Распределенный полнотекстовый поиск в электронных библиотеках: технология и проекты

© С. Х. Ляпин
ООО «Константа»
Архангельск
lyapins@yandex.ru

© А. В. Куковякин
ГБУК АО Архангельский краеведческий музей
Архангельск
magus@softconst.ru

© А. В. Чугунов
Университет ИТМО
Санкт-Петербург
chugunov@egov-center.ru

Аннотация

Важными тенденциями развития современной информационной среды являются: а) создание электронных библиотек, предоставляющих сервисы продвинутого полнотекстового поиска и б) создание распределенных информационных систем, обладающих соответствующим функционалом. В этой связи рассматриваются особенности гибкого тематизируемого полнотекстового поиска в многофункциональной электронной библиотеке, созданной на основе информационной системы T-Libra (ООО «Константа», Архангельск, Россия). Обосновывается его использование для поддержки основной деятельности организаций сфер образования, науки и культуры. Описывается эксперимент по реализации распределенного полнотекстового поиска в децентрализованной распределенной среде под управлением пользовательского браузера, обращающегося к множеству серверов, а также некоторые ведущиеся в этом направлении региональные и междисциплинарные проекты.

Доклад подготовлен при частичной поддержке гранта РГНФ № 14-03-12017.

1 Введение. Состояние дел и постановка вопроса

К числу крупных международных инициатив, определяющих основные тенденции развития современного информационного общества относятся:

– создание открытых электронных архивов научных публикаций (Open Archives Initiative) [1]. Эта инициатива поддержана Берлинской декларацией об открытом доступе к научному и гуманитарному знанию, 2003 г. [2], а также Международным соглашением «Берлин-3», 2005 г. [3];

– создание Европейской библиотеки — (The European Library, сокр. TEL) — Интернет-портала, открывающего доступ к ресурсам 48 национальных библиотек Европы и многих исследовательских библиотек [4]. Портал дает возможность поиска как библиографических записей, так и цифровых объектов;

– создания Europeana [5] — европейской цифровой библиотеки, цель которой — обеспечить доступ к отсканированным страницам книг, отражающих различные аспекты европейской культуры.

Эти инициативы и соответствующие проекты ориентированы на предоставление доступа к файловым ресурсам, представленных в случае TEL и Europeana, как правило, в графическом или мультимедийном формате. В проектах Open Archives Initiative ресурсы — научные публикации — представлены и в символьной форме. Но поиск в рамках этих инициатив осуществляется только по метаданным, в том числе в распределенной среде.

Другая группа инициатив и проектов ориентирована на создание полнотекстовых информационных систем, и предусматривает различные формы аккумуляции полнотекстовых ресурсов, поиск и навигацию по ним.

В их числе — Фундаментальная электронная библиотека «Русская литература и фольклор» (ФЭБ) [6] — полнотекстовая информационная система, созданная с целью аккумулировать разновидовую (текстовую, звуковую, изобразительную и т.п.) информацию о русской литературе XI–XX вв., а также фольклоре и истории русской филологии [7].

Подходом, созвучным предлагаемому в настоящем докладе, является создание виртуальных исследовательских инфраструктур, основанных на электронных библиотеках и процедурах коллективной работы, реализованных в программной среде [8].

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

Мы предлагаем, в рамках вышеназванных тенденций, их конкретизацию и развитие с учетом следующих важных моментов.

Во-первых, электронная библиотека должна обеспечивать возможности не только поиска по каталогу, но и продвинутого многофункционального полнотекстового поиска, — то есть позволять в автоматизированном режиме формировать тематические подборки материала из разных документов, причем с точностью до произвольных единиц полнотекстовой информации, а также обеспечивать различные формы презентации результатов поиска [9].

Сами эти единицы информации, тематически связываемые пользовательским запросом, могут находиться в разных «документах» информационной системы, а для распределенной библиотеки — и на разных пространственно удаленных серверах.

Электронные библиотеки с такого рода сервисами должны обеспечивать взаимодействие с существующими АБИС («электронными каталогами») на уровне импорта метаданных, в том числе и с прикрепленными к ним файловыми ресурсами.

Архитектура такой библиотеки должна быть ориентирована на работу в среде Интернет / Интранет и допускать возможность интеграции на основе унифицированных Веб-сервисов как в локальной сети, так и в распределенной среде.

Аналогичные требования по развитию поисковых сервисов и презентационных возможностей информационных систем выдвигаются в рамках различных проектов «Электронного правительства» [10].

Аналогичные требования по развитию поисковых сервисов и презентационных возможностей информационных систем выдвигаются в рамках различных проектов «Электронного правительства» [10].

Во-вторых, электронная библиотека с ее базовыми поисковыми сервисами должна быть адаптирована к условиям работы в децентрализованной распределенной информационной среде, обеспечивающей независимость ресурсной базы каждой из ее организаций-участников и транспарентность пользовательского запроса [11, 12].

В докладе описывается и демонстрируется одна из возможных реализаций такого подхода, технологической основой которой является информационная система T-Libra (разработка ООО «Константа», Архангельск).

В настоящее время библиотеки на основе ИС T-Libra функционируют в Музеях Московского Кремля [13]; в Архангельском краеведческом музее [14]; в ООО «Константа» [15]; в общедоступных библиотеках (ЦГПБ им. В.В. Маяковского, Санкт-Петербург; библиотека «Дом А.Ф. Лосева» (Москва); библиотека киноискусства им. С.М. Эйзенштейна (Москва); научно-образовательных организациях России (Университет ИТМО, Санкт-Петербург), участвуют в различных проектах по созданию современной сервис-ориентированной информационной среды.

В п. 2 рассматриваются основные сервисы полнотекстового поиска в ИС T-Libra; в п.3 описан алгоритм обработки распределенного полнотекстового запроса, объединяющий несколько удаленных электронных библиотек, функционирующих на основе ИС T-Libra, и приведен пример эксперимента по его реализации; в п. 4 дано краткое описание трех текущих проектов, развертываемых на предлагаемой организационно-технологической платформе.

В заключении обозначены перспективы использования идеологии и технологии децентрализованных распределенных электронных библиотек для качественного развития информационного пространства.

2 Сервисы полнотекстового поиска в ИС T-Libra

Информационная система T-Libra позволяет реализовать следующие типы полнотекстового поиска:

А. Абзацно-ориентированный: в документах, включенных пользователем в поисковую область («корзина ресурсов»), находится множество абзацев, удовлетворяющих условиям запроса (тем самым эксплицируется «горизонтальный» микроконтекст, в котором в составе абзаца находятся искомые термины). Авторский абзац выбран в качестве естественной единицы смыслового членения текста.

Б. Частотно-ориентированный: создает частотно-ранжированный список терминов (имен существительных) из документов на заданную глубину с указанием абсолютной и относительной (в промилле) частоты встречаемости термина.

Каждый из этих типов поиска может включать в себя несколько разновидностей.

В рамках распределенной информационной среды в настоящее время реализованы абзацно-ориентированные запросы, наиболее востребованные для тематической обработки документов.

Простой («однослойный») тематический поиск, с одним комплексным полем для ввода терминов и использованием для этих терминов операторов логического объединения, обязательного исключения или обязательного включения термина в запрос. Результатом поиска является список абзацев, удовлетворяющих заданным условиям.

Каждый из абзацев, входящих в результаты запроса, может быть одним «кликом» мышки раскрыт до своего полного вида. Используя опцию «Контекст» в левом меню, можно последовательно раскрыть абзацы до и после найденного — вплоть до кластера из семи абзацев (три абзаца «до», три абзаца «после», плюс сам абзац — результат запроса).

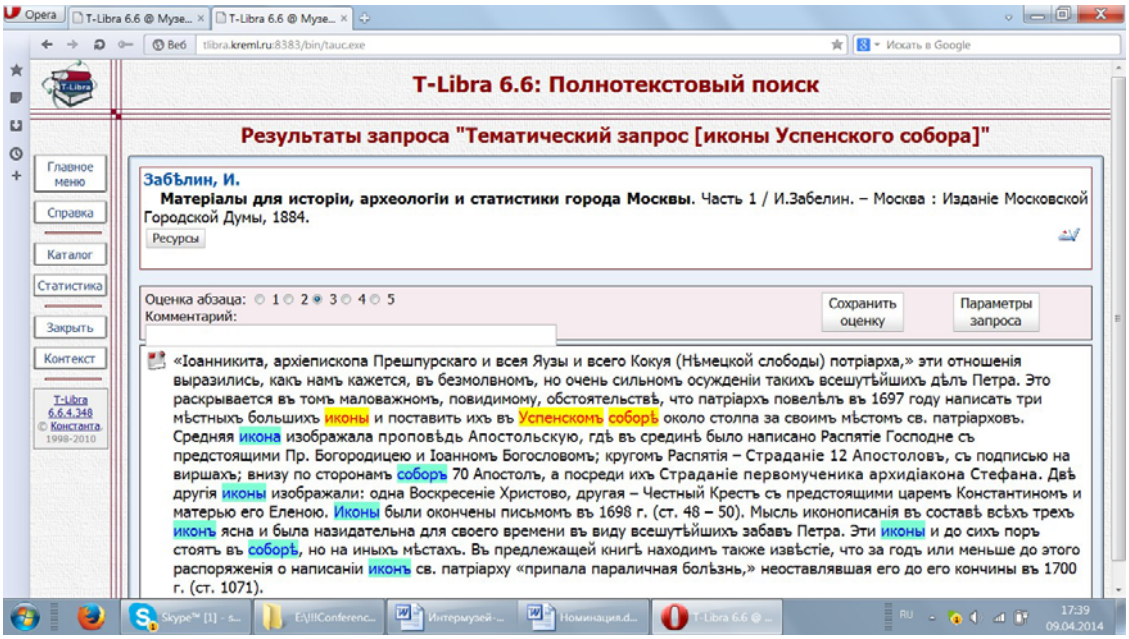

Иллюстрация 1. Один из результатов многослойного запроса «иконы Успенского собора»

Расширенный («многослойный») тематический поиск. Этот вид поиска сохраняет весь функционал простого тематического поиска и обладает дополнительными возможностями тематической фокусировки запроса. Соответствующий инструментарий включает в себя: а) формирование нескольких поисковых полей («слоев») и б) включение в запрос дополнительных количественных параметров его фокусировки.

Поисковое поле «слой» представляет собой технический инструмент для выделения того или иного содержательного «аспекта» интересующей пользователя «темы»; всего может быть сформировано от 2 до 8 слоев. Между слоями действует операция логического пересечения (оператор «AND»), внутри слоя — операция логического объединения (оператор «OR») заданных терминов. Имеется возможность комбинировать актуально используемые слои.

Еще более точная тематическая фокусировка запроса достигается за счет выполнения дополнительных условий: а) указания минимально необходимого количества поисковых слоев (от 2 до 8); б) указания максимального расстояния между терминами, принадлежащими разным слоям: от 0, когда слова из двух разных слоев запроса в составе абзаца примыкают друг к другу, до произвольной величины.

Поисковые термины «иконы Успенского собора» были включены в состав трехслойного запроса (каждый термин вводится в отдельном слое; задано расстояние между поисковыми терминами (в абзаце) не более 8 слов. Всего найдено при этих условиях 22 абзаца в 8 документах. Это — вполне разумное количество для дальнейшего «ручного» анализа. Поиск ведется с учетом дореволюционной орфографии и соответствующей морфологии (хотя сам запрос делался в современной орфографии). Дата осуществления запроса: 09 апреля 2014.

На илл. 1 показан один из 5 абзацев из книги И. Забелина, удовлетворяющий условиям этого многослойного тематического запроса. Красным цветом на желтом фоне покрашены термины, удовлетворяющие дополнительным условиям фокусировки запроса (расстояние между терминами, находящимися в разных слоях, не более 8 слов). Синим цветом на голубом фоне — все остальные термины запроса.

Можно обеспечить и более жесткую фокусировку запроса, задав, например, расстояние между терминами = 1, т.е. все поисковые термины в абзаце должны быть расположены рядом друг с другом. В нашем случае этому условию будет соответствовать всего 1 абзац в 1 документе.

3 Распределенный полнотекстовый поиск и эксперимент по его реализации

3.1 Организация распределенной среды для полнотекстового поиска

При выборе модели организации распределенной среды для сервисов полнотекстового поиска мы ориентировались на Веб-сервисы и Интернет-протоколы. В этой связи была выбрана модель децентрализованной среды под управлением пользовательского браузера, обращающегося к множеству независимых серверов. Этот подход вполне укладывается в парадигму распределенных информационных систем [16], достаточен для наших целей и позволяет не рассматривать более сложные варианты построения распределенных систем, связанные, например, со взаимодействием унифицированного протокола Z39.50 [17, 18] с Интернет-протоколами [19].

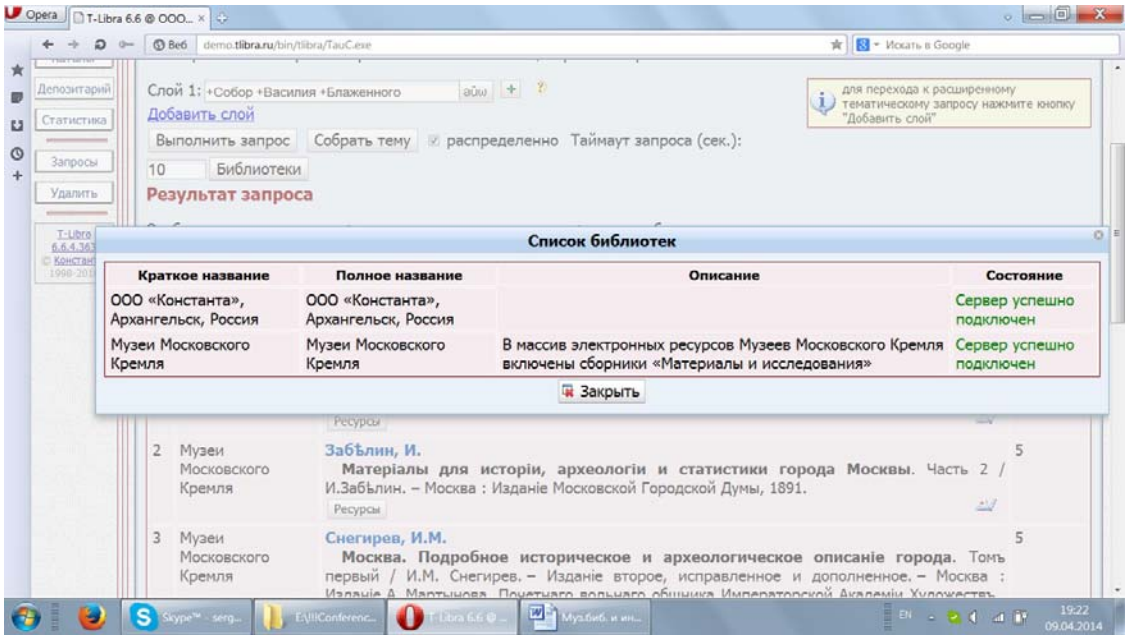

Иллюстрация 2. Страница с результатами распределенного полнотекстового запроса (+Собор+Василия +Блаженного)

Видимо, этот подход может быть рассмотрен как вариант metasearch engine [20]. Инструментом машины метапоиска является в этом случае пользовательский браузер.

3.2 Эксперимент по реализации распределенного полнотекстового поиска

5 апреля 2011 года, в рамках ежегодной конференции «Музейные библиотеки в современном мире», проходившей в Музеях Московского Кремля, был осуществлен эксперимент по реализации распределенного полнотекстового поиска. В нем участвовали электронные библиотеки 5 организаций: 2 в Архангельске и 3 — в Москве.

Вход в распределенную среду осуществлялся с любого из серверов-участников (фактически были использованы 3 из 5 возможностей). Тайм-аут для отклика серверов был установлен в 10 секунд (параметр регулируется пользователем при формировании запроса), этого оказалось достаточно для успешного ответа всех библиотек. Были осуществлены оба варианта абзацно-ориентированного поиска (однослойный и многослойный полнотекстовые запросы), продемонстрирована связь найденных абзацев с релевантными графическими страницами соответствующего документа, а также реализована опция Собрать тему (оценивались найденные абзацы и собирались темы по запросам «Собор Василия Блаженного» и «Иконостас Успенского собора»).

Этот эксперимент был воспроизведен во время видеоконференции, состоявшейся через несколько дней в Российской ассоциации электронных библиотек [21], а в 2012–2014 гг. продемонстрирован (с различным составом участников) еще на нескольких крупных конференциях (Президентская библиотека, октябрь 2012; Крым, июнь 2013; Телематика, июнь 2013; Телематика, июнь 2014).

3.3 Алгоритм обработки распределенного полнотекстового запроса

Замечание. «Базовый» сервер — это сервер, который сформировал страницу, на которой пользователь нажал кнопку «Создать распределенный запрос». Базовым он является функционально, то есть для этого конкретного запроса. Практически это тот сервер, с которым пользователь начал работу с T-Libra в одной из библиотек, включенных в распределенную среду.

1) клиентская часть T-Libra (функционирующая в браузере пользователя), работающая с каким-либо сервером T-Libra («базовый» сервер), получает от базового сервера список адресов других серверов, которые будут участвовать в полнотекством запросе (этот список ведётся администратором базового сервера T-Libra).

2) после заполнения формы полнотекстового запроса пользователь инициирует выполнение запроса (кнопкой «Выполнить запрос»).

3) браузер рассылает http-запросы (на выполнение полнотекстового запроса) на базовый сервер и прочие, перечисленные в списке.

4) каждый сервер выполняет запрос, формирует результат и отсылает в клиентскую часть 10 лучших ответов и общий размер результата.

5) из полученных частичных «топ 10» (от каждого сервера) клиентская часть гарантированно строит и отображает пользователю «топ 10» ответа в

целом (первая страница результата) и суммарный размер результата.

6) в ожидании реакции пользователя клиентская часть рассылает http-запросы (на пересылку остатков результата) на все серверы, участвующие в выполнении запроса.

7) клиентская часть принимает ответы от них, объединяя частичные результаты и позволяя пользователю просмотр следующих страниц результата распределенного полнотекстового запроса.

3.4 Пример осуществления распределенного полнотекстового поиска

На илл. 2 показаны результаты запроса с указанием, из каких библиотек получены ресурсы. На странице активировано всплывающее окно с перечнем библиотек, успешно откликнувшимися на запрос (в данном случае это библиотеки Музеев Кремля и ООО «Константа»). Всего по запросу найдено: 18 документов, в них 51 абзац. Дата запроса: 09 апреля 2014.

4 Текущие проекты

В настоящее время осуществляется несколько региональных и междисциплинарных проектов, основанных на вышеизложенных организационно-технологических подходах.

4.1 Межмузейная распределенная библиотека

Партнерский проект по созданию межмузейной электронной библиотеки с распределенным полнотекстовым поиском стартовал в 24–25 октября 2012 года на региональном семинаре «Музей в современном информационном пространстве» в Архангельске. Его инициаторы и непосредственные участники: Музеи Московского Кремля (научно-справочная библиотека) и Архангельский краеведческий музей (научная библиотека). Часть этого проекта вошла в областную целевую программу сферы культуры Архангельской области на 2013–2015 гг.; в ее рамках предусмотрено создание межмузейной распределенной библиотеки Архангельской области с участием библиотек не менее 10 областных и муниципальных музеев. Функционал каждой из библиотек и распределенной среды в целом будет развиваться в направлении многофункциональности и мультимодальности, что необходимо для поддержки основной деятельности музеев: информационного сопровождения экспозиций и выставок (отбор и подготовка материала для музейных этикеток и аналитических описаний экспонатов), подготовки и проведения экскурсий, обеспечения научно-методической и научно-исследовательской работы в музее.

4.2 Корпоративная сеть муниципальных библиотек

Под руководством ЦГПБ им. В.В.Маяковского (Санкт-Петербург) создается распределенная полнотекстовая корпоративная сеть общедоступных библиотек (КСОБ) Санкт-Петербурга. В нее поэтапно будут включены районные централизованные библиотечные системы. Первый рабочий фрагмент распределенной КСОБ (3 участника) планируется к пуску в эксплуатацию осенью 2014 года, в рамках мероприятий традиционной конференции «Электронные ресурсы библиотек, архивов, музеев». Ресурсно-сервисная ориентация этой среды — поддержка краеведческой работы библиотек, прежде всего в рамках Петербурговедения, предоставление соответствующих сервисов специалистам, научно-образовательным сообществам и гражданам Санкт-Петербурга и России.

4.3 «Humanitariana»

Речь идет о создании виртуального информационно-ресурсного центра для извлечения знаний из гуманитарных текстов на основе продвинутого полнотекстового поиска и функциональной интеграции ресурсов и сервисов в распределенной среде, проект поддержан грантом РГНФ № 14-03-12017 и рассчитан на 2014–2016 гг.

В его рамках при координирующей роли Университета ИТМО (г. Санкт-Петербург) поэтапно создается междисциплинарная информационная распределенная среда с открытым доступом. В рамках проекта разрабатывается типология задач автоматизированного извлечения контекстного знания из гуманитарных текстов, создаются методики составления запросов разного типа и вида для типовых задач извлечения знаний, некоторые из них будут реализованы в технологиях запроса в ходе реализации проекта.

5 Заключение

Предлагаемая организация сервисов полнотекстового поиска и децентрализованной распределенной среды может быть использована, как в описанных выше проектах, для создания кластеров информационно-библиотечного поиска междисциплинарного и межведомственного характера, тем самым — для масштабирования технологии, качественного развития информационного пространства, увеличения количества и повышения качества доступных цифровых ресурсов и предоставляемых поисковых и презентационных сервисов.

При этом предполагается развитие функционала электронных библиотек и распределенной среды в целом в направлении интеллектуализации поиска (семантический анализ текста, разработка и включение тезаурусов в полнотекстовый запрос, разработка и реализация каскадных и гибридных запросов, использование технологии программных

агентов и т.д.) и мультимодальности — в плане предоставления нетекстовых ресурсов и соответствующих поисковых сервисов (включение, в дополнение к полнотекстовым, аудио- и видеоматериалов, осуществление голосового поиска, обработка потокового видео).

Литература

[1] См.: http://www.openarchives.org

[2] http://informika.ru/text/magaz/newpaper/messedu/2003/cour0311/200.htm

[3] Berlin 3 Open Access: Progress in Implementing the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Feb 28th — Mar 1st, 2005, University of Southampton, UK – [Эл. ресурс] URL: http://www.eprints.org/events/berlin3/outcomes.html

[4] http://www.theeuropeanlibrary.org/tel4

[5] http://www.europeana.eu

[6] http://feb-web.ru/feb/feb/about1.htm

[7] http://ru.wikipedia.org/wiki/Фундаментальная_электронная_библиотека

[8] М. Е. Шварцман. Электронная библиотека как основа виртуальной исследовательской инфраструктуры. [Эл. ресурс] URL: http://www.aselibrary.ru/blogs/archives/1209/

[9] С. Х. Ляпин. Сервисы электронной полнотекстовой библиотеки для образования, науки и культуры // Научная периодика: проблемы и решения — 2(14), 2013 март–апрель. URL: http://www.dilib.ru/journal/articles/50865.php; Публикация в eLibrary [Эл. ресурс] http://elibrary.ru/item.asp?id=19013565/

[10] И. А. Мбого, А. В. Чугунов. Электронная коллекция «Электронное государство»: технологические аспекты // Информационные системы для научных исследований: Сборник научных статей / НИУ ИТМО. Труды XV Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург, 10–12 октября 2012 г. — СПб., 2012. — С. 345–347.

[11] С. Х. Ляпин. Как пройти в распределенную библиотеку? // Современная наука: актуальные проблемы теории и практики. Серия «Гуманитарные науки», № 7–8, июль–август, 2012. — С. 17–21. URL: http://www.vipstd.ru/nauteh/index.php/--gn12-07/595; Публикация в eLibrary [Эл. ресурс] http://elibrary.ru/item.asp?id=18279699

[12] А. А. Андрианова, И. А. Груздев, А. В. Куковякин, С. Х. Ляпин. Распределенная ЭБС Российской ассоциации электронных библиотек // Новые информационные технологии и менеджмент качества (NIT&QM'2013). Материалы межд. научной конференции / Редкол.: А.Н. Тихонов (пред.) и др.; ФГАУ ГНИИ ИТТ «Информика». — М.: ООО «АРТ-ФЛЭШ», 2013. — С. 118–123.

[13] http://www.kreml.ru → наука → библиотека → T-Libra

[14] http://gd.kraeved29.ru/tlibra

[15] http://demo.tlibra.ru

[16] Э. Таненбаум, М. ван Стеен. Распределенные системы. Принципы и парадигмы / пер. с англ. В. Горбунков. — СПб.: Питер, 2003. — 877 с. (С. 23).

[17] http://www.iso.org/iso/catalogue_detail?csnumber=27446

[18] О. Л. Жижимов, Н. А. Мазов. Модель распределенной информационной системы Сибирского отделения РАН на базе протокола Z39.50 // Электронные библиотеки. — 1999. — Т. 2, Вып. 2. [Эл. ресурс] http://www.elbib.ru/index.phtml?page=elbib/rus/journal/1999/part2/zhizhimov

[19] Н. В. Максимов, М. А. Сысойкина. О реализации электронной библиотеки с использованием протоколов HTTP и Z39.50 // Электронные библиотеки. — 2002. — Т. 5, Вып. 1. [Эл. ресурс] http://www.elbib.ru/content/journal/2002/200201/MS/MS.ru.html

[20] http://en.wikipedia.org/wiki/Metasearch_engine

[21] http://www.aselibrary.ru/conference/conference43/conference432039

Distributed Fulltext Search in the Digital Libraries: Technology and Projects

Sergey Lyapin, Alexey Kukovyakin, Andrey Chugunov

An important development trends of the modern information environment include: a) the creation of digital libraries providing services for advanced full-text search; b) the creation of distributed information systems with relevant functionality. The paper presents the features for flexible full-text search in the multifunctional digital library, created on the basis of information system T-Libra ("Constanta" Ltd., Arkhangelsk, Russia). Its use to support the core activities of organizations in the spheres of education, science and culture is justified. The experimental implementation of a distributed full-text search in a decentralized, distributed environment running a custom browser that accesses multiple servers, as well as some ongoing in this direction regional and interdisciplinary projects are described.