Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections

Kalinichenko L.A., Briukhov D.O., Skvortsov N.A., Zakharov V.N.
Institute for Problems of Informatics
Russian Academy of Sciences
Vavilova 30/6, Moscow, V-334, 117900
E-mail: {leonidk,brd,scvora}@synth.ipi.ac.ru

Abstract

This work is oriented on creation of integrated virtual digital libraries mediating heterogeneous distributed digital collections of scientific information. A required infrastructure of the subject mediators aiming at semantic interoperability of heterogeneous digital library collections is presented. The diversity of information models that should be uniformly represented at the canonical level is analyzed. The canonical model and an approach for various information models homogenization in the canonical paradigm are introduced. The mediator's infrastructure is defined as a set of functionally-oriented frameworks: collection registration framework, information extraction framework, personalization framework. Mediator scalability measures are discussed.

1 INTRODUCTION

Digital Libraries are considered to be a critical component of the emerging distributed knowledge environments that should provide people with access to virtually all areas of human knowledge, with an intention of improving standards of health, education, and economic well-being as well as the quality of life. As such, the field of digital library research and technology encompasses information creation, acquisition, access, distribution, evaluation and processing. Major applications of digital library research and technology include education, science, commerce, medicine, and culture. 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 [6, 36], Rutgers CIMIC DigiTerra [3]).

The TerraServer [6] 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. A work together with the UCB Digital Library team is intended for building a client application which will display TerraServer projected data-sets that are in the same projection as a layered map set.

The next-generation astronomy digital archives [36] will 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 (see http://www.sdss.org/). 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.

DigiTerra [3] is an Environmental Digital Library that is in the process of developing in Rutgers under the sponsorship from NASA and HMDC, a New Jersey State government agency. DigiTerra objective is to provide continuous land monitoring, fire detection, water and air quality testing, urban planning, as well as supporting research and instructional activities in related areas of science. Vast array of environmental data collected in DigiTerra should include images from a variety of space-borne sattelites, ground data from continuous monitoring weather stations, maps, reports and data sets from federal, state and local government agencies, and serve diverse user commuinities.

Numerous forms of digital collections representations can be included into Digital Libraries as distributed repositories of knowledge. Until some uniformity can be imposed on the available forms, the Digital Library 'readers' will feel themselves in much uncomfortable condition than in conventional libraries. The problem facing researchers and developers in Digital Libraries is fundamental: how to map huge variety of digital collections into their uniform representation and how to support the basic library function of providing access to

©Вторая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
26-28 сентября 2000г., Протвино

the integrated collection of heterogeneous information ?

The project[1] is oriented on building large heterogeneous digital repositories interconnected and accessible through global information infrastructures. New models, theories and frameworks are to be developed in order to understand the complex interactions between various components in a globally distributed digital library.

To provide for interoperability of heterogeneous information objects [35] it is required to establish a global, uniform view of the underlying digital collections 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 the relevant collections, query each one in isolation, and combine manually the information from the different collections. 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. Examples of "semi-structured" information include those found in XML documents, repositories used in the bio-molecular data, Web sites, etc.

In this particular project *subject mediators* are emphasized that support representation and access to various subject domains. Mediators should provide modelling facilities and methods for conversion of unorganized, nonsystematic population of collections registered by different collection 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 collections and the canonical information model making possible to query such collections and 'compute' the response. This model is needed to express the structure and semantics of the integrated data as well as the available DL services.

Each mediator supports the process of systematic registration and classification of collections providing the uniform ontological knowledge and metainformation to improve discovery and compositions of existing resources. This process is planned as a semi-automatic. It is expected that collection providers (the original capital investors) will be interested in registering their collections in a common pool in mediators to optimize the investments.

The mediator's metainformation is intended to be shared by information consumers, collection providers and subject mediators. The paper provides brief analysis of the broad range of information models that should be uniformly represented in mediators. The canonical information model intended for uniform representation of heterogeneous metainformation is introduced. An approach for equivalent representation of different kinds of information models in the canonical one is considered. Multilevel metainformation representation and modularization in mediators are defined. Creation of the metainformation on the interoperation level is specifically emphasized.

The metainformation registry system is planned that will use the canonical model constructs to link diverse contexts and representation of heterogeneous metadata among themselves. Acquisition and integration of metadata defining the information sources' content and capabilities in each domain are the basic functions of mediators. Metainformation assists in the selection of sources relevant for a query, and the creation and optimization of queries against the source.

The paper starts with an analysis of diversity of information that should be presented at the mediator's level and with brief characterization of the canonical metainformation model that is required for the mediator. An approach for uniform representation of heterogeneous collections in the canonical paradigm is presented. Structuring of the interoperation level of the mediator's metadata is discussed. Basic functions of the mediator are introduced in a specific section. The mediator's functions are structured into several frameworks: collection registration framework, information extraction framework, personalization framework. The mediator scalability issues are specifically discussed. Finally, related projects on heterogeneous information resource mediators are briefly surveyed.

2 DIVERSITY OF INFORMATION IN HETEROGENEOUS SOURCES

The broad range of metainformation modeling facilities relevant to the Digital Libraries collections should be considered for their respresentation at the mediator's level, including those for textual and multimedia information, heterogeneous databases, ontological information, unstructured and semistructured information, heterogeneous object components:

- Semistructured data modeling facilities emerging to model the Web itself, structure of Web sites, internal structure of Web pages, and contents of Web sites (such as the Object Exchange Model (OEM) [2], Araneus Data Model (ADM) [5], OQL-doc [1], WebSQL and WebOQL models [4], models expressible in Extensible Markup Language (XML)[20]);
- Digital Library data content description standards (such as Dublin Core as a core element set that provides adequate data for Web resource description [37] as well as those provided in Z39.50 profiles and Warwick Framework that introduced a concept of a container for aggregating multiple sets of metadata);
- Metadata for 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 [26]);
- Metadata expressible in metamodels (such as the World Wide Web Consortium's Resource Description Framework (RDF) [27] designed for exchanging machine-understandable metadata describing Web resources);

[1] This work has been undertaken for the project "Creating integrated virtual digital libraries mediating heterogeneous distributed digital collections of scientific information" being developed in frame of the Digital Library Program supported by the Russian Foundation for Basic Research (RFBR grant N 98-07- 91061).

- Knowledge representation models expressible in well-known notations including the language for knowledge communication based on the predicate calculus semantics (KIF [14]), a model for maintaining ontologies portably in a form that is compatible with multiple representation languages (Ontolingua [18]), a common knowledge model of various knowledge bases (OKBC [15]);

- Heterogeneous object component modeling facilities including interface specifications providing for technical interoperability (IDL [32]), and definitions providing more semantics for component-based development (BOF, CDL [10]);

- Object models for the Web representing a document as a hierarchy of objects which are derived (by parsing) from a source representation of the document (HTML or XML) – Document Object Model (DOM) [16];

- Object and heterogeneous database models charaterized by the basic standards for object modeling (ODMG ODL [30]), object-relational modeling (SQL:1999 [12]), as well as by the heterogeneous multidatabase modeling (IRO-DB [13], Garlic [11]).

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 data model refinement* [22].

Main idea of the approach consists in creation of extensions of the canonical model core for each data/knowledge model that may be used for a digital collection representation (in sequel such models will be called the *local* ones). These extensions should be formed so that the local models should become their justifiable refinements. Satisfying this condition guarantees preservation of information and operations while mapping various models and respective metadata into the common paradigm. This approach is planned as the basic one for the uniform representation of different digital collections representation. Fig. 1 shows canonical model core extensions formed for various information models considered above.

3 CANONICAL INFORMATION MODEL FOR THE MEDIATOR'S ENVIRONMENT

We base the mediator's canonical model on the SYNTHESIS language [21] that has been elaborated for semantic interoperation and component-based information systems development in the wide range of pre-existing heterogeneous information resources. The language possesses hybrid capabilities providing for integration of structured as well as semi-structured data models [23]. Uniform representation of diverse metainformation representation in the canonical one has been investigated.

A set of the canonical model facilities used for the uniform representation of the information resources includes the following:

- Frame representation facilities. Frames are treated as a special kind of abstract values introduced mostly for description of concepts, terminological and weakly-structured information. In particular, information resource metainformation (schema) is represented using the frame language. Frame representation facilities provide for expressing of arbitrary semantic associations of frames, for representation of unstructured, textual and temporal associations. All specifications in canonical model have a form of frames that become a part of the metabase.

- Unifying type system. A universal constructor of arbitrary abstract data types as well as a comprehensive collection of the built-in types are included into a type system. For types a type specialization (subtyping) relationship is defined. 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 hierarchies and type inheritance mechanisms make possible to define the generalization / specialization relationships. Class instances (objects) have specific types. Metaclasses provide for introducing different classification relationships orthogonal to the class generalization relationship.

- Multiactivity (workflow) representation. These are used for the specification and implementation of interconnected and interdependent application activities, for the specificaton of declarative assertions and concurrent megaprograms over the information resources. These facilities provide for specification of concurrent and asynchronous behaviour of application systems and of interoperable resource environments as of dynamic discrete events systems.

- Facilities for the logical formulae expressions. A multisorted object calculus (typed first-order language) is used for querying the integrated set of digital collections as well as for specification of constraints and behaviour.

Schematically basic entities of the canonical model and their relationships are shown on Fig. 2.

Information on the entities and situations observed in a real world is represented in the information resource base as a collection of abstract values that can be immutable or mutable, uniquely identified values (objects). In this range we can differentiate between:

- collections of self-defined objects or collections of frames (worlds);

- worlds with pre-defined frame associations;

- classes containing partially typed objects that may be characterized by their own individual attributes that were not specified in a type of the class instance;

- strictly typed classes (a set of instance attributes is strictly fixed);

- classes of aggregates (associations of objects and frames).

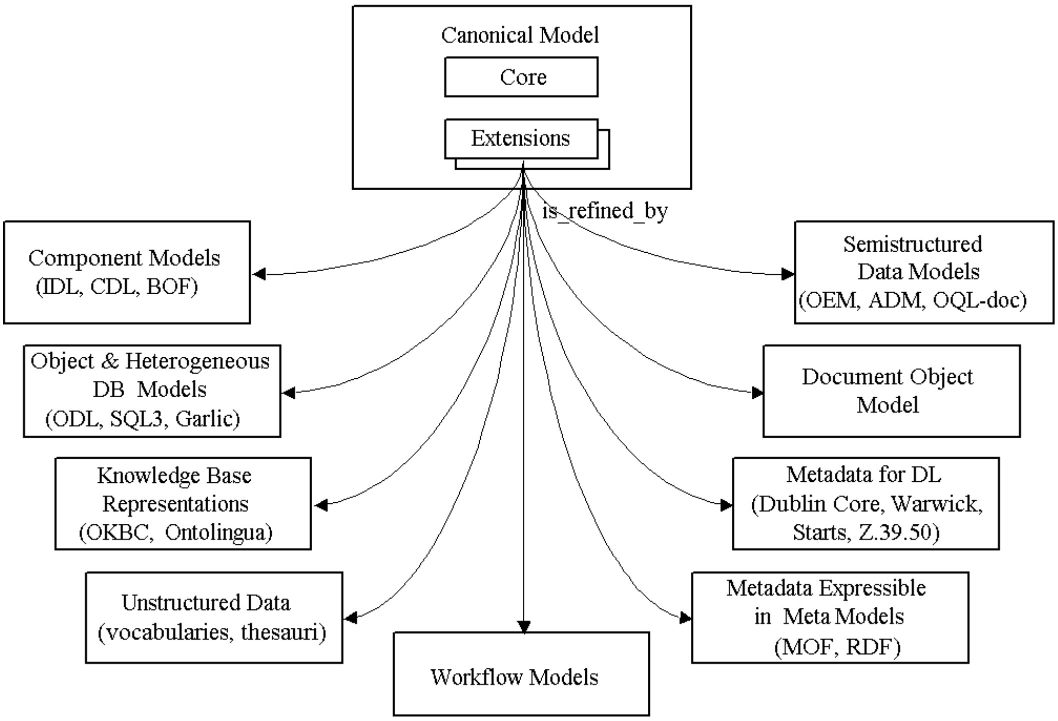
Canonical Model
Core
Extensions
is_refined_by
Component Models (IDL, CDL, BOF)
Object & Heterogeneous DB Models (ODL, SQL3, Garlic)
Knowledge Base Representations (OKBC, Ontolingua)
Unstructured Data (vocabularies, thesauri)
Workflow Models
Semistructured Data Models (OEM, ADM, OQL-doc)
Document Object Model
Metadata for DL (Dublin Core, Warwick, Starts, Z.39.50)
Metadata Expressible in Meta Models (MOF, RDF)

Figure 1: Heterogeneous information models absorbed by the canonical model

Benefits of both models - object-oriented and frame-based - are available in SYNTHESIS. Treating frames as partially typed data allows queries to be written without complete knowledge of a schema. At the same time all object-oriented properties of the model (functions, subtyping, etc.) are available. Path expressions in SYNTHESIS may correspond to database path that may involve structured or semistructured data only, or there may be crossover points to navigate from structured to semi-structured data or vice-versa [23].

4 MEDIATOR'S METAINFORMATION REPRESENTATION IN THE CANONICAL MODEL

4.1 Uniform representation of object model metainformation facilities

Facilities of the canonical model has been checked to represent equivalently various data model facilities in the canonical model. IDL specifications, ODMG model, the models used by the multidatabase projects (like IRO- DB [13] and Garlic [11]) can be equivalently and uniformly represented in the canonical model. The SYNTHESIS model goes beyond that: SQL:1999 schema and RDF uniform mapping into the canonical model is possible.

The Dublin core and Z39-50 profiles metadata are also considered to be of the structured kind that can be easily represented in the canonical model. Modularization features of the canonical model is applicable to represent the Warwick concepts.

The approach of mapping heterogeneous data and object models into the canonical model preserving information and operations has been investigated and developed [22].

4.2 Uniform representation of knowledge base and ontological metadata

The OKBC Knowldge model [15] is representable in the canonical model. Collection of built-in types of OKBC is a subset of the collection of the SYNTHESIS built-in types. Class frame of OKBC can be mapped into the frame of a class specification of SYNTHESIS. Individual frames of OKBC are representable by the individual frames of SYNTHESIS. Slot facets are interpreted by metaslots. All collection kinds of OKBC are included into the canonical model. Thus mapping of the OKBC metadata into the canonical model consists in representation of various *meta-kbs* of OKBC in SYNTHESIS.

Mapping of Ontolingua [18] into the canonical model has been checked to show that it is also possible.

4.3 Uniform representation of conceptual and terminological information

For the mediator that is oriented on a certain subject domain we assume an existence of the pre-defined thesaurus for this

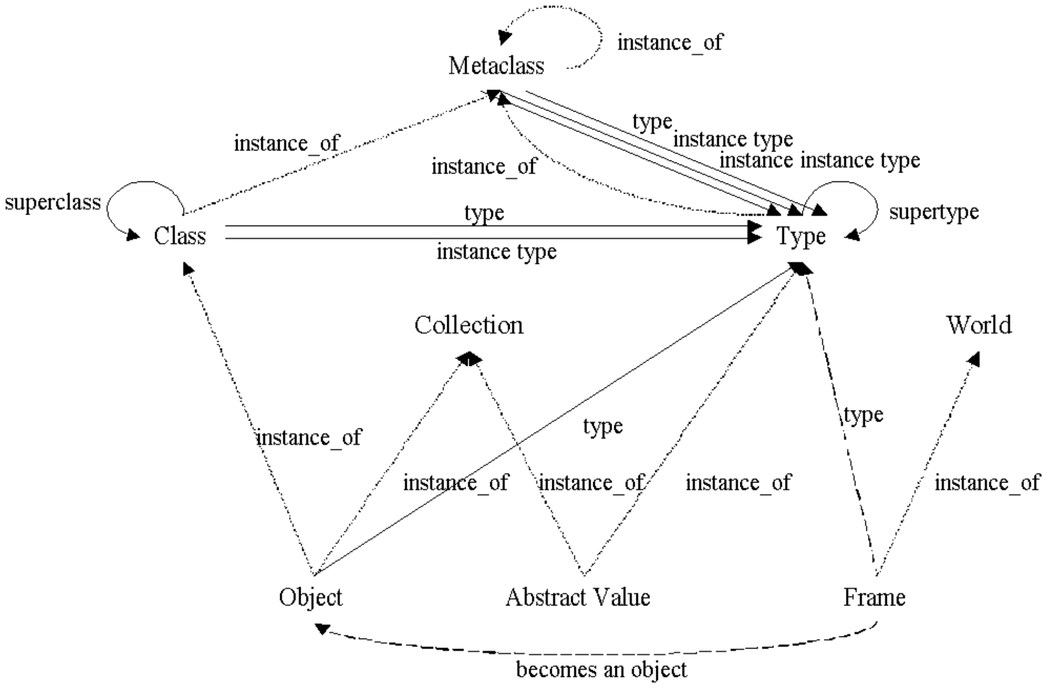

Figure 2: Canonical Model Entities

specific domain [25]. 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 as classes form class/subclass hierarchy. Rubrics from the thesaurus are mapped into the categories. Subrubrics form subclass hierarchy of categories. Categories contain concepts and lexical units as instances. Concept supplies the metabase with conceptual 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).

We assume the following procedure of the common thesaurus formation preserving autonomy of the source thesauri (vocabularies) of collections. During registration of a collection a subset of its thesaurus that is not included into the core is identified. This subset is represented on the mediator's level as an indexing extension of the core. For the remainder of the collection thesaurus a mapping of the respective lexical units of the core into the lexical units of the remainder is formed and is keeped at the mediator's level. The procedure is applied for each new registration considering the core together with all extensions formed so far. As the result, we obtain thesaurus federation (loosely integrated thesauri). The common core thesaurus with its extensions will be further referred to as the COMMON Thesaurus.

At the same time, the core theasaurus and its extensions are keeped classified into the categories using the metainformation structure discussed above.

4.4 Uniform semistructured metainformation representation

To work with semistructured data the following facilities of the canonical model can be used:

- collections of self-defined objects or collections of frames (worlds) without pre-defined associations;

- worlds with pre-defined frame associations;

- classes containing partially typed objects self-defining their own individual attributes that were not specified in a type of the class instances;

- classes of aggregates (associations of objects and frames).

For the Web, each page may be represented by a frame embedding sub-frames representing the fragments of the page. An instance of a world is a collection of frames that can be reached through hypertext-based frame links from the 'root' frame referenced by URLs. These frames can be structured differently (as frames with slots or without slots or combining both approaches). Hybrid object/semistructured data can be represented as it is explained in [23].

5 MULTILEVEL METAINFORMATION REPRESENTATION AND MODULARIZATION

5.1 Structuring homogeneous metainformation for various digital collections in the canonical model

The unifying canonical layer of digital collection specification is splitted into three sublayers - local, interoperation and personalized ones (Fig. 3). Local sublayer provides metainformation corresponding to each digital collection in homogeneous, canonical model form (collections corresponding to the canonical model specifications will be called *virtual*). Interoperation level specifies federated schemas intended for unified access to multiple collections with interrelated data as to a subject domain. Personalized level represents subsets of the interoperation level metainformation and resource abstractions reflecting interests of specific users and user groups. For instance, specific views can be introduced on this level: users can prefer the Dublin Core view above MARC actually used in specific collections, or they can personalize digital picture galleries crossing boundaries of real muzeum collections.

Modules of digital collection descriptions are defined on the local and interoperation levels. Module specifications are given by means of the canonical model providing for the definition of different kinds of module sections: the frame section, the type section, the function section, the information resource specification section (specific collections are defined in the latter section). Arbitrary combinations of sections are allowed. 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 specifications of types are defined. The canonical model supports a comprehensive type system based on a recursive composition of type constructors. Functions are used for support of methods, assertions, derivable entities.

Types are organized into a subtyping hierarchy providing for multiple inheritance of type specifications. Classes (that are introduced in the resource specification section) combine properties of a type and of a set and form a multiple hierarchy introducing generalization/specialization relationships. Orthogonal to the subtyping relationship for types and to the generalization relationship for classes is the classification relationship. Being object themselves, types and classes can become instances of another, more general classes called metatypes for type classification and metaclasses for class classification. Thus multilevel classifications can be formed. It is essential that one and the same object can belong to several classes (one class to several metaclasses). Metaclasses are useful for introducing generic concepts common for several attributes, types, classes. Thus metaclasses provide for better structured descriptions of the application domains and of the digital collections.

A kind of a collection (e.g., database collection, knowledge base collection, multi-media collection, unstructured data (textual) collection that is represented by a class or a world) can be declared by metaclasses. If required the attributes of an object can be declared to possess all properties of objects themselves. These objects become instances of the attribute categories introduced by means of metaclasses of associations and of generic attributes of metaclasses. Such attributes can take concrete forms in classes that are instances of a given metaclass.

A set of digital collection specification modules can be combined into schemas. A specification of any resource in the canonical model takes form of a frame. An information resource specification module is a named collection of frames (a world). For the context formation it is sufficient to include into a context an arbitrary world of frames alongside with the information resource specification modules. Using the canonical language facilities overviewed above it is possible to apply various approaches of structuring information resource specifications.

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 (federated, 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 created by integration and registration of schemas of subject domains of lower levels. Syntactically a hierarchy of subject domains is established by import of the respective schemas.

5.2 Scalability measures

To reach the mediator's scalability with respect to the number of collections that can be potentially registered, two separate phases of the mediator's functioning are distinguished: *consolidation* phase and *operational* phase.

The *consolidation phase* is intended for definition and integration of metadata coming during the process of registration of well-established, *representative collections* in the mediator's subject.

First assumption behind this idea is that the number of representative collections is *small* comparing to the total number of collections in the mediator's subject and such representative collections are identifiable. *Saturation of metainformation* is assumed: after consolidation, new collections in the subject to be registered do not contribute significantly to the metadata consolidated so far.

The consolidation phase is a mixture of the mediator's *metainformation design and integration* in process of representative collection registration. Collaboration of the groups (providers) supporting the representative collections is essential.

Second assumption is that the metainformation defined during the consolidation phase is *fixed and remains valid* for significant period of time - the operational phase of the mediator. At least the metainformation cannot 'shrink' during this period of time though new metadata can be added (if required).

During *the operational phase* the burden of the registration process is imposed on the collection providers. They formulate collections capabilities (schemas, vocabularies, query languages) in terms of the subject mediator's metainformation and develop the required wrappers. This is the way how the scalability is planned to reach.

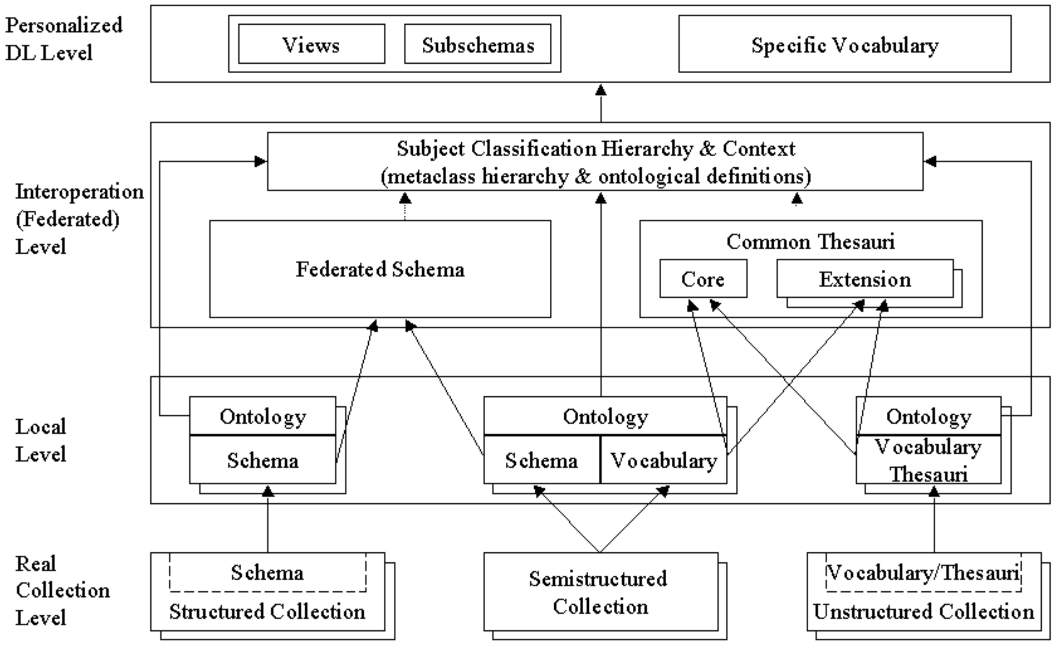

Figure 3: Mediator's Metainformation Layering

5.3 Formation of the interoperation (federated) level of metainformation

The metainformation of the interoperation (federated) level is formed during consolidation phase above the local level (Fig. 3).

Subject classification hierarchy The subject classifying hierarchy is formed to structure the subject domain of the mediator. The hierarchy is assumed to be an acyclic graph. Each class defines certain subject category. Instances of a class could be either other categories or specifications of concepts:

- lexical units of the COMMON Thesaurus;
- type specifications in various modules of the local level;
- class and world specifications in various modules of the local level introducing specific digital collections;
- frame difinitions of the local level modules introducing knowledge base components or software services.

Local context reconciliation Another important action to form the metainformation of the interoperation level is the reconciliation of local name definitions used in independent local collections. We assume that on the local level we have natural language definitions of all names provided in metainformation (in schemas or thesauri (vocabularies)). More formal ontological definitions related to the names can be also introduced. We assume existence of man-machine reconciliation procedure applied during the collection registration phase (similar to that of defined in [9]).

Interoperation level type hierarchy On the interoperation level the common type hierarchy is formed integrating the type hierarchies of the individual collections.

Interoperation level class generalization hierarachy is formed to constraint class extensions belonging to different local collections in terms of classes of the federated level.

5.4 Personalization

Personalized metainformation level is needed to present the personalized virtual digital libraries formed for specific users and groups. Specific modeling facilities are needed to formulate the requirements for the personalized digital libraries. These facilities will be based on the canonical model.

Design of the personalized digital libraries is planned as a process of looking for the interoperation level mediator's specifications and their fragments that can be composed into a refinement of the requirements for the personalized digital library.

Personalized information resources can be formed applying various approaches, e.g.:

- taking a subset of the interoperation level metainformation reflecting the interests of the respective user group;
- introducing specific vocabularies (ontologies) more closely related to specific user groups, providing mapping of the vocabularies into the COMMON Thesaurus;
- introducing views above the interoperation level reflecting information needs of specific user groups.

6 BASIC FUNCTIONS OF THE MEDIATOR

Basic functions of mediators include:

DL Personalization: a process of communication of a group of persons with the subject mediator to arrange a personalization service; the process is supported by a generic facility providing for creation of personalized DL specification,

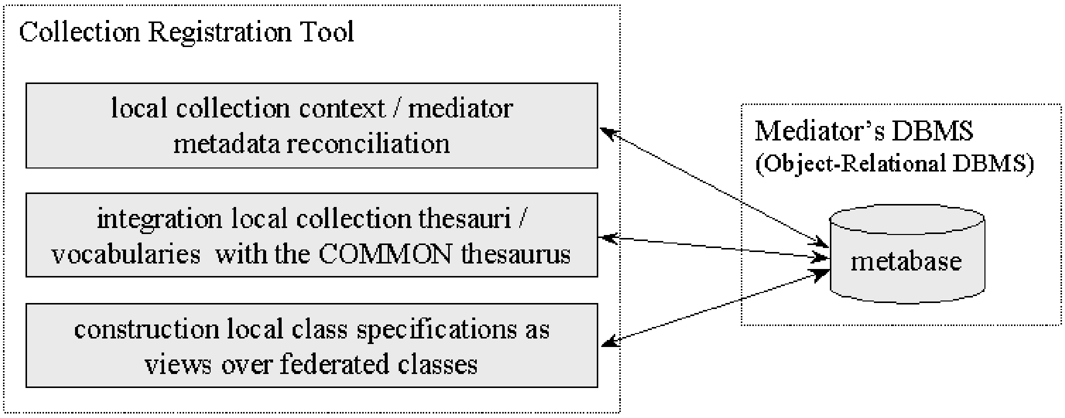
Collection Registration Tool
local collection context / mediator metadata reconciliation
integration local collection thesauri / vocabularies with the COMMON thesaurus
construction local class specifications as views over federated classes
Mediator's DBMS
(Object-Relational DBMS)
metabase

Figure 4: Structure of the collection metainformation contextualizing tool

metainformation definition, query formulation facilities, information processing and editing facilities and for definition of their interpretation by canonical models, languages, digital collections and services known to the mediator.

Collection Registration: a process of communication of a digital collection provider with the subject mediator during the operational phase to advertise the collection information and capabilities, to support processes of the local collection contextualizing, integration, reconciliation, wrapper generation. The registration process is supported by a facility providing for mapping of a collection data model and metainformation into the canonical ones, for integration of the new metainformation with the mediator's one, for inferring of the required information (e.g., classification hierarchy (re)definitions based on the actual collection texts, providing term/frequency/weight information for lexical units, qualitative ontology formation), for semi-automatic construction of a wrapper, for connecting the wrapper to the interoperation environment.

Collection Contextualizing: a process of local collection context/mediator metainformation reconciliation; introducing new definitions into subject classifier hierarchy, COMMON thesaurus, ontologies; specifying the collection types/ classes in terms of the mediator's type/class hierarchy. The process is supported by the specific federated level collection contextualizing tool.

Information Extraction: a process of canonical query planning and execution supported by the mediator's query engine that includes functions of identification of relevant collections, query decomposition, query planning and monitoring facilities.

Outcome Presentation: a process of preparation of the information extracted as a result of the federated level or a personalized DL query supported by services for information objects segmentation, aggregation, composition.

Tools and services supporting these mediator functions are arranged into several mediator frameworks:

1) Collection registration framework;

2) Information extraction framework;

3) Personalization framework.

These frameworks are considered in more details in the next section.

7 MEDIATOR'S FRAMEWORKS

7.1 Collection registration framework

The framework facilities are intended to support functions of collection contextualizing:

- constructing mapping of a collection data model and metadata into the canonical ones,

- representation of the new metainformation in terms of the federated mediator's level,

- inferring from the collection the required information for the federated level,

- semi-automatic construction of a collection wrapper,

- connecting the wrapper to the interoperation environment (e.g., CORBA).

In particular, the process of registraion is planned to be supported by the metainformation contextualizing tool (Fig. 4) that should provide:

- local collection context / mediator metainformation reconciliation,

- introducing new definitions into mediator's subject classifier hierarchy, COMMON thesaurus, ontologies,

- representation of the collection type/class specifications in terms of the mediator's type/class hierarchy.

The tool is based on the prototype for component-based information system development in the heterogeneous interoperable environment [9].

During the registration, a local collection class is modelled as a set of instances (objects) of the class instance type, and the description of the collection in terms of the federated schema specifies the constraints on the instances that can be found in this class. Formally, the content of a local collection class is described by a canonical model formulae simplified as $C(z) \subseteq \exists \overline{x}((C_1(\overline{x}_1)$ & $C_2(\overline{x}_2)$ &...& $C_n(\overline{x}_n)$ & $Con)$ where C is a local collection class, C_1, ..., C_n are federated schema classes, z is a reduct of the local class instance type being a concretization of a reduct of resulting instance type of the conjunctive formula, Con is additional constraints imposed by formula. That is, the defined reduct of an instance obtained from the local collection class should satisfy the constraint expressed by the formula. Of course, this description does not imply that the local collection contains all the instances that satisfy the formula. Obtaining of the definition of a local collection class as federated level queries means that it is not required to add the local class to the federated level on registration.

General idea of such representation of the local classes is similar to those proposed in [29]. Main differences of the current approach consist in taking into account issues more relevant to real environments, such as using general type model and type specification calculus, applying of the refining mapping of the local specific data models into the canonical model of the mediator, resolving ontological differences between federated and local concepts, systematic resolving structural, behavioral and value conflicts of local and federated 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 classes registered on a local level. 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 collection. 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 query decomposition into the child's mediator subqueries.

Thus, basic decisions to be incorporated into the collection registration framework include:

1. Support of the metainformation canonical model suitable for the uniform representation of broad spectrum of heterogeneous information contained in various digital collections and libraries (structured, semistructured and unstructured data, behaviours, knowledge);

2. Extending mediator's contextual metainformation representing lexical unit categories and relationships in thesauri, ontological and classifying definitions, typing information (including constraints and behaviours if required);

3. Basing mapping of a collection data model and metadata into the canonical ones using the *principle of the data model refinement* [22] that leads to preservation information and operations of the original digital collections while forming their homogeneous local metadata representations during the mediator's registration process;

4. Representation of the collection type specifications in terms of the mediator's type hierarchy using specific tool.

7.2 Information extraction framework

The framework facilities are intended to support the information extraction functions (Fig. 5) from multiple collections and presentation production for users, including:

- graphical user-friendly facilities providing integrated DL users support applying the canonical query language directly on the federated level bypassing the personalized level;

- mediator's query engine support including functions of identification of the best relevant collections, query planning and monitoring facilities;

- services for suitable presentation of the outcome including information objects segmentation, aggregation, merging.

Basic decisions to be incorporated into the information extraction framework include:

1. rich canonical query language for the mediator should support querying of various kinds of information in an integrated form (including textual models, semistructured models, structured and object-oriented models), the language should be well agreed with the canonical information model chosen;

2. best relevant collection identification based on the class subsumption principle taking into account term frequency/weight vector functions [17];

3. query planning based on representation of local collections in terms of the federated level and query containment reasoning;

4. approaches for the information objects segmentation, aggregation, ranking, composition, adequate to the rich information model used (in particular, merging the query results from multiple sources into a single collection, establishing meaningful rank term statistics and document score; aggregation of structured and unstructured resulting segments into integrated information objects should be established).

7.3 Personalization framework

The framework facilities are intended to support functions of DL personalization (we assume that personalized metainformation models, query formulation facilities, information processing and presentation facilities as well as personalized DL requirements (including ontologies, schemas, thesauri) in chosen metadata models are specified outside of the mediator), including:

- definition of the DL requirements interpretation by canonical models, languages, metainformation, digital collections and services known to the mediator,

- semi-automatic generation of interpretors supporting the personalized level models and functionalities by the federated level of the mediator,

- establishing a connection between a specific personalized DL clients and the mediator through the interoperation environment.

Basic principle of the interpretation to be established consists in reaching the refinement of the personalized DL requirements by the collections and metadata of the federated and local mediator's levels. More details on how personalization framework is formed can be found in [24].

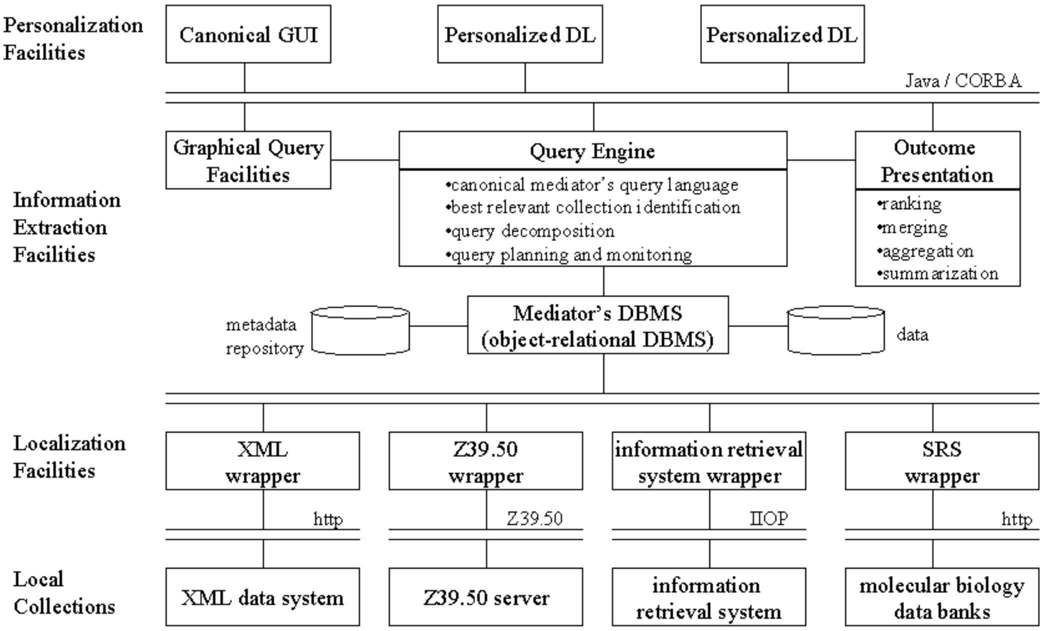
Personalization Facilities
Canonical GUI
Personalized DL
Personalized DL
Java / CORBA
Information Extraction Facilities
Graphical Query Facilities
Query Engine
•canonical mediator's query language
•best relevant collection identification
•query decomposition
•query planning and monitoring
Outcome Presentation
•ranking
•merging
•aggregation
•summarization
metadata repository
Mediator's DBMS (object-relational DBMS)
data
Localization Facilities
XML wrapper
Z39.50 wrapper
information retrieval system wrapper
SRS wrapper
http
Z39.50
IIOP
http
Local Collections
XML data system
Z39.50 server
information retrieval system
molecular biology data banks

Figure 5: Information Extraction Framework

8 HETEROGENEOUS INFORMATION MEDIATORS: STATE OF THE ART

Here main trends of distributed DL development related to mediation of heterogeneous information sources are traced. Among approaches for networked digital libraries are:

- wide network of digital libraries which enhance connectivity across document repositories and provide a quasi standardised access;

- community-oriented digital libraries which appear as a common need;

- broker-based architectures;

- heterogeneous information mediators.

INFOBUS: The Stanford University DL Project, "Interoperation Mechanisms among heterogeneous services" [33] was focused mainly on the core level of interoperability. The CORBA distributed object framework is used as the basic interoperation mechanism [33] to form the information bus (InfoBus). Various data sources conform to various protocols (e.g., HTTP, Z39.50, SQL). *Wrappers* (proxies according to the Stanford projects) are incorporated above any source to provide uniform protocols. E.g., for information services the uniform interactions include login, query submission, result transmission and so on.

The Networked Computer Science Technical Report Library (NCSTRL) [28] is an operational digital library with a distributed, component-based architecture based on the Dienst federated digital library architecture developed as part of the DARPA Computer Science Technical Reports Project. In Dienst one of the sites run an (implicit) mediator service thus defining the set of sites that make up the collection and dispatching queries to appropriate index servers.

Implicit query mediators serve as intermediaries between the user interface and indexer services of digital libraries, translating queries to indexer protocols, choosing indexers to field the query, aggregating search results, and adapting to operational conditions.

Community-oriented digital library is a new direction providing for development of collections of documents built by a community of users which aims at observing or studying a specific subject (e.g., in an environmental or socio-environmental context). WISEDL (Web Integrated System for Environmental Digital Libraries, France) is a digital library architecture project addressing the needs of adaptability of digital libraries built and maintained by communities which have the same interest.

Subject/Information Gateways: global digital library approach leads to gateways which provide access to collection of metadata and eventually data. Academic libraries and institutions are currently looking for ways to help their users discover high quality information on the Internet in a quick and effective way. As Internet publishing and communication become more commonplace low quality of Web search for information could disadvantage researchers as they will miss valuable information and communication resources.

Subject gateways are facilities that allow easier access to network-based information resources in a defined subject area. Subject gateways offer a system consisting of a database and various indexes that can be searched through a Web- based interface. Each entry in the database contains information about a network-based resource, such as a Web page, Web site or document. A description of each resource is provided to help users assess its origin, content and nature, enabling them to decide if it is worth investigating further.

Information gateways are quality controlled information services with the following characteristics:

1. an online service that provides links to numerous other sites or documents on the Internet;

2. selection of resources in an intellectual process according to published quality and scope criteria;

3. intellectually produced content descriptions, in the spectrum between short annotation and review. A good but not necessary criterion is the existence of intellectually assigned keywords or controlled terms;

4. intellectually constructed browsing structure/classification (this excludes completely unstructured lists of links);

5. at least partly, manually generated (bibliographic) metadata for the individual resources.

Renardus project [34] is a project for Subject Gateway Service Europe. Broker service (a middleware enabling the integration of distributed and heterogeneous information resources) is investigated. Federated resources from underlying heterogeneous resources and mediating access to it should be provided. The Renardus project implements a Europe wide Internet information gateway service based on a generic broker architecture and data model that will allow the integrated searching and browsing of distributed resource collections. Eighteen broker architectures that have been developed for existing services and projects are taken into account. Renardus Models Information Architecture (MIA) is a layered architecture with five layers: presenter, coordinator, mediator, communicator and provider.

Renardus partners include subject gateways DuchESS (The Netherlands), NOVAgate (Nordic countries), EELS (Sweden), DEF (Denmark), DAINet (Germany), FVL (Finland), Les Signets (France), RDN (United Kingdom), SSG-FI (Germany). All gateways are operational. Records of all gateways are created by subject specialists and/or librarians, guaranteeing the steady growth of high quality resources. Cross-seraching is planned in engineering, humanities, forestry, agriculture, mathematics. The majority of resources are Web sites. In Renardus searching is preferred to browsing. Different metadata schemes are supported - only two gateways use metadata based on DC. All gateways support Title, Creator, Description, Identifier that can be mapped to a common format.

These projects are good for resource discovery where quite limited metadata might be sufficient. For information search in well developed subject domains including mixture of textual, multimedia, semistructured and object-oriented data (as, e.g., we can find in [6, 36]) much more comprehensive metainformation and mediator's middleware capabilties are required. Several advanced architectures are discussed below.

A primary distinction between mediators can be established as:

- integration information from *pre-selected sources* according to the known information needs. When information needs or sources change, a new mediator should be generated;

- integration information from *arbitrary sources* according to the *predefined information needs.* A declarative approach is known (Information Manifold, InfoSleuth). Mediators contain mechanisms to rewrite queries according to source descriptions. A rewritten query should be contained in the original query.

In *TSIMMIS (The Stanford-IBM Manager of Multiple Information Sources)* [19] mediators are built above a given set of sources with wrappers that export OEM self-describing objects. OEM (Object Exchange Model) is used as a unifying data model. The mediators considered provide integrated OEM views of the underlying information (e.g., if a relational source is considered, it is exported as a set of OEM objects.)

Mediators are specified with MSL (Mediator Specification Language) that can be seen as a view definition language and is a logic-based object-oriented language targeted to OEM. A common query language links TSIMMIS components. MSL is used as query language, as the mediator specification language and query language for wrappers.

In the *Information Manifold* [29] a reasoning phase is required for realizing which sources have the data of interest. The user interacts with a uniform interface in the form of a set of global relations (the mediated schema) used in formulating queries. The actual data is stored in external sources. To answer queries, a mapping between the relations in the mediated schema and the sources must be specified. A method to specify these mappings is to describe each source as the result of a conjunctive query over the relations in the mediated schema.

Given a user query formulated in terms of the relations in the mediated schema, the system must translate it to a query that mentions only the source relations and is a maximally contained plan. The collection of available data sources may not contain all the information needed to answer a query. The Information Manifold provides uniform access to structured information sources on the Web. Query containment is reduced to the problem of finding a solution in terms of views that must be contained in the original query. Description logics can be used as a data modeling language and as a query language. This is a tradeoff between complexity and expressive power.

COIN Project (MIT): COntext INterchange Mediator [8] in which *semantic conflicts* are detected and *reconciliated* by a *context mediator* through comparison of contexts. The *shared ontology* is defined as a set of *propositions* and *deductive relationships* between them. Propositions describe things, attributes and states in the first order predicate language. *Deductive rules* define admissible *conversions* between propositions (including generalization and aggregation-based rules, unit and scale conversion rules, time and space hierarchy conversion rules). *Context mediation* according to this approach is equivalent to *transforming propositions from a source, through deduction, to propositions that satisfy the requirements* of the receiver's context. *Semantic interoperability* is considered on a *type signature level* (all definitions are expressed in a Prolog-like manner).

Mediator envirOnment for Multiple Information Sources (MOMIS) has been developed by several universities in Italy [7]. *A common thesaurus* is constructed which has the role of a shared ontology for the information sources. The knowledge in the Common Thesaurus is then exploited for the identification of *semantically related information* in structured and semi-structured sources and for their integration at the global level. Common object-oriented data model combined with ODB-Tools supporting description logics are used for analysis of sources descriptions for generating a consistent common thesaurus.

Standardization is increasing at different levels of information systems architecture for basic classes of heterogeneity – *Information Heterogeneity* (Semantic, Structural, Representation, Syntactic) and *System heterogeneity* (DBMSs, data models, system capabilities, etc.) [31]:

- *System heterogeneity:* IIOP for interactions between distributed objects and components, KQML for interactions between agents;
- *Syntactic heterogeneity:* XML for all forms of Web-accessible data;
- *Structural heterogeneity:* RDF for general purpose description of information sources, various object models for web-based information exchange, KIF for knowledge representation, OKBC for distributed knowledge bases;
- *Semantic heterogeneity:* work is in progress to support limited forms of semantics with identification of contexts, objective requirments and applications.

9 CONCLUSION

Brief analysis of the heterogeneous digital collection mediator infrastructure is presented. The infrastructure is defined as a set of functionally-oriented frameworks: collection registration framework, information extraction framework, personalization framework.

The mediator considered is distinguished from the other known works with the comprehensive information and metadata models aiming at uniform representation of broad range of heterogeneous digital collections, specific approach for homogenization of heterogeneous information models in the canonical paradigm, specific methods for collection registration and federated metainformation consolidation. Collection registration and personalized DL level formation are treated as compositional development methods.

A prototype of the basic mediation functions is being developed using textual document sources (IRBIS), XML databases (Tamino), bio-molecular sources (SRS data banks), Z39.50 collections. Poly-thematic thesaurus for science and technology is used as the core of the common thesaurus of the federated level.

References

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

[2] Abiteboul S. et al. The Lorel query language for semistructured data. International Journal on Digital Libraries, v.1, N 1, April 1997

[3] Adam N.R., Vijayalakshmi A., Adiwijaya I. System Integration in Digital Libraries. CACM, vol. 43, N 6., June 2000

[4] Arocena G., Mendelzon A. WebOQL: Restructuring documents, databases and Webs In: Proceeedings of ICDE'98, February 1998, Orlando, Florida

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

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

[7] S. Bergamaschi, S. Castano and M. Vincini. Semantic Integration of Semistructured and Structured Data Sources. SIGMOD Record, Volume 28, Number 1, March 1999

[8] S. Bressan, et al The COntext INterchange Mediator Prototype. Proceedings ACM SIGMOD Conference, Tucson, 1997

[9] Briukhov D., Kalinichenko L. Component-based information systems development tool supporting the SYNTHESIS design method. Springer LNCS, *Proceedings of the East European Symposium on "Advances in Databases and Information Systems",* September 1998, Poland

[10] Business Object Component Architecture (BOCA), Revision 1.1, OMG Document bom/98-01-07

[11] Carey M., et al Towards heterogeneous multimedia information systems: the Garlic approach. IBM RJ 9911

[12] Eisenberg A., Melton J. SQL:1999, formerly known as SQL3. ACM SIGMOD Record, Volume 28, Number 1, March 1999

[13] G.Gardarin, et al IRO-DB: a distributed system federating object and relational databases. In: Object-Oriented Multidatabase Systems: A Solution for Advanced Applications (O. Bukhres, A. Elmagarmid Eds.), Prentice Hall, 1995

[14] M. Genesereth and R. Fikes, "Knowledge Interchange Format Reference Manual", 1994, <http://logic.stanford.edu/sharing/papers/kif.ps>.

[15] Chaudhri V., et al Open knowledge base commectivity 2.0.2 Stanford University, February 1998

[16] Document Object Model (DOM) Level 1 Specification. Version 1.0. W3C Recommendation, 1 October 1998.

[17] Gravano L. Querying Multiple Document Collections Across the Internet. Ph.D. Dissertation. Stanford University. August 1997

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

[19] J. Hammer, H. Garcia-Molina, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. "Information Translation, Mediation, and Mosaic-Based Browsing in the TSIMMIS System". In Exhibits Program of the Proceedings of the ACM SIGMOD International Conference on Management of Data, page 483, San Jose, California, June 1995.

[20] Extensible Markup Language (XML) 1.0. W3C Recommendation, Feb 1998. http://www.w3.org/TR/REC-xml

[21] Kalinichenko L.A. 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, 1995, 103 p.

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

[23] Kalinichenko L.A., Integration of heterogeneous semistructured data models in the canonical one. In Proceedings of First Russian National Conference on "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Saint-Petersburg, October 1999

[24] Kalinichenko L.A., Skvortsov N.A., Brioukhov D.O., Kravchenko D.V., Chaban I.A. Designing personalized digital libraries over Web-sites with semistructured data. Programmirovanie and Computer Software, Vol. 26, N 3, 2000

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

[26] Kramer R., Nikolai R., Habeck C. Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies International Journal on Digital Libraries, v.1, N 2, September 1997

[27] O. Lassila and R. Swick (eds.), Resource Description Framework (RDF) Model and Syntax <http://www.w3.org/TR/WD-rdf-syntax/>.

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

[29] Levy A.Y., Rajaraman A., Ordille J.J. Querying heterogeneous information sources using source descriptions. Proceedings of the 22nd VLDB Conference, 1996

[30] The Object Database Standard: ODMG 2.0. Ed. by R.G.G. Cattell, D.K. Barry, Morgan Kaufmann Publ., 1997

[31] A. Ouksel and A. Sheth. A brief introduction to the research area and the special section. SIGMOD Record, Volume 28, Number 1, March 1999

[32] Object Management Group, "The Common Object Request Broker: Architecture and Specification". Revision 2.0. July 1995.

[33] Paepcke A., et al. Using Distributed Objects for Digital Library Interoperability. Computer, May 1996

[34] Renardus evaluation report of existing broker models. Http://www.renardus.org/deliverables/D1_1_final.doc

[35] Scheck H.-J., Birmingham B. 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

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

[37] Weibel Sl., Lagoze C. An element set to support resource discovery. The state of Dublin Core: January 1997, International Journal on Digital Libraries, v.1, N 2, September 1997