Building a Heterogeneous Network of Digital Libraries on the Semantic Web

Sebastian Ryszard Kruk and Mariusz Cygan and Piotr Piotrowski and Krystian Samp and Adam Westerski and Stefan Decker*

Abstract. *Ever since computer networks have become popular libraries tend to join into federations. Each federated digital library, apart from delivering discovery and navigation features within their own database allows users to search among other digital library systems. However the variety of users introduces various types of libraries and eventually types of federations of libraries.*

We present the current state of communication between digital libraries. We also present how Semantic Web technologies can help to overcome the problems with creating heterogeneous networks of digital libraries. We finally exemplify possible solutions with existing tools.

1. Introduction

Libraries have always been sources of high quality information in the real world. Digital libraries tend to play a similar role to their real world equivalents. Digital libraries render well organised and certified information. Although the main purpose of digital libraries is to manage digital resources, many "bricks-and-mortar" libraries deliver the content of their resources e.g.: old historic books and documents, paper magazines, etc., through the digital libraries. Many of those resources were previously accessible only to a very narrow group of readers.

Ever since computer networks have become popular libraries tend to join into federations, where each digital library also allows to search among other digital library systems. In this way readers do not have to query every single library themselves.

In the next section we present the motivation to our research on solving the heterogeneity in the networks of digital libraries. In section 3, we introduce the concept of a semantic digital library. We present one of existing implementations - JeromeDL (section 3.1). We describe extensible library protocol (see Sec. 4) and two main components: communication infrastructure and adapters for different communication protocols. In section 5, we present the metadata mediation services. We show the role of the bibliographic ontologies in the mediation service (see Sec. 5.1). Finally we present other protocols used in library communication.

*National University of Ireland, Galway, Digital Enterprise Research Institute, email: firstname.lastname@deri.org
This material is based upon works supported by Enterprise Ireland under Grant No. *ILP/05/203* and partially by KBN, Poland under grant No. 4T11C00525. The authors would like to acknowledge Henryk Krawczyk, John Breslin, the DERI Semantic Web Cluster and the Corrib.org working group for fruitfull discussions.

2. Motivation

The diversity of bibliographic description metadata standards used in (digital) library systems introduces many problems when trying to connect them together. Searching across more than one digital library at once is often restricted to one type of library communication protocol. Communication in a heterogeneous network of digital libraries is hard to establish due to e.g. incompatibility of meta-data standards.

An effort to build a heterogeneous network of digital libraries for managing and sharing information poses many problems. Current technologies, however advanced, might not provide successful solutions to some of interoperability problems.

The Semantic Web aims to deliver technologies that will allow machines to reason on semistructured information. We present how existing Semantic Web solutions can be utilised to build a heterogeneous network of digital libraries.

3. Semantic Digital Library System

The main disadvantage of classic digital library systems is a limited metadata interoperability. Semantic digital libraries (SDL) describe resources with meaningful bibliographic information. Thus, SDLs deliver a solution to one of the problems in a heterogeneous network of DLs. We propose an architecture of a heterogeneous network of DLs (see Fig. 1) where heterogeneous, legacy networks are connected via a fast P2P-based protocol (see Sec. 4.) through edges, adapters (see Sec. 4.2.) and mediation services (see Sec. 5.).

3.1. JeromeDL

JeromeDL is an open-source semantic digital library. Each of its main components benefit from semantic technologies:

- resource management: each resource is described with a semantic description (the JeromeDL structure ontology, MarcOnt bibliographic ontology).
- search and browsing: semantic query expansion for keyword search [9], social semantic collaborative filtering [8] and natural language query answering are the most prominent features delivered by JeromeDL.
- user profile management: JeromeDL utilises the FOAFRealm [7] library to deliver semantic community-aware (Web 2.0) user management
- communication link: the connection to other JeromeDL instances is based on ELP protocol (see Sec. 4.)
- interfaces: apart from user web interfaces, JeromeDL delivers REST [1], web services and mobile [10] interfaces for search and browsing.

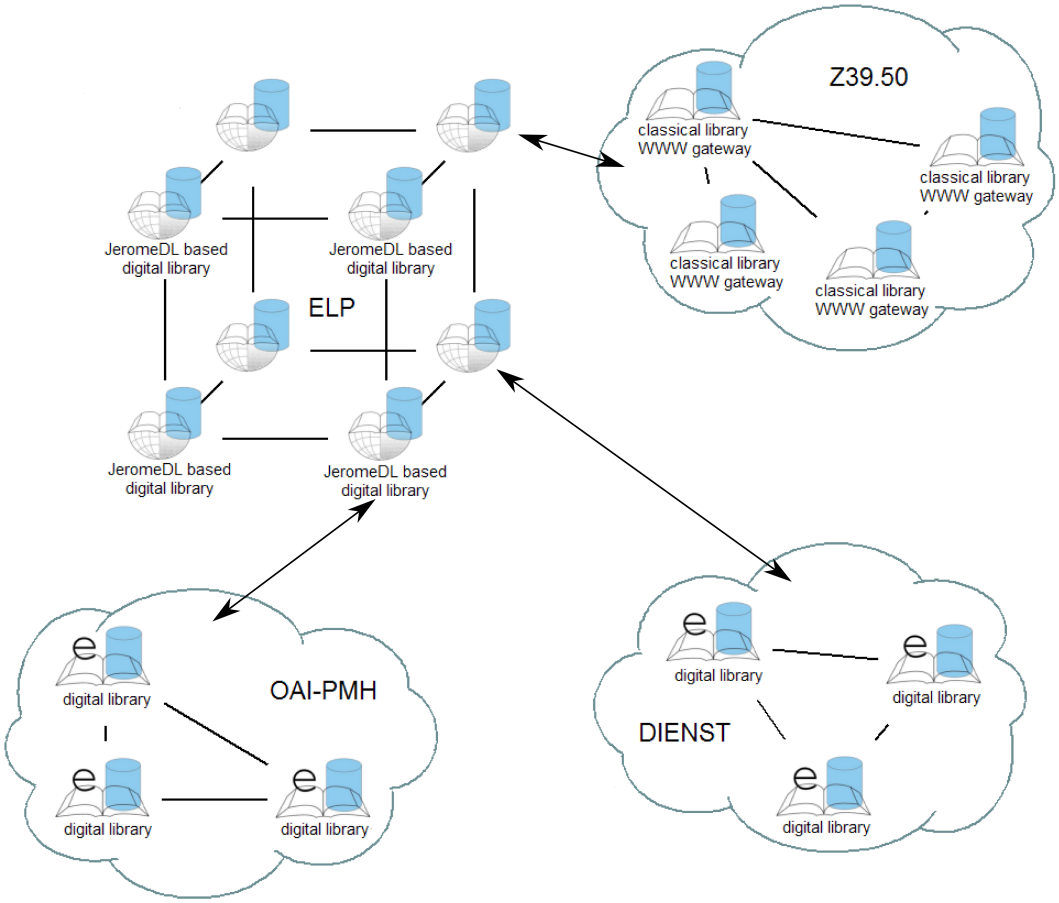

Figure 1. Architecture of a Heterogeneous Network of Digital Libraries

4. Extensible Library Protocol

Communication protocols play the most important role in building library-to-library (L2L) networks [16]. Therefore, the Extensible Library Protocol (ELP) [11] implemented in JeromeDL goes beyond searching by adding, semantic-based features. The meta-data can be easily extended with new concepts, like taxonomies specific for research institutes in a given country. ELP allows the inclusion of supplementary information like a trace of semantic search operations within L2L network. ELP is feasible for solving interoperability issues in heterogeneous L2L networks (see Fig. 2), since it handles messages in a machine understandable way.

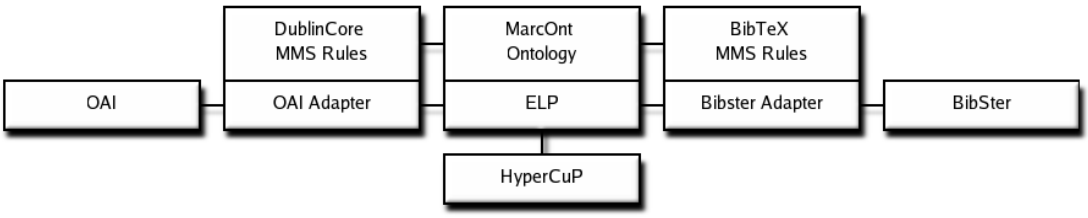

Figure 2. ELP as a mediation protocol in heterogeneous networks

4.1. HyperCuP

The HyperCup Lightweight Implementation [6] delivers a low-level P2P infrastructure for ELP-based communication between JeromeDL instances.

The main advantage of the HyperCuP topology [13, 14] is a very efficient broadcast mechanism. It delivers a solution for transmitting ELP distributed queries across the entire L2L network.

Although the algorithm itself is a result of earlier research [13], it was not previously provided as a ready-to-use independent product. Existing implementations of the HyperCuP protocol created before were built for specific projects. It was the motivation for delivering a multi-purpose lightweight implementation.

4.2. Adapters

An adapter is an element responsible for translating between ELP and other library protocol messages. Different adapters are defined for different protocols. Communication protocol adapters are defined on the lowest level of the interoperability stack in ELP (see Fig.2). The translation process aims to keep as many details of the communication as possible. The search process within a particular L2L network is only limited by the expressiveness of its internal protocol. JeromeDL currently has adapters for OAI[1], OpenSearch[2] and Bibster[3] communication protocols and can easily exchange information with systems using these standards.

5. Metadata Mediation Services

Various libraries use different bibliographic descriptions for their metadata. Therefore, to establish communication between different libraries, the descriptions have to be brought to a "common denominator".

A mediation service converts the input description to the MarcOnt ontology [15] and then to the output description. As opposed to translating directly from one model to another, metadata mediation services requires only two translation rules for each metadata standard. Thus this approach is highly scalable. The direct approach has a quadratic increase in the number of rules required with every new supported standard. The most significant problem is that the intermediate model needs to be able to express all concepts from supported schemata.

Bibliographic descriptions are encoded in a variety of ways. MARC 21 [5] is binary, BibTEX [12] is textual, while Dublin Core [2, 3, 4] is an element set that can be used in XML as well as RDF. Therefore input and output adapters (see Sec. 4.2.) are used to translate between the native format of description and RDF.

The input data is first converted by an appropriate adapter to RDF, a reasoning engine performs the translation of concepts, and then the data is written by an output adapter to an output file. This architecture allows us to easily add support for new formats.

[1] http://www.openarchives.org/
[2] http://opensearch.a9.com/
[3] http://bibster.semanticweb.org/

5.1. MarcOnt Ontology

The MarcOnt ontology is a bibliographic description ontology that aims to express concepts from a number of bibliographic description formats, starting with MARC 21 [5], BibTEX [12] and Dublin Core [2], and later to other formats. Such an ontology can also be used as the native bibliographic description metadata format for semantic digital libraries.

6. Related Work

In order to allow searching in digital library networks, a communication protocol has to be defined. Such a protocol is responsible for sending query requests, responding with results and for other various network communications. Currently, different digital libraries use different metadata standards. This also implies a variety of L2L protocols:

Z39.50 – Z39.50 is one of the oldest protocols used in library environments. It allows queries to be formulated without having to know anything about the target database because its syntax is abstracted from the underlying database structure.

DIENST – DIENST is HTTP based network protocol. It uses HTTP GET queries to embed requests, which later are sent to individually defined services. Each service can support a set of operations, so called verbs. The response from the services can be provided in e.g. `text/xml` DIENST allows i.e. access to resources storage (digital objects), deposition of new resources, discovery and browsing of resources, and user registration. Communication with and among individual DIENST services is established upon an an open protocol.

OAI-PMH – The Open Archives Initiative Protocol for Metadata Harvesting provides an interoperability framework based on metadata harvesting. The OAI-PMH defines two classes of participants: data providers and service providers. The OAI-PMH is based on HTTP requests and XML responses. The response utilises DublinCore as metadata format.

7. Conclusions and Future Work

In this article we sketched the current state of communications between digital library networks. We have identified why it is so hard to create an open heterogeneous L2L network. We elaborated on how Semantic Web technologies can help in building such a network. We pointed out ready to use components like JeromeDL, HyperCuP and MarcOnt mediation services that are being currently used to build the heterogeneous library framework presented in this article. In future work, adapters for DIENST[4] are planned and further MARC21 support is required for the MarcOnt ontology. Together with the BRICKS community[5], the authors are preparing to establish an European heterogeneous network of digital libraries.

References

[1] Roger L. Costello. Building web services the rest way. http://www.xfront.com/REST-Web-Services.html.

[4] http://www.broadcatch.com/dienst.html

[5] http://www.brickscommunity.org/

[2] DublinCore Initiative, http://dublincore.org/documents/dces/. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*.

[3] DublinCore Initiative, http://dublincore.org/documents/2002/07/31/dcmes-xml/. *Expressing Simple Dublin Core in RDF/XML.*

[4] DublinCore Initiative, http://dublincore.org/documents/2003/04/02/dc-xml-guidelines/. *Guidelines for implementing Dublin Core in XML.*

[5] Deborah A. Fritz and Richard Fritz. *MARC21 for Everyone: a practical guide*. The American Library Association, 2003.

[6] Sławomir Grzonkowski and Sebastian Ryszard Kruk. HyperCuP Lightweight Implementation: A Universal Solution for Making Application Distributed. In *ESWC2006 Proceedings*, 2006.

[7] Sebastian R. Kruk. FOAF-Realm - control your friends' access to the resource. In *FOAF Workshop proceedings*, http://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/fp/foaf_realm/, 2004.

[8] Sebastian Ryszard Kruk and Stefan Decker. Semantic Social Collaborative Filtering with FOAF-Realm. In *Semantic Desktop Workshop, ISWC 2005*, 2005.

[9] Sebastian Ryszard Kruk, Stefan Decker, and Lech Zieborak. JeromeDL - Adding Semantic Web Technologies to Digital Libraries. In *DEXA Conference*, 2005.

[10] Sebastian Ryszard Kruk, Sławomir Grzonkowski, Adam Gzella, and Mariusz Cygan. Digime - ubiquitous search and browsing for digital libraries. In *MoSO'2006 Workshop at MDM'2006 Conference*, 2006.

[11] Marcin Okraszewski and Henryk Krawczyk. Semantic Web Services in L2L. In Trojanowski Klopotek, Wierzchon, editor, *Proceedings of the International IIS: IIPWM'04 Conference - Intelligent Information Processing and Web Mining*, pages 349–357. Polish Academy of Science, Springer, May 2004.

[12] Oren Patashnik. Bibtexing. BibTeX distribution, February 1988.

[13] Mario Schlosser, Michael Sintek, Stefan Decker, and Wolfgang Nejdl. Ontology-Based Search and Broadcast in HyperCuP. In *International Semantic Web Conference, Sardinia*, http://www-db.stanford.edu/
s̃chloss/docs/HyperCuP-PosterAbstract-ISWC2002.pdf, 2002.

[14] Mario Schlosser, Michael Sintek, Stefan Decker, and Wolfgang Nejdl. HyperCuP–Hypercubes, Ontologies and Efficient Search on P2P Networks. In *Third International Workshop on Agents and Peer-to-Peer Computing*, 2004.

[15] Kerstin Zimmermann Sebastian Ryszard Kruk, Marcin Synak. Marcont - integration ontology for bibliographic description formats. In *DC2005 Proceedings*, September 2005.

[16] Jan Zielinski, Sebastian Ryszard Kruk, and Henryk Krawczyk. Web services for L2L communication (Uslugi Webowe dla zastosowa L2L). In *Information Technologies*, pages 155–163. Gdansk University of Technology, 2003.