Создание эффективного инструмента формирования личных полнотекстовых коллекций для научной и образовательной деятельности

© Б.В.Олейников

Красноярский государственный университет
oleynik@krasu.ru

Аннотация

В статье обосновывается необходимость создания инструмента эффективного формирования личных полнотекстовых коллекций. Описывается разработанный инструмент TAPIR, связанный с MS Internet Explorer 6 и библиотечной системой Greenstone. Рассматриваются сценарии его использования.

1. Необходимость создания личных полнотекстовых коллекций

Одним из основных моментов ведения эффективной научной и (или) образовательной деятельности является получение необходимых исходных информационных ресурсов. Информационные ресурсы (представленные в том числе и полными текстами) в настоящее время, как правило, размещаются на открытых серверах сети Интернет и в специализированных электронных библиотеках, доступ к которым также обеспечивается посредством Интернет.

Возможность работы с открытыми ресурсами Интернет обеспечивается в основном протоколом HTTP, браузерами и общими поисковиками. Электронные библиотеки же, а также многие другие хранилища информационных ресурсов (репозитарии, базы данных и др.) с позиций доступа к ним, в отличие от открытых ресурсов Интернет, представляют собой некоторые автономные объекты, имеющие свои поисковые механизмы и свой доступ к ресурсам (например, для библиографических систем это технологии доступа, поиска и доставки, определяемые протоколом Z39.50 и внутрибиблиотечными поисковиками).

Необходимость всегда иметь «под рукой» требуемые для работы исходные информационные ресурсы, а также учет фактора негарантированности сохранения открытых ресурсов в сети Интернет вынуждает пользователя создавать личные наборы найденных информационных ресурсов для последующего целенаправленного их использования.

В настоящее время для оперативного ведения этой деятельности пользователь в основном может воспользоваться следующими возможностями:

- Прямое сохранение web-страницы (или выделенной ее части) на своем компьютере в некоторой папке в одном из 3-х форматов, предоставляемое браузером;
- Использование специальных инструментов типа менеджеры закачки файлов (например, FlashGet [13]), менеджеры закачки сайтов в целом (например, WebCopier, Teleport Pro [2]), менеджеры закачки любой части сайта (например, Net Snippets, Onfolio [3,4]), web-закладки и web-органайзеры (например, CyberArticle (WebCatcher), NetCollector, ContentSaverProfessional, обзор и ссылки см. [5]) для сохранения ресурсов в папках и дальнейшей работы с ними;
- Сохранение найденного и выделенного информационного ресурса в определенном исходном формате в виде некоторого файла (например, графического файла);
- Копирование найденных (в Интернет, или в электронной библиотеке) ресурсов в некоторые редакторы (например, Word) и их сохранение в виде документов этих редакторов;
- Некоторые другие аналогичные операции.

Каждый из этих подходов направлен на решение определенных и достаточно узких задач. Поэтому пользователю приходится работать со многими инструментами, тратить много времени на проведение рутинных вспомогательных операций, что значительно увеличивает время создания личных наборов ресурсов, не говоря уже о морально-физических утомительных издержках и нерациональном использовании Интернет времени. К тому же в дальнейшем, когда документов накопится достаточно много, такой пользователь неизбежно столкнется с проблемами некоторого унифицированного описания и представления хранящегося информационного материала, а также эффективного полнотекстового поиска в своих наборах информационных ресурсов.

Учитывая разнородность оформления и представления полученного исходного материала, а

Труды 8ой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - RCDL'2006, Суздаль, Россия, 2006.

также требование некоторой унификации его описания в дальнейшем использовании (например, XML-представление) логичнее было бы обеспечить пользователю удобный инструмент создания личных полнотекстовых коллекций. По определению коллекция – это совокупность документов различных форматов, которые собраны вместе на основании определяемых пользователем критериев и к которым применяется единые механизмы хранения, индексации, поиска, просмотра и представления [8].

Желательно чтобы такой инструмент не только мог поддерживать стандартные функции, связанные с описанием, каталогизацией, хранением, поиском, присущие классическим автоматизированным информационным библиотечным системам, но и обеспечивал бы максимально возможную автоматизацию работы с электронными информационными ресурсами, минимизирующую временные затраты на рутинные операции по работе с ними (поиск, взятие, индексирование, формирование описания и др.), и давал бы возможность полнотекстового поиска. Это в дальнейшем позволит с минимальными затратами эффективно создавать и использовать коллекции, как для исследовательской работы, так и для создания электронных образовательных ресурсов, основным требованием к которым в настоящее время является удовлетворение общепринятым международным стандартам, спецификациям, эталонным моделям (IMS, SCORM, LOM, XML и др. [15]). Очевидно, что такой инструмент должен быть основан на некоторой стандартной свободно распространяемой библиотечной системе.

2. Свободно распространяемые полнотекстовые библиотечные системы

В настоящее время существует достаточно много свободно распространяемых библиотечных систем, в том числе и с открытым кодом, которые могут быть положены в основу создания личной полнотекстовой библиотеки. К их числу можно отнести: Archimede, ARNO, Avanti Circulation System , CERN Document Server Software (CDSware), DLBox, DSpace Durable Digital Depository (DSpace), Eprints Archive Software (EAS), ETD-db, Fedora, GNUTECA, Greenstone, Ibero-american and Caribbean Digital Library Project, i-Tor, Kepler, Koha Open Source Library System, LearningAccess ILS, Museolog, MyCoRe, OCLC SiteSearch Open Source Project (The OCLC SiteSearch™), Open Digital Library toolkit, The Open Source Digital Library System Project (PYTHEAS), OPUS, и др. [8, 10, 11].

Практически все из них обладают определенными достоинствами, однако на сегодня из этого множества библиотечных систем по функциональности, известности, использованию и русской локализации, выделяются две: DSpace [12] и Greenstone [14], в качестве примеров эффективного использования которых можно привести научную библиотеку Уральского государственного университета (DSpace), электронную библиотеку Правительства республики Марий Эл и проект создания распределенной электронной библиотеки LibWeb (Greenstone). С позиций существующих локализаций, возможно, большее внимание отечественными исследователями было уделено системе Greenstone.

Система Greenstone разработана на факультете компьютерных наук университета Вайкато в Новой Зеландии в рамках проекта по созданию цифровых библиотек. Руководитель проекта – Ян Виттен (Ian H. Witten). Разработка проводилась при содействии ЮНЕСКО и неправительственной организации Human info. Распространяется с ноября 2000 года, постоянно дорабатывается. В настоящее время распространяется версия Greenstone 3.

Из основных характеристик системы отметим, что она эффективно работает со многими импортируемыми форматами: текстовые файлы (.txt, text), HTML-файлы (.htm, .html, .cgi, .php, .asp, .shm, .shtml, файлы Word (.doc), файлы формата pdf (.pdf), файлы формата PostScript (.ps), файлы e-mail сообщений, библиографические файлы формата BibText (.bib), библиографические файлы формата refer (.bib), файлы исходных текстов программ (Makefile, .Readme, .c, .cc, .cpp, .h, .hpp, .pl, .pm,.sh), файлы изображений для создания библиотеки изображений (jpeg, .jpg .gif, .png,.bmp, .xbm,.tif, .tiff), файлы базы данных системы FoxBASE (.dbt, .dbf), файлы базы данных системы CDS/ISIS (.mst); использует для описания документов Dublin Core Metadata (Дублинское ядро - Стандарт Z39.85); обеспечивает внутреннее XML-представление документов; имеет развитые возможности атрибутивного и полнотекстового поиска на основе системы Managing Gigabytes; имеет графический пользовательский интерфейс и др.

3. Описание инструмента TAPIR для создания личной полнотекстовой коллекции

Представляемый в данной работе инструмент TAPIR (to Take And Place an Information Resource – взять и разместить информационный ресурс), представляет собой многофункциональный интерфейс, направлен на минимизацию затрат по получению интересующих информационных ресурсов и последующему их размещению в личной электронной полнотекстовой библиотеке, поддерживает работу с двумя типами ресурсов: открытый Интернет (HTTP), и ресурсы, предоставляемые по протоколу Z39.50. В разработке программного обеспечения под руководством автора принимал участие студент В.В.Зинкевич. Общая схема программного обеспечения TAPIR представлена на Рис.1

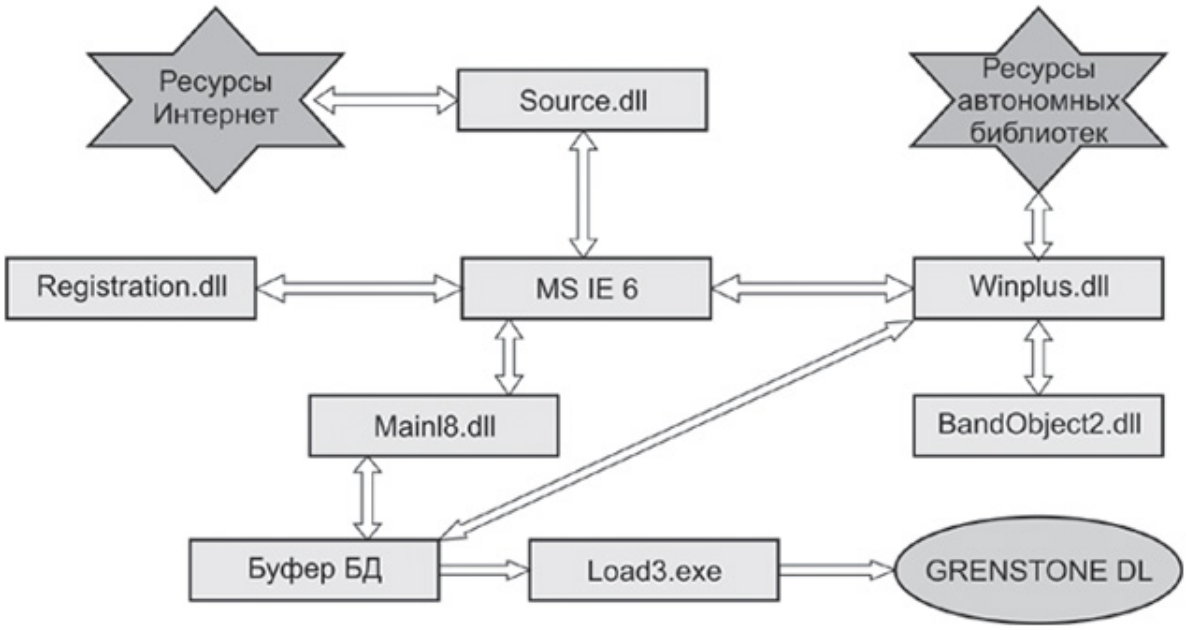

Рис.1 Схема программного обеспечения TAPIR

Программное обеспечение TAPIR включает:

- стандартный браузер Microsoft Internet Explorer 6, дополненный плагином, включающим программные модули **registration.dll, BandObject2.dll, Winplus.dll, Source.dll, main18.dll**;
- специальный **буфер**, предназначенный для предварительной обработки данных;
- загрузчик **Load3.exe;**
- командный модуль **ex.cmd**;
- полнотекстовую библиотечную систему **Greenstone Digital Library.**

Модуль **registration.dll** обеспечивает сбор информации от пользователя, где и что искать. Для диалога с пользователем используется форма, где пользователь может сделать пометки, записи.

В модуле **Winplus.dll** реализуется доступ к ресурсам по протоколу Z3950, осуществляется отображение полученных данных на панели браузера, обеспечивается сохранение записей в форматах MARC, ISO2709, OAI, MARCXML.

Модуль **BandObject2.dll** является дополнением к Winplus.dll. В BandObject2.dll реализованы функции необходимые для добавление новой панели браузера, для этого необходимо было реализовать элементарные низкоуровневые функции по работе с окнами, обработчики событий, и т.д. Главной функцией этого приложения является создание промежуточного окна, в которое можно размещать другие окна.

В модуле **Source.dll** реализуется запрос к поисковым машинам. Первоначально программа обращается к контейнеру объекта WebBrowser2, для того чтоб получить URL поисковых машин, а также выражение запроса, логин, пароль. После этого, выполняется запрос к серверам поисковых машин. Затем полученные данные фильтруются, отбрасываются одинаковые ссылки, формируется новый Html-документ.

В модуле **main18.dll** реализованы следующие функции:

- Получение текущего выделения web-страницы.
- Подсчет статистики слов, на её основании ведется предварительное определение кодов ГРНТИ и УДК.
- Сбор URL изображений, встречающихся в текущем документе.
- Обработка html-документа в целях обнаружения дополнительного описания документа.
- Получение информации о поисковом образе, сохраненном приложением Registration.dll. Для этого текущий COM-объект обращается к объекту WebBrowser2 Object, а через него к контейнеру, хранящего эту информацию. Таким образом, приложение получает выражение запроса.
- Размещение всей полученной информации в XML-файле и последующая отправка ее в буфер.

Модуль **Load3.exe** подготавливает файлы необходимые для помещения в электронную библиотеку Greenstone. Так как кроме полнотекстового поиска осуществляется ещё и атрибутный, то в целях возможности выполнения этих функций для каждого импортируемого ресурса подготавливается файл metadata.xml, содержащий описание ресурса. Учет того, что пользователь может просматривать файлы в Word-формате, требует его создания и последующего конвертирования в него данных,

полученных от приложения Winplus.dll. Затем модуль запускает командный скрипт ex.cmd. Так как загрузка данных из буфера в Greenstone автономна, то запуск модуля Load3.exe может быть осуществлен в любое назначенное время. Это делается с помощью приложения “Назначенные задания”.

Буфер представляет собой обычную выделенную папку в которой помещаются подготовленные для ввода в Greenstone информационные ресурсы в виде XML-файлов. Его наличие обеспечивает возможность двухэтапного подхода при работе с информационными ресурсами, который обусловлен тем, что операция индексации информационных ресурсов при помещении в Greenstone требует значительных временных затрат. Поэтому наличие буфера позволяет разнести во времени основные операции и тем самым оптимизировать по времени импорт новых данных в электронную библиотеку.

Модуль **ex.cmd** предоставляет командный скрипт**.** Его цели:

– Импортирование и построение коллекции
– Запуск скриптов (setup.pl, third.pl), занимающихся редактированием конфигурационного файла коллекции collect.cfg , а также копированием файлов.

Использование плагина для нахождения и взятия требуемых ресурсов осуществляется с помощью специальных кнопок, размещаемых на стандартной панели инструментов браузера (см. рис.2) и набора пользовательских форм, соответствующих сценариям работы с информационными ресурсами.

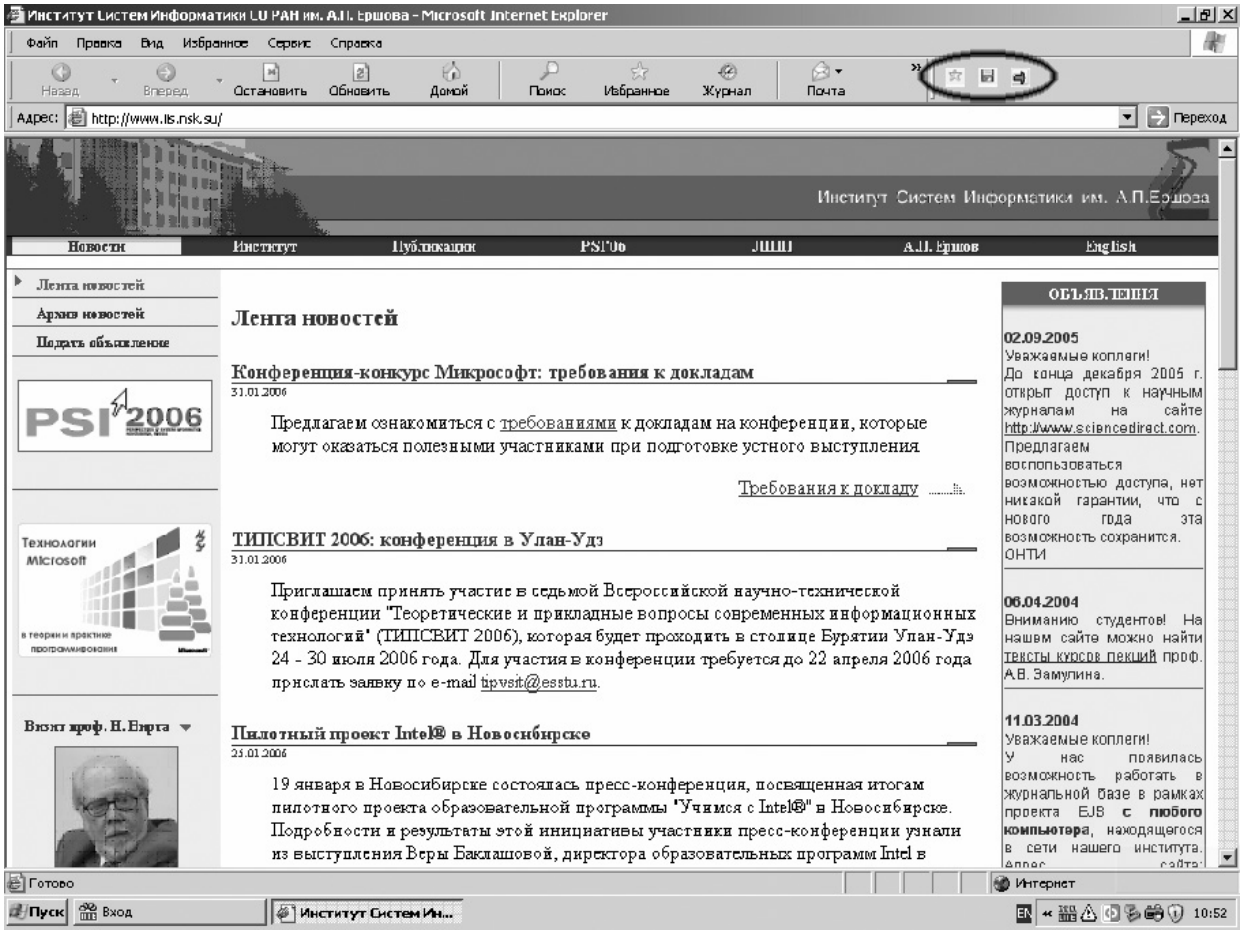

Рис.2 Указание дополнительных кнопок плагина (системы TAPIR) на панели браузера

4. Сценарии работы пользователя

4.1 Открытые Интернет ресурсы.

Сценарий работы пользователя с открытыми ресурсами Интернет включает нахождение с помощью браузера требуемого ресурса в Интернет и, при необходимости, выделение части ресурса, которую необходимо сохранить в библиотеке. Для выполнения остальных операций пользователю достаточно только щелкнуть по соответствующей специальной кнопке на панели браузера. При этом, соответствующие модули плагина в автоматическом режиме обеспечат все

действия, связанные с сохранением URL, и, если возможно, определением атрибутов выделенного Интернет-ресурса: название, автор, время публикации.

Кроме этого также автоматически генерируется полный набор ключевых (наиболее часто встречающихся) слов текста (с указанием ограничений на длину), который может быть визуализирован и в дальнейшем либо полностью, либо частично использован для полнотекстового поиска в системе Greenstone и предварительного определения УДК. Получение предварительного значения УДК осуществляется путем прямого сопоставления полученной частотной таблицы ключевых слов информационного ресурса с общим списком кодов ГРНТИ и соответствующим им кодам УДК [7, 9]. Очевидно, что таким образом определенный код УДК может оказаться не единственным и носит предварительный характер, он просто может быть предложен пользователю в качестве первоначального для последующего уточнения и принятия решения.

Полученные данные используются при описании ресурса и представлении ресурса в виде XML-файла, который автоматически помещается в буфер.

Результат работы с открытыми Интернет ресурсами отображается на специальной отдельной панели (отражающей метаописание ресурса) для визуализации и принятия окончательного решения пользователем (см. рис.3).

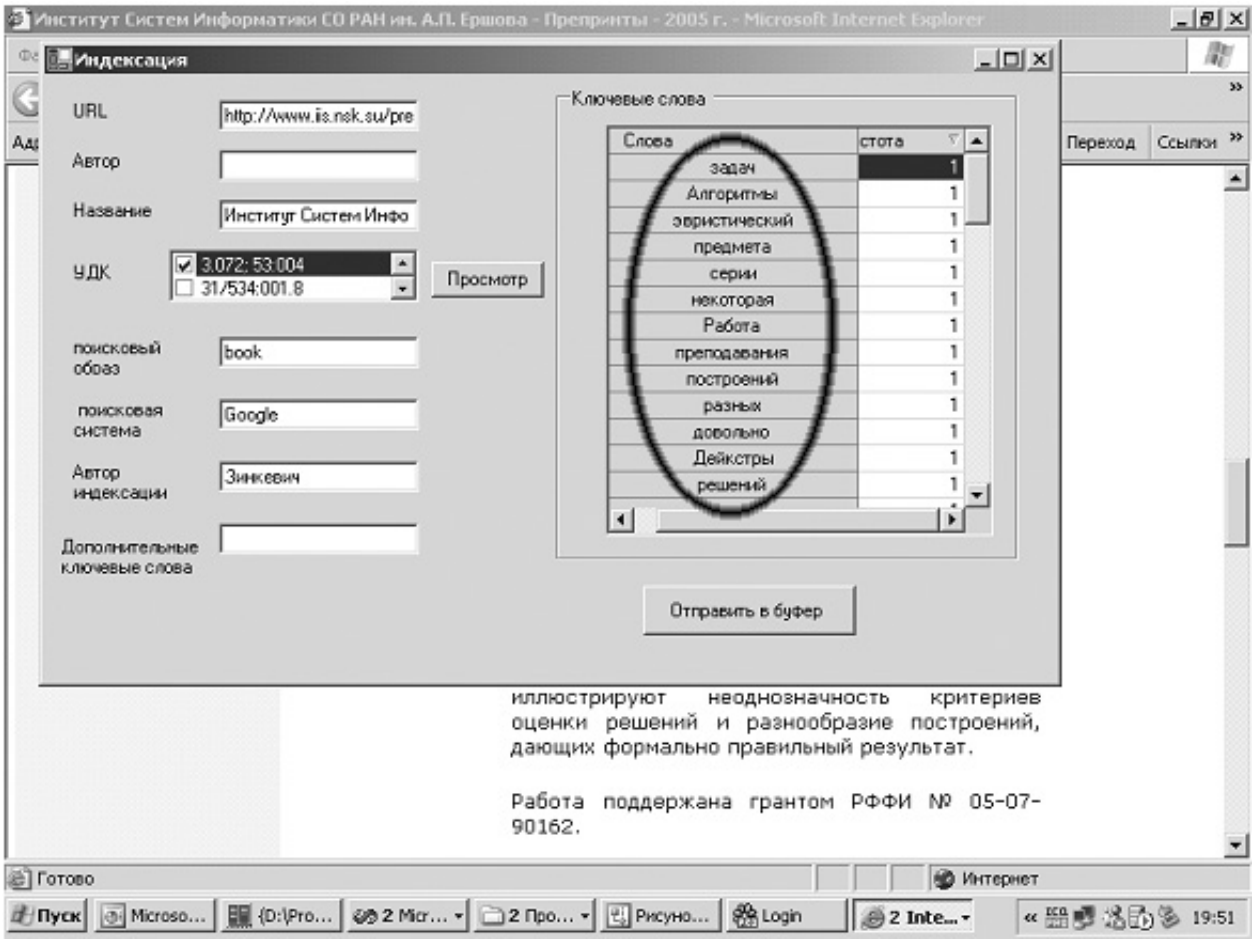

Рис.3 Панель результата работы с открытыми Интернет ресурсами

Если какие-то поля панели по каким-то причинам оказываются не заполненными (например, плагин не может найти информацию об авторе), то пользователь может перед отправкой ресурса в буфер либо сам ввести недостающую информацию или подправить предоставленную (например, УДК), либо отправить в буфер как есть.

В идеальном случае, когда пользователь всем удовлетворен, при взятии ресурса из Интернет ему нужно будет только нажать две кнопки.

4.2 Ресурсы автономных библиотек.

Сценарий работы пользователя с ресурсами автономных библиотек предусматривает возможность взаимодействия с другими электронными библиотеками по протоколу Z39.50. Для обеспечения этого взаимодействия в настоящее время в виде XML-файла представлено около двух тысяч наименований российских и зарубежных библиотек, заимствованных из базы данных библиотечного браузера LibNavigator [1] и представленных в виде древовидной структуры.

Нажав вторую дополнительную кнопку на панели браузера, пользователь начинает работать со специальной панелью (см. рис.4), на которой он первоначально отмечает требуемые библиотеки, затем формирует запрос, вводит поисковые слова, связанные логическими конструкциями, дополнительно могут быть указаны выходные атрибуты документа (автор, название, издательство, и другие) и запускает поиск

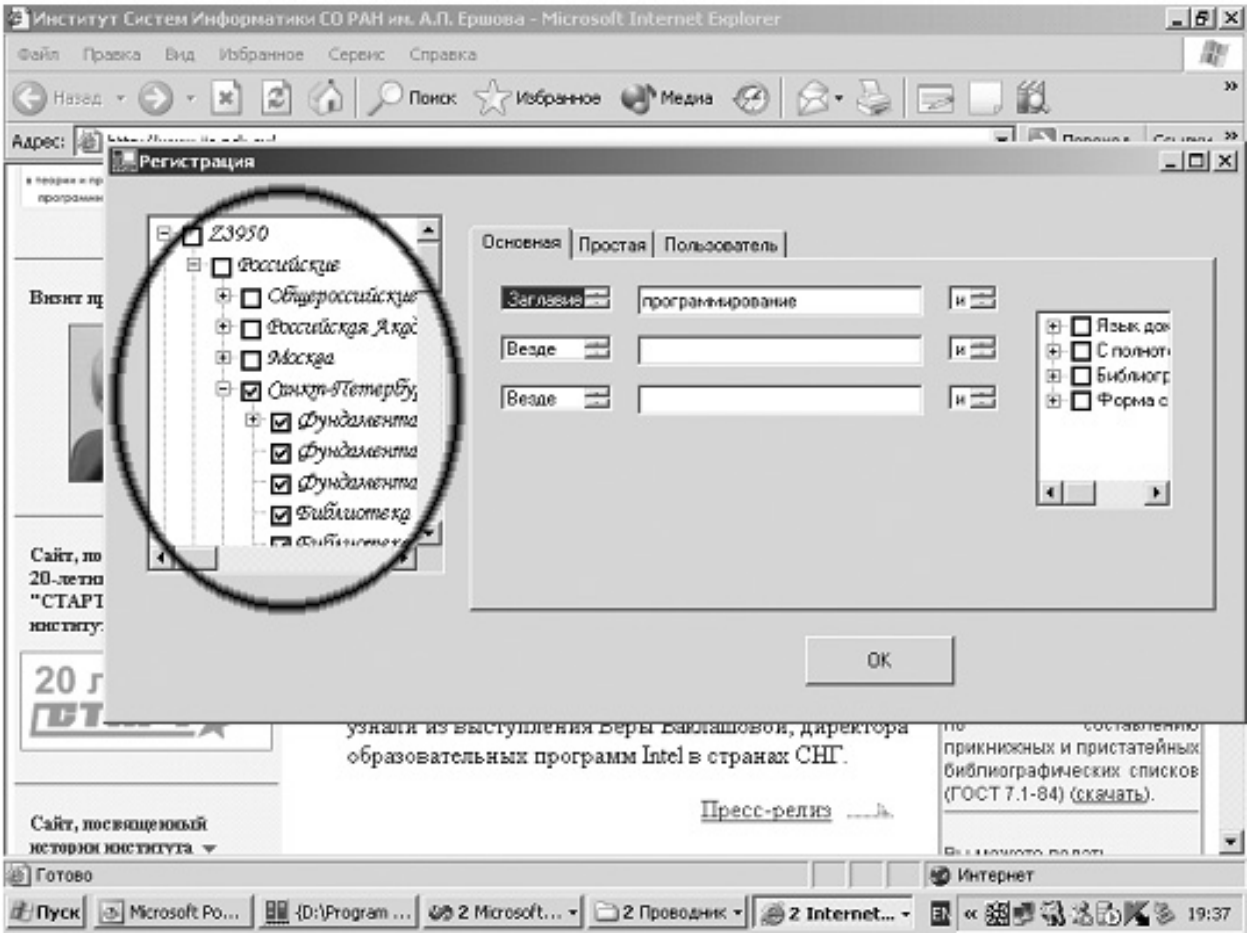

Рис.4 Панель работы с автономными библиотеками

Пользователь может просмотреть полученные результаты с помощью специальной панели браузера (см. рис.5), и отправить их на сохранение в буфер обычным нажатием кнопки на панели инструментов браузера.

Полученное в результате поиска в автономных библиотеках библиографическое описание при сохранении в буфере подвергается такой же обработке, как и любой другой ресурс. Если в библиографическом описании присутствует URL полного текста, то соответствующий полный текст может быть получен и также сохранен в буфере на обычных условиях. Дополнительно может быть установлена взаимная ссылка полнотекстового ресурса на его библиографическое описание.

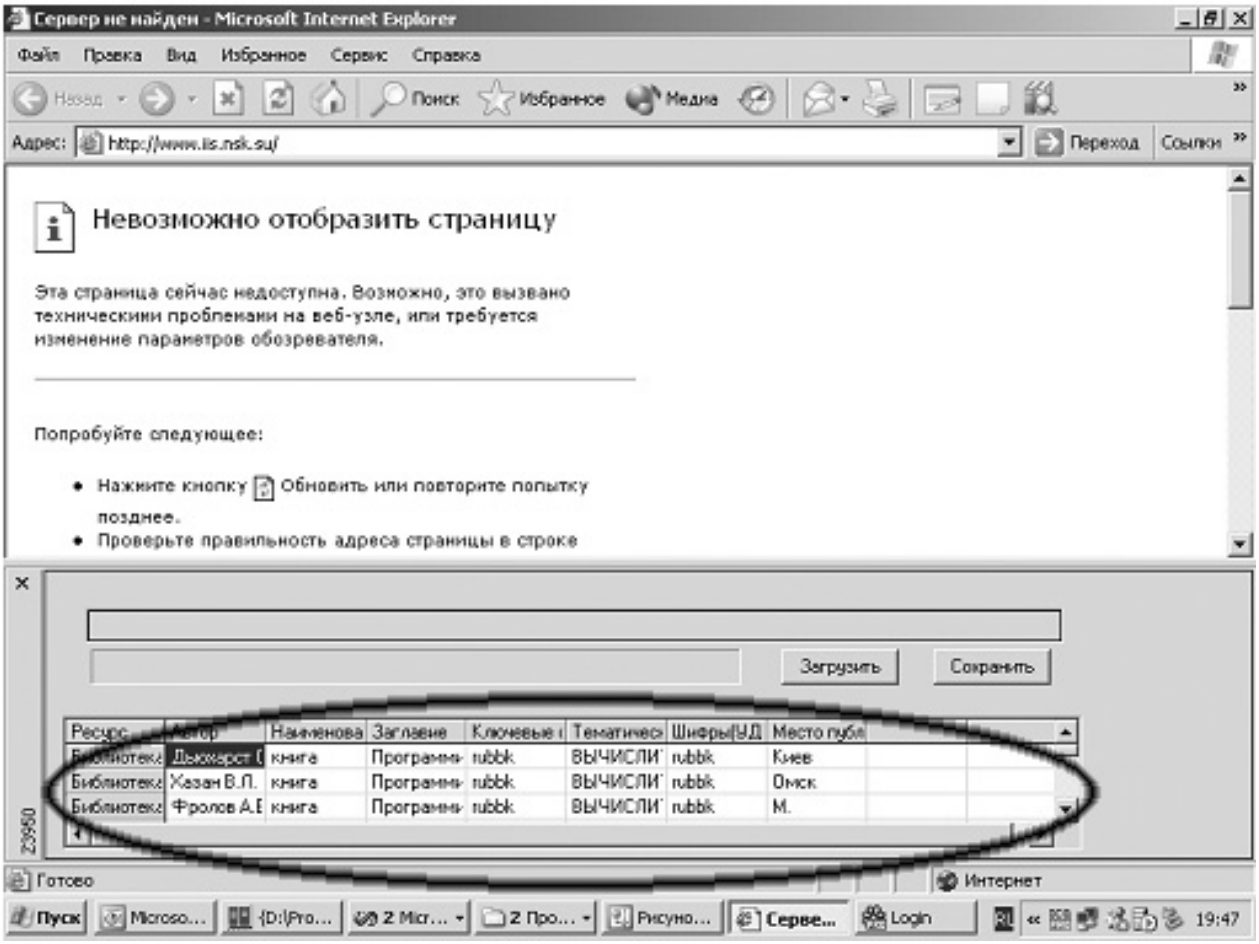

Рис.5 Панель просмотра результатов поиска

В программе предусмотрен контроль над загрузкой данных и из других источников (например, отдельные Word-документы и т.п.) с помощью индикатора загрузки.

4.3 Загрузка в Greenstone.

Программа периодически проверяет содержимое буфера, и в случае его заполнения или по заранее заданному расписанию, определенному пользователем автоматически осуществляет запуск загрузки накопленных ресурсов в Greenstone. В программе также предусмотрены функции непрерывного сохранения данных, рассмотрены отключения компьютера, сбои.

После работы с полнотекстовым ресурсом TAPIR создает описание ресурса в формате ISO 2709, и связывает его с сохраненным ресурсом. Полученное описание может быть использовано для пополнения базы документов в библиографических автоматизированных библиотечных системах.

5. Возможный дополнительный эффект создания личных полнотекстовых коллекций

Помимо основной цели, создание личных полнотекстовых коллекций, удовлетворяющих некоторым общепринятым стандартам, подкрепленное соответствующими инструментальными средствами, может быть направлено и на решение такой задачи, как распределенное формирование центральных полнотекстовых библиотек (некоторый аналог GRID в области всеобщей каталогизации информационных ресурсов).

В настоящее время количество электронных информационных ресурсов (ЭИР) в сети Интернет растет огромными темпами. Большую деятельность в области создания ЭИР проводят и библиотеки. Количество ЭИР и в сети Интернет и в электронных библиотеках исчисляется уже пентобайтами. Причем достаточно очевидна тенденция превалирующего роста ЭИР в открытой сети Интернет, существенным

недостатком которой (в отличие от электронных библиотек), как уже говорилось, является негарантированность сохранения ЭИР в течение продолжительного времени.

Для обеспечения сохранности информационных ресурсов и их эффективного использования в настоящее время реализуется достаточно много разнообразных национальных и международных программ и мероприятий, в том числе:

– Мероприятия по ретроконверсии (массового перевода в электронный вид и сохранения в электронном виде традиционных документов, включая любые издания);
– Мероприятия по создания национальных архивов электронных информационных ресурсов, при этом архив понимается не только как источник информации, но и как набор некоторых связанных с ней услуг [6].

С учетом того, что все большая доля информационных ресурсов приходится на открытый Интернет, усилий одних только библиотечных работников для каталогизации всех появляющихся электронных информационных ресурсов, становится явно недостаточно. Использование роботов web-архивирования также не обеспечит решение задачи тематической каталогизации, так как с одной стороны, не все Интернет ресурсы имеют ценность, а с другой, на одной web-странице может находиться множество разнотематических ресурсов, в том числе и различной ценности. В силу ограниченности интеллектуальных способностей роботов «разгрести» эти ресурсы опять же без помощи специалистов, включая библиотекарей, вряд ли удастся.

Поэтому выход один - привлечь к делу тематической каталогизации электронных ресурсов как можно больше участников (знаменитый и очень широко распространенный принцип «делать руками тех, кому это нужно»). Это можно сделать через создание личных полнотекстовых тематических коллекций, удовлетворяющих общепринятым стандартам, которые в соответствии с определенными правилами могут просматриваться в сети и в дальнейшем «сливаться» в некие централизованные хранилища

Литература

[1] Библиографический браузер LibNavigator. http://www.libnavigator.ru/
[2] Б. Великий Как загрузить сайт целиком: обзор менеджеров закачки сайтов http://soft.mail.ru/article_page.php?id=83
[3] А. Лозовюк Веб-инструменты агента 007 http://hostinfo.ru/htmltree/internet/utilities/snippets
[4] Компания Onfolio выпускает пакет для упорядочивания информации http://mywebsearch.adelite.com/webnews/842.html
[5] Компьютерный форум Ru.Board http://forum.ru-oard.com/topic.cgi?forum=5&topic= 5881&start=280
[6] С.А Нудель, Д.И.Верещака К вопросу об архивировании электронных ресурсов http://conf.cpic.ru/upload/eva2004/reports/doklad_387.doc
[7] Общий список всех кодов, содержащий номера ВАК и УДК, соответствующие кодам ГРНТИ http://mineral.spmi.ru
[8] В.А. Резниченко, Г.Ю. Проскудина, О.М. Овдей Создание цифровой библиотеки коллекций периодических изданий на основе Greenstone.// Российский научный электронный журнал «Электронные библиотеки» 2005, Том 8, выпуск 6 http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2005/part6/RPO
[9] Справочник по УДК http://www.teacode.com/online/udc/
[10] Digital Library http://www.unesco.org/cgi-bin/webworld/portal_freesoftware/ cgi/page.cgi?d=1&g=Software/ Digital_ Library/index. shtml
[11] Dlbox http://dlbox.nudl.org
[12] DSpace http://www.dspace.org/
[13] FlashGet 1.72 build 128 http://soft.softodrom.ru/ap/p287.shtml
[14] Greenstone Digital Library http://www.greenstone.org/
[15] Open Source и e-learning www.cpk.mesi.ru/news/2005/release008/11.ppt

Creation the Effective Tool for Formation Personal Text-full Collections for Scientific and Educational Activity

B.V.Oleynikov

The necessity of creation the tool for effective formation personal text-full collections is proved in article. The developed tool TAPIR connected with MS Internet Explorer 6 and library system Greenstone is described. Scripts of its use are considered.