Сигла: портал доступа к библиографической информации

© Хохлов Александр Юрьевич

Московский государственный университет
alex@lib.msu.ru

Аннотация

В настоящее время в библиотеках очень популярны порталы доступа к библиотечным ресурсам. В статье описывается текущее состояние развития системы распределенного поиска библиотечных ресурсов Сигла, разработанной в Научной библиотеке МГУ. Особое внимание уделяется тем моментам реализации, которые не встречаются в других аналогичных разработках.

1. Введение

В настоящее время почти все библиотеки имеют электронные каталоги своих фондов. Однако кроме стандартных каталогов библиотеки также начинают создавать собственные ресурсы, а также подписываются на электронные издания. В результате у библиотеки для обслуживания появляется целый комплекс ресурсов, каждый из которых имеет свои специфические особенности.

Для упрощения и улучшения обслуживания пользователей в последнее время стали очень популярны готовые продукты, известные как библиочные порталы. Под ними подразумевается целый класс систем, которые позволяют создать единый интерфейс доступа к разнородным библиотечным ресурсам и управлять доступом к ним.

В данной статье будет описано текущее состояние развития системы Сигла [1], которая была разработана в Научной библиотеке МГУ для интеграции доступа к библиотечным ресурсам по протоколу поиска и извлечения информации Z39.50 [2].

2. Библиотечные порталы и сводные каталоги

Развитие библиотечных порталов идет параллельно с развитием физических сводных каталогов. Библиотечные порталы представляют собой системы распределенного поиска, образующие виртуальные сводные каталоги. Физические сводные каталоги собирают информацию из библиотек в одном месте.

Физические сводные каталоги в том или ином виде присутствуют почти в каждой стране, причем как правило они создаются при государственной поддержке и в одном экземпляре. Однако чем больше территория страны, тем сложнее и дороже оказывается создание подобной организационной структуры. Требуется выработка и принятие единых стандартов работы каждой библиотеки, разработка и внедрение механизмов взаимодействия. Становление сводного каталога происходит в течении десятков лет и его использование ориентировано в первую очередь на оптимизацию работ библиотек по стране за счет устранения дублирования операций над одной и той же копией книги в различных библиотеках.

Крупнейшим сводным каталогом на данный момент является OCLC WorldCat, объединяющий библиографические ресурсы библиотек по всему миру. В России физический сводный каталог создает Центр Либнет на базе двух национальных библиотек. Оба упомянутых сводных каталога являются коммерческими и рассчитаными на использование в первую очередь библиотеками.

Для пользователей библиотеки традиционно представляют бесплатный доступ к своим электронным каталогам. По мере того, как присутствие библиотек в Интернет становится больше, появляется желание объединить возможность поиска по нескольким каталогам на одном веб-сайте. Для этих целей также необходимо разработать единые правила взаимодействия. Однако в данном случае правила становятся проще, а обязательные треботвания – более слабыми.

Как следствие развития этого направления и появились библиотечные порталы, выполняющие роль посредников для организации распределенного поиска. Они существуют либо в виде отдельных веб-сайтов, которые обеспечивают всю необходимую функциональность, либо в виде отдельных продуктов, которые можно поставить и настроить в зависимости от текущих потребностей библиотеки.

3. Общее описание системы Сигла

Сигла представляет собой библиотечный портал для доступа к библиографической информации.

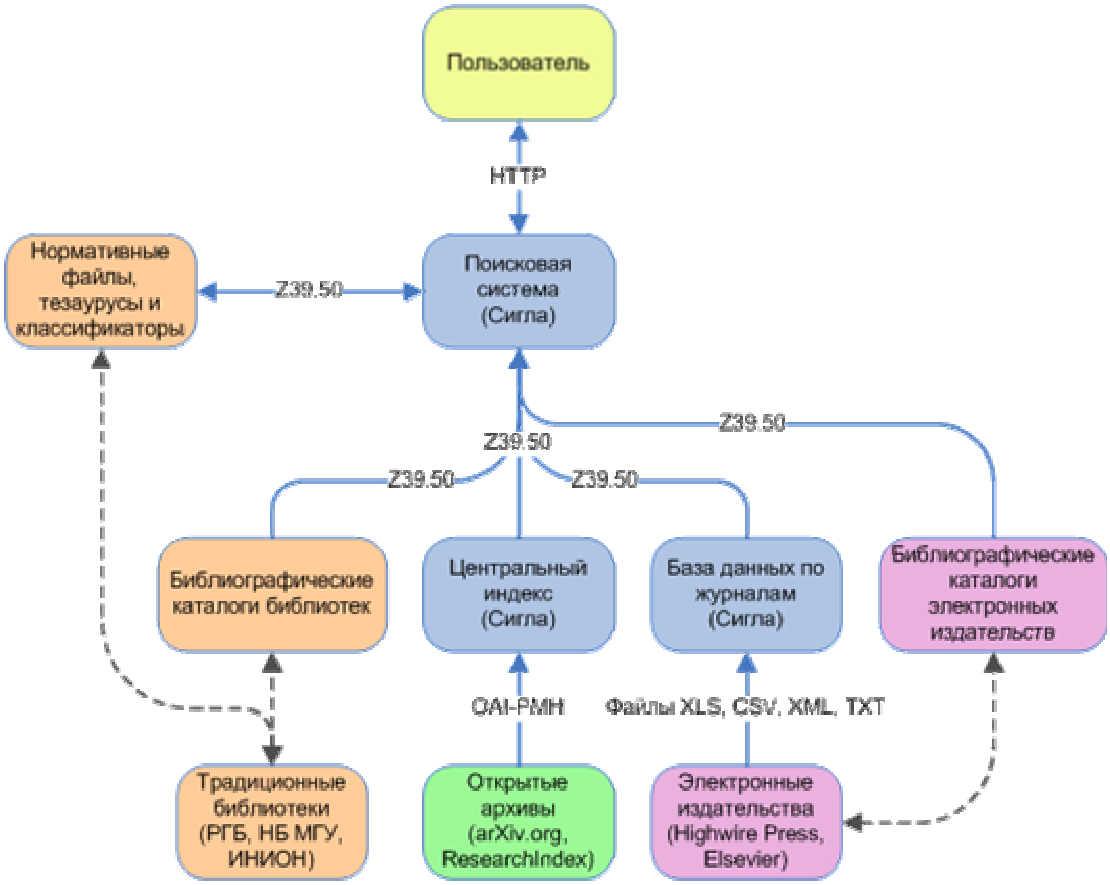

Рисунок 1. Архитекрука поисковой подсистемы

Основными задачами, которые выполняет система, являются распределенный поиск ресурсов в разнородных источниках, показ и работа с результатами поиска, идентификация пользователя и предоставление ссылок на конечные сервисы, релевантные для данного пользователя.

4. Поисковая подсистема

Так как самым распростаненным и проработанным протоколом доступа к каталогам библиотек на сегодняшний момент является Z39.50, то именно он и был выбран для реализации поиска в удаленных каталогах . На данный момент в системе Сигла зарегистрировано более 1500 каталогов, все из которых поддерживают этот протокол.

Однако, для отдельных случаев были созданы специализированные адаптеры доступа. Это было необходимо из-за существенной значимости данных, хранимых в данных каталогах, и отсутствию планов на реализацию стандартных поисковых протоколов для внешнего доступа. Адаптеры выполняют роль преобразователей некоторого заранее определенного API интерфейса доступа к удаленному каталогу в тот формат, который понимает данный каталог. Так как все каталоги из данного класса имеют тот или иной веб-интерфейс, то это сводится к написанию эмитатора веб-браузера, автоматически производящего дейтсвия по поиску и извлечению необходимых результатов.

В последнее время наряду с поисковым протоколом Z39.50 быстро развивается протокол сбора метаданных OAI-PMH [3]. В отличие от протокола Z39.50, его основной целью является предоставление автоматических средств сбора метаданных внешними программными продуктами. Этот протокол начинает широко применяться для организации открытых архивов как средство для организации централизованного поиска, а также для повышения видимости ресурсов. Примером реализации центрального индекса, основанного на OAI-PMH для сбора метаданных, может служить проекты OAIster [4].

На данный момент в системе Сигла реализована отдельная подсистема сбора метаданных по протоколу OAI-PMH с их дальнейшем представлением по протоколу Z39.50 в виде различных каталогов, поэтому работа с ними сводится к работе со стандартными библиотечными каталогами.

Для сбора информации по журналам и статьям была создана отдельная база данных журналов, также представленная в виде отдельного каталога по протоколу Z39.50. Данная база составляется путем сбора информации из электронных издательств и агрегаторов полных текстов.

Общая схема функционирования поисковой подсистемы представлена на рисунке 1.

4.1 Проблемы использования протокола Z39.50

При использовании протокола Z39.50 для построения портала доступа существует несколько проблем. Основными из них являются проблема скорости поиска, а также проблема сохранения семантики поискового запроса в различных релизациях протокола в библиотечных системах.

Для решения проблемы семантики обычно используются профили протокола Z39.50, которые обязательны для реализации всеми каталогами-участниками библиотечного портала, однако в случае одновременного поиска по всем существующим в мире каталогам такой подход неприемлем, так как нет контактов с администратором каждой отдельной системы.

Для решения данной задачи обычно используются настроечные файлы, которые изменяют поведение поисковой системы в зависимости от каталога. Такой подход требует большой ручной работы с каждым каталогом по отдельности. Возможно поставить и настроить прозрачный шлюз протокола Z39.50, который будет преобразовывать запросы в требуемый вид, однако это только решает задачу программирования данной функциональности, но не снимает задачу ручной настройки.

В Сигле был выбран профиль NISO Z39.89 [5] как базовый, и реализована подсистема динамической модификации запросов для автоматической адаптации к различиям в реализациях [6]. Смысл динамической модификации запросов сводится к автоматической модификации поискового запроса при получении поисковой ошибки от сервера Z39.50 и повторной посылке запроса. Правила модификации выбираются таким образом, чтобы результат выполнения нового запроса не противоречил исходному запросу, но, возможно, был бы немного шире.

Практика использования динамической модификации запросов показала, что реализованные правила покрывают бóльшую часть встречающихся разночтений в реализации протокола и не требует дополнительной ручной настройки поиска. Это позволяет легко подключать новые каталоги в систему поиска, не предъявляя минимальных требований к каталогам в виде профиля и не проводя тестирования каталоги на совместимость.

4.2 Оптимизация скорости выполнения запросов

Многие реализации виртуальных сводных каталогов столкнулись с проблемой скорости работы системы. Это было вызвано тем, что поиск в данных системах был реализован синхронно и скорость поиска была не выше, чем самый медленный из участвующих каталогов.

Однако, как показано в практическом исследовании группы разработчиков Indexdata [7], при асинхронной реализации поисковой подсистемы скорость работы системы не зависит от количества участвующих в поиске каталогов и может быть равна скорости ответа каждого конкретного сервера.

Для оптимизации скорости работы поиска в Сигле был реализован механизм асинхронного параллельного выполнения запросов и полностью независимой работы с каждым отдельным каталогом [8]. Это позволило снять ограничение на количество каталогов, одновременно участвующих в поиске, предоставив тем самым возможность поиска во всех каталогах одновременно.

В Сигле был также реализован механизм кеширования последних результатов поиска, что позволяет повысить производительность операций перезагрузок страниц и возврата пользователя к уже полученным в процессе работы результатам.

Кеш записей тесно связан с пулом соединений Z39.50, который используется независимо от текущего пользователя. Это позволяет создавать небольшое количество соединений с каталогом даже при большом количестве одновременно работающих пользователей, что снижает нагрузку на конечный сервер и повышает производительность.

На текущий момент система позволяет производить одновременный поиск в более чем 1500 каталогах. Производительность системы не претерпевает существенного снижения при увеличении количества одновременных пользователей и поисковых запросов.

5. Многоязычие и транслитерация

Библиотеки начали создавать свои электронные каталоги еще задолго до того, как появились общие стандарты представления символов из международных алфавитов и стандарт Unicode [9]. Для упрощения процесса ведения каталога каталогизаторами и упрощения доступа к нему пользователей библиотеки обычно выбирают один базовый язык и один алфавит для создания библиографических описаний. Как следствие, в электронном каталоге библиотеки все описания представлены на одном языке и в одной, максимум двух различных кодировках символов.

Стоит отметить, что в библиотечных системах на данный момент наиболее используемой кодировкой для передачи расширенных символов служит все еще не Unicode, а ANSEL. Поэтому в системе Сигла наряду со стандартными кодировками Unicode, ISO и Windows, была дополнительна реализована кодировка ANSEL, а также ее аналоги для греческих и китайских каталогов.

Для представления иностранных слов в базовом алфавите используются различные правила преобразования букв иностранного алфавита. В основном используется стандарты транслитерации: ISO 9-95 [10], LOC Romanization Rules [11], ГОСТ 7.79 2000 [12]. Однако для полей автора, сведений об ответственности и заглавия часто применяют

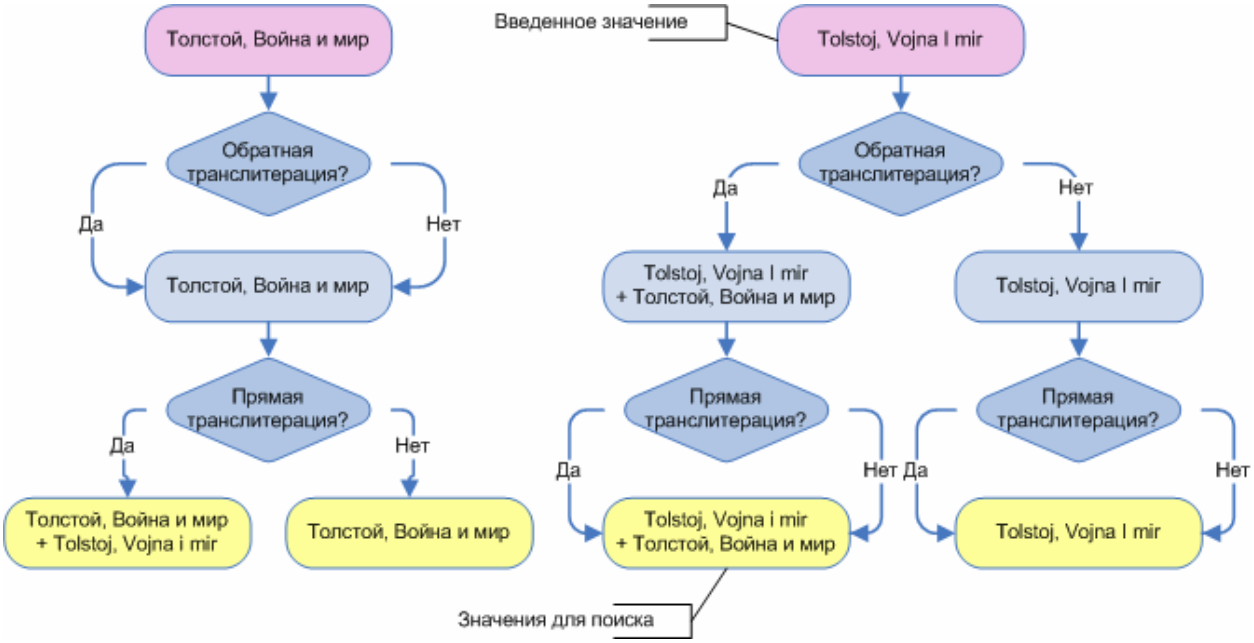

Рисунок 2. Примеры преобразований поискового запроса процедурами прямой и обратной транслитерации

общепринятые формы в родном языке вместо правил транслитерации.

Для повышения качества поиска и возможностей идентификации одного и того же документа в каталогах различных стран мира, в Сигле был реализован прозрачный мезанизм прямой транслитерации введенного запроса из оригинального алфавита в тот алфавит и по тем правилам, которые используются конкретной библиотекой. Это позволяет вводить поисковый запрос в оригинальной форме и не заботиться о различных стандартах каталогизации. Однако ответ от каталога, очевидно, придет именно в той форме, в которой он хранится в конкретной базе данных.

Например, при поиске значения "*Толстой, Война и мир*" в каталоге Библиотеки конгресса США Сигла находит и показывает запись со значением "*Tolstoj, Vojna i mir*". То же верно для других иностранных каталогов, имеющих русскую литературу, описанную при помощи латинской транслитерации.

Также в системе Сигла реализован механизм обратной транслитерации введенных запросов, что позволяет пользователю вводить поисковый запрос на латинице вместо оригинального алфавита. Система сама переведет запрос в оригинальную форму и произведет поиск. Необходимо отметить, что это возможно только при условии, что правила транслитерации обратимы.

Например, при поиске значения "*Tolstoj, Vojna i mir*" в каталоге Российской государтсвенной библитеки при включенной обратной транслитерации Сигла находит и показывает запись со значением "*Толстой, Война и мир*". То же верно для любого другого поискового запроса, в котором указан вид обратной транслитерации.

В общем случае, если для каталога на русском языке включена обратная транслитерация с латиница на русский язык, то поисковые запросы можно вводить как в оригинальной, так и в транслитерированной форме. То же верно для каталога на латинице, для которого установлены используемые правила прямой транслитерации. Примеры работы процедур транслитерации ввода приведены на рисунке 2.

Система также позволяет произвести автоматическую транслитерацию вывода результатов.

Если использовать транслитерацию ввода (поискового запроса) и вывода (записей на экран) одновременно, то получится полный аналог каталога, но в транслитериванном виде на латинице. Т.е. пользователь вводит поисковые значения на латинице, и результат видит тоже на латинице. При этом никаких изменений исходного каталога не требуется – все происходит автоматически. Примером такой прозрачной реализации англоязычной версии каталога служит Электронная библиотека диссертаций РГБ [13] (в англоязычной редакции веб-сайта).

6. Подсистема обработки и представления результатов поиска

Система по мере отклика на поисковый запрос последовательно выдает количество найденных записей в каталогах. После получения от каталога общего количества найденных записей, по требованию пользователя происходит загрузка страницы с первыми 20 записями.

При регистрации каждого отдельного каталога в системе указывается тот формат данных, в котором каталог хранит свои записи. Запрос на предоставления записей всегда происходит в оригинальном для каталога формате для уменьшения потерь информации и

предостращения ее возможного искажения при преобразованиях.

На данный момент реализована поддержка форматов данных, основанных на MARC, таких как MARC21 [14], UniMARC [15], RusMARC [16], а также SUTRS. Для каждого из форматов записей поддерживаются заранее определенные формы вывода.

При наличии в каталоге библиотеки функции заказа документа пользователь может перейти к форме заказа.

Пользователь может отбирать необходимые записи из результатов поиска, и производить с ними различные действия. Поддерживается отсылка записей по e-mail, печать, просмотр записей в виде списка, а также их загрузка на локальный компьютер в форматах MARC с учетом различных кодировок.

При сохранении записей в MARC формате, отличном от формата оригинальной записи, поддерживается преобразование форматов. Это позволяет получить любую запись в любом необходимом формате и кодировке.

Таким образом функциональность системы Сигла не зависит от различий в используемых форматах данных в каталогах.

7. Вспомогательные поисковые средства

В существующих библиотеках накоплено огромное количество документов, поиск и использование которых требует определенных правил их описания. Для систематизации и предметизации документов используются различные тезаурусы, предметные каталоги и классификационные индексы.

7.1 Рубрикаторы и нормативные файлы

В системе Сигла реализована поддержка форматов данных для описания вспомогательных средств. Поддерживаются классификационный и нормативный форматы MARC.

В Сиглу добавлены существующие и доступные по протоколу Z39.50 рубрикаторы и классификационные схемы. В них можно произвести поиск необходимой рубрики или термина, а потом автоматически передти к соответствующим записям в выбранных библиотечных каталогах, использующих данный рубрикатор. Переход осуществляется путем проведения операции поиска термина в каталогах, поэтому не требует физического наличия рубрикатора в каждом из каталогов – только наличие термина в записях и соответствующей настройки поискового сервера.

7.2 Многоязычные тезаурусы

Для предметного поиска в каталоги разных стран по одной и той же тематике требуется создание многоязычных тезаурусов, где одни и те же термины представлены на нескольких языках. Это позволяет формулировать запрос на языке пользователя и находить документы независимо от языка, на котором была проведена их предметизация.

Сигла поддерживает многоязычные тезаурусы, а также позволяет работать с многоязычными нормативными файлами. Для поиска документов, соответствующих выбраному термину, формируется запрос, включающий все формы термина на всех доступных языках. Тем самым будут найдены все документы, в которых встречается хотя бы одна форма на одном из языков многоязычного тезауруса.

8. Подсистема идентификации пользователя

Многие из каталогов, зарегистрированных в системе, являются закрытыми и/или платными. Поэтому в системе Сигла была необходима реализация подсистемы идентификации пользователей и разграничение прав доступа к каталогам. На данный момент реализована система определения принадлежности текущего пользователя к зарегистрированным библиотекам по их IP-адресам. Каждой библиотеке определены права доступа к тем закрытым каталогам, на которые она подписана.

Кроме разграничения доступа к каталогам, идентификация пользователя также используется для разрешения ссылок и показа релевантных сервисов, о которых будет сказано ниже.

Система для каждого сеанса пользователя хранит информацию о выбранных каталогах, а также текущих настройках поискового запроса. Данная информация сохраняется на компьютере пользователя в виде cookie, что позволяет сохранять настройки пользователя между сеансами его работы в системе.

Так, зайдя в систему снова на этом же компьютере, пользователь увидит систему с теми же настройками, которые были им сделаны во время его последнего сеанса работы.

9. База данных журналов

Одним из видов запросов пользователей является поиск статьи из журнала по точным выходным данным. Этот вид поиска не требует наличия электронного каталога для своего выполнения. Его можно обработать путем создания вспомогательной базы данных по журналам и их наличию в библиотеках или электронных базах данных.

В системе Сигла была собрана база данных из более чем 30.000 журналов со ссылками на электронные базы данных, имеющих их полные электронные версии, аннотации или просто постатейную роспись.

Данная база позволяет определить где именно можно найти полный текст журнала и получить к нему доступ.

Если библиотека предоставляет сведения о своей подписке на журнальные электронные базы данных, то, используя эту информацию и идентификацию пользователя, Сигла показывает только те источники, которые доступны текущему пользователю. Это позволяет эффективно заменить те списки заглавий журналов, на которые подписана библиотека, составляемые традиционно вручную. Также библиотека получает в свое распоряжение единый интерфейс доступа к различным агрегаторам полных текстов и электронным издательствам.

10. Дополнительные сервисы и разрешение ссылок

В последние несколько лет очень популярными стали системы разрешения ссылок, которые позволяют по ссылке с метаданными статьи или книги автоматически определить ее местонахождение и показать пользователю соответствующие ссылки. Для передачи метаданных используется протокол OpenURL [8], а системы делятся на 2 типа: те, которые генерируют OpenURL ссылку (источники), и те, которые ее принимают и обрабатывают (целевые системы).

Система Сигла в данном случае является как генератором ссылок OpenURL, так и службой разрешения ссылок. Напротив каждой библиографической записи показывается автоматически формируемая OpenURL ссылка с метаданными, которая указывает либо на заранее определенный сервер разрешения ссылок библиотеки, либо на соответсвующую службу системы Сигла.

При получении OpenURL ссылки система пытается идентифицировать текущего пользователя и показывает ему только те ссылки, которые ему доступны. Основные реализованные виды сервисов в Сигле: поиск полного текста статьи, поиск экземпляра документа в библиотеке, поиск документа в интернет с помощью популярных средств поиска, заказ документа в электронной службе доставки.

12. Заключение

Система Сигла представляет собой хорошо масштабируемый портал доступа к библиографической информации в мире. За более чем год существования в Интренет система зарекомендовала себя как надежное и полезное средство не только для пользователей библиотеки, ищущих интересующих их документ, но и для профессиональных каталогизаторов, осуществляющих поиск редких или сложных документов для уточнения их библиографической информации.

На сегодняшний момент система Сигла является не только веб-сайтом проекта Научной библиотеки МГУ, но и развивается в самостоятельный продукт, который библиотеки могут установить у себя для интеграции их собственных разнородных ресурсов.

На данный момент система обслуживает ресурсы Научной библиотеки МГУ, Российской государственной библиотеки, а также была выбрана для реализации проекта TACIS по объединению ресурсов 5 крупных библиотек России [17].

Литература

[1] Сигла: поиск в библиотеках, 2004 http://www.sigla.ru

[2] ANSI/NISO Z39.50 - 2003 Information Retrieval : Application Service Definition and Protocol Specification http://www.niso.org/standards/standard_detail.cfm?std_id=465

[3] Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0. http://www.openarchives.org/OAI/openarchivesprotocol.html

[4] “OAIster” Web Site, 2004 http://www.oaister.org

[5] ANSI/NISO Z39.89 - 2003 The U.S. National Z39.50 Profile for Library Applications http://www.niso.org/standards/standard_detail.cfm?std_id=734

[6] Хохлов А.Ю. Организация адаптивного распределенного поиска по библиотечным каталогам с использованием протокола Z39.50. *Электронные библиотеки*, том 6 (2), 2003

[7] Sebastian Hammer, 2002. Issues in Z39.50 Parallel Searching http://indexdata.dk/paraz/ParaZ.pdf

[8] A. Khokhlov. Scalable federated search engine using Z39.50 protocol. In *Proc.SYRCoDIS*, pages 50-52, 2004 http://syrcodis.citforum.ru/2004/khokhlov.pdf

[9] The Unicode Consortium. The Unicode Standard, Version 4.0.1, defined by: *The Unicode Standard, Version 4.0* (Reading, MA, Addison-Wesley, 2003. ISBN 0-321-18578-1), as amended by Unicode 4.0.1 http://www.unicode.org/versions/Unicode4.0.1/

[10] ISO 9:1995 Information and documentation -- Transliteration of Cyrillic characters into Latin characters -- Slavic and non-Slavic languages

[11] ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts. (1997) http://www.loc.gov/catdir/cpso/roman.html

[12] ГОСТ 7.79-2001 «Правила транслитерации кирилловского письма латинским алфавитом»

[13] «Электронная библиотека диссертаций» - веб-сайт, 2004 http://diss.rsl.ru

[14] "Marc Standards: Network Development of MARC Standards Office" Web Site, 2004 http://www.loc.gov/marc/
[15] “IFLA Universal Bibliographic Control and International MARC Core Activity (UBCIM)” Web Site, 2004 http://www.ifla.org/VI/3/
[16] РБА. «Российский коммуникативный формат представления библиографических записей в машиночитаемой форме». (2004) http://www.rba.ru:8101/rusmarc/rusmarc/format1.htm
[17] ANSI/NISO Z39.88 – The OpenURL Framework for Context-Sensitive Services http://www.niso.org/standards/standard_detail.cfm?std_id=783
[18] «Российский информационно-библиотечный консорциум» - веб-сайт, 2004 http://www.ribk.net

Sigla: access portal to library resources

Khokhlov Alexandre Yurievich

Library portals for accessing heterogeneous resources are very popular amongst library community. This article describes the current state of development of Sigla – a distributed search system for bibliographic data developed by MSU Scientific Library. Special attention is given to those implementation features that are missing in similar projects.