Хранилище полных текстов для доступа пользователей через электронный каталог поступлений ВИНИТИ

Full-text Storage for User Access through the Electronic Catalog of the VINITI Acquisitions

Сховище повних текстів для доступу користувачів через електронний каталог надходжень ВІНІТІ

Федорец О. В., Фишер А. М., Батюшко А. А.
Всероссийский институт научной и технической информации РАН, Москва, Россия

Oleg Fedorets, Alexander Fisher, and Alexandra Batyushko
Russian Institute of Scientific and Technical Information
of the Russian Academy of Sciences (VINITI), Moscow, Russia

Федорець О. В., Фішер О. М., Батюшко О. А.
Всеросійський інститут наукової та технічної інформації РАН, Москва, Росія

Представлены результаты работы, выполненной в 2005 г., по созданию и наполнению хранилища полных текстов, интегрированного с электронным каталогом поступлений, который доступен через Интернет-сайт ВИНИТИ. Описываются основные принципы построения хранилища, технология обработки и загрузки документов, характеризуется текущее наполнение и перспективы развития.

The authors present the results of the study performed in 2005 and aimed at generating a storage of full-text documents. The storage has been integrated with the electronic catalog of new acquisitions and is accessible via VINITI's Web-site. The basic principles, the processing and downloading technology and the current content and prospects of the storage are described.

Представлено результати роботи, виконаної у 2005 році, зі створення і наповнення сховища повних текстів, інтегрованого з електронним каталогом надходжень, який доступний через Інтернет-сайт ВІНІТІ. Описано основні принципи побудови сховища, технологія обробки і завантаження документів, характеризується поточне наповнення і перспективи розвитку.

1. Электронный каталог поступлений ВИНИТИ

С 1999 г. в ВИНИТИ ведутся работы по созданию электронного каталога поступлений научно-технической литературы в ВИНИТИ. Именно в этом году была введена в эксплуатацию «Автоматизированная система комплектования и регистрации входного потока» (сокращённо АСКР), информационные массивы которой послужили основой для создания Web-каталога [1, 2].

При обработке входного потока информационные массивы АСКР пополняются элементами данных, которые можно извлечь до глубокой содержательной обработки публикаций в отраслевых отделах ВИНИТИ. Таким образом, по сравнению с реферативными базами ВИНИТИ, каталог поступлений предполагает повышенную оперативность отражения библиографической информации. Помимо библиографической информации на монографическом уровне, массивы данных АСКР пополняются описаниями статей, авторскими аннотациями и результатами первичной тематической разметки. Также поддерживается технологическая информация о канале поступления, месте хранения и технологическом маршруте экземпляра НТЛ.

Web-каталог поступлений был модернизирован в 2004 г. и переведён на технологию ASP и Web-сервер Microsoft IIS. При этом пользовательский интерфейс был существенно переделан и приведён к общему стилевому оформлению, принятому для Интернет–сайта ВИНИТИ. В настоящее время каталог доступен по адресу http://catalog.viniti.ru/.

В процессе развития в 2004-2005 гг. увеличилась функциональная нагрузка на каталог. Сегодня он не только объединил спектр разрозненных информационных услуг, в той или иной степени

существовавших до его появления, но и значительно его расширил [3]. Можно выделить следующие достоинства каталога:

- Предоставление полных библиографических указателей с актуальной технологической информацией по всем изданиям, обработанным в институте;
- Заказ копий страниц первоисточника для абонентов ВИНИТИ;
- Показ оглавлений выпусков с авторскими аннотациями для наиболее популярных среди ученых публикаций;
- Поиск по библиографическим полям, кодам тематической разметки и авторским аннотациям;
- Хранение персонифицированных пользовательских настроек, обеспечивающих более эффективную работу с каталогом.

Наряду с показом авторских аннотаций для статей и депонированных научных работ, в каталоге была также реализована возможность показа собственно текстов статей. Изначально это задумывалось главным образом для трёх журналов, издаваемых в ВИНИТИ: «Научно-техническая информация» Серия 1 и 2, и «Международный форум по информации». Тексты этих журналов с полугодовой задержкой были и раньше доступны на сайте ВИНИТИ http://www.viniti.ru/. Метод показа полных текстов был оптимизирован, и сейчас каталог поступлений взял на себя ещё и эту задачу.

В электронном каталоге появился новый поисковый критерий «наличие полных текстов», позволяющий искать только те документы, которые имеют полный текст в базе данных. Если в списке найденных документов присутствуют записи, связанные с полнотекстовой базой, рядом с ними появляется пиктограмма «показать страницы статьи».

2. Форматы документов

Текст в чистом виде можно встретить в художественной литературе или гуманитарных областях, в научно-технических областях знаний мы имеем дело с электронными документами, обильно приправленными формулами, таблицами, рисунками, математическими и прочими специальными символами в различных кодировках. Кроме этого, на читабельность текста влияют элементы форматирования: выделение заголовков, нумерованных списков, колонок, подписей к рисункам и т.д.

В лучшем случае мы имеем дело с документами в символьном представлении в форматах PDF, HTML, RTF. В худшем случае с графическими образами страниц, полученным путём сканирования литературы на бумажном носителе.

В настоящее время подавляющую часть потока электронных документов хранилища составляют графические образы страниц, полученные с участка сканирования ВИНИТИ. Поскольку главная цель разрабатываемой информационной системы – предоставить доступ к страницам через электронный каталог поступлений ВИНИТИ, то выбор форматов невелик. На сегодняшний день любой Web-браузер в состоянии корректно и без установки дополнительных программных компонентов показывать только три графических формата: GIF (Graphics Interchange Format), PNG (Portable Network Graphics) и JPEG (JPEG File Interchange Format).

Формат JPEG рассчитан на цветные фотографии и предусматривает сжатие с потерей качества. В результате для рисунков, схем, математических формул потеря качества приводит к «размыванию» мелких элементов изображений и, таким образом, к искаженному восприятию информации читателем.

Графический PDF мог бы оказаться неплохим вариантом, визуально он выглядит несколько лучше других графических форматов благодаря более совершенному средству просмотра – утилите *Acrobat Reader*. Пользователь может получить не выбранную им страницу, а целиком статью, которая загружается в *Acrobat Reader* на его компьютере. С одной стороны, это удобно. С другой стороны, время отклика системы и трафик через Интернет увеличиваются в несколько раз. Во многих случаях пользователю достаточно посмотреть одну или две страницы статьи и не нужно ждать, пока запустится *Acrobat Reader* и статья целиком будет загружена с Web-сайта на его компьютер. Учитывая, что объем некоторых статей может оказаться весьма значительным (более 10 страниц), этот фактор нельзя не учитывать. Поэтому было принято решение хранить в базе данных и передавать пользователю через Интернет постраничные образы документов.

Выбор для хранения отсканированных черно-белых изображений был сделан в пользу формата GIF и разрешающей способности 300 точек на дюйм. В перспективе, по мере роста вычислительной мощности клиентских компьютеров, можно перейти к формату PNG. Пока что на слабых клиентских компьютерах (с тактовой частотой процессора менее 1 ГГц) время распаковывания браузером сжатого PNG-файла слишком велико.

Однако возможности полнотекстового хранилища не ограничены этими форматами. Как мы покажем ниже, модель данных хранилища рассчитана на первоисточники в различных форматах, в том числе на использование адресов документов (гиперссылок) для обращения к внешним ресурсам.

3. Полнотекстовая база данных

По содержательному признаку полнотекстовые документы могут быть трех видов:

1. Издание в целом.
2. Статья или другая часть издания.
3. Отдельная страница издания.

Во всех трёх случаях документ можно идентифицировать кодом выпуска издания и диапазоном номеров страниц. Поэтому в базу данных можно загрузить файл любого формата, представляющий любой вид документа. При этом необходимо указать формат файла (PDF, GIF, PNG и т.д.) и содержательный признак (издание, статья, страница) в соответствующих полях базы данных, чтобы программное обеспечение Web-каталога могло настраивать пользовательский интерфейс в зависимости от формата первоисточника.

Если документ хранится в графическом виде и доступ к нему планируется осуществлять через Интернет, то наиболее правильным является разбиение его на образы отдельных страниц.

Если документ хранится в символьном виде (обычно в формате PDF), то наиболее часто используемой практикой в Web-каталогах является представление одной статьи из журнала или сборника в виде одного PDF-файла.

При проектировании хранилища важным является вопрос выбор системы для хранения документов. Очевидно, что библиографические записи должны быть в базе данных, а где должны размещаться связанные с ними полнотекстовые документы? Возможны два варианта:

1. Документ хранится в файловой системе, в базе данных содержится адрес файла.
2. Документ хранится в базе данных.

Выбор варианта хранения является ключевым при разработке полнотекстового хранилища, по поводу достоинств и недостатков обоих методов хранения можно найти множество дискуссий на Интернет-форумах, посвящённых разработке баз данных. Этот вопрос носит узкоспециальный характер, поэтому мы сообщим только результат выбора. Был выбран второй вариант – документы загружаются непосредственно в базу данных, а именно, в поля типа *image* СУБД Microsoft SQL Server, предназначенные для хранения больших двоичных объектов.

Вместо документа в полнотекстовой базе может храниться адрес документа на внешнем полнотекстовом ресурсе. В качестве внешнего ресурса может выступать любой Web-сервер. С помощью словаря «дескрипторы ресурсов» можно описать любой внешний ресурс, указав его адрес в локальной или глобальной сети. Таким образом, полнотекстовая база данных выступает в качестве шлюза, через который электронный каталог поступлений может обращаться к любому доступному полнотекстовому ресурсу, а не только к документам, хранящимся внутри системы.

4. Технология загрузки полных текстов

Технология подготовки и загрузки полных текстов в хранилище должна обеспечивать решение трёх основных задач:

1. Установление связи полнотекстового документа с библиографическим каталогом.
2. Обеспечение точности при разметке страниц оглавлений журналов и сборников.
3. Обеспечение качества сканирования.

Связь между библиографическими и полнотекстовыми записями в базе данных устанавливается на этапе загрузки благодаря соглашениям по наименованию папок и файлов на файловом сервере, с которого происходит загрузка. К моменту сканирования каждый первоисточник имеет

уникальный идентификатор в базе данных – числовой штрих-код и связанный с ним идентификатор библиографической записи. Штрих-код приклеен на обложку первоисточника. В названии папки для хранения результатов сканирования также присутствует уникальный идентификатор библиографической записи. Оператор считывает штрих-код ручным сканером, и результаты сканирования (графические файлы) автоматически копируются в нужную папку.

На участке сканирования создаются не только графические файлы образов страниц, но также и текстовый файл оглавления. Файл оглавления получается путём оптического распознавания страниц оглавления программой *FineReader Professional 6.0* с последующим ручным редактированием и разметкой полей. Файл оглавления размещается в той же папке, что и образы страниц первоисточника, и затем загружается в массив описаний статей.

Таким образом, программные модули загрузки оглавлений и образов страниц «узнают» идентификатор библиографической записи из названия папки на файловом сервере, что позволяет им устанавливать необходимые связи в базе данных.

Оператор контролирует качество сканирования и номенклатурное количество страниц. Имя файла должно совпадать с номером страницы в первоисточнике.

Правильность разметки страниц оглавлений контролировать труднее, но такие ошибки возникают редко и они не столь критичны, так как при наличии полнотекстового документа в базе, корректно связанного с библиографической записью, исправить номера страниц можно без повторного сканирования первоисточника. Хотя неправильные номера страниц в описаниях статей могут доставить пользователю неудобства. Действительно, щёлкая мышкой пиктограмму «полный текст» рядом с названием статьи, он ожидает увидеть именно эту статью, а не другую.

Автоматизированный контроль нарушений ссылочной целостности возможно реализовать только для документов в символьных форматах, из которых можно извлечь текст и сравнить первую страницу, обычно содержащую название и список авторов, с библиографической записью. Однако на сегодняшний день в хранилище загружаются в основном графические образы страниц, для которых реализовать такой контроль не представляется возможным. Поэтому приходится полагаться на точность операторов участка сканирования и загрузки данных.

5. Текущее наполнение хранилища и перспективы развития

В настоящее время в полнотекстовую базу данных загружены выпуски следующих журналов, издаваемых в ВИНИТИ:

- Научно-техническая информация, сер. 1. – с 1997 по 2005 г.
- Научно-техническая информация, сер. 2 – с 1997 по 2005 г.
- Международный форум по информации – с 2004 по 2005 г.
- Международный форум по информации и документации – с 2000 по 2004 г.

Свободный доступ к перечисленным журналам открывается через полгода после выхода в свет.

В 2005 году в порядке эксперимента были отсканированы и загружены в полнотекстовую базу более 2 тыс. депонированных работ за 2004 год, которые заняли в базе данных 7,4 Гбайт дискового пространства.

В 2006 г. в ВИНИТИ разрабатывается программно-технологическое обеспечение для формирования электронной полнотекстовой библиотеки депонированных научных работ. Библиографические описания и аннотации депонированных работ вводятся в производственном режиме по мере поступления. Сами депонированные работы можно сканировать, а можно получать от авторов как в печатном, так и в электронном виде, что определяет технологию формирования полных текстов. По соглашению с авторами можно устанавливать степень доступность электронной версии депонированной работы через Интернет. Эти вопросы должны быть прописаны в новой редакции «Положения о депонировании». Для обеспечения сохранности ретроспективного фонда депонированных научных работ планируется отсканировать и загрузить его в полнотекстовую базу данных. В каталоге поступлений по состоянию на март 2006 г. присутствует более 30 тыс. описаний депонированных в ВИНИТИ работ, из них полные тексты пока загружены примерно для 7% документов.

Другой возможный источник наполнения полнотекстовой базы данных – издания ВИНИТИ обзорного характера, проблемно-ориентированные реферативные, информационные сборники и бюллетени. Их можно было бы с задержкой выставлять в открытый доступ через каталог поступлений, а также открыть доступ подписчиков к свежим номерам. Решение этого вопроса находится в компетенции редакционной коллегии и руководства ВИНИТИ, технические возможности уже имеются. Оригинал-макеты обзорных изданий, подготовленные в ВИНИТИ, доступны в электронном виде, поэтому эти печатные издания не придётся сканировать.

Литература

1. Егоров В. С., Шапкин А. В. Каталог поступлений НТЛ как источник новых форм обслуживания потребителей информационных ресурсов ВИНИТИ // НТИ-2002. Информационное общество. Интеллектуальная обработка информации. Информационные технологии. Материалы 6-й международной конференции (16-18 октября 2002 г.). – М.: ВИНИТИ, 2002. – С. 130-132.
2. Шапкин А. В. Автоматизированная система комплектования и регистрации входного потока ВИНИТИ. Ч. 1, 2 // Научно-техническая информация. Сер. 1.– 2005. – №3, – С. 8-19, № 4. – С. 16-31.
3. Фишер А.М. Электронный каталог поступлений: новый информационный ресурс ВИНИТИ // Научно-техническая информация. Сер. 1. – 2006. – №2, – С. 17-26