FULL TEXT ELECTRONIC COLLECTION CREATION TECHNOLOGY. WEB ACCESS TO ELECTRONIC COLLECTIONS.

V.A. Antropov, N.M. Kuzmina
Saratov State University
Address: Saratov, 410026, Astrakchanskaya, 83.
e-mail: kuzminanm@info.sgu.ru

Present article concentrates on solving the problem of web access to full text electronic collections, which are based on data in the USMARC format and html-files with articles. By full text collections we mean information system providing access to articles as well as bibliographical data on the article.
LDAP-server was chosen for storing data. There are several characteristics that make an LDAP directory very convenient for storing catalog and collections data [2]:

- it is accessed (read and searched) much more than it is updated
- allows different kinds of searches to be performed in large amounts of text data
- allows to create substring indexes for effective substring searches
- allows distributed databases to be created

There are also features that make an LDAP directory convenient for storing catalog data in MARC format [2]:

- arbitrary field length
- arbitrary number of fields in the record and multivalued fields

These features allow to effectively model MARC records.

LDAP database is populated by a Converter-program, which adds records from MARC-file to the LDAP directory. Effective modeling of the MARC record in LDAP directory [1] provides a problem-free way of converting records with different sets of fields.

The article text is divided into several equal parts, which are stored as values of the article attributes. The index is created on the article attributes. Such method of article division allows substring searches to be performed effectively in the large amounts of text data.

Three level client-server technology is used to access data in the LDAP-directory. The presentation level is implemented by WWW browser. The application level is presented by the Servlet System (Java Servlet) on the WWW-server [3]. Data level is represented by LDAP server [4].

The Servlet System performs the following functions:

- gets the search request, constructs an LDAP-filter based on the request, queries the LDAP-server and returns the results to user's browser, formatted according to the specified template.
- allows to work with selected records separately
- sends articles, records in USMARC format and pdf-files with article sources by e-mail.
- forms dictionaries for choosing search criterion.

Articles and article source pdf-files are accessed through hyperlinks.

Povolzskii Regional Center of New Information Technologies programmers solved the problem of web access to article text as well as bibliographical information on the article in electronic collection. Users are provided with opportunities of manipulating the search results: they might select records to work with them separately and send relevant documents by e-mail. Dictionaries are created to help in forming the search request.

One can view the electronic collection at Sararov State University site:
http://library.sgu.ru/win/nbsgu/index.htm.

Literature:

[1] V.A. Antropov, C.B. Tairbekov, D.F. Shapovalov "Remote Access System to the Electronic Catalog of the Saratov State University Library", Telematica99.
[2] Heinz Johner, Larry Brown, Franz-Stefan Hinner, Wolfgang Reis, Johan Westman "Understanding LDAP".
[3] Live Software, Inc. "Java Servlet API version 2.1.1".
[4] Sun Microsystems Inc. "JNDI: Java Naming and Directory Interface".