different interpretations; they evaluate a digital library differently and use different terminologies. On one end of the range, DLs are considered to be related to physical libraries performing similar functions, thus creating a *hybrid library* (combining traditional and digital resources). On the other end, DLs are considered to be knowledge repositories and services, organized as complex information systems.

The Technical Committee on Digital Libraries of the Institute of Electrical and Electronics Engineers Computer Society (TCDL of IEEE-CS) to define what DL is, uses the more general term "(digital) *collective memory*" to emphasize the convergence of sources of various kinds. Collective memory development faces challenges in several areas, including storage, classification, and indexing; user interfaces; information retrieval; content delivery; presentation, administration; preservation, etc.

Diversity of understanding of what digital libraries are leads to a wide range of possible visions for DL frameworks and methodologies of use. According to the framework based on the metaphor of a conventional library, a digital library as a collective memory can be considered a container extending the conventional library (cataloguing) practice. In this case the granularity of the memory is at the level of "information entities"—electronic versions of books, journal articles, images, and videos. Metadata schemas support retrieval focused on information entities (as in the conventional tradition of library bibliographic cards), not on subject structuring and the respective granularity of retrieved items. Such an approach looks reasonable at least because of the large heritage of traditional information entities and the significant difficulty in getting access to proper information items by content.

On the other hand, the analysis of digital libraries for education in different branches of science (Kalinichenko et al 2003a) shows that information in such libraries (acting as collective memories) should be structured differently. Textbooks and courses are not good information entities any longer. “Bibliographic cards” are not suitable for information discovery. Educational domains in different branches of science should be properly structured. More suitable entities would be concept spaces, theories, models, hypotheses, experimental results and measurements, curricula, and educational modules. Scientists have spent centuries to reach well-defined structures, concepts and theories in the various branches of science. These definitions cannot be used following the conventional library metaphor, but are more suitable as a guiding principle for information structuring and search in digital libraries.

For this reason, the gradual evolution of digital libraries from the currently dominated framework based on the conventional library metaphor to more knowledge-based organization is expected. With time and experience, these frameworks will be upgraded with conceptual definitions (ontologies) of subject domains and curricula along with the conventional metadata so that information resources can be registered in accordance with the proper subject definition and granularity. This trend will also lead to a higher level of coherency of the information collected in a specific subject domain, by contrast with metadata use, where collected materials are more diverse though less relevant to the subject.

In this paper we discuss a possibility of creating knowledge-based collective memory in a domain of a natural science. Information related to such domain includes domain terminology and concept definitions, material system description, definitions of various theories and models, observable (measurable) characteristics of real world objects, description of methods and instruments for observation, measurement, observation and experimental data, data analysis results, problem definitions and methods of solution, algorithms and programs, simulations. Integration of such information is driven by scientific and educational needs. The range of information required includes also scientific researches reported both formally in journals and informally in Web sites in the domain, curricula and courseware materials, lectures, access to remote scientific instruments, tools, the results of educational research, raw data for student activities, as well as related multimedia (image, audio, or video) banks.

Digital Earth, Digital Sky, Digital Bio, Digital Law, Digital Art, Digital Music are examples of areas of rapidly developing digital repositories of knowledge (see for instance, Microsoft TerraServer, Multi-Terabyte Astronomy Archives (Barclay 2000) (Szalay 2000)). The TerraServer (Barclay 2000) is the world's largest public repository of high-resolution aerial, satellite, and topographic data. It is designed to be accessed by thousands of simultaneous users using Internet protocols via standard web browsers. TerraServer delivers a set of raster images based on a users search criteria. The TerraServer tiling algorithm cuts tiles so that client applications can identify overlapping tiles from separate themes.

The next-generation astronomy digital archives (Szalay 2000) are created to cover most of the sky at fine resolution in many wavelengths, from X-rays, through ultraviolet, optical, and infrared. The archives will be stored at diverse geographical locations. Several multi-wavelength projects are under way: SDSS, GALEX, 2MASS, GSC-2, POSS2, ROSAT, FIRST and DENIS. Together they will yield a Digital Sky, of interoperating multi-terabyte databases. One of the first of these projects, the Sloan Digital Sky Survey (SDSS) is creating a 5-wavelength catalog over 10,000 square degrees of the sky. The 200 million objects in the multi-terabyte database will have mostly numerical attributes in a 100+ dimensional space. The archive will enable astronomers to explore the data interactively. Data access will be aided by multidimensional spatial and attribute indices.

Numerous forms of digital sources representations can be included into collective memories as distributed repositories of knowledge. Until some uniformity can be imposed on the available forms, the collective memory clients will feel themselves in much uncomfortable

condition than in conventional libraries. The problem facing researchers and developers in collective memories is fundamental: how to map huge variety of digital sources into their uniform representation and how to support the basic memory function of providing access to the integrated collection of heterogeneous information?

In this paper the issues of building heterogeneous digital repositories interconnected and accessible through global information infrastructures are discussed. To provide for interoperability of heterogeneous information objects (Scheck 1998) it is required to establish a global, uniform view of the underlying digital sources and services. It is assumed that specific, intermediary layer is formed by *mediators* providing a uniform query interface to the multiple data sources to free the user from having to locate relevant sources, query each one in isolation, and combine manually the information from the different sources. The mediator architecture (Wiederhold 1992) deals with the problem of integration of heterogeneous information. The sources are "heterogeneous" on many aspects: data model used, types of data, the underlying data units, behavior of objects involved, the underlying concepts, an extent to which a schema that the information may conform can be made rigid in advance.

Subject mediators are emphasized that support representation and access to various subject domains (here – in natural sciences). Mediators should provide modeling facilities and methods for conversion of unorganized, nonsystematic population of sources registered by different source providers into a well-structured set of sources supported by the integrated uniform specifications. The mediator's layer is introduced to provide the users with the metainformation uniformly characterizing subject content of the underlying sources as well as with the canonical information model making possible to query such sources and 'compute' the response. This model is needed to express the structure and semantics of the integrated data as well as the available collective memory services.

Each mediator supports the process of systematic registration and classification of sources providing the uniform ontological knowledge and metainformation to help information discovery and composition. This process of registration is assumed as a semi-automatic. It is supposed that source providers will be interested in registering their sources in common pools defined by mediators to optimize their investments.

The rest of the paper will be focused mostly on various aspects of heterogeneous sources uniform representation and contextualization at the mediator level. After very brief analysis of the broad range of information models that should be uniformly represented in mediators, the canonical information model intended for uniform description of heterogeneous metainformation is introduced. An approach for equivalent mapping of different kinds of information models into the canonical one is considered. Multilevel metainformation representation and modularization in mediators are defined. Structuring of the mediator metainformation (including conceptual and terminological information) for various sources defined in the canonical model is discussed. General schema of a process of an information source registration at the mediator is presented. The intermediator framework required to support heterogeneous mediators interoperability is briefly discussed. A small example of the subject mediator for gene expression regulation in molecular biology is given. Appendix contains definition of the basic operations supporting process of heterogeneous sources registration at the mediator.

2. Subject Mediator Definition and Canonical Modeling

Two separate phases of the mediator's life cycle are distinguished: consolidation phase and operational phase.

The *consolidation phase* is intended for creating a definition of the mediator’s subject. During this phase a consensus in the subject community should be reached on the definition of terminology, concepts, structural and behavioral definitions of the mediator subject types and classes. This process is completely independent on pre-existing information sources. Only a

few well-established representative sources can be taken into account. It is assumed that the metainformation defined during the consolidation phase is stable and remains valid for significant period of time.

During the *operational phase* the sources relevant to a subject mediator can be registered in it. The burden of the registration process is imposed on the source providers. They formulate sources capabilities (schemas, vocabularies, query languages) in terms of the subject mediator's metainformation and develop the required wrappers. Registrations of various sources are independent of each other and can be done at any time. This is the way how the mediator scalability with respect to the number of sources to be registered can be reached. Information consumers should know only the mediator's metainformation to query the mediator at any time of the operational phase.

Broad range of metainformation modeling facilities used in heterogeneous sources should be considered for their representation at the mediator's level, including those for textual and multimedia information, heterogeneous databases, ontological information, unstructured and semi-structured information, heterogeneous object components (Figure 1):

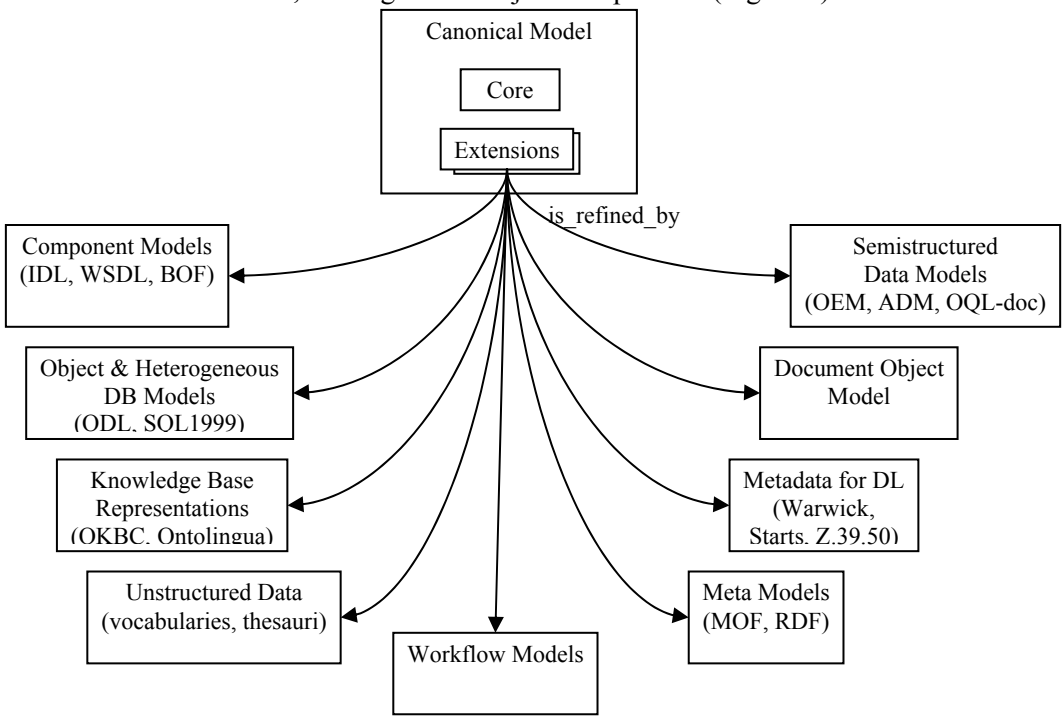

Figure 1: Heterogeneous information models absorbed by the canonical model

- Semi-structured data modeling facilities emerging to model the Web itself, structure of Web sites, internal structure of Web pages, and contents of Web sites (Abiteboul 1997) (Atzeni 1998), models expressible in Extensible Markup Language (XML);
- Content characterization of the unstructured information in a form of sets of natural language lexical units (terms) and their relationships selected for a certain thematic area (thesauri for different subject areas including thematic, poly-thematic, general-purpose, indexing and non-indexing thesauri (Kramer 1997));
- Ontological frameworks (Kalinichenko et al 2003b) designed for exchanging machine-understandable metadata describing Web resources;
- Knowledge representation models, such as the language for knowledge communication based on the predicate calculus semantics (KIF (Genesereth 1994)), a model for maintaining ontologies portably in a form that is compatible with multiple

representation languages (Ontolingua (Gruber 1992)), a common knowledge model of various knowledge bases (OKBC (Chaudhri 1998));

- Heterogeneous structural, object, service modeling facilities providing for interoperable component specifications and development (e.g., (OMG 1997));
- Object models for the Web representing a document as a hierarchy of objects which are derived from a source representation of the document (HTML or XML) – Document Object Model (DOM);
- Heterogeneous object and structural database modeling characterized by the basic standards (ODMG, SQL:1999).

To homogenize such variety of models uniformly representing them in one paradigm a specific approach has been developed providing for the mapping of various data models and metainformation into the canonical one using the principle of commutative data model mapping (Kalinichenko 1900). Each data model is completely defined by data description language (DDL) and data manipulation language (DML) syntax and semantics. Basic principle of mapping of an arbitrary source data model into the canonical one consists in commutative data model mapping according to which the mapping preserves information and operations of the source model provided that the diagram of DDL (schemas) mapping and the diagram of DML (operations) mapping are *commutative*. The second mapping principle consists in the extensibility of the canonical model. According to this principle the mapping should be carried out into the *extension* of the canonical model defined axiomatically so that the target model would become equivalent to the source model (or could be refined by the source model (Kalinichenko 1997)). Finally, according to the third principle – the principle of the *canonical model synthesis*, the core of the canonical model should be fixed, extensions of the core for various source data models should be constructed, all such extensions should be merged to form the result of synthesis of the canonical model. This method has been successfully used for mapping of structured data models (Kalinichenko 1900). For object models the method had been elaborated (Kalinichenko 1997) so that justification of the operation diagram commutativity became possible on the basis of formal model-theoretic notations (Abrial 1996), refinement calculus and respective tools. This approach is used as the basic one for the uniform representation of different digital sources representation in the mediator's canonical model. Figure 1 shows canonical model core extensions formed for various information models considered above.

In a specific project (Kalinichenko et al 2000) the mediator's canonical model is based on the SYNTHESIS language (Kalinichenko 1995) that has been elaborated for semantic interoperation and component-based information systems development in the wide range of pre-existing heterogeneous information sources. The language possesses hybrid capabilities providing for integration of structured as well as semi-structured data models (Kalinichenko 1999b). Uniform representation of diverse metainformation in the canonical one has been investigated.

A set of the canonical model facilities used for the uniform representation of the information sources includes the following (Figure 2):

- Frame representation facilities. Frames are treated as a special kind of self-defined values introduced mostly for description of concepts, terminological and semi-structured information. Frame representation facilities provide for expressing of arbitrary semantic associations of frames, unstructured, textual and temporal associations. All specifications in the canonical model have a form of frames that become a part of the metainformation.

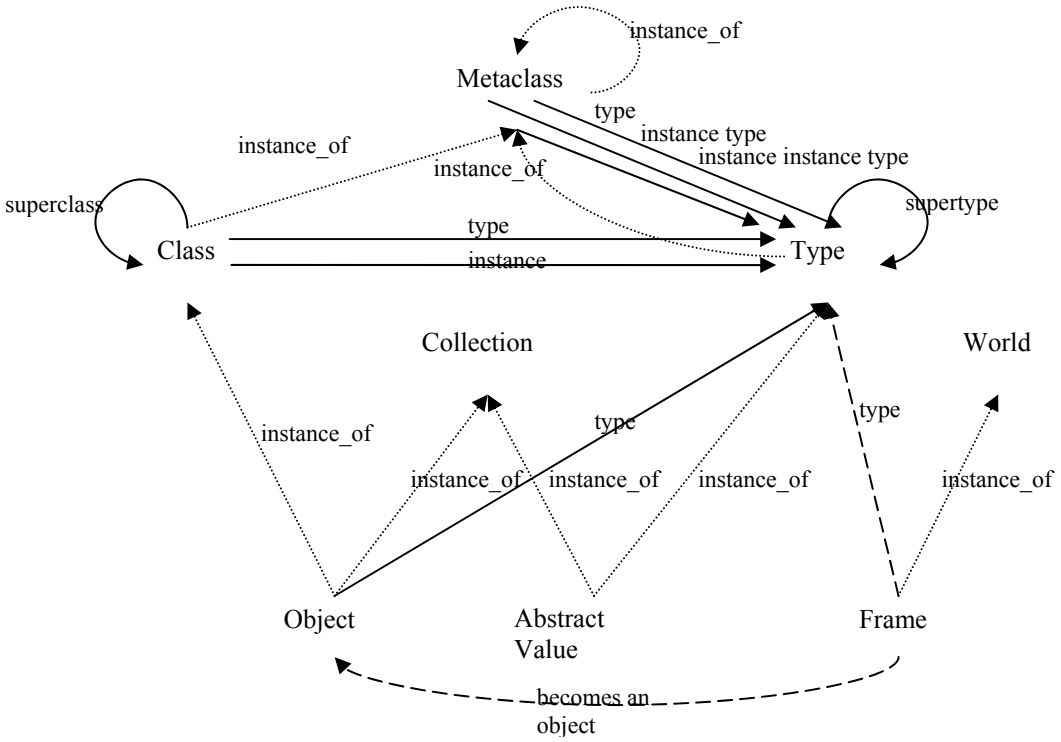

Figure 2: Canonical Model Entities

- Unifying type system. The universal constructor of arbitrary abstract data types as well as a comprehensive collection of built-in types and subtyping relationship are included into a type system. Types are values themselves. Metatypes provide for classification of the type hierarchy. Type expressions are introduced providing for type compositions that are required to type the results of queries and of heterogeneous component compositions.
- Class representation. Classes provide for representing of sets of homogeneous entities of an application domain. Class instances (objects) have specific types. Metaclasses provide for introducing different classification relationships orthogonal to the class generalization relationship.
- Process (workflow) representation. These are used for the specification and implementation of interconnected and interdependent problem solving and data analysis activities, for the specification of declarative assertions and concurrent megaprograms over the information sources. These facilities provide for specification of concurrent behavior of application systems.
- Facilities for the logical formulae expressions. A multi-sorted object calculus (typed first-order language) is used for querying the integrated set of digital sources as well as for specification of constraints and behavior.

Facilities of the canonical model have been checked to represent equivalently capabilities of various data models. Mapping of Ontolingua (Gruber 1992) as well as DAML+OIL (McGuinness 2002) into the canonical model has been also developed (Kalinichenko and Skvortsov 2002).

3. Structuring Homogeneous Metainformation for Various Sources in the Canonical Model

The unifying layer of digital sources specification at the mediator is split into three sublayers - local, mediator and personalized ones. Local sublayer provides metainformation

corresponding to each digital source and expressed in the canonical model. The mediator sublayer specifies the subject mediator. Personalized sublayer represents subsets of the mediator sublayer redefined in a way reflecting interests of specific users and user groups.

All descriptions are given in the form of the canonical model modules that may include definitions of different kinds of module sections: the frame section, the type section, the function section, the information source specification section (specific local sources are defined in the latter section). Any module can import an arbitrary set of other modules containing specifications of types, classes, frames constituting a context of a module. In the type section types and classes being object themselves can become instances of another, more general classes called metatypes for type classification and metaclasses for class classification. Metaclasses (metatypes) are useful for introducing generic concepts common for several attributes, types, classes. Thus multilevel classifications can be formed.

A set of specification modules can be combined into schemas. A specification of any constituent of the canonical model takes form of a frame. For the context formation it is sufficient to include into a context an arbitrary world of frames. A specification of one subject domain in a mediator is represented by means of the respective schema. The schema may include modules of different kinds defining structure, ontology, thesaurus and its extensions, rubric definitions. Specifications of one subject domain belonging to different levels (mediator, local) are included into one and the same schema. A notion of a subject domain is a relative one: specifications of subject domains of the higher levels are formed by integration and registration of schemas of subject domains of lower levels. A hierarchy of subject domains is established by import of the respective schemas.

For the mediator that is oriented on a certain subject domain we assume an existence of the pre-defined thesaurus for this specific domain (Kazakov and Vovchenko 1998). This non-indexing thesaurus is used as the core for the common thesaurus of the mediator. The terminological classifying hierarchy is formed to structure the subject domain of the mediator. Subject categories treated as classes form class/subclass hierarchy. Rubrics from the thesaurus are mapped into the categories. Sub-rubrics form subclass hierarchy of categories. Categories contain concepts and lexical units as instances. Concepts provide with modeling facilities of the mediator for uniform representation of ontologies, rubricators, thesauri. Class definition in the canonical model is quite expressive to provide precise meaning of the category (up to its ontological definition).

The following procedure of the common thesaurus formation preserving autonomy of the source thesauri (vocabularies) is assumed. During registration of a source a subset of its thesaurus that is not included into the core of the common thesaurus is identified. This subset is represented at the mediator's level as an indexing extension of the core. As the result, thesaurus federation (loosely integrated thesauri) is obtained. The common core thesaurus with its extensions is referred to as the Common Thesaurus.

Ontologies are used for context definition in a subject domain. Determination of exact difference between contexts helps to solve a problem of viewing onto an information source from another context or changing of a source moving from one context to another. Using of shared ontologies, establishing correspondence of data to ontological definitions, enhancing ontologies for new tasks allow to achieve correct interoperation between different information systems (Sure 2002) . In the mediator framework ontologies are used for semantic interrelation of the mediator and source specifications during registration of heterogeneous information sources at the mediator. A subset of the mediator’s canonical model is used to define ontologies.

Here we rely on an approach for extensible ontological model construction in a mediation environment (Kalinichenko and Skvortsov 2002) . A mediator ontological language (MOL) may depend on a subject domain and is to be defined at the mediator consolidation phase. On

the other hand, for different information sources different ontological models (languages) can be used to define their own ontologies. Reversible mapping of the source ontological models into MOL is needed for information sources registration at the mediator. An approach for such reversible mapping (Kalinichenko and Skvortsov 2002) has been developed for a class of the Web information sources assuming that such sources apply the DAML+OIL ontological model. A subset of the hybrid object-oriented and semi-structured canonical mediator data model is used for the core of MOL. Construction of reversible mapping of DAML+OIL into an extension of the core of MOL has been investigated. Such mapping is a necessary pre-requisite for contextualizing and registration of information sources at the mediator. The mapping showed also how extensible MOL can be constructed.

To enrich an expressive power of the ontological model there might be a need to use mixed facilities of verbal, logic-based or structural models in one and the same ontology. Ontologies in the mediator environment use a hybrid model (Skvortsov 2002). In such environment an ontology integration is the process based on verbal and type specifications of ontological concepts. Preliminary interrelation of ontologies is detected by linguistic methods including name analysis, verbal definition analysis, semantic relationship analysis. This process is called weak integration. The tight integration process may follow, which considers formal specifications of concepts. It uses reasoning similar to that of subsumption in a description logic.

We do not expect here integration of ontologies created by different working groups automatically. Semantic distinctions in similar concepts may be clarified sometimes only in discussions, applying ontologies to one and the same domain and revealing which real world objects can or cannot correspond to given concepts. For example the DELOS working group for harmonization of the CIDOC and the ABC ontologies has discovered many differences in understanding concepts that looked similarly on the first glance (DELOS 2003).

4. General Schema of a Process of an Information Source Registration at the Mediator

The process of registration is supported by the metainformation contextualizing tool (Briukhov 2001) that provides:

- local source context / mediator metainformation reconciliation and introducing new definitions into mediator's subject classifier hierarchy, thesaurus, ontologies,
- representation of the source type/class specifications in terms of the mediator's type/class hierarchy.

During the registration, a local source class is modeled as a materialized view over the mediator classes. Formally, the content of a local source class is described by a canonical model formula simplified as $C(z) \subseteq \exists \bar{x}(C_1(\bar{x}_1)\ \&\ ...\ \&\ C_n(\bar{x}_n)\ \&\ Con)$ where C is a local source class, $C_1,...,C_n$ are mediator classes, z is a reduct of the local class instance type being a refinement of a reduct of resulting instance type of the conjunctive formula in the right hand part of the rule, *Con* is additional constraints imposed by formula. Obtaining such view definition is considered as a process of compositional development in which local source specifications are considered as specifications of requirements, and mediator classes as components (Briukhov 2001).

General idea of such representation of the local classes is similar to those proposed in the query rewriting using views approach (Halevy 2001). Main difference of the current approach consists in taking into account issues more relevant to real environments, such as using rich type model and type specification calculus (Kalinichenko 1999a), applying the refining mapping of local data models into the canonical model of the mediator, resolving ontological differences between mediator and local concepts, systematic resolving structural, behavioral and value conflicts of local and mediator types and classes.

A representation of the local classes explained above provides for scalability of the mediator architecture: the representation of a specific class does not depend on other registered classes. This representation is used to generate sound and relevant query plans. A plan is sound if all the answers it produces are guaranteed to be answers to the original query. A plan is relevant if it contains answers to the original query.

Generally, a subject information infrastructure may consist of arbitrary number of mediators functioning in various subject domains. The structure of the middleware is recursive in a sense that a mediator is its building block that can be registered at any other mediator as its local, underlying source. Thus a mediator's metainformation can be represented in terms of its parent mediators. Queries submitted to the parent can be resolved in it by decomposing query into the child mediators subqueries.

As a preliminary step, an ontological integration of an information source specification and the mediator level specification is provided. Each element (class, type, attribute, function) of local source as well as of mediator schema specifications is related to a respective ontological concept. Establishing relationships between specification elements is provided on the basis of relationships between respective ontological concepts. Such relationships are defined during the ontological integration. An element of the source specification is ontologically relevant to the element of the mediator specification of the same kind (class, type, attribute, function) if the respective ontological concepts are linked by a proper association (e.g., positive association, hypernym association). Thus ontological specifications are used for identification of mediator classes that are semantically relevant to a source class. More details on the process of the ontological integration can be found in (Skvortsov 2002).

After the ontological integration, for each source class the following steps are required:

1. relevant mediator classes identification

Find mediator classes that ontologically can be used for defining source class as a view over the mediator classes. To a source class, only those ontologically relevant mediator classes may correspond that have common reducts (see appendix for more details) for their instance types.

2. most common reducts construction

For an instance type of each identified mediator class construct most common reducts (see appendix for more details) for instance type of this mediator class and source class instance type to refine (partially) such mediator instance type. Most common reduct may include mediator class attributes that can be derived from the source type instances. In this process for each attribute type of the common reduct a concretizing type, concretizing function or their combination should be constructed (this step should be recursively applied). Thus value, structural and behavioral conflicts are resolved.

3. partial source view construction

For each relevant mediator class construct a partial source view expressing a constraints in terms of the mediator class that should be satisfied by values of respective most common reducts of source class instances. Thus partial views over all relevant mediator classes will be obtained.

4. partial views composition

Construct compositions of the source type most common reducts obtained for instance types of all mediator classes involved (see appendix for more details). Construct a source view as a composition of partial views obtained above. This is an expression of a materialized view of an information source in terms of the mediator classes. An instance type of this view is determined by the most common reducts composition constructed above (see appendix for the class composition for more details).

5. Generalized Intermediator Protocol for Information Sources Registration

Note that mediators form a recursive structure 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.

The diversity of existing architectures exhibiting certain features of mediation platforms can be observed in distributed digital libraries area where different approaches for networked digital libraries are in use, such as z39.50 architecture; wide network of DL which enhances connectivity across document repositories and provides a quasi standardized access (NCSTRL); community-oriented DL which appear as a common need, subject gateways and broker-based architectures (Renardus); federated architectures and mediators (e.g., MIND). It is likely that for one subject domain an information can be needed that has been previously registered at heterogeneous mediator platforms (Figure 3). The *intermediator framework* is required to support heterogeneous mediators interoperability.

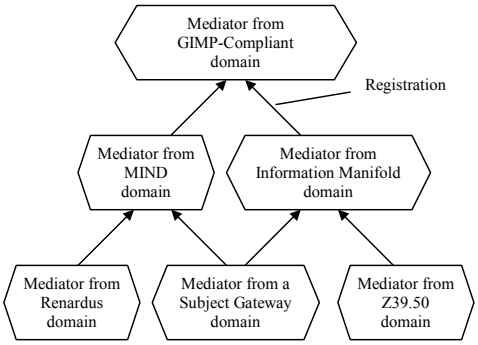

Figure 3: Subject Domain Registration Structure

In the *intermediator framework* (Kalinichenko et al 2001) 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.

Mediator services and facilities can be specified as normative, platform independent in frame of the canonical model. The intermediator framework is formed on the basis of such normative procedures independent on the mediator platforms. For example, a registration procedure is such normative procedure independent on the models of participating mediators. Only one such registration procedure is required for the intermediator framework.

The *Generalized Intermediator Protocol* (GIMP) (Kalinichenko et al 2001) is required for supporting of the intermediator framework and identifying main compliance points (Figure 4). Here we mention a part of this protocol providing for the support of sources registration process in the framework and on the Open Archive Initiative Protocol (OAIP) (OAIP 2001). OAIP provides facilities for metainformation exchange during the registration process. Only part of GIMP, supporting process of registration, is considered here.

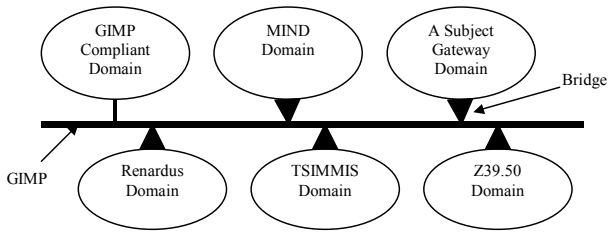

Figure 4: Intermediator Framework

A *mediator metainformation repository* is a network accessible server to which GIMP protocol requests, embedded in HTTP, can be submitted. Such metainformation is output in the form of *records*. The GIMP protocol provides access to metainformation from GIMP-compliant repositories. 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.

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. A collection of schema entity specifications related to a specific context are determined by the canonical model and contained in a namespace. 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 5). In terms of OAIP, 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.

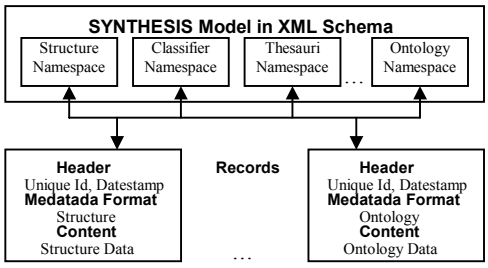

Figure 5: GIMP Items and Metadata Formats

6. Initial Experience

Initial experience in subject mediation has been obtained in the domain of gene expression regulation where it was required to arrange a unified representation of information sources and services by means of a mediator (Kalinichenko et al 2002). The model of the subject domain has been developed by the Institute of Cytology and Genetics of RAS. The model included ontological definition of the related concepts and thesauri, definition of information structuring, types of experiments, data analysis methods, as well as the related models of the respective theory. The mediator is oriented on a broad class of problems. A representative example of this class is preparing of the training samples of regulatory regions, which may be used by recognition programs: to output the set of transcription factor binding sites sequences, which have a definite type of DNA-binding domain, search for transcription factors corresponding to the proteins found, search for transcription factor binding sites; search for the sequences of pre-ordered length including relevant transcription factor binding sites. A set of information sources to be registered at the mediator includes: the database TRRD (contains information about structural and functional organization of extended transcription regulating regions of eukaryotic genes and their expression), the database SWISSPROT (contains information about the structure and functions of genes, their domain structure, sequences), the database EMBL/GenBank (accumulate information about the sequences DNA, RNA, their exon-intron structure, and other functional layout), the database Medline/PubMed (stores bibliography that is required for supporting and verifying the presented data). Uniform representation, contextualization and registration of such sources have been experienced to support solving of the problems mentioned. Consolidation phase of the mediator definition appeared to be one of the most difficult obstacles for the mediator development.

Currently the subject mediation approach is attempted for the Russian Virtual Observatory (RVO). RVO is planned as a collection of interoperating data archives and software tools to form a scientific research environment in which astronomical research programs can be conducted. The first mediator is planned for the problems of the distant galaxy discovery.

7. Conclusion

This paper is an attempt to attract attention to the fact that information in digital libraries for science and education (acting as collective memories) should be structured differently than in accordance with the conventional library metaphor. Papers, journals, books, textbooks, courses are not good information entities any longer. Scientists have spent centuries to reach well-defined structures, concepts and theories in the various branches of science. These definitions are more suitable as a guiding principle for information structuring and search in digital libraries as collective memories.

Subject mediator technology to support collective memories in a domain of a natural science is discussed. It is assumed that specific, intermediary layer is formed by *subject mediators* providing a uniform query interface to the multiple data sources to free the user from having to locate the relevant sources, query each one in isolation, and combine manually the information from the different sources. The paper is focused on various aspects of heterogeneous sources uniform representation and contextualization at the mediator level. The paper briefly reports on a progress in the area, including:

- an approach for equivalent mapping of different kinds of information models into the canonical one and for synthesis of the mediator canonical model as a result of constructing extensions of the canonical model core for various source data models (so that these extensions could be refined by the respective source data models) and merging them to form the mediator's canonical model;
- principles of the subject mediator context definition, multilevel metainformation representation and modularization, structuring of the metainformation (including conceptual and terminological information), mediator ontological modeling are discussed;
- an approach for mediator/source contexts reconciliation;
- general schema of process of an information source contextualization and registration at the mediator that is supported by the respective tool is presented. The process is based on source ontology contextualization, type specification composition and type refinement techniques;
- the intermediator framework and protocol required to support heterogeneous mediators interoperability is briefly discussed.

Initial experience shows that on the current level of technology the mediation approach is applicable to narrow subject domains in natural science where a consensus on subject definition during the consolidation phase can be reached.

References

Abiteboul S. et al. 1997
Querying documents in object databases
International Journal on Digital Libraries, Volume 1, N 1, April 1997

Abrial J.-R. 1996
The B Book: assigning programs to meaning
Cambridge University Press

Atzeni P., Mecca G., Merialdo P. 1998
Semistructured and structured data in the Web: going back and forth
ACM Sigmod Record, N 1

Briukhov D.O., Kalinichenko L.A., Skvortsov N.A. 2001
Information sources registration at a subject mediator as compositional development
In *Proceedings of the Conference on Advances in Databases and Information Systems (ADBIS)*, Springer, LNCS, Vilnius, September 2001

Barclay T., Gray J., Slutz D. 2000
Microsoft TerraServer: A Spatial Data Warehouse
In *Proc. of the 2000 ACM SIGMOD Conference*, ACM Press, May 2000

Chaudhri V., et al 1998
Open knowledge base connectivity 2.0.2
Stanford University, February 1998

DELOS 2003
Building Core Ontologies
White Paper of the DELOS Working Group on Ontology Harmonization. April 2003

Genesereth M. and Fikes R. 1994
Knowledge Interchange Format Reference Manual
http://logic.stanford.edu/sharing/papers/kif.ps

Gruber T.R. 1992
Ontolingua: a mechanism to support portable ontologies
Stanford University, June 1992

Halevy A.Y. 2001
Answering queries using views: a survey
VLDB Journal, 10(4): 270 – 294, 2001

Kalinichenko L.A. 1900
Methods and tools for equivalent data model mapping construction
In *Proc. of the EDBT'90 Conference,* Springer Verlag

Kalinichenko L.A. 1995
SYNTHESIS: the language for desription, design and programming of the heterogeneous interoperable information resource environment
Institute for Problems of Informatics, Russian Academy of Sciences, Moscow

Kalinichenko L.A. 1997
Method for Data Models Integration in the Common Paradigm
In *Proceedings of the First East European Conference on "Advances in Databases and Information Systems",* St. Petersburg, September 1997 (also Springer Workshop in Computing, Electronic Publication)

Kalinichenko L.A.. 1999a
Compositional Specification Calculus for Information Systems Development
In *Proc. of the East-West Conference on Advances in Databases and Information Systems (ADBIS'99)*, Maribor, Slovenia, September 1999, Springer Verlag, LNCS

Kalinichenko L.A. 1999b
Integration of heterogeneous semi-structured data models in the canonical one

In *Proceedings of the First Russian National Conference on "Digital Libraries: Advanced Methods and Technologies, Digital Collections"*, Saint-Petersburg, October 1999

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

Kalinichenko L.A., Briukhov D.O., Tyurin I.N., Skvortsov N.A. 2001
Intermediator framework protocol for information sources registration at heterogeneous mediators
In *Proceedings of the DELOS Workshop Interoperability in Digital Libraries*, September, 2001, GMD-IPSI, Darmstadt, Germany

Kalinichenko L.A., Briukhov D.O., Zakharov V.N., Podkolodny N.L. 2002
Mediation of heterogeneous information resources in the gene expression regulation domain
In *Proceedings of the 3-rd International Conference on Bioinformatics of Genome Regulation and Structure*, Vol.3, Institute of Cytology and Genetics

Kalinichenko L.A., Skvortsov N.A. 2002
Extensible Ontological Modeling Framework for Subject Mediation
In *Proceedings of the Fourth Russian National Conference on Digital Libraries*, Dubna, October, 2002

Kalinichenko L.A. et al. 2003a
Digital Libraries in Education. Analytical Survey
UNESCO Institute for Information Technologies in Education, Moscow, 2003

Leonid Kalinichenko, Michele Missikoff, Federica Schiappelli, Nikolay Skvortsov. 2003b
Ontological Modeling
In *Proc. of the Fifth Russian National Conference on Digital Libraries*, St. Petersburg, October 2003

Kazakov E.N., Vovchenko E.L. 1998
Use of poly-thematic thesaurus for mediators of digital library multicollections supporting their interactions with the users and collections
In *The RFBR DL Workshop*, Moscow, December 1998

Kramer R., Nikolai R., Habeck C. 1997
Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies
International Journal on Digital Libraries, Volume1, N 2, September 1997

McGuinness D., Fikes R., Hendler J., Stein L. 2002
DAML+OIL: An Ontology Language for the Semantic Web
IEEE Distributed Systems Online 3(11)

OAIP 2001
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

OMG 1997
Object Management Group, Meta Object Facility (MOF) Specification
OMG Document ad/97-10-02

Scheck H.-J., Birmingham B. 1998
Summary review of the Working Group on Interoperability
In: *Proceedings of DELOS Workshop on Emerging Technologies in the Digital Libraries Domain,* Brussels, October 1998, ERCIM-98-W004

Skvortsov N.A., Kalinichenko L.A. 2002
An Approach to Ontological Modeling and Establishing Intercontext Correlation in the Semistructured Environment
In *Proceedings of RCDL'2000*, September 2000, Protvino

Sure Y., Staab S., Studer R. 2002
Methodology for Development and Employment of Ontology based Knowledge Management Applications
ACM SIGMOD Record, Volume 31, N 4

Szalay S., et al 2000
Designing and Mining Multi-terabyte Astronomy Archives: the Sloan Digital Sky Survey
In *Proceedings of the 2000 ACM SIGMOD Conference*, ACM Press, May 2000

Wiederhold G. 1992
Mediators in the Architecture of Future Information Systems
IEEE Computer

Appendix

Basic Operations Supporting Process of Heterogeneous Sources Registration at the Mediator

Reducing of Specifications

A *type reduct* R_T is a subspecification of specification of the type T created by a reduced operation set of T. Reducing of specifications leads to decomposition of type specifications into a set of well-defined type specification fragments (reducts). A reduct R_T becomes a supertype of the T type.

Common reducts. A *common reduct* for types T_1, T_2 is such reduct R_{T1} of T_1 that there exists a reduct R_{T2} of T_2 such that R_{T2} is a *refinement* of R_{T1}. Further we refer to R_{T2} as to a *conjugate* of the common reduct.

Most common reduct A *most common reduct* $R_{MC}(T_1,T_2)$ for types T_1, T_2 is a reduct R_{T1} of T_1 such that there exists a reduct R_{T2} of T_2 that refines R_{T1} and there can be no other reduct R_{T1}^i such that $R_{MC}(T_1,T_2)$ is a reduct of R_{T1}^i , R_{T1}^i is not equal to $R_{MC}(T_1,T_2)$ and there exists a reduct R_{T2}^i of T_2 that refines R_{T1}^i .

While constructing most common reducts of the mediator and source types it is required to identify and reconcile various conflicts of such specifications – value conflicts, structural

conflicts, behavioral conflicts. Conflict resolution is based on a combination of two approaches - using a set of predefined rules of structural transformations, and using a high-level language for transformation specifications. A correspondence between elements of specifications (types, attributes, functions) is established as a result of ontological specifications integration.

Most common reduct identification

To identify common signature reducts, we should find for each pair of ontologically relevant types T_m, T_s a maximal collection A of attribute pairs $a_{Tm}^{\ i}$, $a_{Ts}^{\ j}$ that are also ontologically relevant and satisfy the type constraints so that $a_{Ts}^{\ j}$ could be reused as $a_{Tm}^{\ i}$. Analysing attribute pairs, we recognize and resolve various conflicts between the mediator and source type specifications.

Most common reducts composition

Composition of discovered fragments (reducts) into specification concretizing the mediator specifications is realized by applying the operations of type composition (meet and join) and by construction of views over classes (Kalinichenko 1999a). We introduce here definition of only one operation - *join*. Let T_i ($1 \le i \le n$) denotes types.

Type join operation. An operation $T_1 \cap T_2$ produces type T as a 'join' of specifications of the operand types. Generally T includes a merge of specifications of T_1 and T_2. Common elements of specifications of T_1 and T_2 are included into the merge (resulting type) only once. The common elements are determined by another merge - the merge of conjugates of two most common reducts of types of T_1 and T_2 : $R_{MC}(T_1,T_2)$ and $R_{MC}(T_2,T_1)$. The merge of two conjugates includes union of sets of their operation specifications. If in the union we get a pair of operations that are in a refinement order then only one of them, the more refined one (belonging to the conjugate of the most common reduct) is included into the merge. Invariants created in the resulting type are formed by *conjuncting* invariants taken from the original types.

A type T is placed in the type hierarchy as an immediate subtype of the join operand types and a direct supertype of all the common direct subtypes of the join argument types.

Operations of the compositional calculus form a *type lattice* (Kalinichenko 1999a) on the basis of a subtype relation (as a partial order).

Class compositions

Depending on the requirements, such operations over classes as, e.g., *union*, *intersection* or *join* are used. For union the instance type of the constructed view is formed by *meet* operation over instance types of the classes - arguments of the operation. For join the instance type of the constructed view is formed by *join* operation over instance types of the classes - arguments of the operation.