Metadata harvesting in regional digital libraries in PIONIER Network

Authors: Cezary Mazurek, Maciej Stroinski, Marcin Werla
Affiliation: Poznan Supercomputing and Networking Center, ul. Noskowskiego 10, 61-704 Poznan, Poland,
Email: {mazurek, stroins, werla}@man.poznan.pl

Keywords: *digital libraries, open protocols, metadata harvesting, cultural heritage*

Abstract

The national programme "PIONIER - Polish Optical Internet. Advanced Applications, Services and Technologies for Information Society" has been realized in Poland since 2001. One of its main focus was to enrich the content based services in Polish NREN and to reach this goal several digital library installation have been started up. This activity aimed at assistance of librarians and university publishers in digital content management and publishing. However to reflect their expectations (e.g. authorized, local access to academic scripts, identification of the owner of manuscripts, preservation of regional cultural heritage, etc.) PIONIER introduced the concept of regional digital libraries, starting in 2002 with Digital Library of Wielkopolska Region [1]. PIONIER regional digital libraries currently cover the installations based on dLibra software [2]. dLibra is a portable and distributed digital library software prepared to support an entire publication lifecycle. It was developed in cooperation with librarians from a various university libraries and gives users a lot of possibilities from basic library content browsing to RSS [3] based notifications and directory based metadata search. To the end of 2005 there have been four regional digital libraries and additionally several institutional installations deployed in PIONIER[1]. The dLibra Digital Library Framework is a platform developed to be an easily adjustable software basis for digital libraries. The dLibra project has been started in 1999 in the Poznan Supercomputing and Networking Center and for now it has become the most popular Polish digital library framework.

The dLibra platform consists of six specialized portable servers, creating together complex digital library system [4]. Using dLibra client applications a user can store digital objects of any type, such as text (PDF, DjVu, HTML, etc.), images, audio or video. All stored objects can be precisely described with the adjustable set of metadata attributes. There are sophisticated mechanisms supporting creation of the metadata, like dictionaries of attribute values with thesaurus functionality. The dLibra supports many well known standards like MARC[5], RDF[6] and DublinCore[7], which are used for the metadata exchange. When a user wants to access content gathered in the dLibra – based digital library, he/she can use WWW pages generated by the dLibra framework. These WWW pages allows to easily browse and search all digital collections of given digital library and other installations through OAI-PMH interface [8]. Access to all gathered assets is precisely controlled by one of dLibra servers.

In this paper we will address the issue of communication between digital libraries in the sense of the exploration of metadata and information about library structure. The latest functionality provided for all PIONIER digital libraries included implementation of OAI-PMH protocol, which transformed the set of regional digital libraries into the distributed platform where each of digital library became an access point to all regional resources stored in PIONIER digital libraries. The service which handles the communication with other repositories is the dLibra

[1] The complete list of digital libraries in PIONIER is available at http://dlibra.psnc.pl)

Distributed Search Server. It is used to harvest remote dLibra instances by means of the OAI-PMH protocol. It also gives the user a possibility to search through gathered remote metadata. In fact, any OAI-PMH-enabled repository can be harvested and searched using that service. The deployment of OAI-PMH protocol enables communication with other digital library systems – not only those based on dLibra software. In the paper we will reference to other examples of virtual collections [9][10] and mention similar solutions realized in other countries, however neither of them offers such level of unification of access to distributes resources managed by the same kind of digital library framework.

We will conclude the paper with examples of content-based services, which are enabled through the PIONIER platform for distributed regional digital libraries and which are provided for research and education users. There are services such as: virtual collections of regional cultural heritage, distributed exhibitions, scientific comments and annotations for group of digital resources, etc. Another group of complementary services covers also information services provided by Grid environments [11].

PIONIER is currently providing an access to more than 10.000 of digital publications in its regional digital libraries. It is already a huge potential for research and education activity, which is also stimulating the development of regional RENs through building new regional services.

References:

[1] Digital Library of Wielkopolska Region, http://www.wbc.poznan.pl/

[2] dLibra Digital Library Framework, http://www.dlibra.psnc.pl/

[3] Hammersley , B. "Content Syndication with RSS". O'Reilly. 1st Edition. March 2003.

[4] Mazurek, C., Werla, M. – "Distributed Services Architecture in dLibra Digital Library Framework". 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems, 29.03-01.04.2005, Schloss Dagstuhl, Germany. Workshop Proceedings.

[5] MARC Standards at Library of Congress webpage, http://www.loc.gov/marc/.

[6] Klyne, Graham; Caroll, Jeremy J. – "Resource Description Framework (RDF): Concepts and Abstract Syntax", http://www.w3.org/TR/rdf-concepts/

[7] Dublin Core Metadata Element Set ver. 1.1, http://dublincore.org/documents/dces/

[8] Lagoze, C.; Van de Sompel, H. – "The Open Archives Initiative: Building a low-barrier interoperability framework", pages 54-62, Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, VA, USA, June 2001.

[9] Candela, L.; Castelli, D.; Pagano, P. "A Service for Supporting Virtual Views of Large Heterogeneous Digital Libraries" in LNCS 2769, p. 362 – 373, 7th European Conference on Digital Libraries, Trondheim, Norway, August 2003.

[10] Falquet, G.; Mottaz-Jiang, C.; Zisweiler, J. "Ontology Based Interface to Access a Library of Virtual Hyperbooks" in LNCS 3232, p 99 – 110, 8th European Conference on Digital Libraries, Bath, UK, September 2004.

[11] Kosiedowski, M.; Mazurek, C; Werla, M. – „Digital Library Grid Scenarios" in European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, 25-26.05.2004, London, U.K. Workshop Proceedings, p. 189 – 196.

Vitae:

Cezary Mazurek, Ph.D., is the head of the Network Services Department at the Poznan Supercomputing and Networking Center. He received his PhD degree in computer science from the Poznan University of Technology in 2004. His research interests concern a wide variety of advanced network services including portal solutions,

digital multimedia libraries, streaming technologies, distance learning and access to grid services. He has been the manager of numerous projects in those fields coordinated by PSNC. Some of the major ones include the Multimedia City Guide, Polish Educational Portal, Digital Library Framework: dLibra, Nabor 2003-2004, PROGRESS and interactive TV platform. He is author or co-author of over 30 papers in professional journals and conference proceedings.

Maciej Stroiński, Ph.D., received the Ph.D. degree in Computer Science from the Technical University of Gdansk in 1987. Currently he is Technical Director of the Poznan Supercomputing and Networking Center. He is also the lecturer at the Institute of Computing Science of the Poznan University of Technology. His research interests concern computer network architectures and grids. He is the author and co-author of over 130 papers in major professional journals and conference proceedings. He is also a co-author of the "PIONIER: Polish Optical Internet - Advanced Applications, Services and Technologies for the Information Society" programme, which constitutes the basis for building an information infrastructure for science within the years 2001 - 2005.

Marcin Werla started working as a software developer in Poznan Supercomputing and Networking Center in November 2002. After he received his master's degree from the Poznan University of Technology in 2004 he changed his position at the Center and is now a computing systems analyst. He is a technical leader of the dLibra project which is focused on the development of a digital library system framework. The dLibra framework was successfully used as the technological platform of Wielkopolska Biblioteka Cyfrowa and several other digital libraries in Poland. He is also an author or co-author of several papers in conference proceedings.