Mathematical Problems of Computer Science 22, 2001, 91–95.

A Conceptual Model of Digital Libraries Integration, Based on Metadata Architecture

Tigran A. Shahinian

Institute for Informatics and Automation Problems of NAS RA and YSU
E-Mail: artsha@sci.am

Abstract

The metadata models in digital libraries are considered. A new conceptual model of integration is proposed for a class of digital libraries with strongly typed information resources and relationship models.

1 Introduction

One of the most important elements in digital libraries is the metadata model. Metadata are data about information objects and their properties and can have diverse meanings, e.g., search attributes, navigation capabilities, rules for working with resource types, and various administrative information about resources.

Digital libraries are developed mostly for heterogeneous information integration. One of the most important problems in digital libraries is the problem of their integration. There are several approaches to solve this problem. One of the possible solutions is based on the mediation architecture [1-2]. In this paper we propose a new conceptual model of integration, which extends the mediation architecture, but significantly differs from other such models. This model provides mediation on the level of information objects, unlike others providing mediation on the level of user queries.

The proposed model is based on the architecture of the digital library, which acts as the client, interacting with server digital libraries (or information systems). Mainly it depends on the scheme of relationships between information objects (resources), which is implied from the metadata model of the client digital library.

The conceptual model is shown on the Integrated System of Information Resources (ISIR) as the client system [3].

2 Metadata Models

Different digital libraries implement different metadata models. They can differ by their structure and information contents (e.g., the MARC bibliographic metadata model consists of a large plain list of fields, whereby the Dublin Core consists of 15 base elements)[4-5].

The goal of the Dublin Core metadata model is to define a minimal set of descriptive elements, listed below, that facilitate the description and the automated indexing of document-like networked objects.

TITLE: The name given to the resource by the CREATOR or PUBLISHER.

CREATOR: The person(s) or organization(s) primarily responsible for the intellectual content of the resource.

SUBJECT: Keywords or phrases that describe the subject or content of the resource. The intent is to use controlled vocabularies and keywords, so the element might include scheme-qualified classification data (for example, Library of Congress Classification Numbers) or scheme-qualified controlled vocabularies (such as MEdical Subject Headings).

DESCRIPTION: A textual description of the content of the resource, such as document abstracts or content descriptions of visual resources. This could be extended to include computational content description (e.g., spectral analysis of a visual resource). In this case this field might contain a link to the description rather than the description itself.

PUBLISHER: The entity responsible for making the resource available in its present form.

CONTRIBUTORS: Person(s) or organization(s) in addition to those specified in the CREATOR element who have made significant intellectual contributions to the resource.

DATE: The date the resource was made available in its present form.

TYPE: The category of the resource, such as home page, novel, poem, working paper, etc.

FORMAT: The data representation of the resource, such as text/html, ASCII, Postscript file, executable application, or JPEG image (as well as non-electronic media).

IDENTIFIER: String or number used to uniquely identify the resource. Examples for networked resources include URLs and URNs (when implemented). Other globally unique identifiers such as International Standard Book Numbers (ISBN) or other formal names would also be candidates for this element.

SOURCE: The work, either print or electronic, from which this resource is derived, if applicable.

LANGUAGE: Language(s) of the intellectual content of the resource.

RELATION: Relationship to other resources, for example, images in a document, chapters in a book, or items in a collection.

COVERAGE: The spatial locations and temporal durations characteristic of the resource.

RIGHTS: A link (e.g., a URL or other suitable URI as appropriate) to terms and conditions, copyright statements, or similar information.

In addition to these data elements, the DC model specifies a number of principles that apply to the entire core metadata set.

- The core metadata set should be extensible to permit site specific or domain specific data elements.
- All elements in the Core metadata set should be optional.
- All elements should be repeatable allowing, for example, multiple author elements.
- The semantics of each element should be modifiable by either:
  - the use of qualifiers, borrowed from other existing metadata schemes, which allow the use of more detailed or specific semantics from those schemes. For example, a Subject element might be specified as Subject (scheme=LCSH), indicating that the subject terms are taken from the Library of Congress Subject Headings.
  - ad-hoc specializations and extensions developed specifically for use with the Core so as to refine the normal meanings of the core data elements.

The Dublin Core metadata model is ideal for storing information about a set of interrelated types of information resources and is implemented in ISIR(Integrated System of Information Resources) of the Russian Academy of Sciences [3].

3 Metadata Model Implementations in Digital Libraries

As an example of a metadata model implementation we consider ISIR. Metadata in ISIR are information about a set of interrelated types of information resources. Metadata processing in such model allows performing more exact and more complicated search operations. The result of the performed search, for example, could be information not only about publications, but also about related resources and other analytical information [3].

Analysis and structuring of metadata reveal the relationships between resources (fig.1).

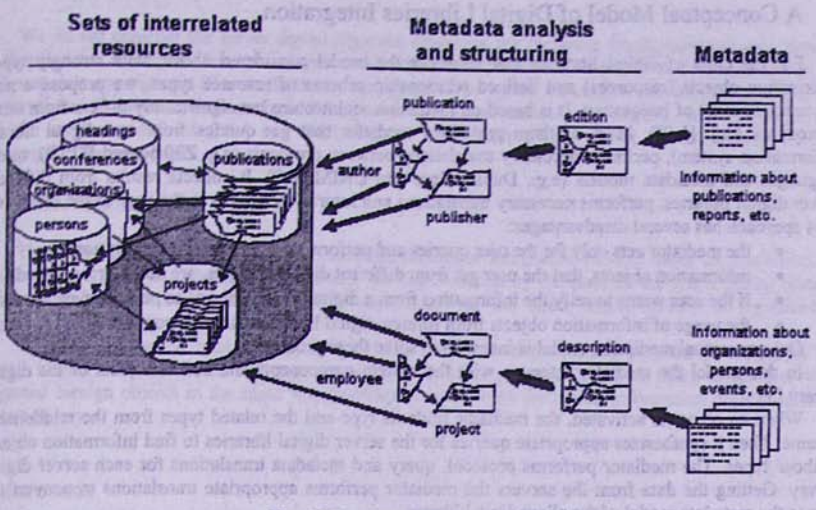

Fig. 1 Metadata analysis and structuring

Adding to this set resource types not pointed in bibliographic descriptions but often concerned to scientific publications, e.g., projects, contests, grants, conferences, sponsors, organizations, we get the resource types relationships scheme of ISIR (fig.2).

Such system of resource types and relationships significantly extends digital library's service facilities. For example, the following queries can be performed:

- find projects, in which the author of the given publication participates and return the materials and publications of that projects;
- find information about conferences, where the author has passed his(her) publications;
- find publications of the author's colleagues;
- find the materials of conference sections, to which the author's colleagues-by- project have reported about their works[3].

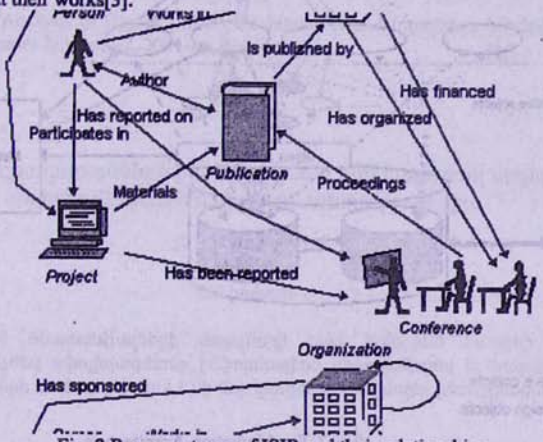

Fig. 2 Resource types of ISIR and their relationships

4 A Conceptual Model of Digital Libraries Integration

For the class of digital libraries that resemble the model considered above, with strongly typed information objects (resources) and defined relationship scheme of resource types, we propose a new conceptual model of integration. It is based on mediation architecture but significantly differs from other approaches (e.g. [1-2]). Most of them provide a mediator that get queries from one digital library (information system), performs necessary translations between protocols (e.g. Z39.50 and HTTP), query languages and metadata models (e.g., Dublin Core and UNIMARC). It collects results from different server digital libraries, performs necessary translations and returns the results to the user of the client [6]. This approach has several disadvantages:

- the mediator acts only for the user queries and performs mostly searching operations,
- information objects, that the user get from different digital libraries, are not interconnected,
- if the user wants to reify the information from a digital library, he has to perform new requests,
- the usage of information objects from foreign digital libraries cannot be optimized.

Our conceptual mediation model is intended to solve these problems.

In this model the mediator interacts with the system components and active objects of the digital library [3].

When an object is activated, the mediator finds its type and the related types from the relationship scheme. Then it synthesizes appropriate queries for the server digital libraries to find information objects of those types. The mediator performs protocol, query and metadata translations for each server digital library. Getting the data from the servers the mediator performs appropriate translations to convert the data to the metadata model of the client digital library.

On the next step the mediator performs a matching of the received information with the metadata of the objects related to the active object. It filters out the «data units» that have same metadata as one of the related objects. The remaining units are being passed to the appropriate component of the digital library to create new objects. The unique identifiers of objects are being returned to the mediator and it stores them in a table of identifiers and manages it.

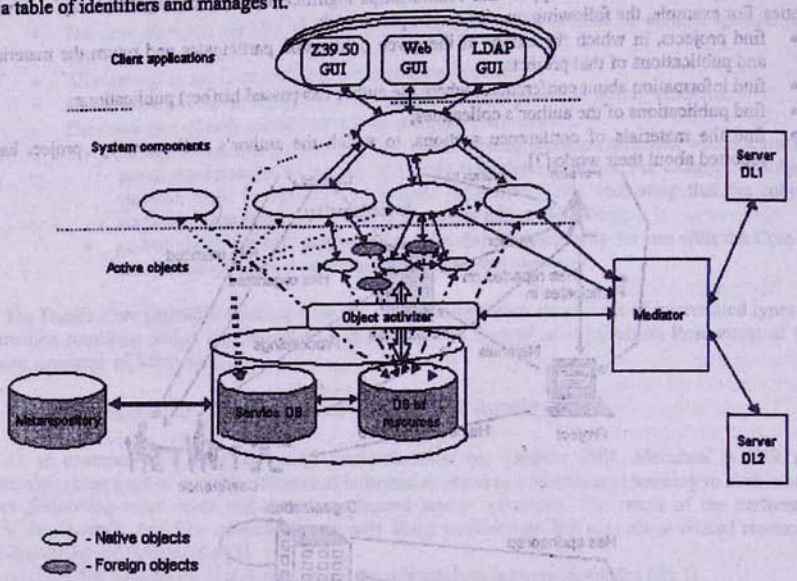

Fig. 3 The Mediation Architecture

We do not consider the server digital libraries, assuming the mediator functionality doesn't depend on them on this conceptual level. It is important to know their properties when performing protocol, query and metadata translations. The mediator is considered to be based on component architecture, and the server digital libraries' specifications will be passed to it as parameters (e.g. in RDF, schematically). We left the servers undefined (DL1 and DL2)(fig. 3). The details of the model and realization will be presented in the next paper.

5 Conclusions

The model proposed is mostly a functional extension of integration models based on mediation. In this model we propose a method of importing "foreign" objects by the need of the client digital library. This model is based on (mostly metadata) architecture of the client digital library. It significantly improves the integration capabilities.

The model allows to integrate digital libraries in the sense of performing all possible actions of the digital library on information resources from all integrated systems. It allows users to interact with imported foreign objects in the same way that they interact with the native information objects of their digital library.

References

[1] Sergey Melnik, hector Garcia-Molina, Andreas Paepcke "A Mediation Infrastructure for Digital Library Services", ACM Digital Libraries 2000.

[2] S. Bergamashi, S. Castano, and M. Vincini "Semantic Integration of Semistructured and Structured Data Sources". SIGMOD Record 28-1, March 1999.

[3] S.V. Agoshkov, A.N. Bezdushni, M.P. Galochkin, M.V. Kulagin, A.M. Medennikov, V.A. Serebryakov, «Integrated System of Information Resources (ISIR) – an Approach to Creation of Integrated Digital Libraries", international conference" Digital Libraries: Perspective Methods and Technologies, Digital Collections", Saint-Petersburg, 1999. (in Russian).

[4] MARC, lcweb.loc.gov/marc/marc.html

[5] Dublin Core, purl.oclc.org/metadata/dublin_core

[6] T.A. Shahinian, "Analysis of Digital Library Architectures and Interaction Models". Mathematical Problems of Computer Science 22, 2001 (in Russian).

Էլեկտրոնային գրադարանների ինտեգրացման կոնցեպտուալ մոդել, հիմնված մետատվյալների ճարտարապետության վրա

S. U. Շահինյան

Ամփոփում

Քննարկվում են մետատվյալների մոդելների հետ կապված հարցեր էլեկտրոնային գրադարաններում: Ուժեղ տիպիզացմամբ ինֆորմացիոն ռեսուրսներով եւ հարաբերությունների մոդելով գրադարանների համար բերվում է մի նոր կոնցեպտուալ մոդել ինտեգրացման համար: