ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК [002:004.658]:51/53

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

База данных и Электронная библиотека русскоязычной литературы по физико-математическим наукам

Описываются технология реализации базы данных по русскоязычным источникам по физико-математическим наукам, а также создание электронной библиотеки по этим же источникам. Полученная База данных входит в единый комплекс подготовки информационных продуктов ВИНИТИ — сигнальной информации, печатных выпусков РЖ и их электронных аналогов, пополнения Банка данных ВИНИТИ.

Задачи электронной библиотеки — предоставление копий статей, что является традиционной услугой, которую предоставляет ВИНИТИ, а также снабжение электронным изображением статьи референтов и редакторов по внутренней сети ВИНИТИ.

ВВЕДЕНИЕ

С развитием электронных средств предоставления информационных услуг большое значение приобретают такие параметры, как скорость обработки информационных продуктов в информационных центрах и время доставки их потребителю. Так, например, в настоящее время в ВИНИТИ временной промежуток от поступления единицы информации (книга, журнал, депонированная рукопись, патент и т. д.) до отражения ее в Реферативном журнале, поступившем к подписчику, составляет не менее 0,5 года, а зачастую и более [1]. Раскроем любой печатный или электронный РЖ середины текущего года и увидим, что он наполнен рефератами из источников прошлого года. Нередко встречаются рефераты из источников позапрошлого и более ранних годов. Это связано с долгим процессом обработки документов на входе, переработки их в отделах научной информации при реферировании и редактировании, а также при печати тиража в типографии и доставке печатной продукции подписчику. То же самое и для электронной продукции за исключением тиражирования. На сокращение времени обработки на каждом участке требуется много сил, и эта работа должна выполняться постоянно. По первой части, касающейся переработки документов в ВИНИТИ, немало уже сделано: разработана автоматизированная система регистрации документов — АС "Вход", используется Единая технологическая база данных (ЕТБД), позволяющая автоматизировать весь процесс подготовки РЖ, в том числе автоматически передавать документы из одного выпуска в другой (выполнять дублирование), генерировать оригинал-макет издания буквально за минуты. Формально-логический контроль встроенный в ЕТБД позволяет тщательно проверять документы по многим заданным параметрам перед окончательной передачей их в печать и в Банк данных ВИНИТИ.

Однако не все возможности ускорения обработки информации еще используются. Наряду с пополнением Информационно-технологического комплекса ВИНИТИ Единой технологической базой данных [2, 3], в этой работе мы описываем новую разработку, касающуюся ускорения обработки русскоязычной литературы по физико-математическим наукам. Делая акцент на физико-математические науки, мы кроме всего прочего имели в виду сложность представления математических формул, отдельных символов и физических и математических понятий в электронной форме. Из-за этой сложности БД по математике появилась в ВИНИТИ лишь в 1997 г., для сравнения — по автоматике и радиоэлектронике, машиностроению, биологии — в 1981 г.

Но и эти БД нельзя назвать полностью отражающими наполнение РЖ ВИНИТИ. Например, символьная информация 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 отсутствуют. Выпуски РЖ Математика, которые набираются в TeX'e, в этой кодировке и предстают перед пользователем. Для просмотра рефератов в такой БД пользователь должен перенести каждый реферат на свой компьютер и только после компиляции текста в пакете TeX и его довольно длительной настройки получить нормальное отображение. Решение этой проблемы найдено при разработке Электронного РЖ для физико-математических наук, что описано в [4].

Возвращаясь к русскоязычным источникам и их представлению в РЖ, следует сказать, что в

ВИНИТИ была сформулирована цель реализации отдельной русскоязычной базы данных со своими собственными задачами, которые сформулированы далее, а также с возможностью использования ее наполнения в технологии подготовки РЖ. Второй частью этой работы было создание электронной библиотеки полнотекстовых изображений по этим же источникам. Задача этой библиотеки — предоставление копий статей заказчикам, что является традиционной услугой, которую предоставляет ВИНИТИ, а также снабжение референтов и редакторов электронным изображением статьи по внутренней сети ВИНИТИ (Интранет ВИНИТИ).

Реализация такого проекта включает выбор структуры БД, определение необходимого набора таблиц с заданными полями и связей между ними, ввод библиографической информации и полное сканирование страниц русскоязычных источников. Состав полей базы данных определяется ПТП 10–2004 ВИНИТИ [5] как наиболее полным в стране на данный момент источником по описанию документов различных видов в базах данных.

Полностью направление исследований в ВИНИТИ, названное Русскоязычной базой данных, состоит из решения нескольких связанных друг с другом задач:

1. Введение монографической, библиографической и аналитической информации в базу данных.
2. Формирование БД по русскоязычным источникам (РуБД), включающей библиографическое описание, резюме или аннотацию с грубой рубрикацией по отраслям знания.
2. Создание Центра оперативного хранения цифровых изображений (ЦОХ).
3. Сканирование и передача полнотекстовых изображений в ЦОХ.
4. Создание Электронной библиотеки русскоязычной литературы и web-интерфейса пользователя, позволяющего проводить поиск издания или его составляющих.
5. Создание Узла связи Центра оперативного хранения цифровых изображений с автоматизированным рабочим местом редактора (АРМ Редактора).
6. Формирование информационных продуктов. Сигнальной информации в печатной форме, Электронного журнала сигнальной информации, Информационных извещений для отделов научной информации (ОНИ) в печатной и электронной формах (форма 33).

Принципы решения этих задач описаны далее.

СТРУКТУРНАЯ И ТЕХНОЛОГИЧЕСКАЯ СХЕМЫ ОБРАБОТКИ ЛИТЕРАТУРЫ ПРИ ФОРМИРОВАНИИ РУССКОЯЗЫЧНОЙ БАЗЫ ДАННЫХ

Структурно комплекс подготовки Русскоязычной базы данных и производства информационных продуктов можно разделить на несколько участков (рис. 1):

- участок ввода библиографической и реферативной информации;
- участок сканирования источников;
- русскоязычная база данных;
- библиотека долговременных копий;
- участок подготовки информационных продуктов;
- участок взаимодействия с отделами научной информации (ОНИ) ВИНИТИ и АС “Вход”.

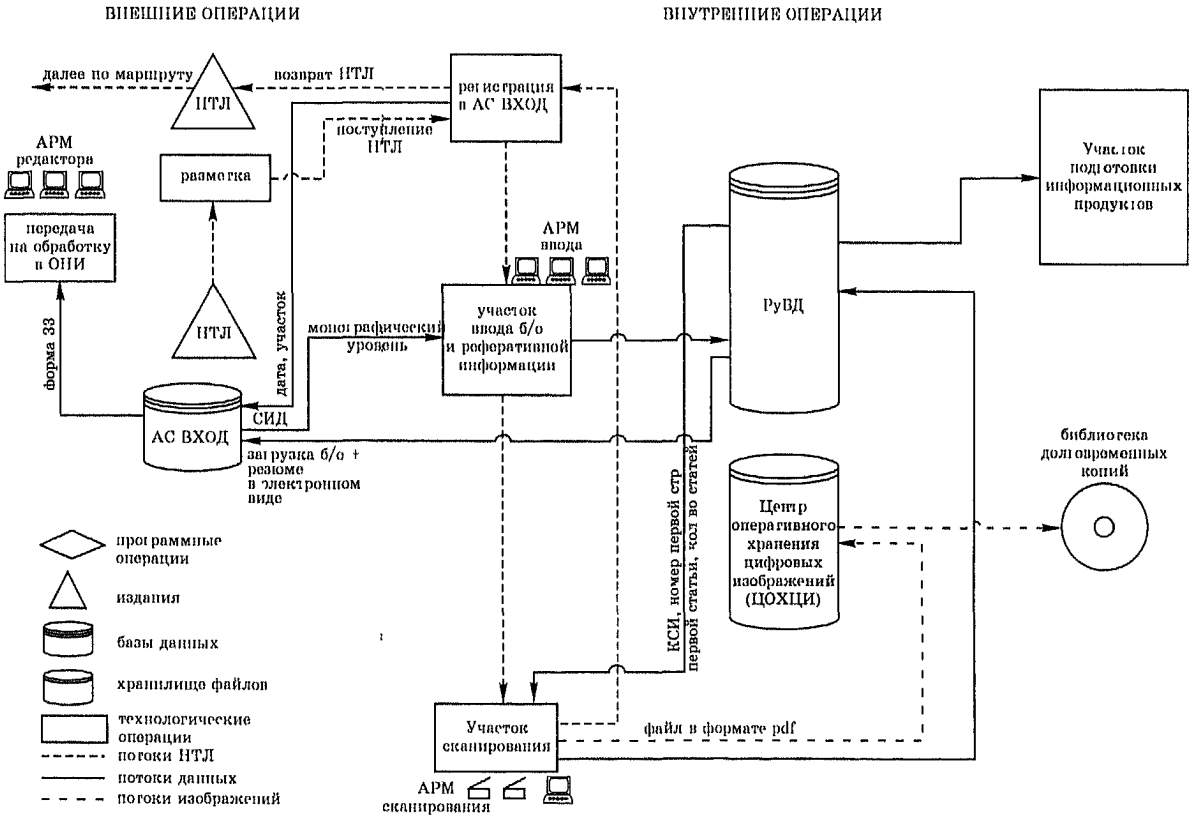

Рис 1. Структурная схема обработки русскоязычной научно-технической литературы

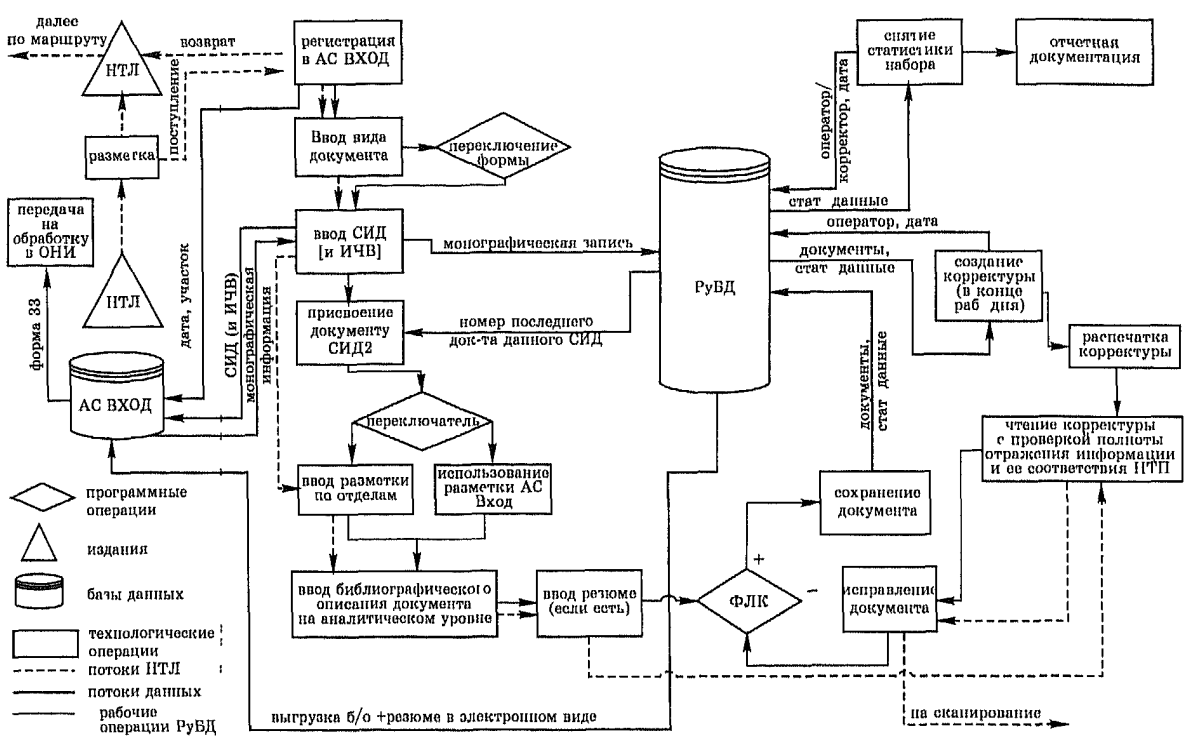

Рис 2. Технологическая схема обработки научно-технической литературы на участке ввода библиографического описания

На участке ввода библиографической информации используются АРМы операторов набора, позволяющие вводить библиографическую и реферативную информацию в зависимости от источника (журнал, статья в журнале, книга, статья в книге, материалы конференций, депонированная работа и т. д.). Монографическая часть библиографии берется из БД Монографического уровня АС "Вход" по системному идентификатору документа (СИД-1). Набор осуществляется прямо с источника. Результаты набора после корректуры и формально-логического контроля помещаются в Русскоязычную базу данных. В отличие от Единой технологической базы данных (ЕТБД), результаты аналитической переработки документов хранятся в РуБД постоянно (хранение документов в ЕТБД ограничено в настоящее время одним годом).

Затем документальный источник поступает на участок сканирования, где выполняется его постраничное сканирование и помещение оцифрованных изображений в Центр оперативного хранения цифровых изображений. Для длительного хранения и, одновременно, создания резервных копий по мере накопления материала информация переписывается на DVD, которые помещаются в Библиотеку долговременных копий

На участке подготовки информационных продуктов формируется оригинал-макет для печатного выпуска Сигнальной информации поступлений в РуБД и снимается Электронный выпуск сигнальной информации.

На участке взаимодействия с ОНИ и АС "Вход" происходит съем с РуБД информации для запуска ее в виде так называемой формы 33 в документальный поток, поступающий в ОНИ, а с помощью АРМ Редактора, установленного в ОНИ, возможен просмотр редакторами полного текста заинтересовавшей их статьи из Центра оперативного хранения. Наличие этого участка позволяет также избегать копирования источников на бумагу и рассылки копий по отраслевым отделам.

АКУСТИЧЕСКИЙ ЖУРНАЛ. 2005, том 51, № 3, с. 342–351

УДК 537.23 (Ф 35) (Рус Рез. англ)

ИССЛЕДОВАНИЕ ЭНЕРГЕТИЧЕСКИХ ХАРАКТЕРИСТИК ПОЛЯ ПРИ РАСПРОСТРАНЕНИИ ЗВУКА В БАРЕНЦЕВОМ МОРЕ

© 2005 г. О.П. Галкин, Л.В. Швачко

Акустический институт им. Н.Н. Андреева РАН
117036 Москва, ул. Шверника 4
E-mail: bvp@akin.ru

Поступила в редакцию 23.12.2003 г.

Приводятся результаты экспериментального исследования энергетической структуры звукового поля в Баренцевом море на трассе протяженностью ~80 км при глубине моря ~220–250 м. Использовалось псевдошумовое излучение в двух третьоктавных полосах частот со средними значениями 1.25 кГц и 3.15 кГц. Источник звука располагался в приповерхностном слое на глубине 10 м и под слоем скачка на глубине 100 м. Прием сигналов производился на горизонтах 15 м, 100 м и 200 м. Экспериментальные результаты сопоставлялись с результатами лучевых расчетов, учитывающих поверхностное волнение и параметры грунта, полученные на основе систематизации имеющихся данных. В результате анализа показана принципиальная возможность прогнозирования энергетических характеристик структуры поля в мелком море с учетом зависимости скорости звука от глубины и параметров его границ.

Рис. 3. Пример разметки статьи из журнала

Далее рассмотрим подробнее технологию обработки научно-технической литературы на отдельных участках. Источники поступают на участок разметки в отдельном потоке, чтобы сократить время, проводимое в АС "Вход" (рис. 2). Здесь происходит оперативная разметка библиографии и

резюме и производится грубое "рубрицирование" по отраслям знания — ставится штамп разметки (пример, разметки статьи из журнала см. на рис. 3).

Далее, уже на участке подготовки РуБД, происходит регистрация источника в АС "Вход" (делается электронная отметка о нахождении источника на данном участке в "путевом листе" прохождения источника по этапам обработки) с использованием ручного сканера. Монографическая информация по СИД-1 источника запрашивается у БД АС "Вход", исключая тем самым повторный набор. На этом же участке каждому документу (статье) присваивается его уникальный идентификатор — СИД-2, отражающий порядковый номер статьи в издании. Далее, если была выполнена уточняющая разметка по отделам научной информации, то она вводится, в противном случае используется разметка АС "Вход".

Затем начинается ввод аналитической информации (библиографии, резюме, разметки по отделам научной информации, что тождественно у нас отраслям знания) согласно полям БД в соответствии с НТП 10–2004 ВИНИТИ прямо с источника. В зависимости от вида источника (журнала, книги, депонированной работы и др.) производится переключение формы ввода. В конце рабочего дня происходит генерация и распечатка корректуры, и в течение следующего дня корректор читает корректуру и проверяет полноту отражения информации и соответствие ее НТП. Параллельно снимается информация для учета выработки операторов набора, которая затем используется при подсчете ежемесячной выплаты заработной платы операторам набора и корректорам. После правки корректора, к которому также поступают источники, в записи вносятся необходимые исправления.

После этих операций документы проходят этап формально-логического контроля (ФЛК) и либо сохраняются в РуБД, либо исправляются с последующим повторением ФЛК и при положительном исходе окончательно сохраняются в РуБД. Монографическая информация для каждого документа снимается с БД Монография АС "Вход". Далее источники поступают на участок сканирования (рис. 4).

На участке сканирования для каждого источника перед сканированием вводится СИД-1, выполняется запрос РуБД на наличие издания, имеющего такой СИД-1 и проводится проверка на отсутствие скана для этого источника (с целью исключения повторного сканирования). При положительном ответе вводятся номер страницы (страниц) содержания, номер страницы первой статьи источника для последующего использования при извлечении документов из хранилища изображений, код пригодности источника для сканирования (определение кода пригодности описывается ниже) и выполняется сканирование, этапы которого подробно описаны в [6]. Проверку качества каждого скана на первом этапе выполняет оператор сканирования сразу после его проведения и, в дальнейшем, тщательно просматривает изображения при комплектовании выходного продукта для передачи в Центр оперативного хранения цифровых изображений.

В РуБД, кроме библиографической информации, хранится также техническая информация, позволяющая в дальнейшем извлекать необходимые страницы издания: признак наличия у документа изображения (документы плохого качества не сканируются), номер страницы оглавления, номер страницы первой статьи (для последующего правильного извлечения других статей источника) и др.

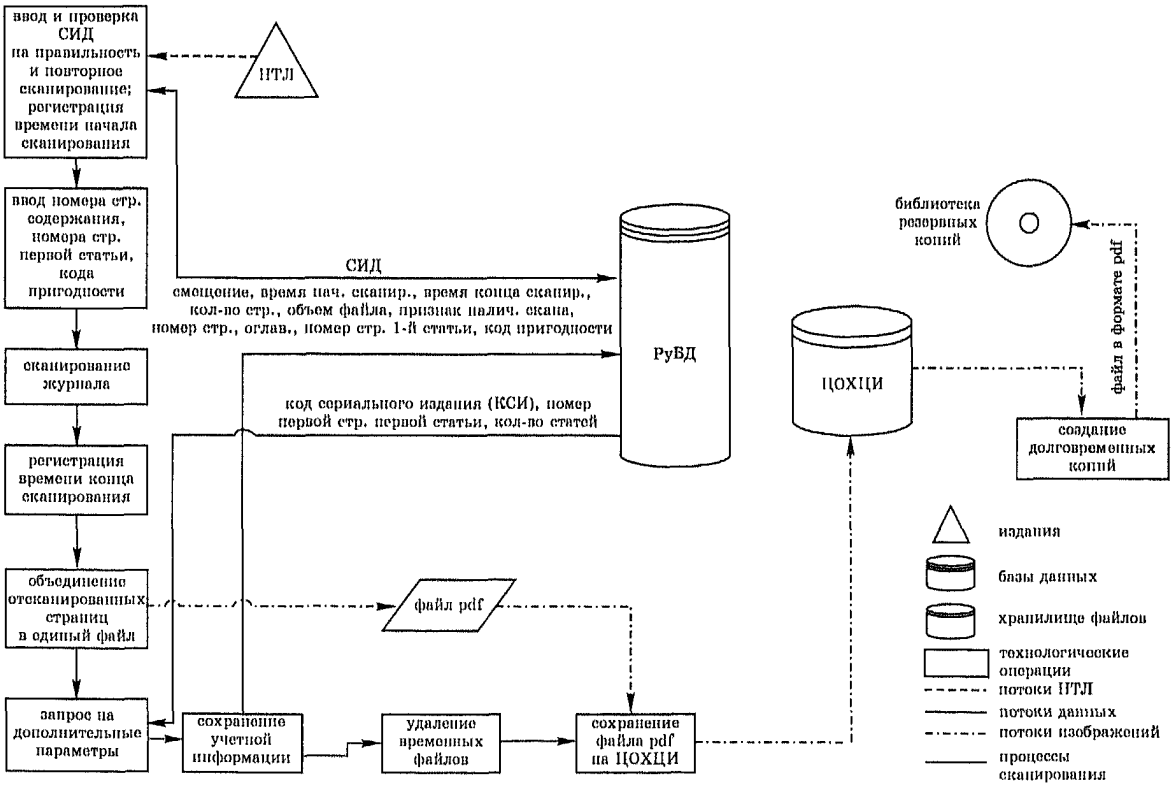

Рис. 4. Технологическая схема обработки научно-технической литературы на участке сканирования

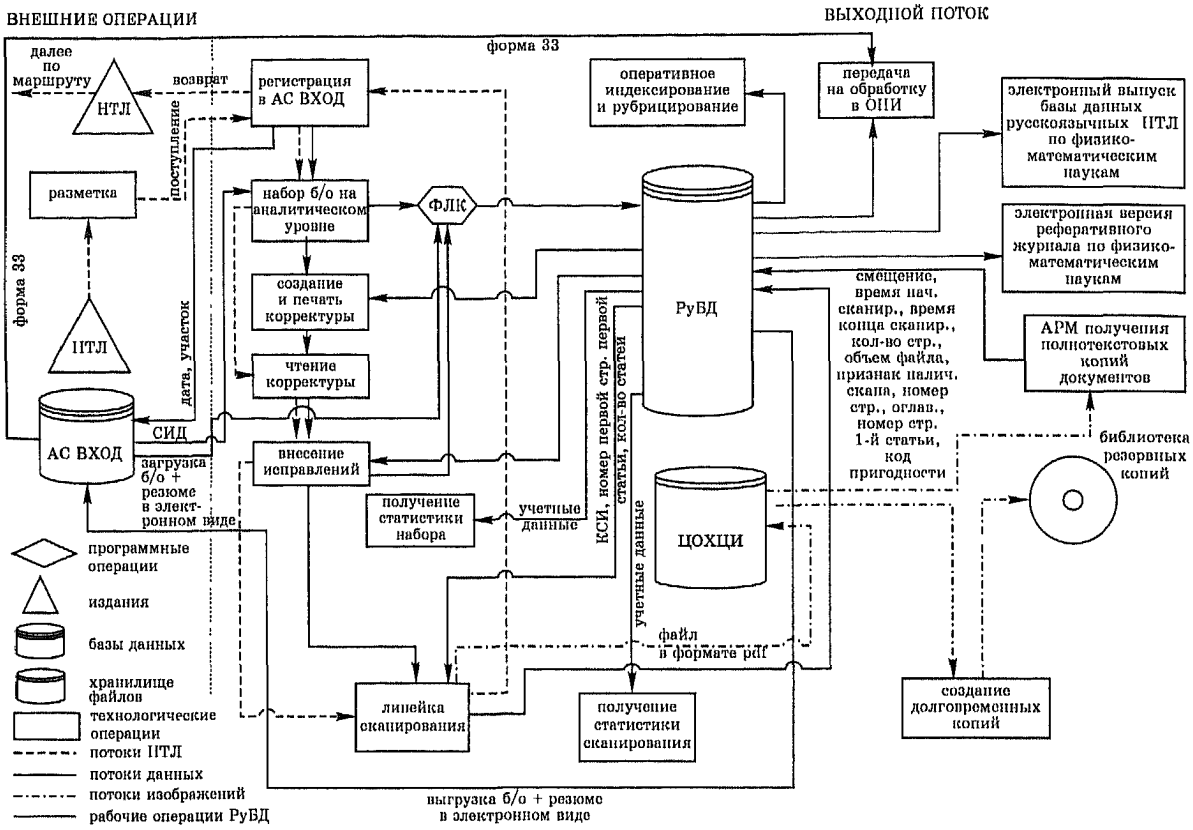

Рис. 5. Схема обработки научно-технической литературы и формирования конечных информационных продуктов. РуБД и Электронная библиотека

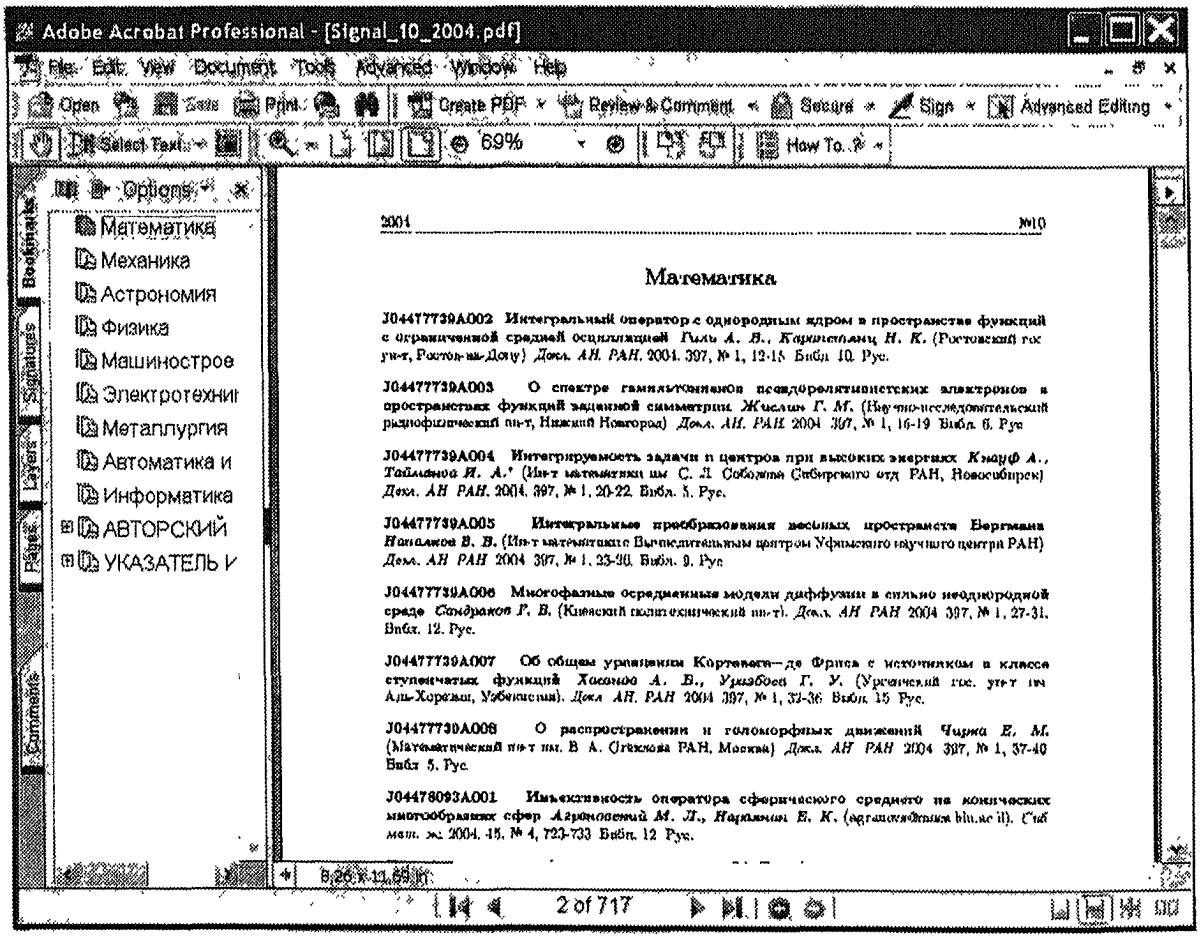

Рис. 6. Adobe Acrobat Professional с преобразованным файлом и открытой вкладкой "Bookmarks", слева — оглавление сигнального выпуска, справа — первая страница, соответствующая указателю сигнального выпуска

Затем, после регистрации времени конца сканирования, отдельно отсканированные страницы источника объединяются в единый файл и записываются в Центре оперативного хранения цифровых изображений. Запись статистической информации в БД и управление процессом объединения изображений страниц в единый файл осуществляется с помощью программы ListToFile. После каждой смены, как и в случае набора информации аналитического уровня, выполняется операция получения статистики по сканированию для каждого оператора с целью последующего учета выработки и составления ведомости выплат.

По мере накопления отсканированных изображений резервные копии этой информации записываются на DVD для создания Библиотеки резервных (долговременных) копий.

Схема полного процесса обработки русскоязычной литературы по физико-математическим наукам приведена на рис. 5.

На схеме опущены некоторые детали, но дополнительно приведены блоки, отражающие получение выходных продуктов:

а) передачи документов на обработку в отделы научной информации;

б) съема печатной версии Сигнальной информации по физико-математическим наукам по русскоязычным источникам (или отдельного выпуска РЖ);

в) съема Электронной версии Сигнальной информации по физико-математическим наукам по русскоязычным источникам (или отдельного выпуска Электронного РЖ).

ОСНОВНЫЕ СТРУКТУРНЫЕ ЭЛЕМЕНТЫ РуБД И ИХ КРАТКОЕ ОПИСАНИЕ

Русскоязычная база данных представляет собой совокупность библиографической и содержательной информации, а также служебных данных, характеризующих обрабатываемый поток научно-технической литературы с точки зрения прохождения определенных технических этапов.

Документы организованы в виде совокупности таблиц:

RuBD_mono — таблица, содержащая системный идентификатор издания (СИД-1), монографическое описание в XML-формате и служебную информацию;

RuBD — таблица, содержащая системный идентификатор издания (СИД-1), системный идентификатор документа (СИД-2), данные о тематической разметке документа, виде документа, содержательную информацию в XML-формате;

RuBD_texno — таблица, содержащая системный идентификатор документа (СИД-2), признак наличия сканированного изображения и служебную информацию.

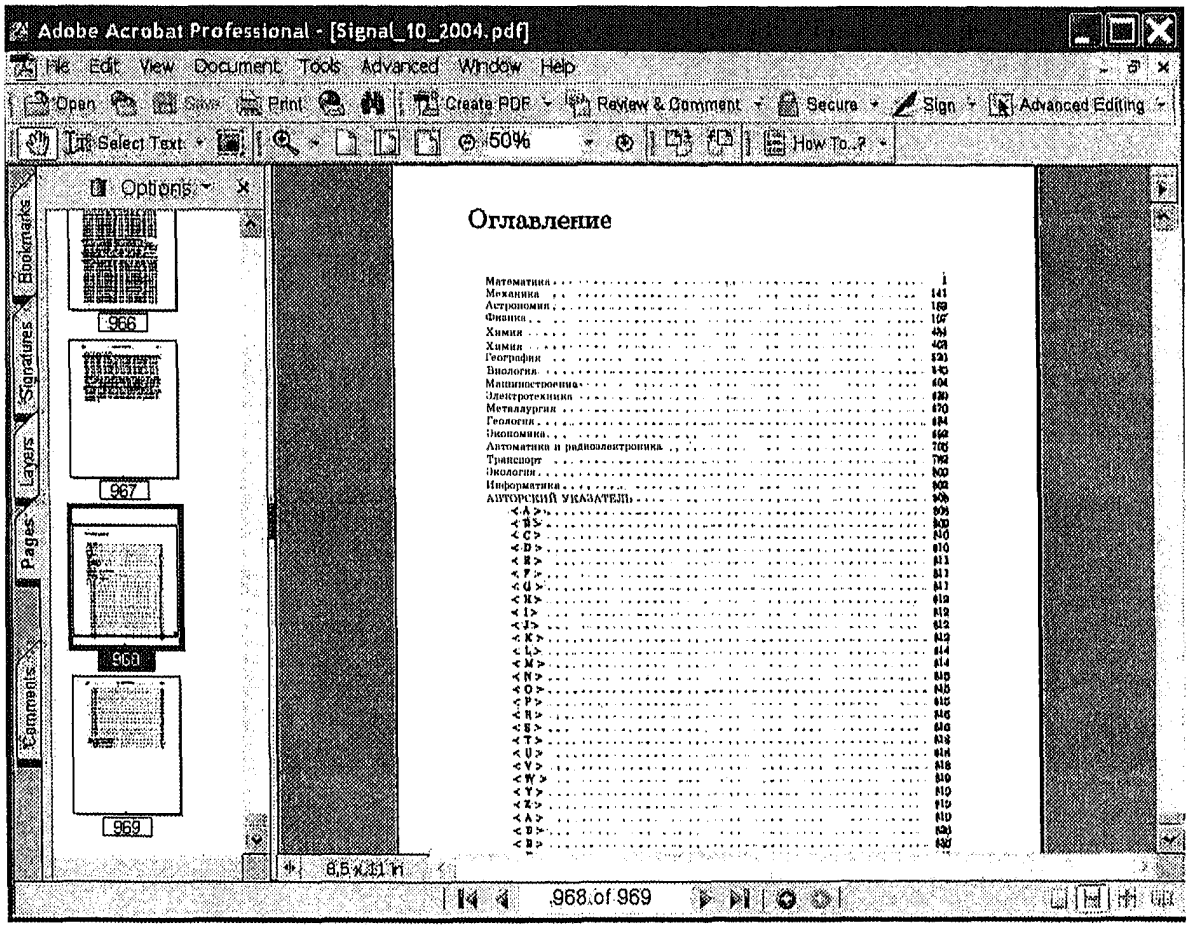

Рис. 7. Adobe Acrobat Professinal с открытой вкладкой "Pages" и выделенной страницей "Оглавление"

Кроме основных таблиц RuBD, RuBD_mono и RuBD_texno, используются вспомогательные таблицы.
NTP_lab — описание полей НТП — используется для проведения формально-логического контроля, операции сборки и извлечения полей в форматах XML:
NTP_doc — распределение полей НТП по видам документов с признаками обязательности и другими параметрами.

ФОРМИРОВАНИЕ СИГНАЛЬНОГО ВЫПУСКА РЖ

Подготовка Сигнального выпуска производится в несколько этапов.

- подготовка файла Сигнального выпуска в формате LaTeX;
- обработка файла Сигнального выпуска в формате LaTeX и получение электронной версии в формате PostScript;
- подготовка электронной версии в формате PDF, удаление временной информации;
- подготовка титульного листа выпуска;
- вставка титульного листа выпуска в файл в формате PDF;
- подготовка и печать оформления CD, подготовка образа диска для записи выпуска на CD.

Подготовка файла Сигнального выпуска РЖ в формате LaTeX производится в АРМ Publisher. Название файла строится следующим образом: Signal_MM_YYYY.tex , где MM — номер выпуска (месяц), YYYY — год. Например, для выпуска сигнальной информации № 4 за 2006 год файл будет сформирован под именем Signal_4_2006.tex.

Файл сигнального выпуска в формате LaTeX обрабатывается компилятором LaTeX, а полученный DVI-файл посредством утилиты dvips преобразуется в файл формата PostScript, готовый к дальнейшей конвертации в формат PDF.

Подготовка Сигнального выпуска в формате PDF производится в программе Adobe Acrobat Professional (рис. 6, 7).

Титульный лист выпуска готовится на основании шаблона, созданного с помощью Microsoft Word (см. рис. 8) путем модификации номера выпуска, года и другой значимой информации

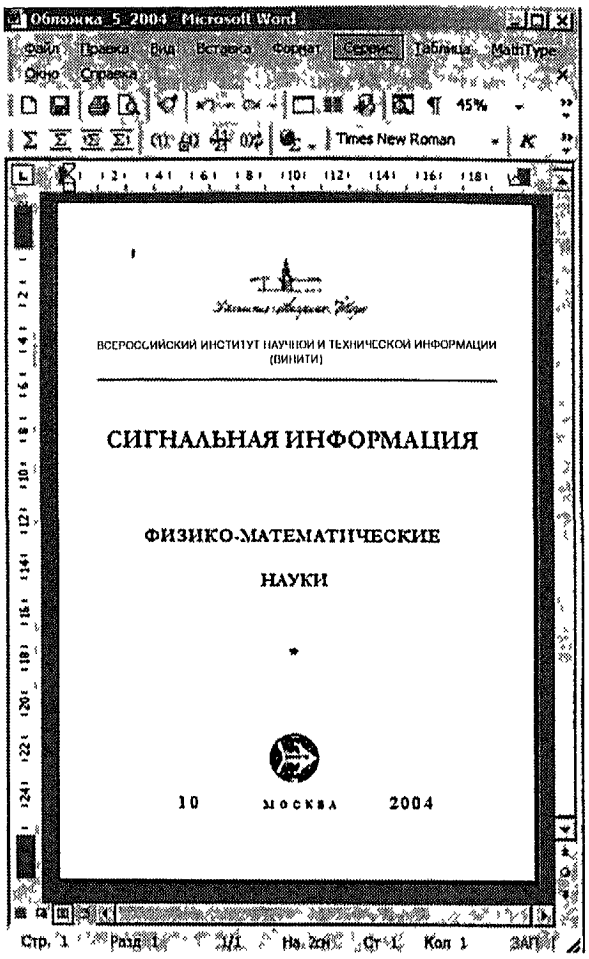

Рис 8 Титульный лист выпуска Сигнальной информации по физико-математическим наукам

ТЕХНОЛОГИЧЕСКИЙ УЧАСТОК ПОТОЧНОГО СКАНИРОВАНИЯ РУССКОЯЗЫЧНОЙ ЛИТЕРАТУРЫ ПО ФИЗИКО-МАТЕМАТИЧЕСКИМ НАУКАМ

Общие положения

Технологическая линейка поточного сканирования является частью проекта разработки технологии обработки входного массива русскоязычных научно-технических изданий по физике и математике. Выходной поток может быть использован для изготовления печатных копий, просмотра текстовых статей на мониторе компьютера, для полного или частичного распознавания текста в изданиях и индексирования и поиска требуемой информации, пересылки образов печатных изданий по локальной сети (Интранет ВИНИТИ) и сети Интернет.

Результатом работы технологической линейки процесса поточного сканирования (ППС) является файл в формате PDF, содержащий индивидуальные изображения отдельных страниц издания, упорядоченные в порядке номеров страниц. Имя файла содержит номер СИД-1, что обеспечивает привязку файла к РуБД ВИНИТИ

Для достижения этого результата выполняются следующие операции:

- оптическое сканирование периодических изданий;
- преобразование изображений в формат хранения;
- объединение файлов изображений отдельных страниц в необходимом порядке в общий файл, соответствующий журналу;
- привязка конкретного журнала к информации в РубД по номеру СИД-1

Требования к технологической линейке

Технологическая линейка должна обеспечивать:

- производительность по потоку сканируемых материалов на уровне, допустимом параметрами оборудования — сканера, канала связи с управляющим компьютером и быстродействием компьютера. Ориентировочная производительность — 200 страниц в час;

• выполнение ППС в рамках регламентных сроков обработки документов в линейке русскоязычной информации;

• качество получаемых изображений, достаточное для воспроизведения всех элементов печатного текста (на уровне 600 dpi), возможности масштабирования изображений для просмотра на экране, а также для распознавания текстовой информации;

• сохранение смысловой информации, содержащейся в источнике, и минимизацию оптических искажений, неизбежно возникающих в процессе обработки;

• связь получаемых файлов изображений с информацией в РуБД ВИНИТИ;

• возможность (факультативной) постобработки сканированной информации с целью выделения отдельных частей документа, преобразования к другому разрешению, компрессии изображений и распознавания текста.

При определении требований к технологической линейке следует иметь в виду ограничения, накладываемые техническими параметрами используемого оборудования. Это касается, прежде всего:

а) параметров качества изображения;

б) процесса коррекции положения и центрирования оригинала;

в) общей скорости (производительности) процесса.

Программное обеспечение, используемое в ППС

• TWAIN драйвер сканера, для управления параметрами сканера и работой сканера (входит в комплект поставки).

• Программа Adobe PhotoShop 7 для оперативной корректировки, управления потоком сохранения, компрессии и преобразования изображений.

• Программа Appligent AppendPro для пакетного объединения файлов с изображениями в один pdf-файл.

• Программа ListToFile для создания списка файлов и файла параметров для AppendPro.

• Программа Adobe Acrobat 6 для посткоррекции общего pdf-файла с изображениями и привязки дополнительной информации.

• Программа Adobe Acrobat Reader 6 для просмотра результирующих файлов.

Технологические операции в процессе ППС

• Регистрация оригиналов печатных изданий при поступлении их на участок ввода русскоязычной библиографии, выделение потока, предназначенного для сканирования.

• Оценка пригодности печатной единицы по визуальному контролю и по списку принятых для сканирования изданий.

• Создание временной директории хранения с именем, соответствующим номеру СИД.

• Коррекция рабочих параметров сканера и выполнение поточного сканирования выбранной печатной единицы (одного номера журнала) с сохранением текущих изображений в оперативной памяти управляющего компьютера.

• Сохранение постраничных изображений в отдельные файлы (формат PDF) с преобразованием формата и компрессией.

• Объединение файлов изображений отдельных страниц в один файл с удалением промежуточных файлов;

• Сохранение результатов сканирования на жестком диске Центра оперативного хранения.

Характеристики входного потока периодических изданий

Входной поток по участку русскоязычной периодики оценивается в 380 наименований журналов, в среднем по 96 страниц. Таким образом, максимальная цифра составляет 36 500 страниц или около 1800 страниц в день (восемнадцать журналов в день).

Распределение поступлений по дням неравномерно, что может приводить к переполнению технологических линеек и задержке в обработке журналов.

В соответствии с техническими характеристиками среди поступающего потока выделяются журналы, непригодные для сканирования по своим оптическим характеристикам. Это журналы, отпечатанные на цветной бумаге, с неоднородным фоном, со значительным количеством цветных фотоиллюстраций. Таких журналов около 10%.

Оставшаяся часть принимается для сканирования и подразделяется на три группы по характеру переплета и качеству бумаги. Этот критерий является ключевым для определения нормы производительности технологической линейки.

ЗАКЛЮЧЕНИЕ

Наполнение РуБД по физико-математическим наукам началось в 2004 г. и в настоящее время насчитывает около 110 000 единиц хранения. Данные по библиографии доступны в открытом режиме по адресу: http://catalog.viniti.ru/ElLibrary/.

СПИСОК ЛИТЕРАТУРЫ

1. Черный А. И. Всероссийский институт научной и технической информации: 50 лет служения науке. - М.: ВИНИТИ, 2005.— 316 с.

2. Шамаев В. Г., Жаров А. В., Горшков А. Б. Единая технологическая база данных для подготовки информационных продуктов ВИНИТИ // НТИ. Сер. 1. 2006.— № 5.— С. 10–15.

3. Шамаев В. Г., Жаров А. В., Батурина О. Н., Горшков А. Б., Лось Е. К., Лукашевич Н. Л., Максимов И. Н., Седякина А. Н., Старцева О. Б., Щербина-Самойлова М. Б., Ягельницая О. А. Разработка технологии создания единой технологической базы данных для подготовки информационных продуктов ВИНИТИ.— М.: ВИНИТИ, 2005.- 72 с. Ил. 41. Библиогр.— 2 назв.— Рус. Деп. в ВИНИТИ 07.11.2005, № 1430-В2005.

4. Шамаев В. Г., Жаров А. В. Электронный реферативный журнал ВИНИТИ по физико-математическим наукам // НТИ. Сер. 1. – 2006.- - № 3.- С. 15–25.

5. Представление элементов данных во внутрисистемном формате ВИНИТИ. Нормативно-техническое предписание НТП ВИНИТИ 10-2004. — М.: ВИНИТИ, 2004. 104 с.

6. Шамаев В. Г., Жаров А. В., Батурина О. Н., Горшков А. Б., Лось Е. К., Максимов И. Н., Старцева О. Б. База данных и Электронная библиотека русскоязычной литературы по физико-математическим наукам.-- М.: ВИНИТИ, 2005.- 84 с.: — Рус.-- Деп. в ВИНИТИ 15.12.2005, № 1682-В2005.

Материал поступил в редакцию 11.04.06.