

Integrated System of Information Resources of RAS

1. System Requirements
2. ISIR as a digital library
3. Digital library basic principles
4. Relations between resources
5. Overview of the ISIR general organization
6. Metamodel and the metarepository
7. Distribution
8. Searching
9. Information interoperability (basis for publication, cooperation, collaboration)
10. Implementation
11. References
12. Appendix A. Dublin Core specification of *Database* resource
13. Appendix B. *Database* resource definition
14. Appendix C. Search parameters

1. System Requirements

Integrated Information System of the CCE/NIS (IIS CCE/NIS) is aimed to collect of various information submitted by EIS project partners. The EIS system architecture is based on the design concepts and framework of the Information System of Integrated Resources (ISIR) [20] elaborated by the partner RAS experienced in developing of such information systems. The ISIR project is developed the conceptual ground and infrastructure for integrating various informational and computational resources into a unified information space.

The project is aimed to provide an integrity of different kinds of information resources and systems employing both existing approaches and open architecture technology of ISIR. The flexibility of distributed information organization, its integrated presentation for end-users, the open architecture of the system are the key points in the system development.

The following requirements have been laid to the system:

- *Logical grouping of data* - the system is to allow to process all the queries over the logical groups of database, thus completely hiding the physical allocation of the latter.
- *Abstract data model* – the information system is build on the basis of abstract data scheme, so that concrete databases are mapped onto it. This enables to join data from various systems into a single logical group.
- *Abstract query system* - the system is not to perform the definite query syntax; it has to deal with its logical essence making use of abstract features of information resources.
- *Metainformation* - the system has to provide full information of itself and about all its resources, its services should be metainformation driven to provide system flexibility and scalability.
- *Access control* - the system is to be capable to provide users with different privilege levels in accessing the information.
- *Registration and management* - the system should be able to collect statistics on users' queries and to manage their budgets.

- *Working with distributed data* - the information system is to enable to operate with data allocated on various physical servers, various hardware and program platforms and kept in different internal formats.
- *Resource integration* - the system is to make it possible to integrate its own resources with resources of other information systems.
- *User-friendly interface* - the system has to supply simplicity of users interfaces to access the information.

2. ISIR as a digital library

We consider ISIR as a digital library because like other digital libraries the system is intended to store and provide for the interaction with large distributed arrays of various resources. Like digital libraries the system is aimed to meet the following user requirements:

1. Storing different information in electronic form (text, scanned text, graphic images, sound, video, programs). The system is to provide users with loading mechanisms for all kinds of information.
2. Distributing the information over the network.
3. Searching over the distributed information.
4. Presenting the information.
5. Keeping access rights and information security.
6. Mechanisms for loading, searching, and displaying the information has to maintain the interaction with various digital libraries (both in the country and abroad).

However, we interpret the concept of a digital library, its subject and features in a wider way. We consider, the contents of library make not only information but as **resources** which may be both informational and program (dataStreams[10], elements[11]), and computational ones. Resources are of different types, rather than being only documents (in a general sense) or even multimedia resources with complicated description and structure. These various resources should make up the integrated space of a library. They are in a close interrelation - they may characterize one another and affect each other. Our standpoint is that the functionality of a library when accessing resources is not exhausted by granting the information (retrieval of the information and its representation). Users may have an opportunity to **interact** with the resources discovered. To put it more exactly, users may interact with the objects of the discovered resources in accordance with their open interfaces. This fact may be employed by the components of a library to extend the services supplied by the system.

The conceptual model of our system does not break the basic principles and concepts of a digital library [1, 2] that are:

- *digital object* - the data structure, principal components of which are *data* and *metadata*;
- *handle* - an identifier globally unique to the digital object;
- *repository* - a network-accessible system storing digital objects for subsequent access and retrieval.

On the one hand the system extends the concepts above and on the other hand it makes them more certain so that they would present not only distributed collections of digital objects but would provide the distributed integrated space of digital resources which are not only stored all together but also *characterize* one another and, moreover, which are not only sequences of bits but also *functional components* that capable to enlarge the standard library services, provide users with program service or computing resources they need.

3. Digital library basic principles

two parts - *metadata* (the information about the resource) and, probably, *data* (the content of the resource). Metadata is information about the resource that enables to manage the resource, to carry out information retrieval and interaction with the resource. Metadata allows information not just of human browsing but for machine understanding: searching, reasoning and analyzing. The resources are identified by globally unique persistent name - handle/URN [3,4,5,6] - it is metadata too. Interpreting resource content is performed on the basis of its type or program interface described in metadata. Library resources are not uniformed. They have different attributes and supply different services. Each resource belongs to a certain type that is called the *resource type* and defines the following:

- the structure and the composition of metadata - the sort of information about the resources of a certain type, their services;
- the way of internal organization of the resource;
- interaction with (relation to) other resources;
- resource search mechanisms and the way of representing the information about the resource;
- the way how the resource interact with users.

Resources are closely *related with* one another that is due to both their mutual characteristics (a person is an author of a publication prepared in a single project framework which is invested by an organization) and mutual actions (a compiler carries out the compilation of a program, a program runs under OS and calls processes while running). *Relations* are not similar and have certain *relation types* like the resources. Choosing a set of resource types depends on orientation of digital library, its subject domain.

The digit library system ought to maintain the following functionality:

- providing the information about resources,
 - searching for the resources on the basis of metadata,
 - navigation in the space of related resources,
 - viewing the resource attributes,
- managing the resources,
 - depositing, storing metadata and resource contents from different sources and formats,
 - exporting metadata and resource contents into different sources and formats,
 - automated metadata preparing,
 - publishing, deleting resources, editing their attributes and, probably, the resource itself,
 - indexing of metadata to provide discovering resources,
 - maintaining the tools for identifying resources,
 - maintaining the actuality of resources,
 - managing the access to resources,
- interacting with resources
 - retrieval resources, providing the content of the resource,
 - maintaining and providing the resource functionality on the basis of open interfaces,
 - processing (possibly, analytical) of resource information,
- maintaining the mechanisms of distributing resources
 - accumulating and caching of metadata,
 - search query routing and broadcasting,
 - integrating and identifying resources from other systems,
 - providing actuality of the side-resources information,

- the conversion of search queries,
- integrating the results come from the respond to search queries,
- maintaining and providing better functionality on the basis of open interfaces.

Some of these functions that are considered the most important, are shown in Figure 1

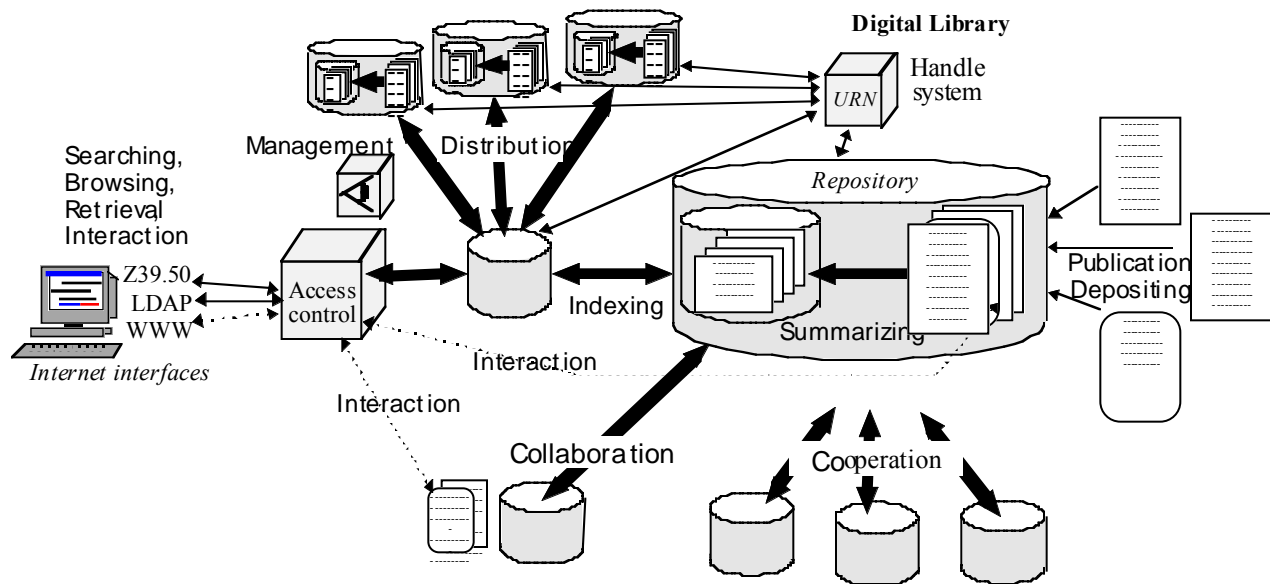


Figure 1

and defined as follows:

- *Publication* - the process of granting of a resource by some user, as a result of which other users can find it and to address to it
- *Depositing* - the process of storing a resource and/or its metadata in a digit library repository,
- *Summarizing* - the process of extracting the representative set of characteristic items of information (metadata) from a resource and/or its environment,
- *Indexing* - the process of converting the information obtained after summarizing into the form enabling the effective search,
- *Searching* - "index" processing aimed to form the respond to the user query according to distributed metadata.
- *Browsing* - process under the user's control aimed to examine distributed information space formed by the system on the base of resources and their relations.
- *Retrieval* - users' action aimed to retrieval the found information they need from the distributed information space.
- *Interaction* - process under the user's control and within the framework limited by the resource providing the user with information or functional service.
- *Distribution* - activity of the system aimed to maintain data processing, with data located on different physical servers, various hardware-software platforms.
- *Control* - activity of the system on providing different user access privileges, data security and on observing users' budgetary limitations.
- *Management* - the action on accompanying the system, observing the information actuality, integrity and safety within the distributed information space.

and effective search processing.

- *Cooperation* - the system has to be capable to co-operate with different systems while processing user search queries; the search queries for those systems are converted and the results of responds to queries are integrated by the system under consideration.

The actions *collaboration* and *cooperation* are to provide interoperability of digital library systems through the definition of open interfaces of digital library services, allowing flexible interaction with existing systems.

4. Relations between resources

For efficiency in query processing, the resources in a digital library should be organized properly. Within the library, resources may have different forms. Metadata enables to manage the resource, to carry out information retrieval and interaction with the resource. There are a lot of metadata formats that describe various kinds of the information - publications, programs, chemical formulas, etc. The issues to be addressed in analyzing these formats, that their elements have the different nature - one of elements represent internal properties of a described resource and others designate other resources or, more precisely, interrelations with them. The ideas in structuring metadata already included

- “**Key metadata** that is used to manage the object in a networked environment” [1].
- “A **data type** that describes technical properties of data, such as format, or method of processing” [11].
- “**Structural metadata** that describes the types, versions, relationships and other characteristics of digital materials” [11].

Furthermore, it is possible to make a conclusion about necessity of a storage and work with metadata not as an aggregate of the properties of *one* type of resources, but as with a set of the items of information about a *series* of the interrelated types of resources. The appropriate metadata processing (Figure 3) opens up many new opportunities for deriving new content from the same set of metadata. It enables

- to carry out more exact and more complex search requests,
- to provide the information not only about publications but also about the resources interrelated with them, and a lot of other information, for example, analytical one.

Adding less traditional resources types, such as organizations, projects, conferences, grants, sponsors, that are not usually included in bibliographic records, but frequently connected with the scientific publications, allows, for example, to receive the resource types scheme shown in Figure 2.

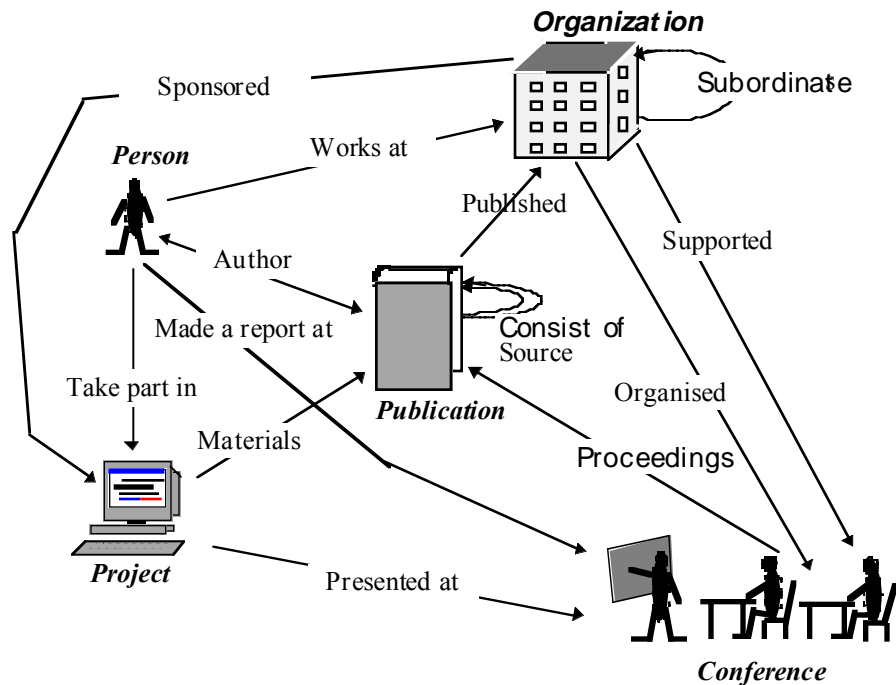


Figure 2

Such scheme of resource types and relations among them essentially extends opportunities of information system on providing various services. For example, it is possible to carry out the following kinds of queries:

- find out the projects in which the author of the found publication took part, and to view materials and publications of these projects,
- find out conferences, on which the author represented his (her) publications,
- find the publications of the colleagues of the author of different backgrounds,
- familiarize with materials those sections of conferences, on which the colleagues of the author reported on their work.

We can see, that the resources of digital libraries are in a close interrelation, which is due to both their mutual characteristics and mutual actions. They use each other and affect each other etc. The digital library infrastructure must exploit this capability. It is necessary to separate the description of relationships between resources from others metadata and to move them in other category of the special characteristics of resources to provide the appropriate interpretation and support. As result, we will have a facility by which we can integrate resources, instead of a facility for a mention of resources. For example, this means the system will have an opportunity to search on exactly what specified, instead of try to enumerate all resources with the same name.

Figure 3 schematically shows the construction and downloading of information on resources and relations between them on the basis of the analysis of traditional formats metadata. With static linking, all relations between resources can be computed, during loading. This is effective for a well-defined, frequently used relations, such as authors, publishers, proceedings. In general, not all relations can be or need to be precomputed. The dynamic linking allows to compute relations only when required for a user.

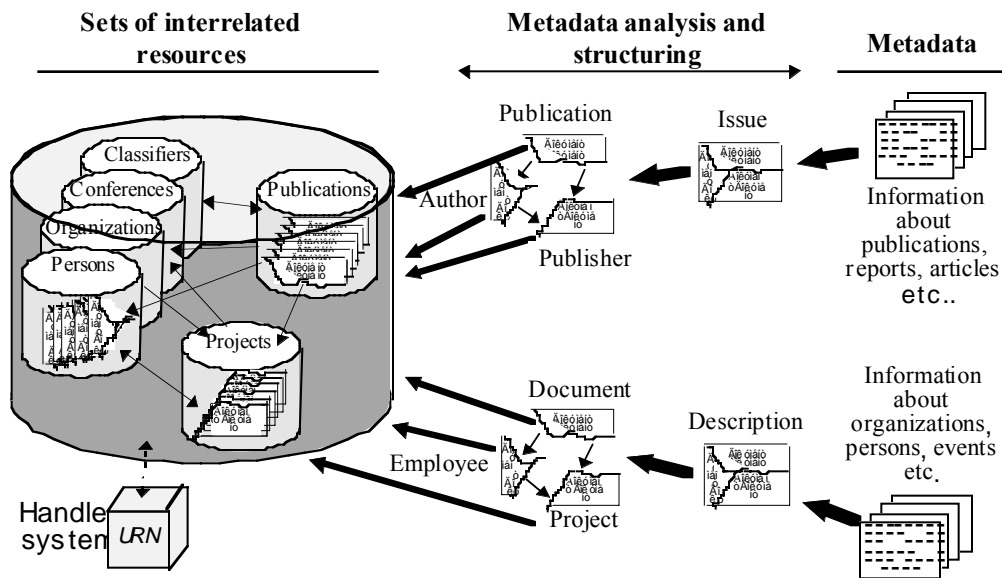


Figure 3

The conceptual data model including only concepts of resource types, properties (attributes) of resources and relations between resources can serve as basis of the implementation. The relations should be defined between types of resources (Figure 4). They peculiar for all resources of some type and give the flexibility to manage resources processing and their communication effectively, particularly in the case of a distributed environment.

The relations between resources, as well as the metadata elements make crossed groups, which are classified as follows:

- the relations, which can be used in search queries, for example, an author of a publication, a participant of a project.
- the relations that support interactive browsing on a resource-by-resource basis, that are used during the interaction with users to show the information about a resource, its relations with other resources, to demonstrate the information context of the resource.
- the relations associated with the management of and administration of the resource. These are relations providing performance of search queries, determining the access rights for supported categories of the users etc.

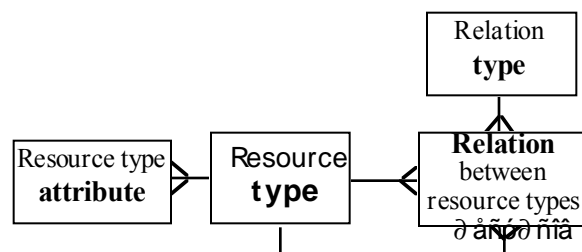


Figure 4

The library system should provide global, distributed search and retrieval of electronic resources. It should support navigation in the space of related resources. Hence, relations should be defined between the metadata required for "look-up" of resource (focused on providing search and navigation services), instead of defining them between resources directly. We can abstract from the

limited" search and connecting resources into a unified information space is not what a particular resource is used, what specific properties and internal structure it has. Resource should remain as "black box " for a search part of system, the rules of management and interaction with which are defined by others metadata (Figure 5). This is reasonably simple to implement, allows to simplify and to make the more effective distributed search and yet provides sufficient flexibility for many situations.

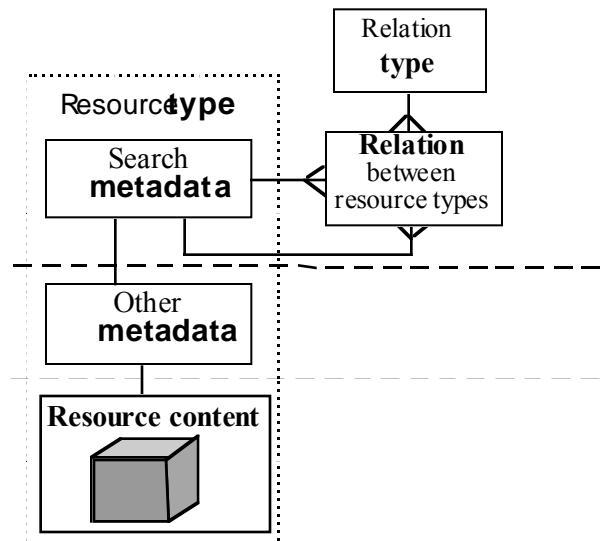


Figure 5

5. Overview of the ISIR general organization

In our implementation of this approach metadata of resources, relations between them are stored in a relation database. The resource content, if it exists, may be stored either in a relation database or in the file system depending on the resource type. The resource may have a few copies of the content, or their versions that may be organized in different formats.

The system supports the following functionality:

- depositing, storing metadata and resource contents from different sources and formats,
- exporting metadata and resource contents into different sources and formats,
- resource searching by metadata and relations between them,
- navigation in the space of related resources,
- editing the attributes of a resource and, probably, its content,
- managing the access to resources for user groups,
- distributing resources over the network,
- maintaining the actuality of resource data,
- maintaining the relations between resources.

The following elements of metadata are used in the system

- search metadata
- resource identification metadata
- viewing metadata for various visual forms
- metadata for managing the edit process
- metadata for managing access to resources

- metadata for supporting the interaction with a resource (standard support of some data formats, program interfaces).

Metadata for searching, viewing, identifying, actualizing are specific for different resources, e.g. each of the resources has its own set of values for the metadata elements mentioned above. Metadata for access control, editing, managing the search process, viewing, interaction support (e.g. setting parameters for standard system mechanisms) are associated with resource type. For most of the elements from the latter category the set of available values are determined, with the semantic supported by the system.

The parameterization of the digital library functionality by control metadata results in a single point of maintenance and deployment. It significantly reduces complexity and ultimately results in a lower total cost of ownership.

6. Metamodel and the metarepository

Information about resources supported by the system (Figure 6), their attributes, and relations between resources is kept in so-called *system metarepository*. This is a special database that contains a description of the information model of an application area and some additional parameters. The metarepository can keep descriptions of a number of different application areas. Metarepository information structure is shown in Figure 7. The digital library database is generated on the basis of the metarepository contents.

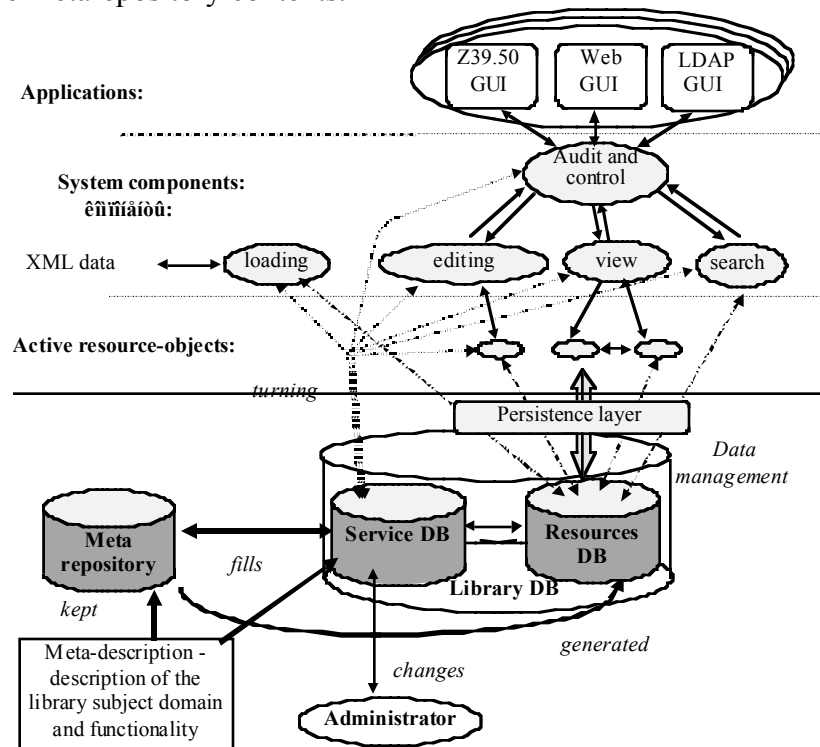


Figure 6

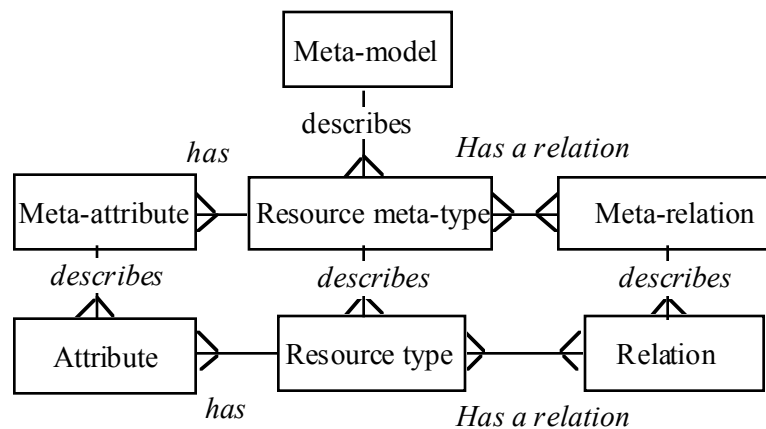


Figure 7

7. Distribution

ISIR is developing as a distributed system. Different servers of the system are intended to provide different kinds of services depending on requirements from organizations they belong to. Some servers keep only resources, some support searching on the basis of metadata, some provide full services including query transport (Figure 8).

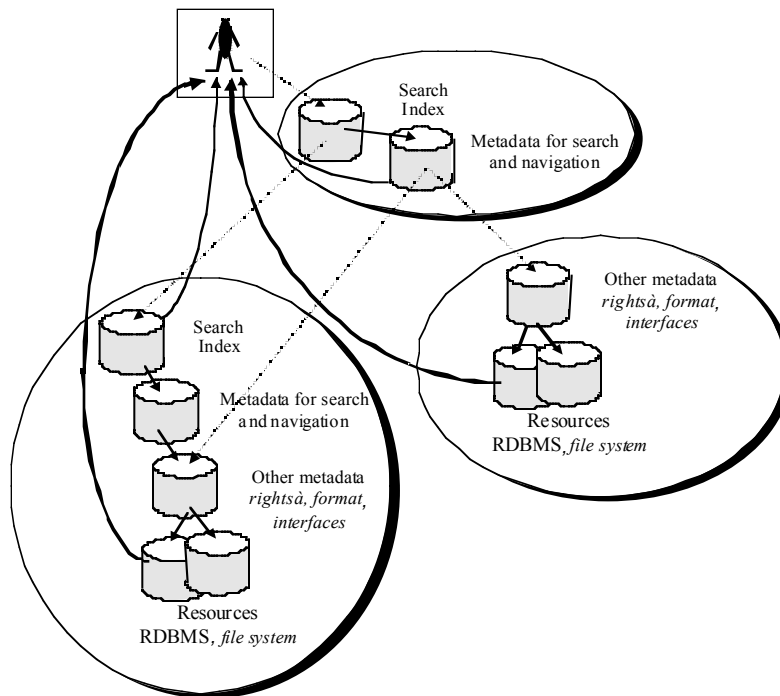


Figure 8

ISIR distribution facilities of storing integrated data and access to them are based on the following key notions: resource, relations between resources, search metadata, and resource unique identifier. Each server has its own segment in the registered name space. Each organization is responsible for uniqueness of all its resources. The top-level organization can give a part of its segment to subordinate organizations. The following agreements are used to support distributed facilities (Figure 9):

- Each resource has its own unique identifier;
- For each resource there is one and only one server that keeps it - its metadata and content. This means that all changes of the resource data (and metadata) are supported by this server. Other

effective search, access and to provide sufficient flexibility.

- Besides its own data and their metadata each server has to keep copies of search metadata of all resources that are *immediately related* to its own data according to relations between resource types. For example if a server keeps information about a publication it has to keep search metainformation about its authors.
- Some servers can keep search metadata of its subordinate servers. This provides more effective execution of search queries and does not require a lot of memory. It enables to create collections of the information similar [12], where resources automatically become members of the collection because they conform to a set of formal collection criteria for selecting resources from the distributed information space.
- Servers can form segments where broadcast distribution of search queries can be supported. A system respond to a query is the sum of the responds that servers selected to search supply about their own resources. Methods like “forward knowledge & query routing”[13] allow to optimize query routing in this case.

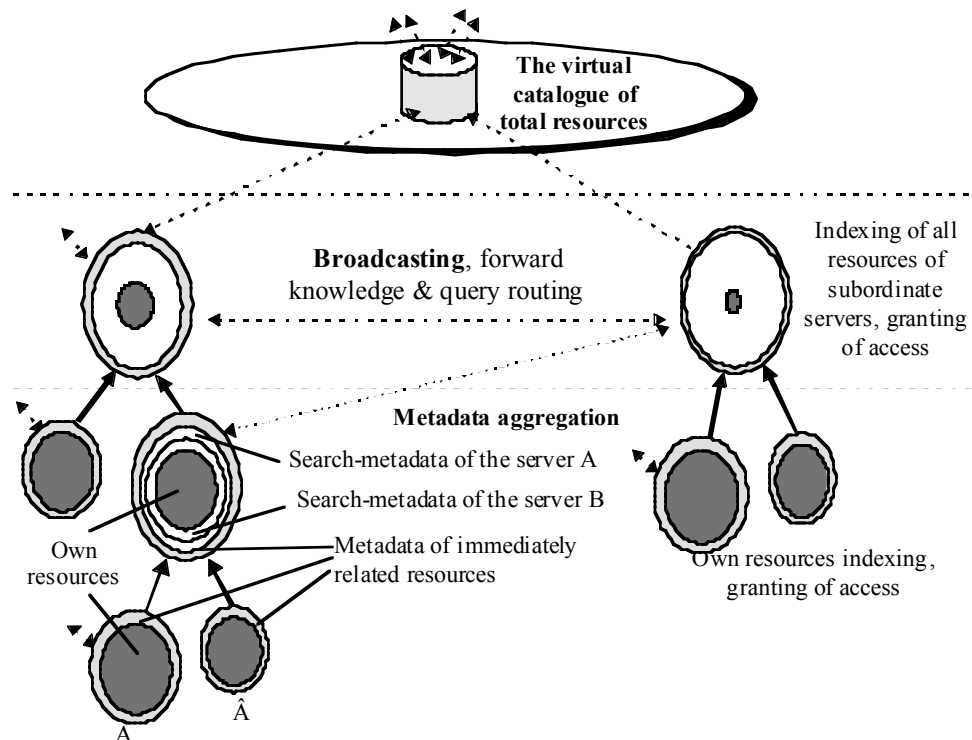


Figure 9

Web standards such as WEBDAV (Web Distributed Authoring and Versioning) and DASL (DAV Searching and Locating) [18,19] allow to implement advanced managing and searching of Web-based content collections.

8. Searching

Searching is made in a distributed environment on the basis of information about resources distribution, their relations and data base structure that is kept in server repositories. The query implementation scheme is shown in Figure 10. The query is firstly analyzed at a search server. On the basis of information about resource distribution a number of queries is generated to some of servers. At each server that stores a metarepository database information about the representation of the resources at this server is kept. On the basis of this information a query compiler transforms this query into SQL to select information from RDBMS. The selected result is then sent to the

about relations between resources it generates the final result.

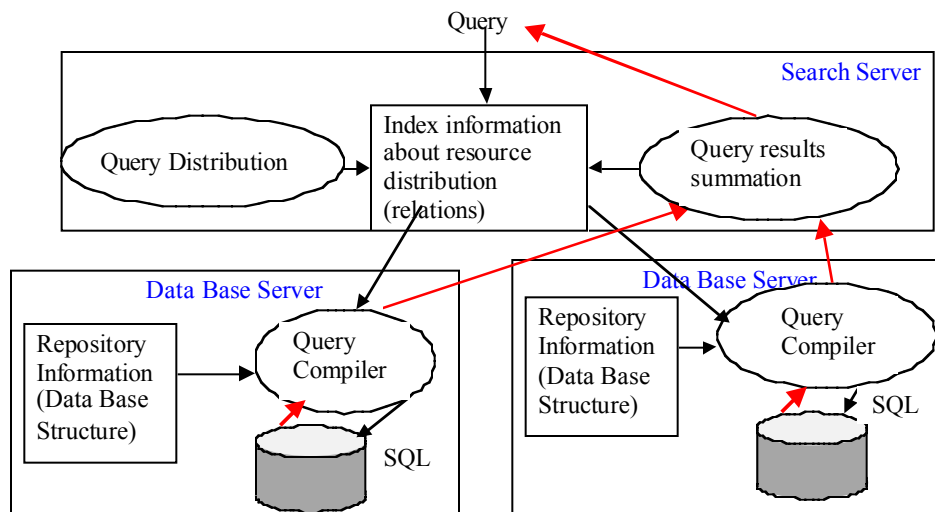


Figure 10

9. Collaboration

Metadata improves discovery of information and access to it. The effective use of metadata among applications requires common conventions about semantics, syntax, and structure. Individual resource description communities define the semantics, or meaning, of metadata that address their particular needs. The common metadata representation syntax and conventions acceptable by communities facilitate the interoperability among separate metadata element sets and the exchange, and use of metadata among multiple applications. These conventions can be expressed by a data model. Currently, the best one is the data model of Resource Description Framework. It resemble an entity-relationship model.

On the other hand, the RDBMS provides advanced possibilities for the storing, management, retrieving and searching structured data. To have adequate data processing opportunities (for example, searching, reporting) the data model should reflects the informational properties of data (the business data requirements - business entities, their attributes and entity relationships), the business and system needs of the application.

So, the system should has the well designed data model for the critical resources, for example, organization, person, publication description. However, it can support more general data schemes for resources that do not require advanced data processing services (for example, contents of XML documents), thus providing flexible resources storing (for example, documents with arbitrary structure).

10. Information interoperability

(basis technologies for publication, cooperation, collaboration)

The opportunity and the quality of the automated registration and the cataloging of the publications essentially depends on how the document is prepared and encoded. Extensible Markup Language (XML) offered IETF (Internet Engineering Task Force) has the potential to give structure and meaning to the information contained in documents or any other data form. It makes such information searchable and structured as the information located into a relation database. It is widely supported and used. There is a number of XML applications (Figure 11). For example,

resources in an electronic environment) and the Warwick Framework (modular approach to metadata) have influenced the design of the RDF. Now Dublin Core is one of RDF applications. The document [21] describes how the DC model may be considered, extended, tested and manipulated within the RDF.

A lot of others Internet standards, for example, such as WEBDAV and DASL, have also adopted XML syntax, because of it provides vendor independence, user extensibility, validation, human readability, and the ability to represent complex structures.

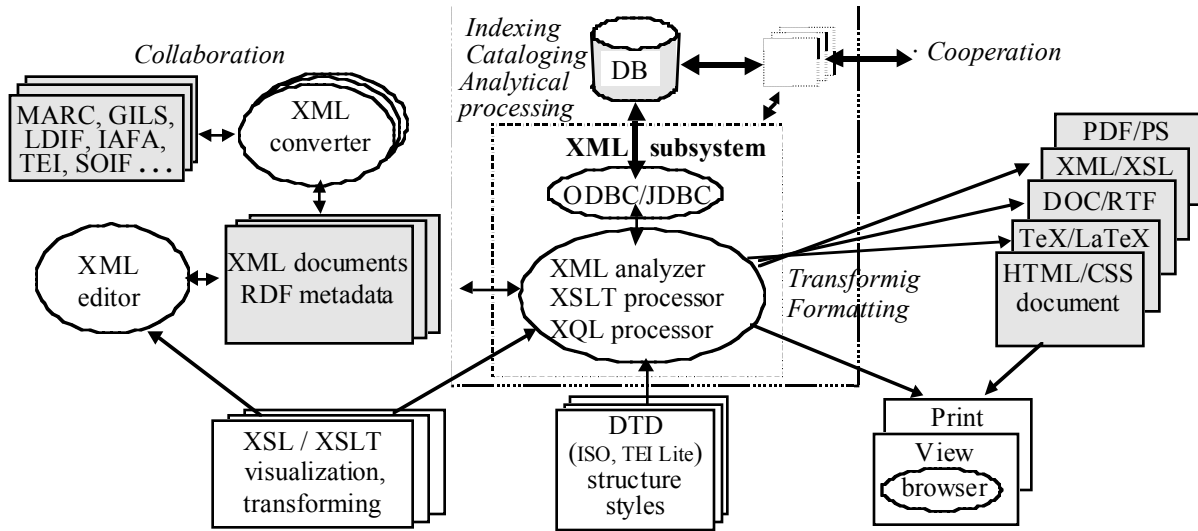


Figure 12

By exploiting the XML features ISIR (Figure 12) enables consistent encoding, exchange, and machine-processing of standardized metadata. Using RDF it provides the unambiguous, standard expression of semantics.

Figure 13 shows the collaboration of the system with various data sources and systems that can be provided with means of the XML- subsystem.

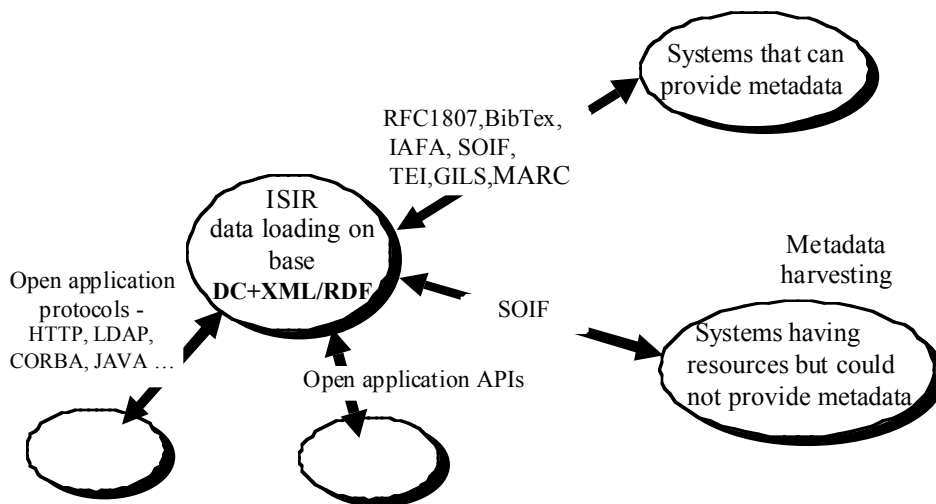


Figure 13

11. Implementation

Directing the system to the certain field of interest is made by a description of resource types, relations between them, and metadata. There are a variety of data type constructors in the system metalanguage defining the structured data for determining the resources of a digital library. Relation declare operator enables to determine relations between resource types. The collection of standard types (which have already been defined in the system) makes it possible to simplify the description of a digital library. The definitions of system functionality as well as objects are performed by declarative descriptions of functional characteristics of resources that are standard for the system. Accessing the system is reached via Web-interface. Web-interface is described by standard templates, which are changeable and may be private not only for a resource type but also for separate resources. Web-access to the system is accomplished by its own technology called Web-SQL. This technology joins HTML and SQL languages allowing to specify the rules for forming dynamic HTML pages in declarative manner (see Appendix D).

The operation of the system is supported by a variety of system program components (groups of main elements) carrying out group operations on resources (such as importing, exporting, searching, presenting search results), activating and maintaining activity of the objects of resources while relevant resources are being served. The object of a resource belongs to the object class determined for the type of this resource. All the resource object classes are successors of the class *Resource* that is pre-determined in the system. Virtual methods of this class support the standard service and system mechanisms. Import /export mechanisms operate with variety of standard metadata exchange formats (DC+HTML, DC+RDF, RDF, XML) [14] and incline to the resource type by its description and mapping onto the database scheme.

12. References

- [1] Kahn, R. and R. Wilensky, "A Framework for Distributed Digital Object Services," May 1995; <http://www.cnri.reston.va.us/k-w.html>.
- [2] William Y. Arms, "Key Concepts in the Architecture of the Digital Library", D-Lib Magazine, July 1995; <http://www.cnri.reston.va.us/k-w.html>.
- [3] Sollins, K., and L. Masinter. "Functional Requirements for Uniform Resource Names", RFC 1737, MIT/LCS, Xerox Corporation, December 1994.
- [4] Berners-Lee, T., Masinter, L., and M. McCahill, Editors. "Uniform Resource Locators (URL)", RFC 1738, CERN, Xerox Corporation, University of Minnesota, December 1994.
- [5] The Handle System Home Page, <http://www.handle.net/>
- [6] International DOI Foundation web site; <http://www.doi.org>
- [7] Stuart Weibel and Eric Miller, "Dublin Core Metadata"; http://purl.org/metadata/dublin_core
- [8] DC Elements, Reference Version; http://purl.oclc.org/metadata/dublin_core_elements
- [9] Dempsey, Lorcan, Rachel Heery, Martin Hamilton, Debra Hiom, Jon Knight, Traugott Koch, Marianne Peereboom, and Andy Powell. "A Review of Metadata: A Survey of Current Resource Description Formats. Work Package 3 of Telematics for Research project DESIRE (RE 1004). Version 1.0"; <http://www.ukoln.ac.uk/metadata/DESIRE/overview/>
- [10] Payette, S. and C. Lagoze, "Flexible and Extensible Digital Object and Repository Architecture (FEDORA)," Second European Conference on Research and Advanced Technology for Digital Libraries, Heraklion, Crete, Greece, September 21-23, 1998, Springer, 1998, (Lecture notes in computer science; Vol. 1513).
- [11] Arms, W.Y., C. Blanchi, and E. Overly, "An Architecture for Information in Digital Libraries," D-Lib Magazine, February 1997; <http://www.dlib.org/dlib/february97/cnri/02arms1.html>
- [12] Carl Lagoze, David Fielding, "Defining Collections in Distributed Digital Libraries"

- Searching Subject Gateways: The Query Routing and Forward Knowledge Approach”;
<http://www.dlib.org/dlib/january98/01kirriemuir.html>
- [14] Tim Bray, Jean Paol and C. M. Sperberg-McQueen, “Extensible Markup Language (XML) 1.0”; <http://www.w3.org/TR/REC-xml>
- [15] Tim Bray, Dave Hollander and Andrew Layman “Namespaces in XML”;
<http://www.w3.org/TR/REC-xml-names>
- [16] Lassila, Ora and Ralph R. Swick, "Resource Description Framework (RDF) Model and Syntax", W3C Working Draft, <http://www.w3.org/TR/REC-rdf-syntax>
- [17] Dan Brickley, R.V. Guha and Andrew Layman, “Resource Description Framework, (RDF) Schemas”, W3C Working Draft; <http://www.w3.org/TR/WD-rdf-schema>,
- [18] Y. Goland , E. Whitehead, A. Faizi, S. Carter, D. Jensen, “HTTP Extensions for Distributed Authoring – WEBDAV”, RFC 2518; <http://ds.internic.net/rfc/rfc2518.txt>
- [19] Saveen Reddy, Dale Lowry, Surendra Reddy, Rick Henderson, Jim Davis, Alan Babich, “DAV Searching & Locating”; <http://www.webdav.org/dasl/protocol/draft-dasl-protocol-00.html>
- [20] S.Agoshkov, A.Bezdushny, M.Galochkin, A.Medennikov, M.Koulagin, V.Serebiakov, “The Integrated System of Information Resources of the Russian Academy of Sciences”, “Software technology” FUSST’99, Tallin, Estonia, 1999.
- [21] Guidance on expressing the Dublin Core within the Resource Description Framework (RDF), <http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/>
- [22] A.Bezdushny, Web-SQL technology description, iR-1999-3, SMO, CC RAS.