ПРОЕКТИРОВАНИЕ, СОЗДАНИЕ И НАПОЛНЕНИЕ ЭЛЕКТРОННОГО АРХИВА

С.В.Антюфеев, А.Г.Марчук, А.Н.Немов, К.В.Федоров, В.Э.Филиппов, М.Я.Филиппова, Н.А.Черемных
Институт систем информатики им. А.П.Ершова СО РАН, Новосибирск
630090, пр-т Академика Лаврентьева, 6
{mag, cher}@iis.nsk.su

DESIGN AND IMPLEMENTATION OF ELECTRONIC ARCHIVE OF DOCUMENTS

S. Antyoufeev, A. Marchuk, A. Nemov, K. Fedorov, V. Filippov, M. Filippova, N. Cheremnykh

A.P.Ershov Institute of Informatics Systems, pr. Akad. Lavrentiev, 6, Novosibirsk 630090, Russia
{mag, cher}@iis.nsk.su

The paper describes the technology of electronic (digital) archive design and development. It is intended for making Web-published versions of existent paper and other documents of heterogeneous information. Implemented programs and interfaces allow archive professionals to build and support document database and provide end-users with a possibility of distant access to the archive information by means of simple and efficient interface. The system consists of several modules, the kernel is constructed using client-server architecture. The system has been used for creating the electronic archive of academician Andrei Ershov (http://ershov.ras.ru).

This paper describes the technology and system of electronic (digital) archive development. The system is intended for making Web-published versions of existent paper and other documents of heterogeneous information. Implemented programs and interfaces allow archive professionals to build and support document database and end users – to have distant access to the archive information with simple, but effective interface. System consists of several modules, the kernel is constructed using client-server architecture. The system was used for building of the electronic archive of academician Andrei Ershov (http://ershov.ras.ru) which is available for users from the beginning of 2001.

Академик А.П.Ершов и его архив

Роль академика Андрея Петровича Ершова в становлении и развитии системного программирования в нашей стране трудно переоценить. Достаточно сказать, что под его руководством и при непосредственном участии был создан первый транслятор с алгоритмического языка АЛЬФА,

близкого к Алголу 60. А.П.Ершов первым начал эксперименты по обучению школьников программированию, ему, в частности, принадлежит известный тезис "программирование – вторая грамотность".

После безвременной кончины А.П.Ершова остался уникальный архив. Это более 500 папок с документами, отражающими жизненный путь академика и историю развития информатики в России. Они были собраны им самим и систематизированы как хронологически, так и тематически. В Институте систем информатики им. А.П.Ершова, с учетом имеющегося опыта создания информационных и библиографических систем [1] и при финансовой поддержке научно-исследовательского подразделения фирмы Microsoft предпринята попытка сделать этот архив общедоступным, ввести его в широкий научный оборот с помощью Интернета.

Постановка задачи, требования к электронному архиву

Перевод "бумажных" архивов в электронную форму представляется важной и насущной задачей. Этим достигается существенно бóльшая доступность документов архива для исследователей всего мира, сохранность исходных документов, появляется возможность разнообразной структуризации содержащейся информации и построения поисковых систем. Важным видится также группирование аналитической деятельности исследователей и связывание документов с их профессиональной интерпретацией.

Мы будем рассматривать архив как набор документов, представленных в разных формах. Типичным документом является текстовый материал (рукопись, письмо, записка и т.д.), состоящий из одной или более страниц, написанный от руки или напечатанный; документ может содержать подписи, пометки и др. Для исследовательской работы важны и форма (визуальное изображение материала), и содержание документа. Документы могут иметь и другой, например, вещественный характер (медали, подарки и т.д.), содержать мультимедийную информацию (фотографии, звуковые и видео записи), но данный класс документов (экспонатов) в настоящей статье рассматриваться не будет.

Первично, архив – это система идентификации документов и их частей, сами документы и каталог документов или единиц хранения. Исходное состояние типичного необработанного архива – только набор документов. Нашей работе над архивом академика А.П.Ершова существенно помогло то, что Андрей Петрович был весьма системным и аккуратным человеком и документы архива были им упорядочены и сгруппированы по некоторым принципам. Эти принципы нашли отражение в тематической структуре каталога электронного архива, что позволило провести начальную группировку, упростило решение идентификационных задач.

Предложенный и реализованный подход к созданию электронного архива в основном состоит в следующем:

- все документы должны быть отсканированы с качеством, удовлетворяющим основные исследовательские потребности;
- на каждый документ должна быть составлена (электронная) карточка, отражающая метаинформацию и основные логические связи формируемой базы данных;
- каталог документов имеет иерархический вид в соответствии с видом документа и принадлежностью его к тематической группе или подгруппе;
- каждый документ также сохраняет ссылку на соответствующие единицы хранения.

В случае больших и очень больших архивов следует обратить особое внимание на поддержку поисковых действий пользователя. Понятно, что для работы с десятками и сотнями тысяч документов будет малоэффективным использование только иерархической структуры каталога без дополнительных, например, ассоциативных средств навигации. Возможность выполнять выборки групп документов по заданным критериям и сбор статистики также весьма важна.

К электронному архиву были предъявлены следующие пользовательские требования:

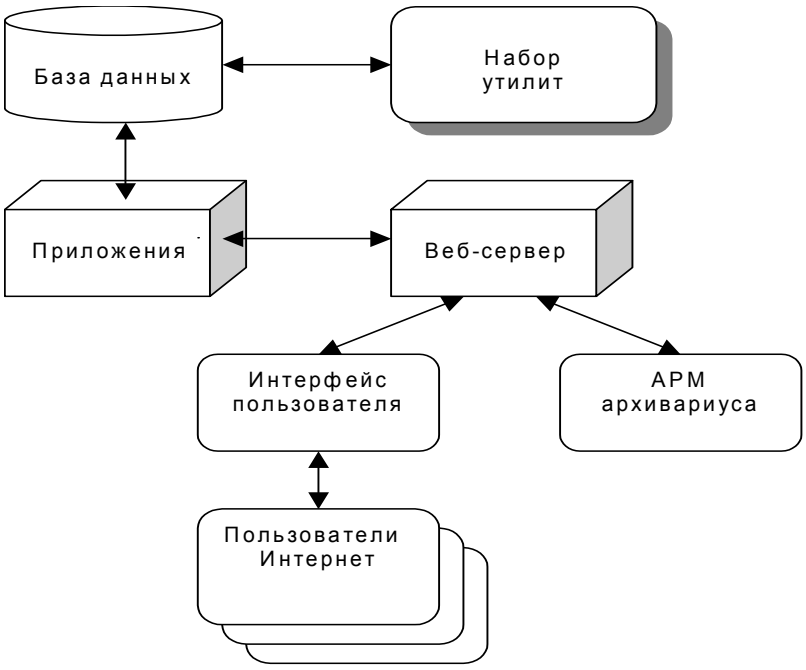
- работа пользователя осуществляется через стандартный Web-браузер;
- просмотр документов, навигация во множестве документов, поиск нужных документов являются главными действиями пользователей;
- необходимо реализовать различные, соответствующие характеру работы пользователя, навигационные стратегии;
- требуются высокая эргономичность при работе с документами и минимизация трафика;
- система должна быть многоязыковой, т.е. обеспечивать возможность работы с архивом как на русском, так и на английском языке; кроме того, наиболее интересные документы должны быть переведены либо на русский, либо на английский язык, соответственно;
- электронный архив должен поддерживать систему комментариев и интерпретаций и предоставлять к ним доступ для пользователей.

Архитектура электронного архива

В рамках работ по созданию электронного архива академика А.П.Ершова была разработана общая концепция создания архива и его архитектура. Модель данных электронного архива, поддерживает различные представления документов (текстовое, графическое, гипертекстовое, аннотационное). Разработана технология и инструментальные средства как для создания (наполнения, редактирования и актуализации данных) и дальнейшей работы над материалами архива, так и для организации работы по

наполнению архива информацией. Для этой работы характерна распределенность во времени и пространстве, и выполняется она достаточно большим коллективом.

Система построена с использованием трехуровневой архитектуры клиент-сервер.



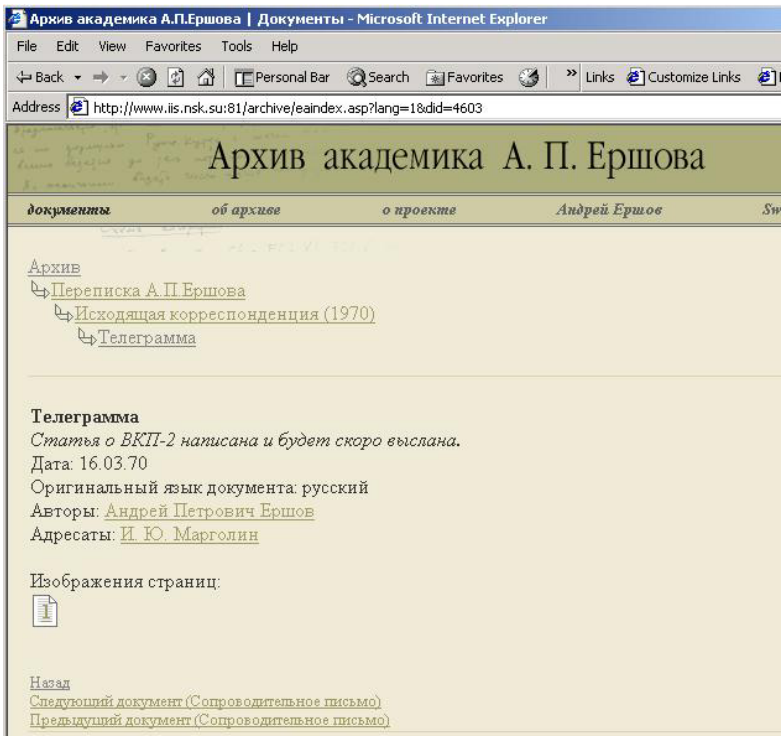
Архитектура электронного архива

Реализация

Основой электронного архива является база данных, в которую входит хранилище данных, в основном это файлы изображений страниц, и реляционная база данных документов. Кроме основной таблицы документов, имеются сопряженные таблицы людей, организаций, стран, городов, языков. Предложенная модель данных, физически состоящая из 54 таблиц, позволяет эффективно работать со всей совокупностью документов, выполнять поисковые и навигационные действия.

Основным пользовательским интерфейсом предоставляется доступ к иерархии документов, сгруппированных по видо-тематическому признаку. На верхнем уровне иерархии выделяются такие разделы, как "Личное дело", "Переписка А.П.Ершова", "Международные конференции", "Всесоюзные конференции", "ИФИП", "Программные проекты", "Теоретические исследования", "Зарубежные командировки", "Школьная информатика", и др. Далее идет углубление по иерархии в детали разделов. Например, в разделе "Всесоюзные конференции" – "ВКП-2" – "Отчет о ВКП-2" содержится 6 документов. Конкретные документы классифицированы по виду, например, статья, письмо, телеграмма. Документ первично описывается

видом, заголовком, датой, автором, адресатом и, по необходимости, некоторыми другими полями. Даны иконки изображения страниц, нажав на которые, можно последовательно просмотреть страницы документа в оригинальном виде.



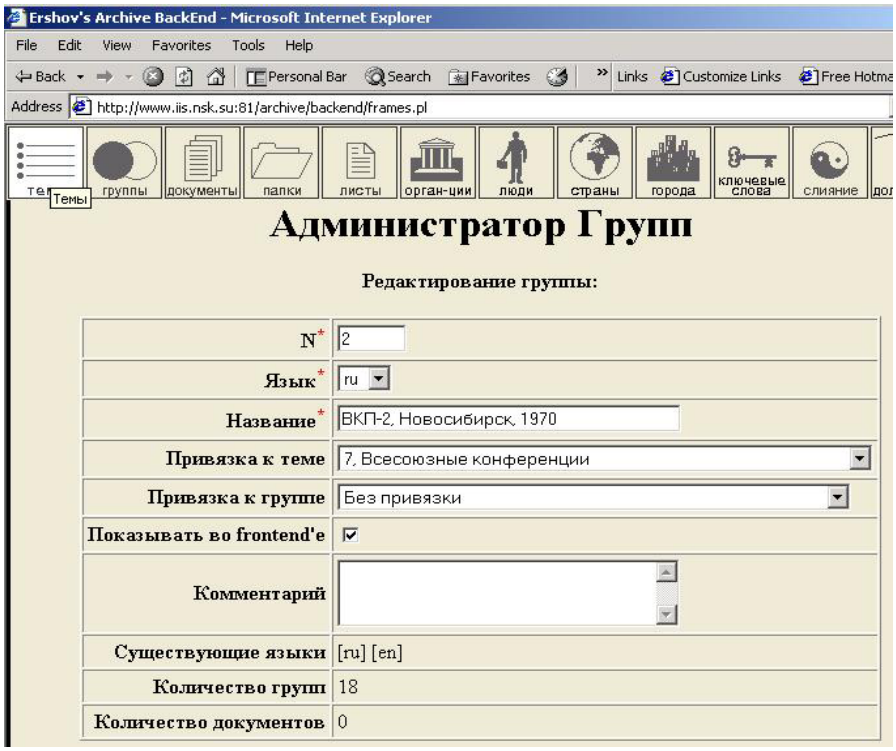
Фрагмент пользовательского вида на документ из архива

Важными элементами карточки документа, отображаемыми в интерфейсе или используемыми для навигационных и поисковых действий, являются упоминаемые в документе фамилии и названия организаций. По представленной гиперссылке можно посмотреть имеющуюся информацию о человеке или организации, найти все документы, в которых они присутствуют, произвести другие выбирающие и фильтрующие действия.

Заполнение карточек на документы является одной из самых трудоемких при формировании информационного наполнения действий. Для работы информационного администратора имеется совокупность специальных интерфейсов, выполненных также в Web-овской технологии. Несмотря на скудость стандартных HTML форм, удалось разработать достаточно эффективный и удобный интерфейс занесения новой информации и редактирования имеющейся.

Аналогичные интерфейсы имеются для редактирования карточек документов, информации о людях, организациях и т.д.

Созданная технология производства электронных архивов является многоязыковой. Языками интерфейса пользователя архива А.П.Ершова являются русский и английский. В письмах же встречаются документы и на многих других языках, что потребовало соответствующей поддержки.



Интерфейс информационного администратора по работе с группами

Отдельные части программного обеспечения (утилиты) разрабатывались для оптимизации ввода и структуризации данных. Была разработана специальная программа сканирования документов. Суть проблемы в том, что использование стандартных сканирующих программ требует многочисленных манипуляций, проводимых мышью и клавиатурой. Кроме того, что работа сканера замедляется на проведение этих манипуляций, оператор быстро устает и начинает ошибаться. Созданная программа ограничивает набор действий минимально необходимым: заправить документ в сканер, нажать клавишу начала сканирования, удалить документ из сканера. Вообще процессу сканирования и работы с полученными изображениями было уделено достаточно внимания. Потребовалось экспериментально определить параметры сканирования, такие как разрешающая способность, яркость и контрастность. Оказалось, что черно-белые изображения теряют слишком много информации, поэтому был выбран режим градаций серого.

Ряд документов архива потребовал создания специальных средств ввода для порождения структурированной информации. Например, в архиве имеется около 700 анкет участников Второй всесоюзной конференции по программированию (ВКП-2). Информация, содержащаяся в анкетах, позволяет порождать интересные аналитические результаты. Простыми средствами (XML, XSLT, ASP) был создан интерфейс ввода данных, позво-

ливший экономно ввести такую специализированную базу данных. В итоге, в электронном архиве теперь имеются и отсканированные образцы анкет, и полная база данных, пригодная для анализа.

Об использованной технологии.

В качестве операционной системы сервера используется Microsoft Windows NT 4.0, SQL-server – Microsoft SQL Server 7.0 [2], cgi-язык на front-end – VBS. На рабочих станциях могут использоваться любые из операционных систем Microsoft – от Microsoft Windows 98 до Microsoft Windows 2000 professional. При написании утилит использовалась MS Visual Studio 6.0 (MS Visual C++ 6.0), а также библиотека MSDN 2000 [3].

В настоящее время полностью разобраны, отсканированы и выставлены на сайте документы, относящиеся к темам “Международный симпозиум в Ургенче” (628 документов), “Всесоюзная конференция по программированию ВКП-2” (1178 документов), ведется обработка документов, относящихся ко многим другим темам. Обработана и доступна на сайте значительная часть переписки А.П.Ершова (3258 документов). Заметим, что документ может быть представлен несколькими страницами, а некоторые документы (например, дневник академика) содержат более 100 страниц. Все документы снабжены комментариями, часть русскоязычных документов переведена на английский язык.

Перспективы

Созданная технология создания электронных архивов выглядит разумной и достаточно универсальной, позволяет получать доброкачественный результат и обеспечивать живучесть и развиваемость как системы, так и архива. Основным недостатком технологии видится относительно высокая стоимость выполнения полного комплекса работ по формированию информационного наполнения. Реально это означает, что архив, состоящий из нескольких десятков тысяч документов, можно полностью обработать силами нескольких человек лишь за два-три года. Эта проблема имеет объективный характер и связана с тем, что основную работу по ведению карточек на документы выполняют информационные администраторы (операторы), анализируя суть каждого из документов и выделяя из него информацию для карточки, а потом внося эту информацию в базу данных.

На дальнейших этапах разработки технологии предполагается решить следующие задачи:

- ускорить сканирование документов путем применения других технологий, например, фотосканирования;
- улучшить обработку изображений, обеспечить возможности динамического балансирования между качеством и объемом изображений;

- применить программы распознавания текстов;
- применить программы автоматического перевода текстов;
- создать и использовать программы автоматической экстракции данных для карточек на документы;
- улучшить структуризацию данных, дать пользователям средства описания модели данных;
- разработать и реализовать адаптер Z39.50;
- разработать систему для поддержки малых архивов, реализуемых на CD-ROM.

В заключение отметим, что созданная программная система обеспечивает устойчивое функционирование и непрерывное пополнение электронного архива в течение почти полутора лет [4]. Она оказалась удобным и надежным инструментом, пригодным для создания и поддержания работы электронных архивов, библиотек и музеев.

Благодарности

Проект выполняется Институтом систем информатики им. А.П.Ершова СО РАН при поддержке фирм Microsoft Research (Великобритания) и xTech (Новосибирск). Выражаем глубокую искреннюю благодарность всем участникам и спонсорам проекта.

Литература

1. Елепов Б.С., Марчук А.Г., Бобров Л.К., Константинов В.И. Новые информационно-библиотечные технологии // Информационные технологии и вычислительные системы - 1997, N 2. С. 83-87.
2. Reference book: "SQL Server Book online". Microsoft Corporation. 1998.
3. Reference book: "October 2000 Release of the MSDN Library", Microsoft Corporation. 2000.
4. Электронный архив академика А.П.Ершова // московский и новосибирский вебсайты: http://ershov.ras.ru, http://www.iis.nsk.su:81