УДК 004.43'23XML:004.738.5.057.3/.4

ИМПЛЕМЕНТАЦИЈА ПРОТОКОЛА ЗА ПРИКУПЉАЊЕ МЕТАПОДАТАКА У МРЕЖИ ДИГИТАЛНИХ БИБЛИОТЕКА ТЕЗА И ДИСЕРТАЦИЈА*

Мирослав Зарић
Факултет техничких наука, Нови Сад
Душан Сурла
Природно-математички факултет, Нови Сад

Сажетак

У раду је описана имплементација PMH - Протокола за прикупљање метаподатака, у његовом делу који дефинише публиковање података у *Dublin Core* формату. Имплементација протокола је завршена и у току је тестирање на тест серверу за размену на Политехничком институту и државном универзитету Вирџиније (*Virginia Polytechnic Institute and State University – VirginiaTech*). Размотрена је могућност примене протокола PMH у библиотечким окружењима у нашој земљи. Дат је списак светских библиотека сталних чланица мреже институција OAI - Иницијативе за отворене архиве које својим корисницима нуде сервис за преузимање и претраживање садржаја путем PMH протокола.

Кључне речи: Dublin Core, OAI-PMH, XML документи

1. Увод

У овом раду приказан је део резултата добијених у реализацији пројекта *Мрежна дигитална библиотека докторских, магистарских и дипломских радова*. [1]. Пројекат финансира Покрајински секретаријат за науку и технолошки развој АП Војводине. Реализација пројекта поверена је Универзитету у Новом Саду, који се 2003. године придружио Мрежи дигиталних библиотека теза и дисертација *(Networked Digital Library of Thesis and Dissertations - NDLTD)* [2]. Ова мрежа базирана је на иницијативи отворених архива (*Open Archive Initiative – OAI, OA иницијатива*) [3]. Ова иницијатива осмишљена је са циљем развоја и промовисања стандарда који би требало да омогуће ефикасну размену садржаја различитих архива.

UDC 004.43'23XML:004.738.5.057.3/.4

IMPLEMENTATION OF PROTOCOL FOR METADATA HARVESTING IN NETWORKED DIGITAL LIBRARY OF THESES AND DISSERTATIONS!

Miroslav Zarić
Faculty of Technical Sciences, Novi Sad
Dušan Surla
Faculty of Science, Novi Sad

Abstract

Implementation of *Protocol for Metadata Harvesting* - PMH protocol, and its use for exposing metadata in Dublin Core format have been described in this paper. Protocol implementation has been finalized, and it is currently being tested, using the test environment provided by the *Virginia Polytechnic Institute and State University – VirginiaTech*. Short disscusion on usage of this protocol in library environments in our country, as well as reference to the list of *Open Archive Initiative* (OAI) member institutions have also been provided.

Key words: DublinCore, OAI-PMH, XML documents

1. Introduction

Partial results obtained during development of the project of *Networked digital library of thesis, dissertations and graduation thesis* have been presented In the paper [1]. Project is financed by the Provincial Secretariat for Science and Technological Development of the Autonomous Province of Vojvodina. The realization of the project has been handed out to the University of Novi Sad which has become a member of the *Networked Digital Library of Thesis and Dissertations - NDLTD* [2] association during the year of 2003. This association is based upon the *Open Archive Initiative – OAI, OA initiative*) [3]. The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content.

* Рад је изложен на *Стручном скупу* "Рад у систему узајамне каталогизације" одржаном у оквиру Девете скупштине Заједнице библиотека универзитета у Србији, 10 - 11. октобра 2003. године у Народној библиотеци Србије.

! Paper presented on the 9th Professional Meeting "Shared cataloguing system" organized by the Serbian Academic Library Association and held in the National Library of Serbia. Belgrade, 10 -11 October 2003.

За размену тих садржаја дефинисан је *Протокол за прикупљање метаподатака* (PMH) [4].

Корени OAI леже у залагању да се унапреди доступност архива електронских издања, при чему је основни циљ био повећати доступност образовних садржаја. Континуиран рад на унапређењу управо овог сегмента електронског издаваштва остаје и даље једна од кључних тачака те иницијативе. Основна техничка решења, као и стандарди који се развијају подршке иницијативи, ипак су независни од типа садржаја који се нуди и, у перспективи, треба да омогуће знатно већу отвореност и лакши приступ широком спектру материјала у дигиталној форми.

Као резултат овог настојања, OA иницијатива тренутно представља заједницу у сталном развоју, а напори који се чине посвећени су истраживању и развоју нових апликација које имплементирају постојеће стандарде и доприносе даљем развоју OA иницијативе. Како се буду прикупљала сазнања везана за опсег примењивости имплементиране основне технологије, како се буду развијали стандарди и како се буде разумевала структура и културне разлике различитих учесника у иницијативи, очекује се континуиран еволутивни развој целокупне иницијативе у свим њеним сегментима (концепти, циљеви, техничка решења).

Одлуке везане за правце развоја OA иницијативе доноси Управни одбор иницијативе. Основне поставке техничке инфраструктуре која омогућава сарадњу разнородних система је развио технички комитет. Технички комитет и даље наставља свој рад са циљем унапређења техничких решења, а на основу досад прикупљених искустава. Приказана је кратка историја OA иницијативе, уз нагласак на развоју PMH протокола.

Централни део рада посвећен је теми имплементације PMH протокола, у оном његовом делу који дефинише публиковање података у *Dublin-Core* формату. Обрађени су минимални технички захтеви које мора имплементирати сваки систем који претендује да се мрежи OAI институција прикључи у својству система који податке из своје архиве публикује за остале чланове OAI (тј. да се појављује у улози *data provider-a*). Такође су размотрене и смернице за имплементацију протокола у системима који у мрежи OAI институција имају улогу клијентских система.

2. OA иницијатива

Први званични почеци OA иницијативе могу се везати за ***Санта Фе конвенцију*** [5] која је

Within this initiative the *Protocol for Metadata Harvesting* – PMH [4] has been developed.

The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the Open Archives program. The fundamental technological framework, as well as the standards that are developing to support this work are, however, independent of both type of offered content and economic mechanisms surrounding that content, and promise to have much broader relevance in opening up access to a range of digital materials.

As a result, the Open Archives Initiative is currently an organization and an effort explicitly in transition, and is committed to exploring and enabling this new and broader range of applications. As we gain greater knowledge of the scope of applicability of the underlying technology and standards being developed, and begin to understand the structure and culture of the various adopter communities, we expect that we will have to make continued evolutionary changes to both the mission and organization of the Open Archives Initiative.

Policy decisions about the Open Archives Initiative are made by a Steering Committee. The interoperability infrastructure was developed by a technical committee, which continues to advise on the infrastructure as experience with it develops. A short history of OA initiative, since the first efforts in promotion of e-print community, up to recent developments has also been presented in this paper, with special concern paid to the development of protocol, and its implementations.

The central section of this paper addresses the topic of implementation PMH protocol, in its section that defines dissemination of metadata in Dublin Core format. The minimum technical requirements that must be fulfilled by any implementor tending to be a data provider have been discussed, as well as the guidelines for implementing protocol in client systems within OA initiative.

2. OA Initiative

The roots of the OA initiative can be traced to the ***Santa Fe Convention*** [5]. This document was a

настала као резултат састанка који се одржао у Санте Феу, у америчкој савезној држави Нови Мексико, октобра 1999 (*Santa Fe, New Mexico, USA*). Ова конвенција је једногласно усвојена од стране свих учесника, представника из организација које су већ имале или планирале увођење архива електронских издања намењених јавном приступу. Такође су је усвојиле и организације заинтересоване за понуду услуга, као што су претраживачи, претраживачи цитата, а који би били базирани на подацима које нуде ове архиве. Конвенција представља једноставан технички и организациони оквир чији је циљ омогућавање основне сарадње међу архивама електронских издања. Учесници су тада изразили намеру да у релативно кратком року имплементирају смернице, како би се у току 2000. г. могле извршити прве пробе међусобне размене података. Истовремено је упућен позив организацијама које нису биле представљене на конвенцији да их и оне што пре имплементирају.

Циљ конвенције је да аутори научних и стручних радова могу да учине своје радове доступним широкој јавности у виду електронских докумената путем архива електронских издања. У моменту усвајања конвенције, број успостављених и функционалних архива електронских издања био је релативно мали. Међу њима је била arXiv.org, архива електронских издања научних радова института у Лос Аламосу, коју је основао Паул Гинспарг, која је постала основно чвориште за размену резултата истраживања у домену физике. Предвиђало се тада и оснивање бројних других архива које би биле доступне путем Интернета: неке од њих би биле уско специјализоване за одређене образовне дисциплине, док би садржаји других представљали области интересовања институција која их одржавају. Да би електронске архиве могле постати признат и прихваћен механизам за размену научно-образовних докумената, мора се подржати одређени ниво међусобне сарадње. OAI управо треба да омогући развој и примену техника које би ово учиниле могућим.

Санта Фе конвенција представља технички и организациони оквир осмишљен са циљем да омогући проналажење садржаја који се налази у дистрибуираним архивама електронских публикација. Техничке препоруке за имплементацију су једноставне и. када буду имплементиране, омогућиће да подаци у архивама постану широко доступни путем њиховог укључивања у различите корисничке сервисе као што су машине за претраживање и системи за повезивање докумената. Додатно, конвенција поред техничког, промовише и организациони оквир препоручен да омогући

result of the meeting held in October 1999. in Santa Fe, New Mexico, USA. This convention has been adopted by all participants, representing organizations that were already using, or organization that were planning to use e-print archives, as publicly accessible archives, as well as organizations that were interested in developing a value added services based upon that archives (search engines, citation search etc). Convention was a simple technical and organizational framework with main goal to support interoperability between e-print archives. All participants stated their intention to implement these guidelines in a relatively short period, thus allowing first tests of the data interchange to take place during the year of 2000. At the same time a public invitation to join the initiative has been issued to all interested parties.

Objective Scholarly authors can make electronic documents available to a global audience by submitting them to e-print archives. At the time of writing this convention (January 2000), the number of established e-print archives was small. One was arXiv.org, the Los Alamos e-print archive created by Paul Ginsparg, which has become a crucial hub for communicating research findings in physics. At that time the creation of many more e-print archives in the coming years was anticipated. Such archives will be distributed across the Internet; some will be oriented to particular scholarly disciplines, others will be based on institutional affiliation. In order to make e-print archives to become an established mechanism for scholarly communication, some level of interoperability among them needs to be supported. This convention was a first step towards such interoperability. The Open Archives Initiative provides a process for growth and development.

The Santa Fe Convention presents a technical and organizational framework designed to facilitate the discovery of content stored in distributed e-print archives. It makes easy-to-implement technical recommendations for archives that – when implemented – will allow data from e-print archives to become widely available via its inclusion in a variety of end-user services such as search engines, recommendation services and systems for interlinking documents. In addition, the convention introduces an organizational framework recommended in order to support desired level of interoperability. Santa Fe convention docu-

жељену сарадњу и размену. Санта Фе конвенција, чији су документи дати су у референцама [5-12], у међувремену је прерасла у ***Протокол за прикупљање метаподатака.***

Основне одреднице ОА иницијативе

Сам термин архива у називу иницијативе рефлектује полазиште ОА иницијативе – заједницу електронских издавача, где је термин *архива* био општеприхваћен као синоним за базу научно-образовних публикација. ОАІ користи термин *архива* у нешто проширеном значењу: као базу информација. Међу најзначајнијим организацијама и удружењима који подржавају ОАІ и учествују у њој јесу и Федерација дигиталних библиотека, Коалиција за умрежене информације, Национална научна фондација, и др. ОА иницијатива је отворена за сарадњу са свим заинтересованим организацијама. Оквир за размену информација је дефинисан Протоколом за размену метаподатака у оквиру ОАІ. Организације могу у ОА иницијативи учествовати у два својства: **понуђачи података** *(Data providers)* и **понуђачи услуга** *(Service providers)*

Понуђачи података администрирају системе који РМН протокол користе као средство путем којег чине своје метаподатке доступним за размену. Понуђачи података могу одлучити да ли желе да се, преко странице за регистрацију, региструју у оквиру ОА иницијативе и тиме објаве усвајање ОАІ-РМН протокола на својим серверима.

Понуђачи услуга користе податке које су прикупили уз помоћ ОАІ - РМН протокола у циљу развоја неке напредније (комерцијалне) услуге. И понуђачи услуга могу да се региструју у оквиру ОАІ, путем одговарајућих страница, и на тај начин учине своје услуге доступним корисницима.

"Отвореност" ОАІ огледа се у архитектури система за размену – дефинишу се и промовишу технички стандарди размене који омогућавају доступност садржаја које нуде најразличитији понуђачи на својим серверима. Термин *отвореност* не мора неопходно значити и неограничен и слободан приступ информацијама свих архива које су у ОАІ. Треба водити рачуна да се термин *отворен* не схвати у погрешном контексту јер ограничења које неке архиве намећу не морају бити везана уз новчану накнаду за коришћење информација, већ је могуће да неке архиве забрањују приступ неким информацијама, или их нуде уз заштиту ауторских права.

ments are listed in references [5-12]. Santa Fe convention has been phased out in the meantime and ***Protocol for Metadata Harvesting*** has developed from its base.

OA Initiative Basics

The term "archive" in the name *Open Archives Initiative* reflects the origins of the OAI – in the E-Prints community where the term archive is generally accepted as a synonym for repository of scholarly papers. The OAI uses the term "archive" in a broader sense: as a repository for stored information.

Most influential members and supporters of OAI are Digital Library Federation, the Coalition for Networked Information, and National Science Foundation, among others. OA initiative is open for cooperation to any interested party. The framework for data exchange is defined by the OAI Protocol for Metadata Harvesting. Any organization can take part in OAI in two roles:

- *Data provider*
- *Service provider*

Data provider A data provider maintains one or more repositories (on web servers) that support the OAI-PMH as a means of exposing metadata. They can, if they want, register their repositories within the initiative and in that way state publicly implementation of the PMH on their servers.

Service providers are using data gathered by the harvester (by the OAI–PMH) to provide some value added services. Service providers can also register their activities within the initiative, and in that manner make their services publicly available.

The term "open" in the OAI should be viewed from the architectural perspective – defining and promoting machine interfaces that facilitate the availability of content from a variety of providers. Openness does not mean "free" or "unlimited" access to the information repositories that conform to the OAI-PMH. Such terms are often used too casually and ignore the fact that monetary cost is not the only type of restriction on use of information – any advocate of "free" information recognizes that it is eminently reasonable to restrict denial of service attacks or defamatory misuse of information.

3. Протокол за прикупљање метаподатака

Постојећа техничка инфраструкутура OAI која је специфицирана Протоколом за прикупљање метаподатака, дефинише начин на који понуђачи података морају изложити своје метаподатке. У OA иницијативи не постоје ограничења која би редуковала могућност употребе само на метаподатке, тако да је могуће предвидети и размену других видова података. Међутим, како је један од основних циљева реализација једноставног и широко примењивог оквира за међусобну сарадњу различитих база података, размена метаподатака представља логичан начин за ефикасно постизање таквог циља.

PMH протокол дефинише механизам за прикупљање метаподатака у форми XML документа. Тренутно не постоји могућност да се подаци размењују у неком другом формату. Сам протокол не захтева постојање директне референце између метаподатака и садржаја базе података која је описана датим метаподацима. Пошто многи клијенти могу затражити приступ садржају који је описан прикупљеним метаподацима, пожељно је да се у метаподацима дефинише линк на сам садржај. За ту сврху у *DublinCore* формату метаподатака користи се елемент *identifier*.

У протоколу је извршена потпуна спецификација формата XML докумената који се размењују, као и дефиниција свих појмова везаних уз сам протокол и његову имплементацију у оквиру OA иницијативе. OAI-PMH протокол је базиран на HTTP протоколу, тј. функционише по принципу захтев – одговор. Захтев шаље клијент који жели прикупити податке, док одговор шаље серверска апликација на страни архиве која се јавља у својству понуђача података. Захтеви се шаљу у виду стандардних параметара HTTP GET и POST захтева, док је одговор у форми XML документа. PMH протокол уводи следеће концепте:

1. **прикупљач** (*harvester)* – клијентска апликација која шаље OAI–PMH захтеве и прикупља метаподатке из архива;

2. **архива** (*repository)* – мрежно доступан сервер коме је могуће приступити путем Интернета и који је у стању да процесира шест дефинисаних OAI-PMH захтева, а ради излагања метаподатака прикупљачима.

3. **јединица података** (*item*) – члан архиве из којег се могу прикупити метаподаци о неком ресурсу;

4. **јединствени идентификатор** (*unique identifier*) – једнозначно идентификује јединицу у оквиру архиве; користи се у захтевима за

3. Protocol for Metadata Harvesting

Existing technical infrastructure of the OA initiative, as specified by the Protocol for Metadata Harvesting, defines the form in which data providers must expose their metadata. There are no restrictions in the roots of OAI that would narrow its usage only to metadata, so other implementations can be expected. However, since one of the basic goals is realization of one easy and widely acceptable interoperability framework, use of metadata exchange is a logical means for achieving that goal.

PMH defines a mechanism for metadata harvesting in the form of XML documents. Currently, no other format of data interchange is supported. The protocol itself does not require a direct reference between metadata and the repository content that is described by that metadata, but since many clients can request access to the content described by the harvested metadata, it is a good idea to include a link to the content. In Dublin Core a field identifier is used for that purpose.

Format of the XML documents to be exchanged, as well as other concepts related to the protocol itself and its implementation within the OAI are fully specified. OAI–PMH protocol is based upon HTTP protocol. It works on request – response pattern. Client application, that wants to harvest some metadata sends a request, while server application (on the archive) sends a response. Requests are sent as a parameters within standard HTTP requests, while the answer is in the form of the XML document.

Following definitions and concepts are introduced :

1. A ***harvester*** is a client application that issues OAI-PMH requests. A harvester is operated by a service provider as a means of collecting metadata from repositories;

2. A ***repository*** is a network accessible server that can process the six OAI-PMH requests in the manner described in this document. A repository is managed by a data provider to expose metadata to harvesters .

3. An ***item*** is a constituent of a repository from which metadata about a resource can be disseminated. Each item has an identifier that is unique within the scope of the repository of which it is a constituent.

4. A ***unique identifier*** unambiguously identifies an item within a repository; the unique identifier is

извлачење метаподатака из одређене јединице података.

5. **запис** (*record)* – представља метаподатке о јединици архиве, изражен у једном од могућих формата; у одговору на OAI-PMH захтев враћа се као XML-кодирани низ бајтова.

6. **скуп** (*set)* – опциони концепт путем којег се могу груписати јединице, с циљем олакшавања *селективног прикупљања метаподатака;

Селективно прикупљање – омогућава прикупљачима података да лимитирају своје захтеве на подскуп метаподатака из архиве.

3.1 Особине протокола

Као што је већ раније наведено, захтеви који се шаљу у оквиру OAI-PMH протокола шаљу се у виду стандардних HTTP захтева. Сервер који имплементира OAI-PMH мора бити у стању да коректно обради захтеве и одговори одговарајућим XML документом, чији је формат специфициран за сваки од захтева. При слању захтева користе се стандардна правила која важе за HTTP захтеве, тј. захтеви морају бити послати у виду GET или POST HTTP захтева. Поред основне URL адресе, сви захтеви садрже и дозвољене парове *параметар=вредност_параметра*. Сваки OAI-PMH захтев мора садржати барем један пар *параметар=вредност_параметра* који специфицира сам OAI-PMH захтев којег је издао прикупљач.

Одговори које сервер шаље клијенту такође су у виду HTTP одговора, при чему је формат самог одговора XML документ. Опционо се може користити и нека од метода за компресију одговора.

Сам формат XML докумената који се креирају као резултат одговора на одређени OAI-PMH захтев мора да задовољи следеће услове:

1. XML документ мора бити добро формиран.

2. Морају се користити референце карактера, а не референце ентитета, што омогућава да се XML документи третирају као самостални, без зависности од декларације ентитета ван самог документа.

3. XML документ мора бити валидан у односу на шему XML Schema дефинисану протоколом.

OAI-PMH подржава размену података у различитим форматима. Захтева се да сваки архив који учествује у OA иницијативи подржи минимално ***DublinCore*** формат метаподатака. Протоко-

used in OAI-PMH requests for extracting metadata from the item.

5. A ***record*** is metadata expressed in a single format. A record is returned in an XML-encoded byte stream in response to an OAI-PMH request for metadata from an item.;

6. A ***set*** is an optional construct for grouping items for the purpose of selective harvesting;

Selective harvesting allows harvesters to limit harvest requests to portions of the metadata available from a repository.

3.1 Protocol features

As it has been mentioned earlier, OAI-PMH requests are expressed as HTTP requests. A typical implementation uses a standard Web server that is configured to dispatch OAI-PMH requests to the software handling these requests. The remainder of this section describes the aspects of the protocol that are specific to the HTTP embedding. OAI-PMH requests must be submitted using either the HTTP GET or POST methods. Repositories must support both the GET and POST methods. In addition to the base URL, all requests consist of a list of *keyword arguments*, which take the form of key=value pairs. Arguments may appear in any order and multiple arguments must be separated by ampersands (&). Each OAI-PMH request must have at least one *key=value* pair that specifies the OAI-PMH request issued by the harvester.

Response sent by the server is in the form of the HTTP response, and the Content-Type returned for all OAI-PMH requests must be text/xml. Optionally, Some of the compression schemes for compressing the response can be used.

The format of resulting XML documents created as the response to some OAI-PMH request must comply to following rules:

1. XML document must be well-formed
2. character reference not the entity references must be used, thus allowing XML documents to be treated as standalone documents, independent of any external entity reference.
3. XML document must validate against the XML Schema specified by the PMH.

OAI–PMH supports different formats for data interchange. It is required that any archive involved in OA initiative should support ***DublinCore*** metadata format. The protocol itself does not create any

лом се не праве претпоставке о томе како неки формат метаподатака треба да изгледа, већ се само дефинише елемент XML документа у оквиру кога се метаподаци уписују у складу са траженим (XML) форматом метаподатака.

OAI-PMH протокол подржава и контролу тока размене података путем механизма тзв. токена за наставак. Смисао овакве контроле је у томе што су одговори на неке захтеве у виду листи дискретних ентитета, које у зависности од величине архиве могу бити и веома дуге. У том случају се као одговор на клијентски захтев шаље само део листе уз одговарајући токен за наставак.

Издавањем новог захтева у коме се као једини параметри налазе тип захтева и токен за наставак, клијентска апликација ће као одговор добити наставак листе. Ови токени за наставак морају обезбедити да клијентска апликација увек када изда исти токен добије исти део листе ентитета као одговор. Или, уколико се у међувремену садржај архиве битније променио, клијентска апликација мора примити унапред дефинисану грешку. Из ових разлога токени могу бити и временски ограничени.

У оквиру OAI-PMH протокола дефинисани су и механизми за управљање грешкама. За сваки од могућих захтева дефинисани су дозвољени параметри и њихове вредности. У случају да захтев не одговара спецификацији, клијентској апликацији се у виду одговора враћа XML документ са елементом који у себи садржи код и опис настале грешке. Тако је могуће анализирати разлог због којег је дошло до грешке и евентуално имплементирати механизме за корекцију грешке на клијентској апликацији.

Основни захтеви који се могу издати путем OAI-PMH протокола овде се наводе уз опис одговора који се на захтев добија:

1. GetRecord – враћа тачно један запис метаподатака из архиве. Обавезни параметри спецификују идентификаторе јединице архиве из које се тај запис захтева и формат метаподатака који би требало да буду укључени у запис.

2. **Identify** – одговор на овај захтев представља опште информације о архиви.

3. **ListIdentifiers** – као резултат се добија листа која представља јединствене идентификаторе јединица у архиви. Обавезни параметар је

assumption about metadata format specification, only the exact placement of the XML element that should contain specific metadata formats is specified, while metadata itself is formatted according to relevant format.

OAI–PMH protocol supports the flow control features. Some of OAI-PMH requests (so called list requests) return a *list* of discrete entities. In some cases, these lists may be large and it may be practical to partition them among series of requests and responses. Flow control allows for partition of list responses. In such cases, as a response to its request client will receive only one part of the complete list plus the *resumptionToken*. If the client wants to continue listing it issues a second request with received resumptionToken as parameter. As a result client application will receive next part of the list. Repositories that implement resumptionTokens must do so in a manner that allows harvesters to resume a sequence of requests for incomplete lists by re-issuing a list request with the most recent resumptionToken. The purpose of this is to allow harvesters to recover from network or other errors that would otherwise mean that the list request sequence would have to be started again. If there has been a significant change in repository structure since last request an error message should be displayed. For these purposes resumptionTokens could be time limited.

In the OAI – PMH error handling routines are also defined. For each valid request a complete list of allowed and required parameters and their values are specified. If a request doesn't conform to the specification, the client will receive XML document containing the *error* element with error code and description. This allows for further analysis of the error, and can lead to implementation of some error correction mechanism.

Six requests that can be issued using OAI–PMH protocol are:

1. **GetRecord** – This verb is used to retrieve an individual metadata record from a repository. Required arguments specify the identifier of the item from which the record is requested and the format of the metadata that should be included in the record.

2. **Identify** – This verb is used to retrieve information about a repository. Repositories may also employ the Identify verb to return additional descriptive information.

3. ListIdentifiers – This verb is an abbreviated form of ListRecords, retrieving only headers rather than

MetadataFormat, док опциони параметри омогућавају селективно прикупљање података.

4. **ListMetadataFormats** - добија се листа формата метаподатака коју архива подржава. Зависно од начина имплементације, не морају све јединице у архиви подржавати објављивање у свим форматима метаподатака које архива подржава.

5. **ListRecords** – добија се листа записа које је могуће преузети из архиве у задатом формату метаподатака. Опциони параметри омогућавају селективно прикупљање.

6. **ListSets** – добија се листа која представља структуру архиве, уколико је архива организована на тај начин. Уколико архива не подржава скупове, враћа се одговарајући код грешке.

Илустрација имплементације протокола

Протокол је имплементиран у Java окружењу. На **сликама 1-3** дат је приказ одговора сервера на неке од захтева. За тестирање одзива сервера коришћен је раније споменути *RepositoryExplorer* OA иницијативе. У овом алату могуће је мануелно послати сваки од захтева, са одговарајућим обавезним и опционим параметрима на жељени сервер и проверити одговор који сервер генерише. Одговор се може погледати у форматираном облику, и у облику XML документа који је примљен од сервера. Сервер који смо тестирали представља сервер на којем је имплементирана развојна верзија PMH протокола за дигиталну библиотеку. За све захтеве заједничко је постојање обавезног параметра ***verb*** који дефинише који захтев се шаље серверу, тј вредност овог параметра представља назив саме акције која се извршава.

На Слици 1 приказан је одговор сервера на ***Identify*** захтев. За овај захтев шаље се само обавезан параметар ***verb*** чија вредност је ***Identify***, остали параметри нису допуштени у овом захтеву. На слици се виде подаци сервера од којега је захтевана идентификација (сервер Дигиталне библиотеке теза и дисертација Универзитета у Новом Саду).

records. Required parameter is MetadataFormat, optional arguments permit selective harvesting of headers based on set membership and/or datestamp.

4. ListMetadataFormats - This verb is used to retrieve the metadata formats available from a repository. An optional argument restricts the request to the formats available for a specific item.

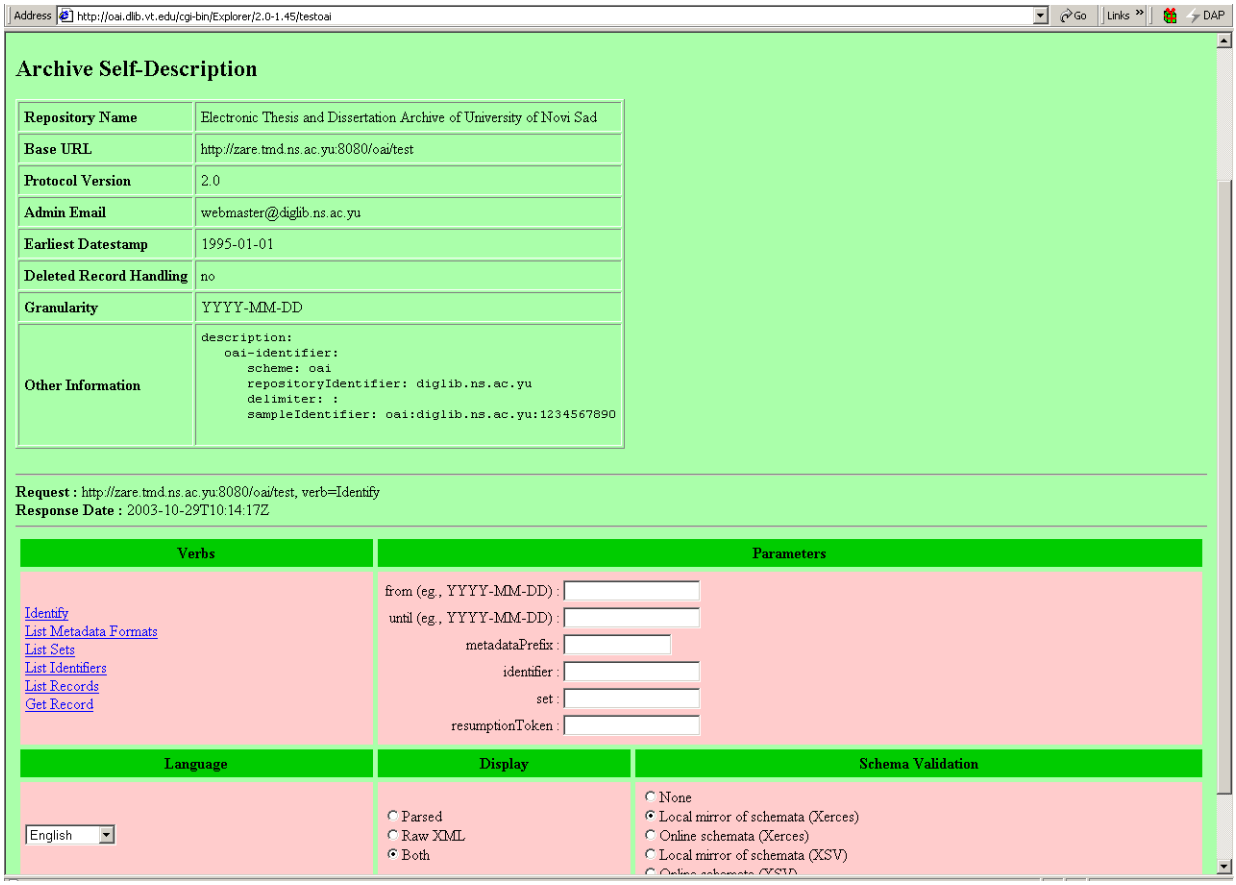
4. **ListRecords** – This verb is used to harvest records from a repository. Optional arguments permit selective harvesting of records based on set membership and/or datestamp.

6. **ListSets** – This verb is used to retrieve the set structure of a repository, useful for selective harvesting.

Protocol implementation examples

Protocol is implemented in Java environment. **Figures 1-3** shows responses received from server as a result of specified requests. For testing purposes OAI Repository Explorer has been used to validate our responses in real world environment. Repository Explorer of OAI initiative allows manual sending of each requests, with form that allows entering all required and optional parameters. The response can be seen as parsed output or as raw XML document. Tested server is server on which we have implemented development version of PMH services for Digital library. Common parameter for each request is ***verb*** which identifies specific request sent by the client application - that is value of this verb represent the action that should be taken.

Figure 1 shows server response to ***Identify*** request. Required parameter is ***verb*** with value ***Identify***. Other parameters are not allowed here. Figure shows basic information returned by the server describing the repository (server of the Digital Library of Thesis and Dissertations of University of Novi Sad).



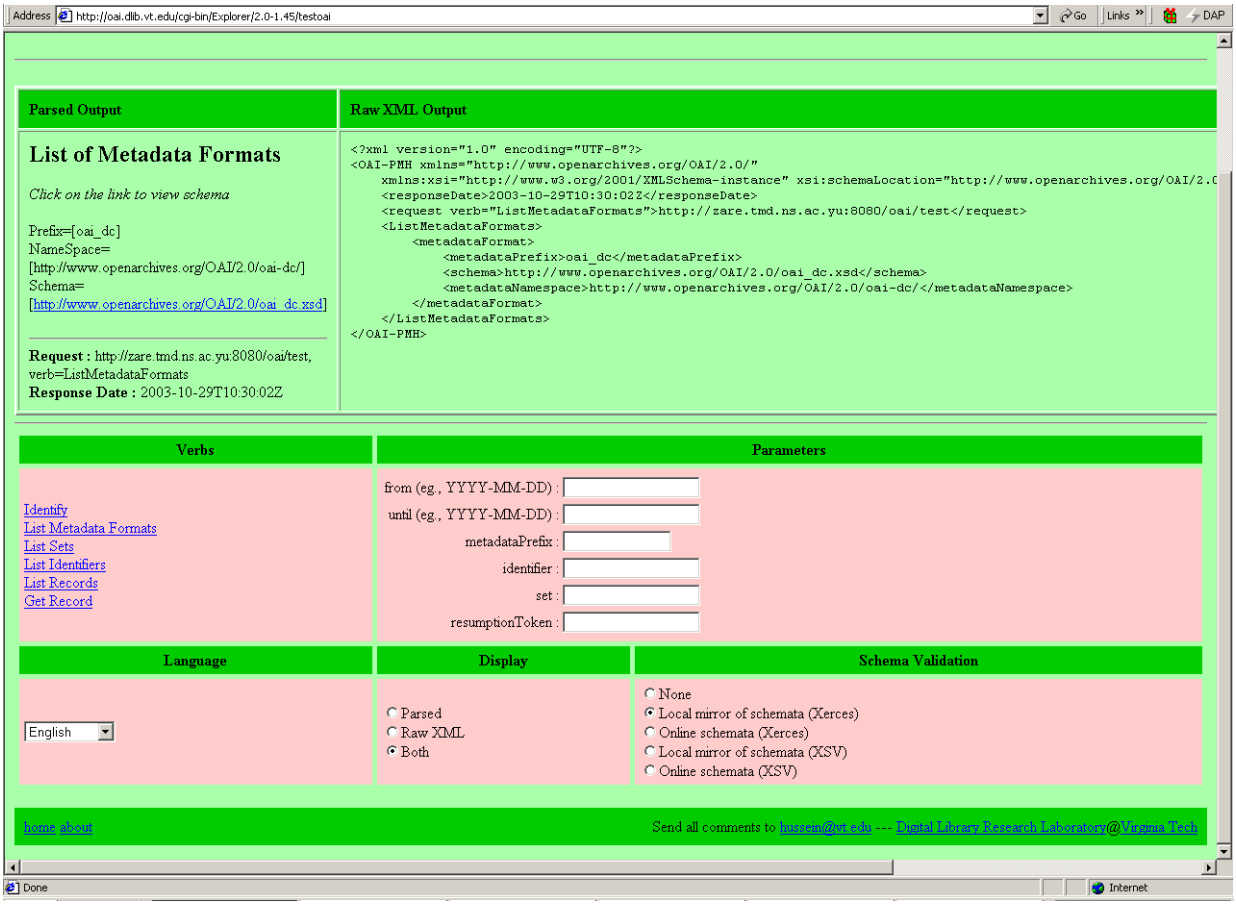
Слика 1 *Одговор сервера на* ***Identify*** *захтев*
Fig. 1 *Server response to* ***Identify*** *request*

У захтеву **ListMetadataFormats** се шаље само обавезни параметар ***verb*** чија је вредност ***ListMetadataFormats.*** На **Слици 2** су дати и форматирани приказ и приказ самог XML документа који сервер шаље као одговор. Са слике се види да је наш сервер у стању да прикаже податке из архиве у *DublinCore* формату.

За **GetRecord** захтев обавезни параметри су, поред ***verb*** и параметри ***identifier*** и ***metadataPrefix***. Као одговор на овај захтев добија се испис садржаја конкретног записа у *DublinCore* формату. На **Слици 3** је приказан један запис добијен као одговор сервера на овај захтев. Како XML документи које сервер генерише по спецификацији морају користити UTF-8 кодирање текста, могућ је и пренос и приказ докумената писаних и ћириличним писмом (приказано на **Слици 3**).

Within **ListMetadataFormats** request only required parameter ***verb*** is sent with value ***ListMetadataFormats.*** **Figure 2** shows parsed and raw response returned by the server. It can be seen that our server currently is able to disseminate metadata only in *DublinCore* format.

Required parameters for **GetRecord** request are ***verb***, ***identifier*** and ***metadataPrefix***. As a response client application will receive content of specified record in specified metadata format (*DublinCore)*. **Figure 3** shows one record received as a result of this request. PMH specifies that XML document must be UTF-8 encoded, thus allowing different languages to be displayed correctly, as shown in **Figure 3**.



слика 2 *Одговор сервера на* ***ListMetadataFormats*** *захтев*
fig 2 *Server response to* ***ListMetadataFormats*** *request*

3.2 Dublin Core

Dublin Core представља једини захтевани формат метаподатака који свака OAI архива мора да имплементира. Спецификација DublinCore формата може се наћи на адреси наведеној у референци [13]. OAI-PMH протокол специфицира коришћење тзв. неквалификованог *DublinCore* формата, а сам елемент XML документа који представља метаподатке описане у *DublinCore* формату мора да се валидира у односу на XML шему [14].

Иако у оквиру пројекта DublinCore постоје специфицирани и додатни елементи, његова употреба у оквиру OAI-PMH протокола специфицира употребу следећих елемената:

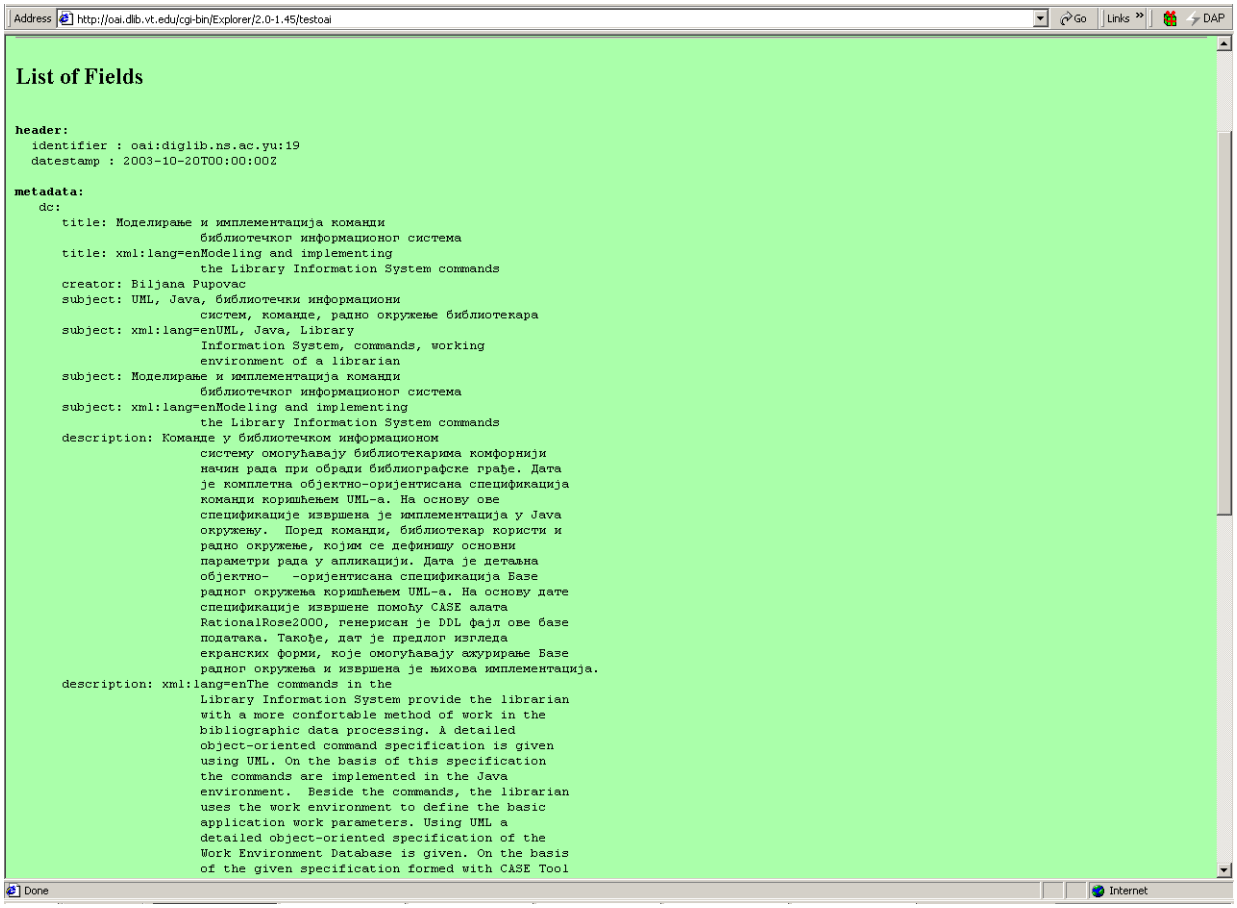
dc:title – Наслов документа тј. назив под којим је документ објављен и познат широј јавности.

3.2 Dublin Core

Dublin Core is the only required format that any OAI archive must implement. DublinCore specification can be found in reference [13]. OAI-PMH protocol specifies usage of unqualified DublinCore format, while the XML element that represents metadata in DublinCore format must validate against specified XML schema [14].

Although DublinCore project consists of specified and additional elements, its usage in OAI-PMH specifies usage of the following elements:

dc:title – Title of the document, the name under which document is published and is generally known.



Слика 3 Одговор сервера на GetRecord захтев
Fig. 3 Server response to GetRecord request

- **dc:creator** – Име особе или особа које се сматрају ауторима документа.
- **dc:subject** – Елемент који описује основну тему документа.
- **dc:description** – Елемент који даје скраћени опис делова и садржаја целокупног документа.
- **dc:publisher** – Елемент који дефинише организацију која се јавља у својству издавача и која је одговорна за објављивање документа.
- **dc:contributor** – Елемент који дефинише особе или организације које су у значајној мери допринеле објављивању документа или неког његовог садржаја.
- **dc:date** – Датум објављивања документа.
- **dc:type** – Описује тип садржаја оригиналног документа (текст у HTML формату, текст у PDF формату, слика итд...). Препоручује се употреба типова наведених у речнику *DublinCore* формата како би се што боље структурно описао садржај.

- **dc:creator** – Name of the person primarily responsible for document content.
- **dc:subject** – A topic of the content of the resource.
- **dc:description** – An account of the content of the resource (an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content).
- **dc:publisher** – An entity responsible for making the resource available.
- **dc:contributor** – An entity responsible for making contributions to the content of the resource.
- **dc:date** – A date of an event in the lifecycle of the resource. Typically, Date will be associated with the creation or availability of the resource.
- **dc:type** – The nature or genre of the content of the resource.

- **dc:identifier** – Јединствени идентификатор који дефинише начин како се може приступити оригиналном документу (уобичајено коришћење подразумева уношење јединствене URL адресе на којој се садржај оригиналног документа може наћи).

- **dc:source** – Извори који су коришћени у документу.

- **dc:language –** Елемент који описује језик оригиналног документа.

- **dc:relation –** Елемент који описује релацију документа са другим документима.

- **dc:coverage –** Елемент који описује област документа.

- **dc:rights** – Елемент који описује права везана за коришћење оригиналног документа.

4. Дигитална библиотека дипломских радова, теза и дисертација Универзитета у Новом Саду

У току 2003. на Универзитету у Новом Саду развијен је и пуштен у пробни рад софтверски систем **Дигитална библиотека дипломских радова, магистарских теза и докторских дисертација** [15]. Систем је и развијан са циљем да буде отворена архива која ће олакшати приступ резултатима истраживања, у виду елек-тронских издања, студентима, научним радницима и свим заинтересованим. При томе су прихваћене смернице које је одредила иницијатива мреже дигиталних библиотека теза и дисертација – NDLTD. Оснивање овакве мреже библиотека је иницирано од стране Техничког универзитета Вирџиније (VirginiaTech University – USA [16]), на основу претходних искустава у размени научних радова у дигиталној форми у оквиру кампуса. Основна идеја јесте да се промовише што лакша размена научних достигнућа, јер су документи смештени у електронској форми у било којој од библиотека које су учлањене у иницијативу доступне за претраживање свим другим члановима иницијативе. Као протокол за размену информација о садржајима који се чувају у дигиталним библиотекама - члановима ове иницијативе усвојен је управо OAI-PMH. (Универзитет у Вирџинији је један од најактивнијих чланова и покретач ОА иницијативе).

Универзитет у Новом Саду се путем своје Дигиталне библиотеке званично прикључио овој светској мрежи библиотека у лето 2003. године. NDLTD иницијатива тренутно броји 189 институција чланица (165 универзитета), са својим дигиталним библиотекама научних радова [17].

- **dc:identifier** – An unambiguous reference to the resource within a given context.

- **dc:source** – A Reference to a resource from which the present resource is derived.

- **dc:language –** A language of the intellectual content of the resource.

- **dc:relation –** A reference to a related resource.

- **dc:coverage –** The extent or scope of the content of the resource.

- **dc:rights** – Information about rights held in and over the resource.

4. Digital Library of Graduation Thesis, Thesis and Dissertations, University of Novi Sad

Software system of Digital Library of Graduation Thesis, Thesis and Dissertaions [15] was developed and deployed in test environment during the year of 2003, at the University of Novi Sad. System is envisioned as an open archive allowing in the form of electronic publications, for easy access, to the results of the scientific researches. System is supposed to be open for access to students, scientists, and all other interested parties. The guidelines proposed by the NDLTD initiative are accepted and implemented. Establisment of such network of libraries was initiated by VirginiaTech University – USA [16], based upon prior experience in document circulation and interchange in electronic form on their campus. The basic idea is to simplify exchange of scientific research results, in the form of electronic documents. The documents stored in one archive, are accessible for all members of the NDLTD initiative. OAI–PMH protocol is accepted as a standard protocol for metadata interchange.

University of Novi Sad, with its Digital library joined the NDLTD initiative in the summer of 2003. NDLTD initiative currently has 189 member institutions (165 univerisities) [17].

Као и OA иницијатива, и NDLTD иницијатива је дала смернице за развој дигиталне библиотеке, док је сама техника имплементације неспецифицирана. У случају Дигиталне библиотеке Универзитета у Новом Саду за софтверску имплементацију је одабрана Јава платформа (JBoss J2EE server, JDK 1.4.2), као систем за управљање базом података искоришћен је SAP DB 7.4, а комплетан софтверски систем инсталиран је на Linux серверима. Приступ дигиталној библиотеци врши се путем Интернета, на адреси http://diglib.ns.ac.yu.

У току развоја имплементиран је и OAI-PMH протокол.Тренутно је подсистем за размену података у пробном раду – тестирање се врши помоћу тест сервера OA иницијативе. OA иницијатива је између осталог својим члановима омогућила и да пре коначног регистровања сервера за размену података, путем тест сервера испитају своје решење. Тест серверу се приступа преко Интернета, а могу се тестирати сви захтеви које би било који клијент PMH протокола могао послати серверу за размену. Ово омогућава да се лако открију могуће грешке у имплементацији, а пре него што се подсистем за размену пријави за размену у оквиру иницијативе.

Последњи корак јесте регистровање архива у оквиру OA иницијативе чиме ће садржај дигиталне библиотеке постати доступан за размену. На овај начин ће и научна достигнућа са Универзитета у Новом Саду бити лакше доступна библиотекама других светских универзитета. При томе је важно напоменути да OA иницијатива не прави никакве претпоставке у погледу права приступа документима, тј. да ли ће неки корисник имати бесплатно право коришћења, или приступа уопште, зависи од политике библиотека, а право коришћења може бити одређено и на бази појединачног документа.

Закључак

Број чланова OA иницијативе се свакога дана повећава, при чему су најбројнији чланови управо универзитети и друге научне институције са својим библиотекама радова у електронској форми. Чланством у OA иницијативи и имплементацијом PMH протокола доступност података из различитих светских архива се вишеструко увећава, што размену научних достигнућа олакшава, а време потребно за примену нових достигнућа знатно скраћује, независно од географског положаја потенцијалних корисника.

As well as OA initiative, NDLTD initiative has given the guidelines for development of digital library, while the technology of the implementation is left to the developers. Java platform (JBoss J2EE server, JDK 1.4.2) has been chosen to be used for development of Digital library of University of Novi Sad. SAP DB has been used as a database management system, and the whole system is run on Linux servers. Users access to the digital library through the Internet on the address http://diglib.ns.ac.yu.

The OAI - PMH protocol has been implemented during the development of the Digital library system Currently, system for data exchange is working in the test environment. OA initiative is allowing its members, before the final registration of their services, to test their implementation. Test server can be accessed through the Internet, and all requests can be tested. This allows for easier error recognition and correction well before the service is publically available.

Last step that needs to be done is to register our server as a data provider within the OA initiative, thus making the content of our digital library available to other members of the inititative. This will make researches and scientific work that took place at the University of Novi Sad easily accessible to othe libraries. The OA initiative does not make any assumption about access rigths concerning some documents. Whether or not some user will gain free access to content of some document is completely left to the libraries.

Conclusion

Number of members of the OA initiative gets increased every day, and most of them are universities and other scientific institutes with their libraries of electronic documents. Members of the OA initiative gain easy access to data of various archives around the world, making scientific data exchange much easier, while the time needed to propagate new achievments is shorten.

РМН није замена за друге комплексније протоколе (нпр. Z39.50), његова намена је да омогући лако преузимање основних метаподатака. Имплементација РМН клијента омогућава библиотечком информационом систему једноставно преузимање записа из других архива у форматима које те архиве нуде, независно од софтверске имплементације. Имплементација РМН сервера омогућава да записи локалног библиотечког система постану доступни другим архивама.

Дигитална библиотека Универзитета у Новом Саду развијана је као интегрални део библиотечког софтверског система BISIS. На тај начин омогућено је преузимање метаподатака у локалне базе записа. У оквиру истог протокола могуће је моделирати и друге формате (нпр. *UNIMARC, MARC, MARC21*). У том случају могуће је реализовати и преузимање комплетне библиографске записе.

PMH is not a substitution for other more complex protocols (such as Z39.50). Its primary goal is to make it easy to implement and to use this framework for metadata interchange. Implementation of PMH protocol client application allows library information systems to retrieve records easily from other archives, in the formats offered by those archives. Implementation of PMH server allows users to gain access to the records stored in your archive.

Digital library of the Novi Sad University has been developed as an integral part of the BISIS. Thus, it allows for metadata from the Digital library to be retrieved to the local database. Using the same protocol other formats (UNIMARC, MARC, MARC21...) can also be used. In that case the retrieval of complete bibliographic records could be achieved.

References

[1] *Мрежна дигитална библиотека докторских, магистарских и дипломских радова* (Покрајински секретаријат за науку и технолошки развој, Аутономна покрајина Војводина, руководилац пројекта Д. Сурла), Нови Сад, 2003.

Digital Library of Disertations, Thesis, and Graduation Thesis (Provincial Secretariat for Science and Technological Development of the Autonomous Province of Vojvodina, Project manager: D. Surla), Novi Sad, 2003.

[2] *Networked Digital Library of Theses and Dissertations*. Virginia Tech, USA. http://www.ndltd.org

[3] *The Open Archives Initiative,* http://www.openarchives.org/

[4] *The Open Archives Initiative Protocol for Metadata Harvesting* http://www.openarchives.org/OAI/openarchivesprotocol.html

[5] *Santa Fe Convention base document,* http://www.openarchives.org/sfc/sfc.htm

[6] *OAI metadata set,* http://www.openarchives.org/sfc/sfc_oams.htm

[7] *Dienst OAI data subset,* http://www.openarchives.org/sfc/sfc_dienst.htm

[8] *OAI Data providers registration template*, http://www.openarchives.org/sfc/data_provider_template.htm

[9] *OAI Service providers registration template*, http://www.openarchives.org/sfc/service_provider_template.htm

[10] *List of metadata formats used in context of OAI,* http://www.openarchives.org/sfc/sfc_metadata.htm

[11] *List of data providers within the OAI,* http://www.openarchives.org/sfc/sfc_archives.htm

[12] *List of service providers within the OAI,* http://www.openarchives.org/sfc/sfc_services.htm

[13] *DublinCore format specification,* http://www.openarchive.org/OAI/2.0/oai_dc

[14] *XML Schema for validation of metadata in DublinCore format* http://www.openarchives.org/OAI/2.0/oai_dc/oai_dc.xsd

[15] Д. Сурла, З. Коњовић, Б. Милосављевић, Г. Сладић, З. Протић, С. Комазец, Д. Окановић. Приказ реализације мрежне дигиталне библиотеке докторских, магистарских и дипломских радова. *Инфотека: часопис за информатику и библиотекарство*, 5(2004)1-2, стр. 75-86.
D. Surla, Z. Konjović, B. Milosavljević, G. Sladić, Z. Protić, S. Komazec, and D. Okanović. An Overview of the Implementation of the Networked Digital Library of Theses and Dissertations. *Infotheca: Journal of Informatics and Librarianship*, 5(2004)1-2, pp. 75-86

[16] Virginia Polytechnic Institute and State University (VirginiaTech), http://www.virginiatech.com

[17] *NDLTD member institutions list*, http://tennessee.cc.vt.edu/~lming/cgi-bin/ODL/nm-ui/members/index.htm