ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК [004.658:51]:002.6 ВИНИТИ

В. Г. Шамаев, А. В. Жаров, А. Б. Горшков

Разработка технологии создания ретроспективных реферативных баз данных ВИНИТИ РАН по физико-математическим наукам

Описывается технология создания ретроспективной реферативной базы данных по физико-математическим наукам на основе печатных выпусков РЖ ВИНИТИ, начиная с 1953 г. Обосновывается особый характер такой БД, связанный с необходимостью отображения сложных физико-математических текстов. Проектирование базы данных выполнено под СУБД Microsoft SQL Server 2000. Приводится структура БД, кратко сообщается о работах по наполнению и работе с БД с представлением АРМов набора библиографического описания и администратора. Полученная База данных является составляющей единого комплекса Автоматизированного банка данных ВИНИТИ РАН.

ВВЕДЕНИЕ

Реферативный журнал ВИНИТИ РАН всегда выходил ограниченными тиражами. Но последние полтора десятка лет перекрыли все даже самые пессимистические прогнозы. Например, Сводный том РЖ "Математика" в 1955 г. имел тираж 3500 экземпляров, в 1985 г. – 1282, в 1993 г. — 667, в 2006 г.— всего 84 экземпляра; Сводный том РЖ Физика в 2006 г. — 48; Сводный том РЖ Химия — 55, Сводный том РЖ Механика — 58. Выпуски Реферативного журнала далеко не всех годов издания есть даже в крупнейших научно-технических библиотеках. В свою очередь, реферативные журналы по физико-математическим наукам прежних лет издания не теряют своей актуальности до настоящего времени.

Параллельный выпуск реферативных журналов, баз данных и электронных РЖ и современный уровень развития информационных коммуникаций частично снимают проблему получения информации. Доступ к этим ресурсам через Интернет дает возможность пользоваться этой информацией в каждом подсоединенном к Интернету учреждении. К тому же базы данных и электронные РЖ обладают дополнительными сервисными возможностями в виде встроенного поискового аппарата.

Базы данных — как тематические, так и политематические — начали появляться в ВИНИТИ с 1981 г. [1, 2]: по автоматике и радиоэлектронике, химии, машиностроению, электротехнике, а также биологии — в 1981 г., по физике — с 1983 г., по механике — с 1985 г., по геофизике — с 1986 г., по астрономии — с 1989 г. Вне этого прогресса была только математика, базы данных по которой начали формироваться лишь в 1997 г., когда Производственно-издательский комбинат ВИНИТИ принял в качестве стандарта набора программный пакет LaTeX.

Эти БД нельзя назвать полностью соответствующими РЖ ВИНИТИ. Например, символьная информация в подавляющем количестве БД отражается с использованием так называемого алфавита ВИНИТИ: греческая буква α записывается как _a, β как _b, математический символ x^2 как x_2 и т. п., рисунки и схемы, а также даже несложные формулы отсутствуют. БД Математика, которая набирается в TeX'е, в этой кодировке и предстает перед пользователями. Для просмотра рефератов в такой БД на компьютере пользователя должны быть установлены соответствующие средства компиляции.

Поэтому проблема предоставления пользователю баз данных ВИНИТИ по физико-математическим наукам разделяется на две задачи:

1) совершенствование средств визуализации уже существующего электронного массива данных, который, как правило, представлен в алфавите ВИНИТИ. Эта задача связана с разработкой новых или применением уже существующих конверторов для приведения к какому-либо стандартному средству отображения (визуализации) информации. В качестве конечной формы может быть TeX или XML как средство представления физико-математической информации и PDF как средство отображения;

2) перевод в электронную форму бумажных РЖ, не имеющих сопутствующих БД. Сопутствующих потому, что технология подготовки информационных продуктов ВИНИТИ до последнего времени не ставила перед собой задачу полного соответствия БД и РЖ. На определенном этапе технология их подготовки имела разветвление — отдельно для РЖ и отдельно для БД. Современная технология заключается в тщательной подготовке БД и снятии с нее как оригинал-макета РЖ для печатной версии, так и электронного журнала для подписчиков на электронную версию РЖ.

И здесь мы подходим к необходимости пополнения БД не только текущей, но и ретроспективной информацией, отраженной в РЖ до начала подготовки БД, т. е. перевода на электронные носители всего массива РЖ с 1953 г.

Сделаем небольшое отступление, чтобы пояснить важность ретроспективной информации.

В настоящее время только в небольшом количестве крупнейших научных библиотек России хранятся ретромассивы бумажных РЖ. Спрос читателей на этот вид литературы значителен и часто превосходит спрос на другие специальные издания. Как уже указывалось ранее, РЖ по физико-математическим наукам прежних лет издания востребованы и сегодня. Большинство специалистов, особенно в начале работы над новой тематикой, обращаются, прежде всего, к вторичной информации, которая кратко по содержанию и достаточно полно по числу публикаций представлена в РЖ Однако полный комплект РЖ едва ли доступен.

Кроме того, следует принять во внимание, что в выпусках РЖ ВИНИТИ (а именно на их перевод в цифровую форму, и направлено наше внимание) наиболее полно представлены работы советских ученых, составляющие "золотой фонд" мировой науки. Достаточно вспомнить о крупных научных школах в области физики, химии, астрономии, математики, механики, геофизики и т д Большинство работ ученых СССР 1950 1980 гг. было опубликовано только на русском языке и поэтому слабо представлено в зарубежных РЖ и БД.

Бурное развитие информационных технологий позволяет взглянуть по-новому на проблемы, сопровождающие создание и ведение ретромассивов реферативных изданий в электронном виде. Таким образом, оцифровка бумажных изданий сегодня – одна из актуальнейших задач.

Необходимость оцифровки ретромассивов РЖ ВИНИТИ вызвана проблемами поддержки и сопровождения научных исследований, включающими:

- поиск информации для определения новизны планируемой работы на стадии подачи заявок;
- выявление публикаций, полезных научным или практическим интересам исследователя (сообщения о работах в смежных дисциплинах или узкотематических направлениях, содержащих необходимые фактографические данные, результаты опытов или наблюдений и др);
- большое количество так называемой "серой" литературы (сборников трудов научных учреждений, учебных заведений, трудов конференций, препринтов и т. д.), которая во многих случаях малоизвестна;
- проведение наукометрического анализа, позволяющего получать данные по состоянию и перспективам развития какой-либо дисциплины.

Неоцененные ранее публикации при новом взгляде могут послужить в качестве основы для развития новых направлений работ.

Существующие в настоящее время программные средства позволяют начать разработку информационных технологий по широкомасштабному переводу накопленной информации в электронную форму [3].

Перевод всего накопленного информационного массива печатных изданий ВИНИТИ по физико-математическим наукам в электронную форму (в первую очередь из-за стоимости проекта) на первый взгляд не представляется бесспорным. Это если говорить о далеких от собственных интересов научных дисциплинах. Однако как только дело касается конкретной области, то необходимость такого проекта кажется очевидной. Уж очень заманчиво иметь под рукой в электронной форме всю доступную ретроспективу выпусков Реферативного журнала. Каждый, кто пользовался Интернетом, знает, как трудно остановиться в поиске, так как появляются все новые интересные сведения, порой уже далекие от того, что ищется, но сильно завлекающие. Нечто аналогичное происходит и при поиске по базе данных, и чем она имеет больший объем, тем интереснее. Первые базы данных ВИНИТИ, как уже указывалось выше, появились в 1981 г., а Реферативный журнал ВИНИТИ издается с 1953 г. Вот этот массив информации в бумажном виде, собранный для одних дисциплин за 30 лет, а для других и за 40, планируется перевести в электронный вид

Задача перевода в электронную форму выпусков бумажных РЖ доцифровой эпохи по физико-математическим наукам, т. е. подготовка Ретроспективной реферативной БД (РеБД), в технологическом плане не представляет непреодолимых трудностей, но очень трудоемка. Один из способов решения заключается во вводе всей информации в "сегодняшнюю" БД в TeX'овском наборе, как это делает Производственно-издательский комбинат ВИНИТИ (ПИК) начиная с 1997 г. Однако анализ показал, что извлечь из печатной версии РЖ всю структурированную информацию, необходимую для наполнения текущей БД, невозможно. К тому же, за десятки лет структура подаваемой в печатном издании информации хоть и не сильно, но менялась. Поэтому было принято решение о создании отдельной ретроспективной базы данных, которая и заполнит пробел в подготовке БД, начиная с 1953 г. (*Отметим, что, на наш взгляд, реализация технологии наполнения БД по математике, выполненная в ПИК ВИНИТИ, недооценена как в ВИНИТИ, так и во всем математическом сообществе. Проделана огромная работа и отсутствует только последний штрих, не зависящий от ПИК, визуализация на экране. Собственно об этом и говорится в первом пункте выше. Правда, существует мнение, что настоящие математики легко читают тетховские символы как обычные формулы, но базы данных существуют не только для "настоящих" математиков*).

Выполнение второго пункта упирается в проблему необходимых трудозатрат. Рассмотрим это на примере РЖ Математика. Ежемесячный Сводный том РЖ Математика состоит из четырех выпусков:

13А. Общие вопросы математики. Математическая логика. Теория чисел. Алгебра Топология. Геометрия

13Б Математический анализ

13В. Теория вероятностей и математическая статистика

13Г. Вычислительная математика. Математическая кибернетика.

Суммарный их объем – 2500 рефератов, что при средней плотности 38 рефератов на 1 учетно-издательский лист составляет 65 уч.-изд. л. в месяц. При структурированном наборе (а только такой набор позволит наполнить базу данных) это потребует работы трех профессиональных наборщиков в течение месяца. Следовательно, набор годичного выпуска РЖ Математика это 3,25 чел./лет при отсутствии периодов нетрудоспособности, по

с учетом отпусков. Всего для введения в БД информации из РЖ с 1953 по 1996 гг. потребуется 140 чел./лет. Коллектив из 20 высококвалифицированных наборщиков сможет выполнить эту работу за 7 лет. Нельзя сказать, что особых средств не потребуется! А если учесть, что на каждых трех наборщиков требуется два корректора со знанием TeX'овского набора и с такой же, если не большей, квалификацией, то проблема переходит в разряд несбыточных.

Применение технологии сплошного сканирования печатных источников с использованием средств распознавания текстовых изображений и последующим структурированием распознанного материала, что на обычном библиографическом материале успешно демонстрирует фирма "Электронный архив" [4] также потребует огромных трудозатрат. Они связаны не только с качеством исходного материала и необходимостью его распределения по многочисленным полям БД, но и со спецификой физико-математического текста (особенно математического), суть этого текста находится зачастую не в словах, а в формулах, спецсимволах, графике и т. п.

На наш взгляд [5, 6], перспективнее ставить задачу решения проблемы электронного вида и связанную с ней задачу поиска в этом электронном массиве информации, используя комбинированный путь. Он заключается в создании специализированной БД с необходимым набором полей для ввода библиографической информации и классификационных индексов (рубрикаторы, ключевые слова и пр.) и (!) полного сканирования страниц РЖ. В результате может быть реализован поиск по всем имеющимся полям и предоставление при необходимости реферативной информации в виде изображения страницы, на которой имеется текст запрашиваемого реферата. Затем для унификации средств хранения и поиска эта БД передается в Автоматизированный банк данных ВИНИТИ откуда и осуществляется обслуживание.

По трудозатратам приведем следующий расчет: библиографическое описание составляет примерно четверть всего документа и, таким образом, на ввод всей необходимой информации потребуются те же 20 наборщиков и 1,75 года, что уже более реально. Сканирование же можно выполнить на планетарном сканере Minolta PS 7000, два экземпляра которого имеются в ВИНИТИ. По времени это займет не более 0,5 года. При выполнении пилотного проекта "Ретроспективная база данных" нами были использованы решения, в том числе и касающиеся технологии сканирования и связи изображений с записями в БД, которые были получены в проекте "Русскоязычная база данных" в части связи ее с Электронной библиотекой [7].

ПОДГОТОВКА РЕТРОСПЕКТИВНОЙ РЕФЕРАТИВНОЙ БАЗЫ ДАННЫХ

Этапы подготовки

1. Анализ наполнения РЖ с 1953 г. и выбор полей, которые послужат наполнением БД.
2. Создание структуры базы данных.
3. Разработка технологии наполнения базы данных (состав полей, кодировка спецсимволов, АРМ оператора набора, АРМ администратора и т. д.).
4. Разработка технологии сканирования РЖ с последующей постатейной сборкой отдельных изображений страниц в отдельные pdf-файлы.
5. Разработка технологии съема электронного издания, включающей вид издания на экране компьютера (интерфейс пользователя), средства поиска и т. д. (для локальной БД).
6. Разработка технологии "склейки" созданного электронного документа с pdf-файлом, содержащим страницы РЖ, на которых расположен реферат.

Основные характеристики

Ретроспективная реферативная БД по физико-математическим наукам обеспечивает:

1) просмотр и редактирование библиографических описаний публикаций РЖ в соответствии с требованиями представления данных во внутрисистемном формате ВИНИТИ;

2) просмотр и редактирование текстов рефератов (при их наличии в БД не в виде изображения, а в виде текста);

3) возможность постоянного пополнения базы данных до окончания проекта.

Библиографическое описание представлено здесь в виде, максимально приближенном к структуре библиографического описания, принятом в АБнД ВИНИТИ, что позволяет производить поиск документов по набору параметров, привычных и удобных пользователям АБнД ВИНИТИ. Набор полей, используемых для описания документов в РеБД, ограничен, по сравнению с сегодняшним составом, в силу ограниченности набора данных в печатном издании РЖ.

Представление реферативной части в виде полнотекстовых изображений позволяет снизить трудоемкость работ по загрузке информации в БД при полной адекватности данных, включая математические формулы. Реферативная часть хранится в виде pdf-файлов. Для просмотра текстов рефератов, хранящихся в виде изображений, используется свободно распространяемая программа Acrobat Reader.

ОБРАЩЕНИЕ К БД И ПОИСК ДАННЫХ

Поиск документов и обращение к ним пользователей реализованы с помощью динамического WEB-интерфейса БД, что позволяет организовать доступ к РеБД как через Интранет ВИНИТИ, так и через Интернет.

В настоящее время разработана технология перевода в электронную форму печатных изданий архивных РЖ Астрономия, Математика, Физика и доступа пользователей к РеБД, а также загрузки ретроспективной БД в АБнД ВИНИТИ.

Для работы с РеБД в АБнД ВИНИТИ пользователь выбирает на сайте www.viniti.ru в разделе "Услуги" пункт "База данных в режиме online", далее — пункт "Для пользователей Intranet ВИНИТИ вход здесь", затем — "Доступ к Банку данных ВИНИТИ", где после ввода имени пользователя и пароля открывается страница поиска в БД (рис. 1).

Выполнив запрос, пользователь получает в нижней части страницы количество найденных документов. Перейдя по ссылке "Показать", можно посмотреть заголовки найденных документов (рис. 2), а затем, выбрав пункт "Смотреть реферат", и сам документ в виде отсканированного изображения (рис. 3).

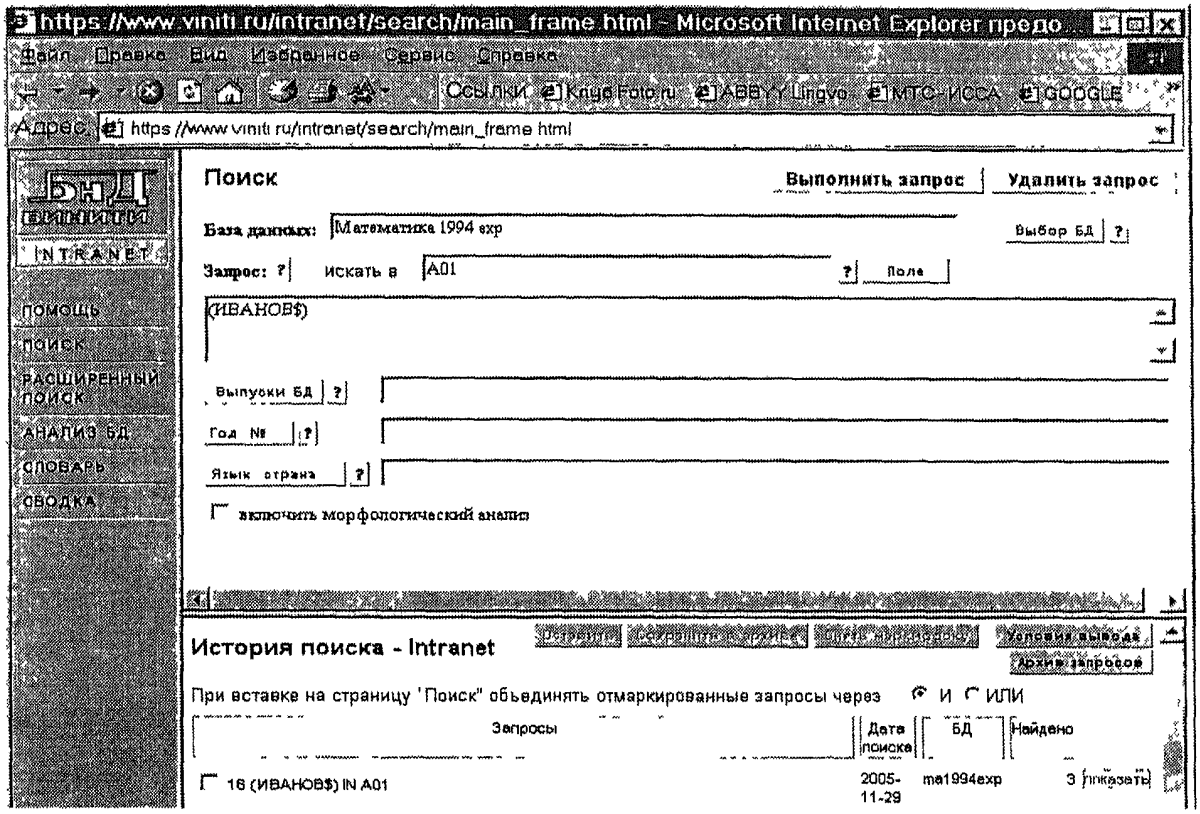

Рис 1 Страница поиска в АБнД ВИНИТИ

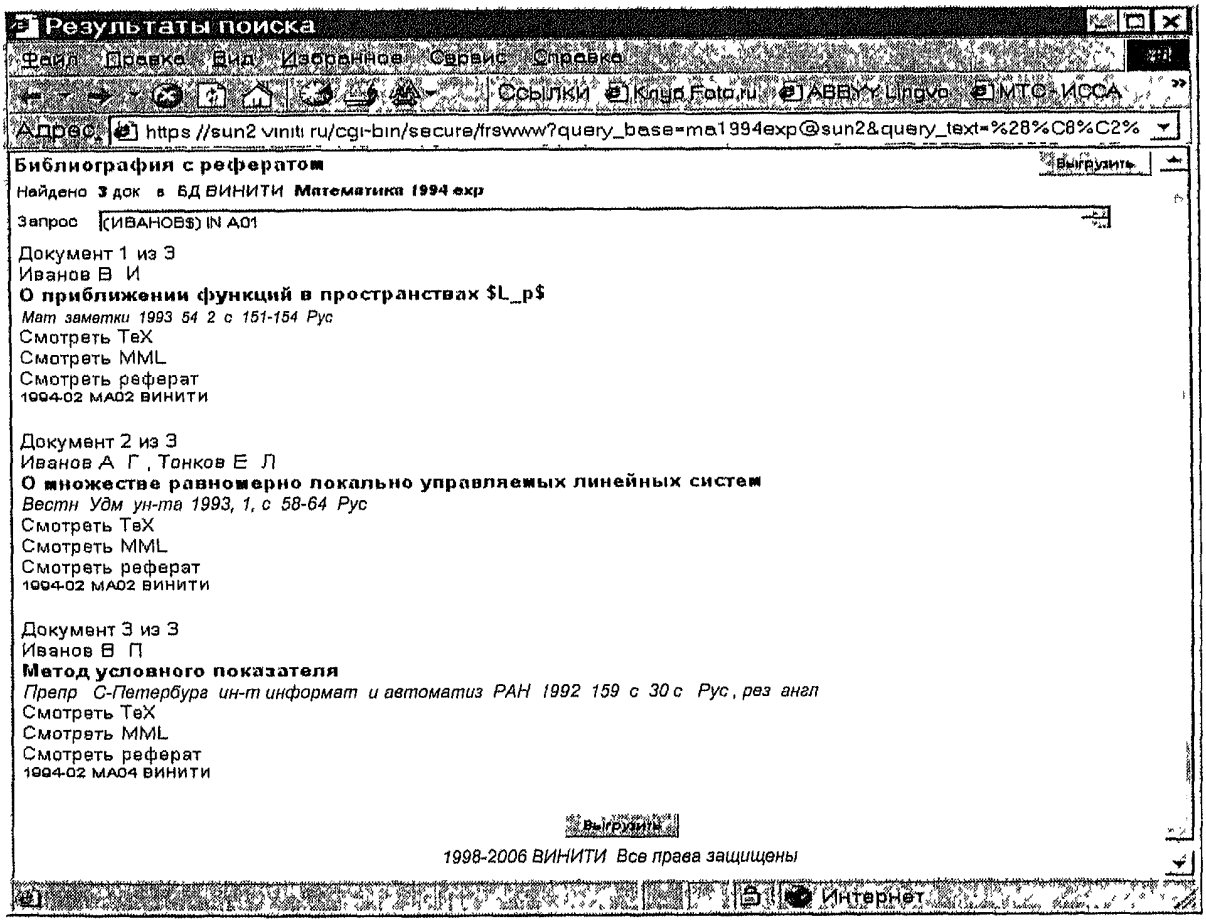

Рис 2 Список документов, найденных в АБнД ВИНИТИ по запросу на рис 1

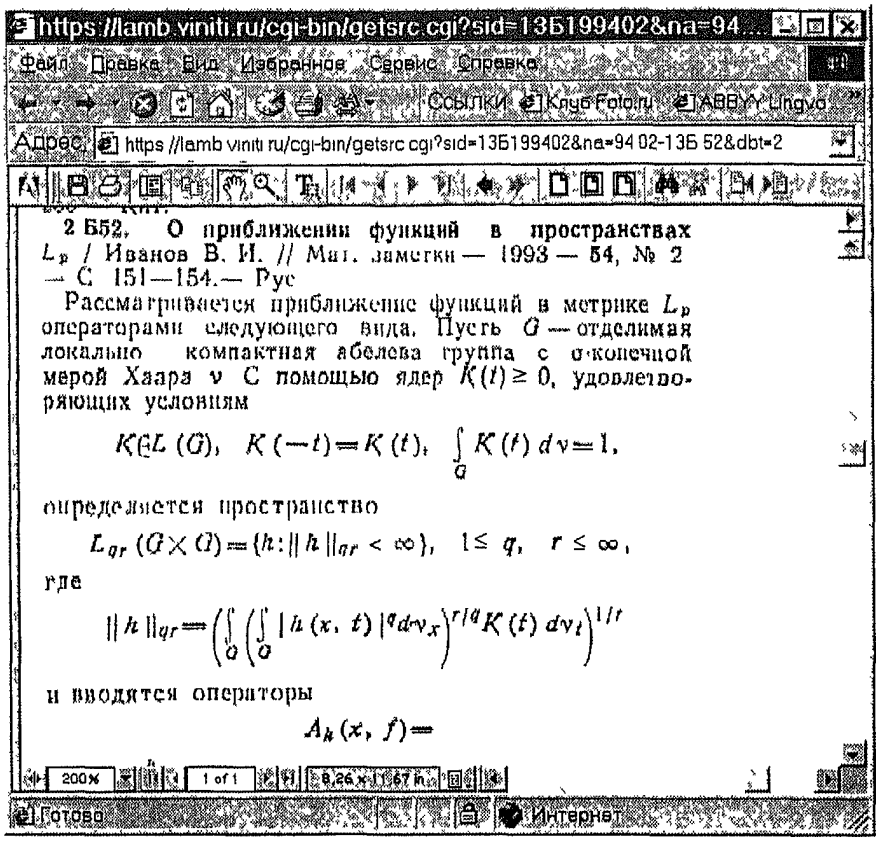

Рис 3 Реферат первого документа из рис 2 в формате PDF

Выполненные работы

1 Проектирование базы данных под СУБД Microsoft SQL Server 2000. Разработка хранимых процедур и пользовательских хранимых функций.

2. Разработка и согласование схемы кодирования математических формул с использованием стандарта LaTex.

3. Создание комплекса программ ведения базы данных.

3.1. Рабочее место оператора ввода библиографии предназначено для ввода:
библиографического описания документа (БО),
рефератов,
корректуры;
классификационных индексов.

3 2. Рабочее место администратора РеБД предназначено для:
просмотра документов и служебной информации в БД;
просмотра документов, содержащихся в файлах формата ISO-2709,
выгрузки заданной выборки документов из БД в файл ISO-2709 и, при необходимости, в файл LaTex;
администрирования учетных записей пользователей АРМ оператора набора и АРМ администратора. добавления и удаления пользователей, смены паролей, задания прав доступа к БД и т. д.;
сбора статистики по работе операторов набора и корректоров за выбранный период;
создания служебных файлов, задающих соответствие между документами в БД и номерами страниц pdf-файла, содержащего отсканированный образ заданного выпуска РЖ

4. Разработка методики получения электронных образов рефератов. Подготовка аппаратного комплекса на основе сканера Minolta PS7000.

5. Пилотный проект формирования РеБД по Математике за 1994–1996 гг.

5.1. Ввод библиографических описаний и классификационных индексов документов.

5.2. Редактирование данных.

5.3. Формирование электронных образов рефератов

5.4. Анализ результатов пилотного проекта Формирование рекомендаций по оптимизации структуры БД и требований к аппаратному обеспечению сервера РеБД.

В 2005 г функциональность Ретроспективной реферативной базы данных была расширена (см. [8, 9]) за счет добавления возможности просмотра на экране и вывода на печать диапазона документов, выбранных по коду тетради РЖ, году и месяцу РЖ, а также возможности экспорта выбранных записей в коммуникативном формате ISO 2709

В настоящее время РеБД эксплуатируется в рабочем режиме, постепенно наполняясь выпусками печатных версий РЖ более ранних годов.

СТРУКТУРА РЕТРОСПЕКТИВНОЙ РЕФЕРАТИВНОЙ БАЗЫ ДАННЫХ

В основе РеБД лежит таблица ABD01, в которую заносится содержательная часть каждого

документа, его идентификаторы, ссылка на PDF-файл с отсканированным образом выпуска РЖ и служебная информация (табл.).

Таблица ABD01

Ключ	Поле	Примечание
key	RZHYear	Год выпуска РЖ
key	RZhIssue	Номер выпуска РЖ
	RZhPage	Номера страниц с рефератом
key	SID2	Номер реферата
	KindDoc	Вид документа
	Doc_M	Монографическая часть документа в формате XML
	Doc_X	Библиографическая+ реферативная часть документа в формате XML
	Doc_Extra	Поля документа, не вошедшие в Doc_X и Doc_M, в формате XML
	InputDate	Дата набора документа
	InputOperator	Оператор набора
	CorDate	Дата ввода корректуры
	Corrector	Идентификатор корректора
	ScanPath	Место расположения отсканированных страниц РЖ с данным рефератом
	NeedToCheck	Признак наличия неясностей в документе

Кроме того, в РеБД входят служебные таблицы: list_CHARACTER, list_COUNTRY, list_LANGUAGE — содержат списки допустимых значений характера документа, страны и языка, соответственно; Users_ACT содержит имена, идентификаторы, зашифрованные пароли и права доступа пользователей АРМ; Correctors содержит имена и идентификаторы корректоров; NTP_Lab и NTP_Doc — содержат описание элементов данных и их распределение по типам документов согласно Нормативно-техническому предписанию ВИНИТИ 10-2004 [10].

КОДИРОВАНИЕ СПЕЦСИМВОЛОВ И МАТЕМАТИЧЕСКИХ ФОРМУЛ

Для кодирования спецсимволов и математических формул при создании Ретроспективной базы данных на основе анализа имеющихся свободно распространяемых и коммерческих средств был выбран пакет MiKTeX — современный пакет программ для работы с документами в формате TeX.

При выборе учитывались достоинства пакета MiKTeX:

- является открытым свободно распространяемым ПО и широко используется мировым научным сообществом для подготовки научно-технической литературы;
- удовлетворяет всем потребностям кодировки сложных математических формул;
- обеспечивается технической поддержкой, имеет необходимые технические описания и руководства для пользователя, средства автоматизации установки обновлений ПО;
- поддерживается LaTeX;
- обладает сравнительной простотой установки и настройки;
- имеет стандартизованные средства генерации библиографических и других указателей, классификационных индексов и пр.;
- имеет поддержку русского языка.

К приведенным характеристикам следует добавить, что программа просмотра Yap, входящая в пакет, позволяет оптимизировать цикл "редактирование — компиляция — просмотр", что является существенным преимуществом MiKTeX, поскольку технологии языка TeX не предусматривают реализации принципа WYSIWYG.

ВЫВОДЫ

1. Выполнено проектирование базы данных под СУБД Microsoft SQL Server 2000. Разработаны хранимые процедуры и пользовательские хранимые функции.
2. Разработаны и согласованы схемы кодирования математических формул с использованием стандарта LaTeX.
3. Создан комплекс программ ведения базы данных.
4. Разработана методика получения электронных образов рефератов. Подготовлен аппаратно-программный комплекс на основе сканера Minolta PS7000.
5. Выполнен пилотный проект формирования РеБД данных по Математике (1994–1996 гг.).
6. Работа по дальнейшему наполнению РеБД выполняется уже в рабочем режиме.

СПИСОК ЛИТЕРАТУРЫ

1. Черный А. И. Всероссийский институт научной и технической информации.-- М.: ВИНИТИ, 2005. 316 с.
2. Арский Ю. М. и др. Банк данных ВИНИТИ. Состояние и перспективы развития. М.: ВИНИТИ, 2006. 242 с.
3. Гиляревский Р. С. и др. Информатика как наука об информации: Информационный, документальный, технологический, экономический, социальный и организационный аспекты. - М.: ФАИР-ПРЕСС, 2006. - 592 с.
4. Сайт фирмы "Электронный архив". http://www.elar.ru.
5. Шамаев В. Г., Жаров А. В. Проблемы создания ретроспективных реферативных баз данных ВИНИТИ по физико-математическим наукам. Деп. в ВИНИТИ 24.05 2005 № 737-В2005.
6. Шамаев В. Г., Жаров А. В, Батурина О. Н., Горшков А. Б., Максимов И. Н., Старцева О. Б. Описание технологии подготовки ретроспективных реферативных баз данных ВИНИТИ по физико-математическим наукам. Деп. в ВИНИТИ 24.05 2005 № 739-В2005.
7. Шамаев В. Г., Жаров А. В., Горшков А. Б. База данных и Электронная библиотека русскоязычной литературы по физико-математическим наукам // НТИ. Сер. 1. - 2006 - № 7 - С. 6 13.
8. Горшков А Б., Жаров А. В., Лось Е. Л., Максимов И. Н., Старцева О. Б., Шамаев В. Г., Щербина-Самойлова М. Б. Расширение функциональности Автоматизированного рабочего места администратора Ретроспективной реферативной базы данных ВИНИТИ по физико-математическим наукам. Деп. в ВИНИТИ 07.11.2005 № 1431-В2005.