ТЕХНОЛОГИЧЕСКИЕ АСПЕКТЫ РАБОТЫ С ПОСТОЯННО ДЕЙСТВУЮЩИМИ ЗАПРОСАМИ В АВТОМАТИЗИРОВАННОЙ БИБЛИОТЕЧНОЙ СИСТЕМЕ

(на примере электронного каталога Библиотеки Ботанического института им. В.Л. Комарова РАН)

Рубцов В.В., ведущий инженер НИОИА БАН

Панкратова Г.Н., заведующая отделом БАН при БИН

Батурина М.Г., ведущий библиотекарь, отдел БАН при БИН

Рубцова Т.Ф. библиотекарь, отдел БАН при БИН

Библиотека Российской академии наук (БАН), Санкт-Петербург, Россия

Библиотеке Ботанического института им. В.Л. Комарова РАН, Санкт-Петербург, Россия

Повышение значимости информации, превращение ее в стратегический ресурс общества, зависимость успешности деятельности организации от уровня информированности – всё это в настоящее время не вызывает сомнений. Мы расскажем об использовании современных информационных технологий на примере электронного каталога в отделе БАН при Ботаническом институте им. В.Л. Комарова РАН.

В Библиотеке Российской академии наук и академических библиотеках – подразделениях БАН, функционирующих при научно-исследовательских учреждениях, создана автоматизированная библиотечная информационная система, включающая комплекс взаимосвязанных библиографических, технологических, нормативных и справочных баз данных. Все эти базы данных объединены в многобазовый комплекс. Возможности нашего электронного ресурса сочетают в себе принципы одновременного и однопространственного построения алфавитного, систематического, предметного каталогов, работающих в интерактивном режиме поиска. Технологическая цепочка подготовки и обработки электронных библиографических записей включает несколько последовательных этапов:

- создание электронных записей собственной генерации
- использование готовых записей Всероссийской Книжной Палаты (РКП)
- служебную обработку данных и пополнение основного электронного каталога

Принципиальная схема документопотока и обработки данных показана на рис. 1.

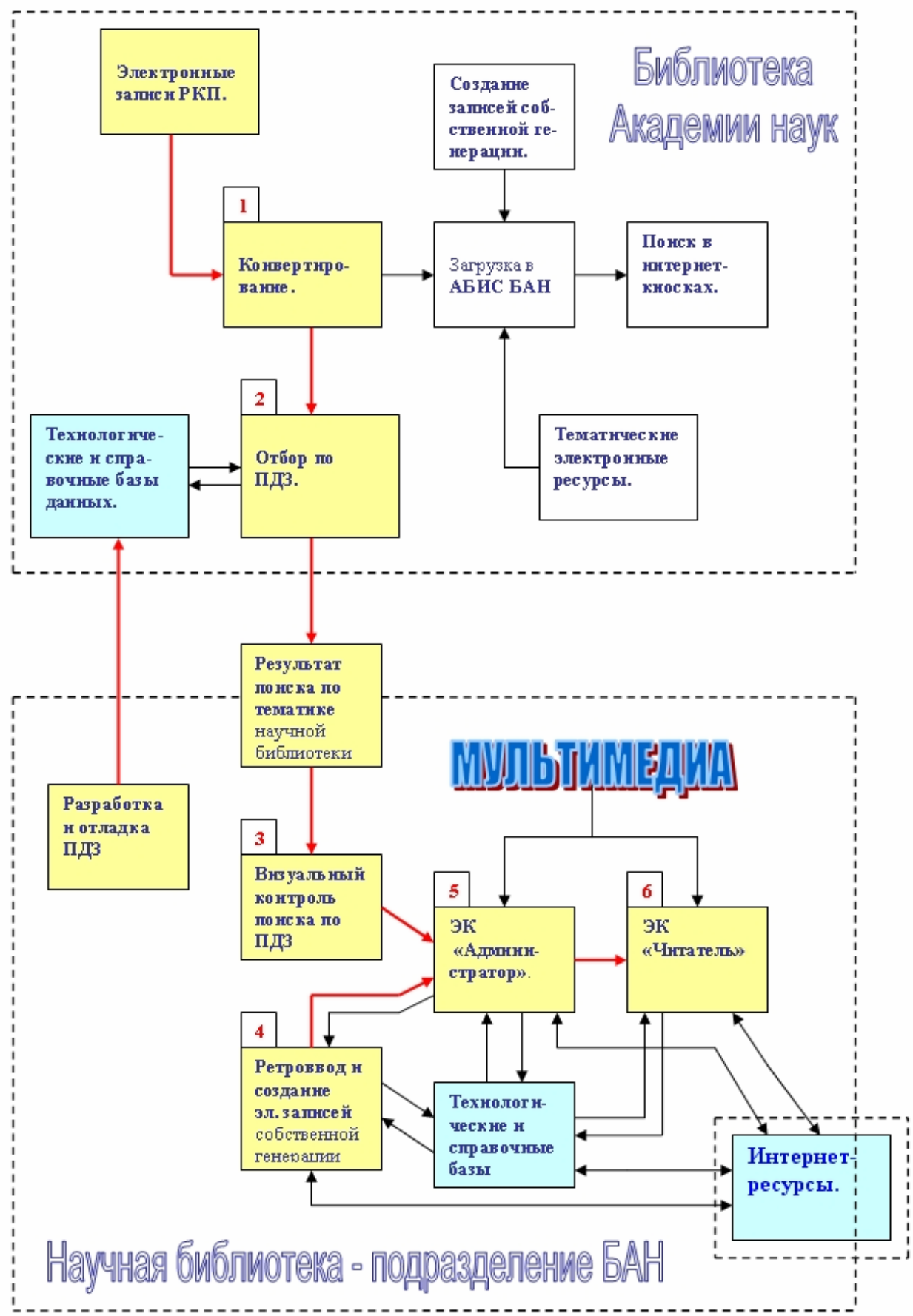

Рис. 1. Общая схема информационных потоков при создании ЭК БИН.

Наше решение согласуется с современными тенденциями, рассматривающими информационное обслуживание с точки зрения сервисного подхода. Созданная в библиотеке Ботанического института многобазовая поисковая система (на схеме обозначена как «ЭК Читатель») представляет собой сложный информационный ресурс. Он ориентирован на работу как с библиографическими массивами (раскрывающими состав фонда), так и со справочными и фактографическими электронными ресурсами (предназначенными для технологических и консультационных целей). Справочно-фактографический компонент содержит обязательные справочные базы, которые используются другим базами в автоматическом режиме, и самостоятельные локальные справочно-фактографические БД. Обязательными являются справочники кодов стран, языков и кодов отношения к авторству, локальными – справочник аббревиатур библиографических описаний, в том числе специальных латинских ботанических, и БД сведений о ботанических журналах. Последняя из упомянутых база данных включает служебную информацию для работников библиотеки и интернет-адреса (URL) для доступа к содержанию отечественной и зарубежной периодики в сети Интернет. Таким образом, с точки зрения ресурсов для оказания услуг нами используются как собственные фонды, так и совокупные ресурсы общества.

В рамках описываемой технологии Библиотека РАН покупает в РКП электронные записи для несколько видов документов, это: отечественные монографии, статьи из отечественных журналов и газет, авторефераты диссертаций. Записи конвертируются и далее действует технология специального **отраслевого тематического отбора**. Многобазовый комплекс Библиотеки РАН включает технологическую БД, в которой хранятся постоянно действующие запросы (ПДЗ) отраслевых библиотек. Каждая вновь полученная порция электронных записей РКП обрабатывается поисковой программой, и результаты отбора накапливаются в целевых каталогах.

Для библиотеки Ботанического института ПДЗ составляются специалистом (ведущим библиотекарем) библиотеки, а при необходимости тематика запроса уточняется с учёными института. Для монографий, статей из журналов и авторефератов созданы самостоятельные ПДЗ. Разработка ПДЗ выполняется в несколько этапов:

- создание списка поисковых элементов, отвечающих тематике института (соответствует элементам поисковых словарей БД: рубрики ББК, УДК, предметные рубрики РКП, слова из заглавий, названия источников для статей из журналов)

- проработка общей логики отбора, т.е. последовательности выполнения операций поиска по аналогии с обыкновенным поиском в базе данных
- «перевод» запроса на язык поиска базы
- отладка ПДЗ на тестовых массивах
- передача ПДЗ в БАН для использования в отборе данных
- текущее внесение изменений в ПДЗ (корректировка) с целью снижения информационного шума при поиске, добавление или изъятие тех или иных поисковых предписаний

Цель этих этапов работы – максимально широкий и точный отбор релевантных записей и снижение уровня информационного шума. Созданные ПДЗ включают десятки строк взаимосвязанных последовательных поисковых предписаний.

Для работы ПДЗ в технологической базе БАН создан поисковый аппарат из элементов, наиболее полно отражающих содержание информационного массива. К сожалению, электронные записи, получаемые нами из РКП, не соответствуют требованиям, предъявляемым к электронным записям научных библиотек. Например, записи включают ББК, используемые в массовых библиотеках, а такой важный для нас реквизит, как код специальности ВАК, не выделен в самостоятельный поисковый элемент и включен в состав одного из полей в подбор, как часть текста. Предметные рубрики зачастую представлены в неструктурированном виде, что снижает их информативность, например:

- «**Биологические науки** – На степень кандидата наук»
- «Российская Федерация, **биол. науки**»

Понятно, что недостаточно отобрать в соответствии с ББК и УДК только такие разделы, как, к примеру «Ботаника», а необходимы и более общие разделы: «Общая биология», подраздел «Лесоводство» раздела «Лесное хозяйство, Лесохозяйственные науки» и др. Вместе с тем многие документы, которые отражаются в общих разделах, не соответствуют тематике нашей библиотеки, то есть при поиске станут одной из причин появления информационного шума.

Специализированный тематический отбор мы усложнили характерными словами из заглавий, набрался довольно внушительный список. Из них был составлен «словник» при просмотре рубрикаторов и библиографических описаний из номеров Реферативного журнала «Ботаника». Слова из заглавий брались в усеченном виде (отбрасывались изменяемые части слов). Необходимо учитывать, что есть немало обычных в ботанике

терминов, таких как «карт$» (карты [геогр.]), «спор$» (споры [растений]), «лист$» (листья растений), которые употребляются в других значениях и в других науках (символ «$» означает усечение термина справа). Так, слово из заглавия «карт$» приведет к появлению документов, связанных с игральными картами, «спор$» – к появлению документов, связанных с юридическими спорами, а слово «лист$» – документов, описывающих «обработку листового материала без образования стружки» или «погружения и вложения листа Мебиуса в евклидовы пространства». Отсюда в результирующем списке появляются «шумовые» записи, например с такими предметными рубриками:

- **для «лист$»: «технология обработки без$» и «геометрия$»**
- **для «спор$»: «криминол$»**
- **и т.д.**

Для удаления «информационного шума» в нашем ПДЗ применен такой алгоритмический приём: сначала мы отбираем электронные записи «с избытком», используя предметизацию и систематизацию РКП, суммируем списки фраз запросов и помещаем эти записи в промежуточный избыточный целевой список. Параметры отбора имеют в общей форме такой вид:

- **#1: Ботани$ +Растен$ + Лист$ + Словари по ботанике$ + Сосудистые растен$ + Альгология$ + Охрана природы$ и т.д.**
- **#2: УДК:58 и т. д.**
- **#3: ББК:28.5$ + ББК:28.0$ и т.д.**
- **#4: заповедн$ + фотосинт$ + хлоропласт$ +хлорофилл$ и т.д.**
- **#5: растен$ + лишайник$ + флор$ + гербар$ + мох$ + спор$ и т.д.**
- **#6: #1 + #2 + #3 + #4 + #5 и т.д.**

В последней строке (**#6**) выполнено сложение предыдущих результирующих списков. Далее следует отбор по специализированному списку предметных рубрик и слов из заглавия (по смежным или более общими направлениям):

- **биология$**
- **фармакология$**
- **биофизика$**
- **палеогеография$**
- **и т.д.**

Полученные записи добавляются к ранее созданному «избыточному» множеству, а для отсечения информационного шума в самостоятельное множество собираются записи из

смежных областей: это записи, содержащие явно нетематические рубрики, слова из заглавия, к примеру такие, как:

- **криминол$ + заболеван$ + ветеринар$**
- **зоол$ + литератур$ + теория$ + деревообраб$**
- **и т.д.**

В конце работы ПДЗ из «избыточного» тематического списка записей производится вычитание записей множества результирующего списка, содержащего заведомо нетематические записи.

Как видно из сказанного, создание и отладка научного тематического запроса – процесс творческий и, несомненно, один из ключевых этапов описываемой информационной технологии. От качества выполнения этого этапа работы зависит полнота получения сведений отраслевой библиотекой. Отладка и доводка запроса выполняется в специальной базе на тестовых массивах, после чего ПДЗ передается в основную технологическую базу в БАН. Хранение ПДЗ в базе высоко технологично: записи в БД «Запросы» снабжены вспомогательными поисковыми и пояснительными полями, следовательно имеют свой поисковый аппарат (облегчен поиск и доступ к ПДЗ при внесении изменений и дополнений), администратор БД использует единый механизм создания резервных копий как для библиографических, так и для технологических баз данных.

Кратко об алгоритме программы отбора данных. Поступившие из РКП записи загружаются в БД для отраслевого тематического поиска, программа отбора работает одновременно с двумя базами данных:

1. **с базой данных запросов**
2. **с библиографическим массивом (с электронными записями РКП)**

Программа автоматически находит нужный ПДЗ, переключается на целевую библиографическую базу и выполняет отбор записей. Если в течение дня обрабатывается несколько поступивших из РКП порций, то программно осуществляется контроль имен файлов во избежание потери информации предыдущих поисков. Результат отбора размещается в специальном каталоге, закрепленном за конкретной научной библиотекой. Данные хранятся в обменном формате ISO-2709, передаются «хозяину» в отраслевую библиотеку по мере накопления.

В состав технологической цепи в качестве обязательного этапа обработки включена БД для визуального контроля результата отбора. Любой автоматический поиск, как бы он не был отлажен, всегда оставляет шумовые записи. Поэтому окончательно в научной библиотеке информационный шум отсекается квалифицированным специалистом-предметизатором. Чистка 200-300 поступающих в одной порции записей занимает не более 10-20 минут, после чего «остаток» загружается в базу «Администратор», где проводится проверка на дублетность, выполняется предметизация и прочие действия: вносятся шифры, инвентарные номера, год учета в библиотеке, исправляется ряд реквизитов РКП в соответствии с локальными требованиями библиотеки. Проверенные и отредактированные записи переносятся в основной электронный каталог для читателей.

База данных «Администратор» имеет расширенный набор функций, к ней подключены оригинальные сервисные программы, специально разработан технологический поисковый аппарат, облегчающий контроль новых поступлений. В этой же базе создаются электронные записи собственной генерации для авторефератов, которые поступили в библиотеку, но еще не имеют электронного аналога, а также выполняется их ретроввод. Обязательная предметизация авторефератов является залогом получения уникального специализированного тематического ряда в поисковом аппарате ЭК. Интерфейс ввода выполнен с учетом эргономичности и адаптирован для ввода конкретного вида документа. Все реквизиты размещены на одной «странице», легко обозримы, следуют в том порядке, в котором можно максимально быстро создать электронную запись и при этом требуется минимальное листание бумажного документа. Особое внимание уделено размеру шрифта в полях ввода и просмотра данных: оператор может настроить размер шрифта с учетом особенностей зрения, это снижает утомляемость и существенно влияет на скорость и качество работы. Интерактивный доступ к справочникам с кодированной информацией при вводе данных снижает ошибочность и является инструментом контроля ранее введенных данных. Пример доступа к списку названий и кодов языков приведен на рис. 2.

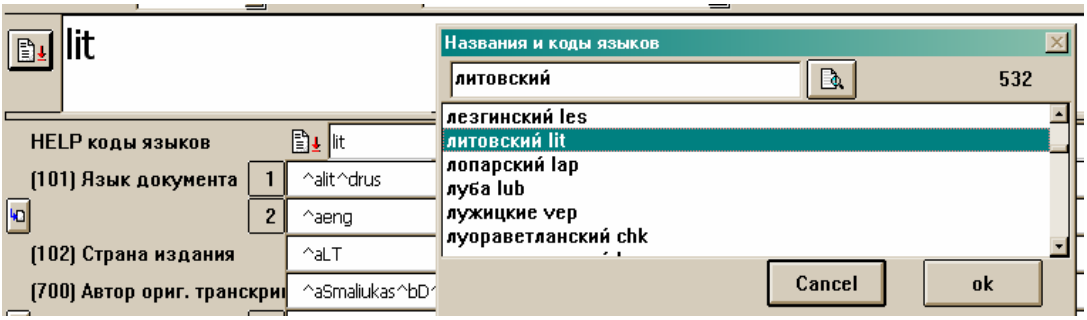

Рис. 2. Фрагмент листа ввода данных и список названий и кодов языков.

Нами уже упоминалось об использовании перспективных направлений. Для повышения качества обслуживания читателя предусмотрена возможность включения в электронную запись ссылок на мультимедийные компоненты (поддерживаются фото-, аудио-, видеоформаты, текстовые файлы: bmp, jpg, djvu, tiff, pdf, mp3, mp4, avi, htm, html, txt, rtf, doc, docx и пр., интернет-ссылки URL). Ввиду ограниченности ресурсов библиотеки мы не можем создавать электронный каталог как полнотекстовую базу данных. Однако, фрагменты текста авторефератов, признанных ветхими или поврежденными, частая выдача которых нежелательна, могут быть включены в базу в виде сканов (обычно в формате PDF). Это повышает сохранность фонда и обеспечивает быстрый доступ к тексту. Такой подход особенно актуален для ретроввода авторефератов. Внедрение технологической цепи, обеспечившей наполнение БД с помощью электронных записей РКП, позволило нам практически полностью описать в электронном виде поступления авторефератов последних лет. Описание новых авторефератов теперь занимает относительно небольшую часть времени, что позволило нам начать работу по ретроспективному вводу данных, глубина ретроспекции в настоящее время достигает 1985 года.

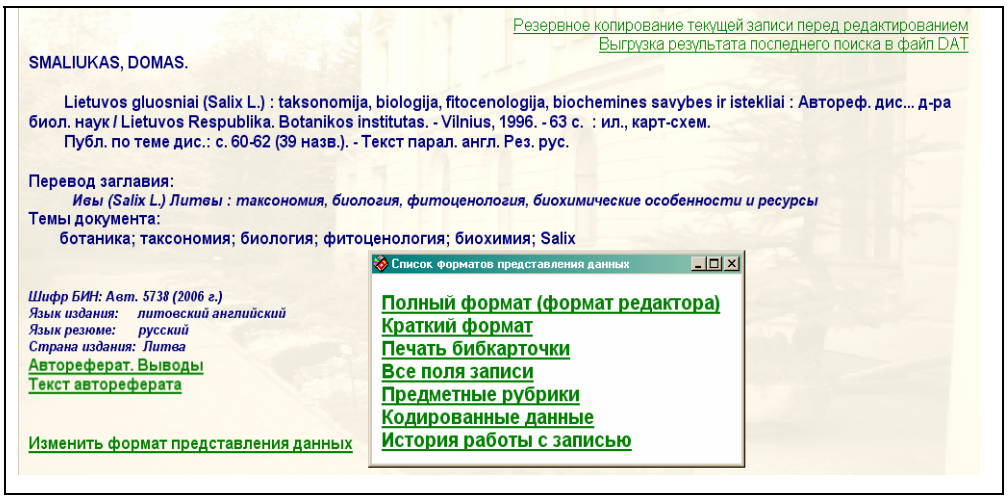

Рис. 3. Окно просмотра документа (автореферат) и дочернее окно со списком доступных форматов вывода.

В базе «Читатель» доступ разрешен только в режиме «поиск и просмотр», также применен оригинальный дизайн с изображением библиотеки и Ботанического института, т.е. информационный продукт библиотеки оформлен в фирменном стиле (дизайн и цвет). Поисковый аппарат базы проработан так, что напоминает привычные для читателей специализированные, традиционные каталоги библиотеки – это облегчает ориентацию в поисковом аппарате БД, но в то же время в полной мере охватывает особенности

машинного поиска. Подключены справочные и фактографические базы данных и инструктивные тексты: описание правил работы в них учитывает специфику тематики библиотеки. По аналогии с современными интернет-технологиями по типу закладок сохраняется статистика выполненных запросов. Специализированный поиск в БД нередко оказывается многовариантным, многоступенчатым и разветвлённым. В результате список выполненных фраз поиска и полученные количественные характеристики сами по себе приобретают научный интерес. Специальная утилита, вызываемая по гиперссылке с экрана, сохраняет все поисковые фразы в файле. Эти материалы могут использовать как читатели, так и сотрудники библиотеки.

Оригинальный интерфейс позволяет быстро изменять формат просмотра документов, не засоряя экран лишней информацией: механизм смены форматов вывода позволяет как бы «пролистывать» варианты представления данных, двигаясь по гиперссылкам списка в дочернем окне. Читателю предоставлена возможность изменять размер шрифта на экране, работать с найденной информацией в многооконном интерфейсе, т.е. видеть одновременно информацию двух и более баз данных. Указанные характеристики повышают комфортность и привлекательность системы.

Интерфейс читательской мультибазы позволяет выходить в сеть Интернет по гиперссылкам, размещенным на экране в текущей записи базы или в списках целевых гиперссылок в дочерних окнах. Количество и наполнение таких «контейнеров» для доступных Интернет-ресурсов практически не ограничено. Пример дочернего окна «контейнера» с гиперссылками имеет вид, приведенный на рис. 4.

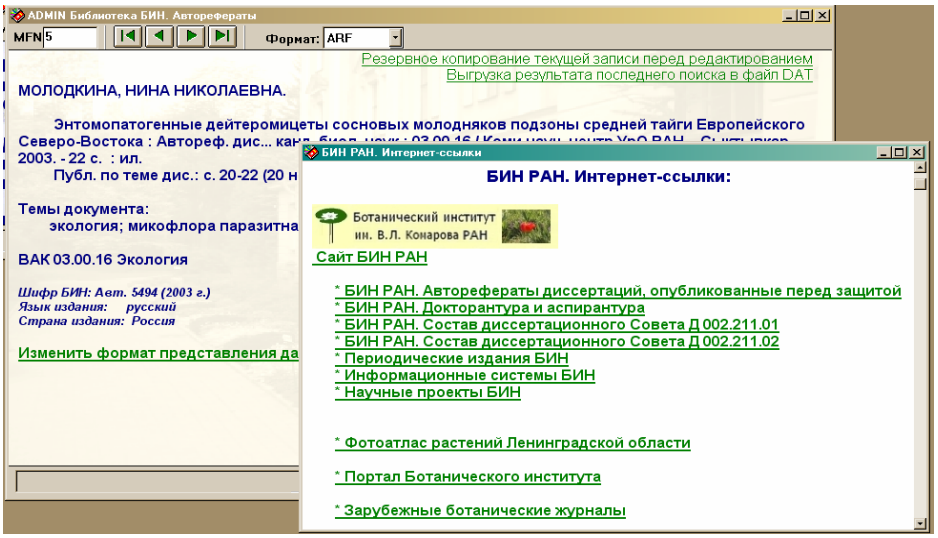

Рис. 4. Дочернее окно «контейнер» с Интернет-ссылками.

Щелчок мышью на любой строчке в окне «контейнере» вызывает переход к конкретному, обозначенному гиперссылкой Интернет-ресурсу.

Учитывая тот факт, что публикация в Интернете – обязательное условие при защите кандидатской или докторской диссертации, то база данных представляет собой удобный инструмент для работы с полными текстами авторефератов диссертаций, подготовленных к защите. Из базы данных можно перейти на сайт Ботанического института и сразу ознакомиться с новыми Интернет-публикациями и новостями БИН. Включенная в многобазовую систему справочная база данных ботанических журналов, содержащая Интернет-адреса, позволяет легко осуществить тематический поиск и точный целевой выход на нужный, заранее подготовленный специалистами библиотеки Интернет-адрес, в связи с чем отпадает необходимость долгого самостоятельного поиска.

Таким образом, благодаря созданной технологии в библиотеке БИН создан и функционирует мощный информационный ресурс, присущий современной научной библиотеке. К нему могут обращаться и постоянные читатели, и молодые ученые, только начинающие свою научную карьеру. С помощью разработанной технологии и автоматизированной библиотечной системы электронный ресурс выполняет не только чисто библиотечные функции по информационному поиску в библиографическом массиве, но и существенно расширяет диапазон предоставляемых научной библиотекой услуг, включая современные интермедийные и Интернет-технологии.

Литература

1. CDS/ISIS for Windows Reference manual (version 1.5). - Paris : UNESCO, 2004.
2. Winisis 1.5 (Build num. 3) Changes and News since Winisis 1.311.
3. Бакстон, Э. Руководство по CDS/ISIS для Windows / Э. Бакстон, А. Хопкинсон ; Пер. с англ. твор. коллектива Ассоциации ЭБНИТ; Под общ. рук. Я. Л. Шрайберга ; Отв. ред. Е. М. Зайцева. — М. : Ассоциация ЭБНИТ, 2002.
4. Леонов, В.П., Захаров В.П. Новые информационные технологии в централизованной системе БАН / В.П. Леонов, В.П. Захаров // Информационное общество, информационные ресурсы и технологии телекоммуникации : Материалы конференции "НТИ-2000", Москва, 22-24 ноября 2000 г. - М. : ВИНИТИ, 2000. – С. 206-208.
5. Захаров, В. П. Автоматизированная система обработки новых поступлений в БАН / В. П. Захаров, Н. Д. Зиганшина, Т. В. Ипполитова, А. Ц. Масевич, Н. А. Панченко, Ю. Б.

Пугач, С. В. Реуф, В. В. Рубцов // Информационное обеспечение науки : новые технологии : сб. науч. тр. – М., 1997. – С. 102-107.

6. Рубцов, В. В. Что может ISIS? Взгляд программиста на библиотечные информационные технологии // Петербург. библ. шк. – 2006. - № 1/2. – С. 74.

7. Емельянов, Александр. Возведение в степень: полномочия разделили. URL: http://www.aspirant.istu.ru/info/mnenie/step1/ (дата обращения 25.08.2009).

8. Брежнева, В.В., Минкина, В.А. Информационное обслуживание: продукты и услуги, предоставляемые библиотеками и службами информации предприятий / В.В. Брежнева, В.А. Минкина; СПбГУКИ. – 2-е изд., перераб. – СПб.: Профессия, 2006. – 304 с. – (Серия «Библиотека»).