ПРИНЦИПЫ ПОСТРОЕНИЯ ЦИФРОВОГО ФОТОАРХИВА РОССИЙСКОЙ АКАДЕМИИ НАУК

В.Н. Карнаухов, Н.А. Кузнецов, Н.С. Мерзляков, Л.И. Рубанов
Институт проблем передачи информации РАН, Москва

В 1999 г. российская и мировая общественность отметила 275-летие Российской академии наук (РАН), основанной по указанию Петра I в феврале 1724 г. За годы существования академии она собрала уникальные документы и располагает разнообразными богатейшими архивами, являющимися неотъемлемой частью мирового культурного наследия и представляющими широкий интерес. Они хранятся в центральном Архиве РАН, в различных академических музеях, научных институтах, других организациях Академии, а также частных собраниях и включают в себя художественные портреты и фотографии деятелей науки, их рукописи и прочие материалы, относящиеся к деятельности Российской академии и истории мировой и российской науки.

Уникальность разнообразных печатных, рукописных и графических документов, а также забота об их сохранности не позволяют обеспечить к ним широкий доступ. Более того, многие материалы уже сейчас требуют срочной реставрации и консервации, иначе они будут безвозвратно утрачены. Это в первую очередь относится к фотодокументам – стеклянным и пленочным негативам, кинолентам, фотографиям, срок жизни которых даже в идеальных условиях хранения измеряется десятилетиями из-за необратимых физико-химических изменений в структуре применяемых фотографических материалов. Положение усугубляется тем, что большинство материалов в настоящее время известны и доступны только узкому кругу архивных работников и специалистов, публикуются выборочно и в крайне малом объеме. Поиск необходимых документов также затруднен отсутствием централизованных каталогов и указателей; особенно это касается архивов фотодокументов и подобных графических материалов, индексация которых сама по себе представляет достаточно сложную задачу. Практикуемые в архивах способы обычно основываются на составлении текстовых описаний (аннотаций), ведении журналов учета, картотек и тому подобных бумажных носителей, которые не являются неотъемлемой частью фотодокументов, а хранятся отдельно и лишь сопоставлены им с помощью системы шифров или нумерации, что потенциально угрожает целостности информации.

Таким образом, на передний план выдвигаются следующие **задачи**, характерные, по-видимому, не только для архивов РАН, но и для других крупных архивов изображений:

а) реставрация и надежное сохранение фотодокументов и изобразительных материалов;

б) разработка и создание методов и средств индексации и поиска архивной информации;

в) реализация многоуровневой системы широкого доступа к содержимому архивов.

Современные информационные технологии позволяют с высоким качеством решить большинство из перечисленных задач при умеренных затратах, и тем самым изменить к лучшему ситуацию, сложившуюся в рассматриваемой сфере. Очевидное решение состоит в постепенном переходе к цифровым архивам, в которых архивные документы представлены в цифровой форме, в которой могут сохраняться бессрочно. Работа по наполнению цифрового архива складывается из следующих общих этапов. Вначале производится массовый высококачественный цифровой ввод графической и текстовой архивной информации с запоминанием ее на оперативных носителях. Затем по мере необходимости и строго в индивидуальном порядке выполняется цифровая реставрация введенной информации, и окончательное размещение ее в архивной базе данных в соответствии с выбранной формой представления. В зависимости от размеров архива он может базироваться на большой ЭВМ или локальной сети малых машин. Дальнейшее использование этого **первичного цифрового архива** развивается по следующим направлениям.

Во-первых, создается необходимое количество копий для повышения надежности хранения и размещения в регионах. В зависимости от характера фонда и с учетом требований сохранности интеллектуальной собственности, к первичным архивам предоставляется ограниченный доступ или же они могут полностью открываться для контролируемого доступа через оборудованные локальные рабочие места.

Во-вторых, на основании первичного цифрового архива формируется система **вторичных архивов**, предназначенных для широкого круга пользователей. Информация представляется в них не в самом полном объеме или с пониженным качеством (в частности, изображения имеют меньшее разрешение), что препятствует несанкционированному коммерческому использованию и в тоже время снижает затраты на хранение. Такие архивы могут полностью размещаться на одном или нескольких CD-ROM или DVD-ROM, выпускаться массовыми тиражами и распространяться по невысокой цене или бесплатно (например, передаваться в публичные библиотеки). Другой очевидной формой размещения вторичного архива может выступать сетевой узел (сайт) Интернет, что открывает глобальный доступ к содержимому архива.

По существу, подобный вторичный архив в Интернет или на компакт-диске играет роль иллюстрированного **каталога** содержимого первичного архива, который позволяет провести детальный поиск необходимой информации, а также знакомство с ней (с образовательными и культурными целями). Он также позволяет провести необходимый предварительный отбор информации и сформировать конкретный запрос в первичный цифровой архив на получение исходных данных, которые предполагается

использовать в научно-исследовательских или коммерческих целях. Важно отметить, что создание вторичных цифровых архивов является не предметом специальной разработки, а результатом выполнения раз спроектированной **автоматической** или автоматизированной процедуры. Это позволяет оперативно или на периодической основе выпускать новые редакции таких архивов по мере наполнения первичного архива (которое может длиться годами) и в дальнейшем при новых поступлениях.

Таким образом, в обозримом будущем основная задача по сохранению архивов и включению их в орбиту мирового культурного наследия видится в переводе на цифровые носители, а основным каналом ознакомления и научной работы с этими материалами будет представление архивной информации в специализированных первичных базах данных с контролируемым доступом и автоматически формируемых сжатых вторичных архивах, размещаемых в Интернет и тиражируемых на компакт-дисках для автономной работы с ними при отсутствии соединения с глобальной сетью. Такой триединый (база–сайт–диск) подход позволяет в конечном итоге ускорить разработки, повысить их качество и избежать ненужного дублирования в работе.

Такова общая методика, которая была разработана и успешно применена в Институте проблем передачи информации РАН (ИППИ РАН) в ходе построения тексто-графического цифрового фотоархива Российской академии наук – одной из первых работ в данном направлении. В ИППИ РАН накоплен более чем 25-летний опыт в области обработки изображений и цифровой оптики, реализованы многочисленные масштабные проекты космической, медицинской и культурной тематики. В течение ряда последних лет ведется работа по сохранению отечественного и мирового культурного наследия, сосредоточенного в больших собраниях изображений, в числе которых Рукописная картотека древнерусского словаря, фотоархив ЛАФОКИ РАН, Международная база данных водяных знаков в западноевропейских древних рукописях и актах, и др. [1-6].

С самого начала работы стало ясно, что на полный перевод имеющихся архивов Академии наук в цифровую форму с минимальной реставрацией требуются годы, поэтому был выделен круг документов и материалов, подлежащих обработке в первую очередь – это документы **портретного фонда**, где хранятся негативы, фотографии, графические и живописные портреты деятелей российской и мировой науки. С одной стороны, интерес к этому фонду был всегда, он усилился в связи с приближением 275-летнего юбилея Российской академии наук, а сейчас, после юбилея он стал просто постоянным. С другой стороны, сохранность фотоматериалов, многие из которых относятся к началу века и даже ранее, внушала и внушает опасения, что мы можем их потерять навсегда.

Реляционная модель данных была принята для создания первичного цифрового архива. Такое решение обеспечивает необходимые возможности наращивания структуры по мере постепенного охвата архивных фондов, обеспечивает функционально полные средства поиска информации и облегчает перенос первичного архива на новые программно-аппаратные платформы по мере прогресса вычислительной техники. Хотя связанная с фотодокументами текстовая информация (описания, аннотации, биографии) слабо структурирована и типична скорее для полнотекстовых баз данных, были предприняты усилия по выделению ключевых фрагментов этих текстов и представлении их в табличной структуре записей базы данных, чтобы можно было применять дескрипторный поиск, а не только полнотекстовый. Отдельную задачу составляло сведение в единую базу данных чрезвычайно разнородной информации, охватывающей почти трехвековой исторический период, за который радикально менялись не только наименование, структура и задачи Академии, но и само общественно-политическое устройство государства. Приведем только один пример: точная датировка биографических событий известна далеко не всегда, так что пришлось разработать специальные способы совместного представления в базе данных и использования (скажем, при сортировке в хронологическом порядке) одновременно точных дат и датировок типа "не позднее середины 1768 г.".

Для хранения изображений в первичном цифровом архиве использованы стандартные графические форматы файлов, обеспечивающие умеренное сжатие без потери первичной информации. Файлы изображений, оцифрованных с высоким разрешением, увязаны в общую структуру реляционной базы данных с помощью специально разработанных интерфейсных модулей, которые обеспечивают вывод изображений на экран и печать в рамках обычных функций манипулирования данными. На сегодняшний день первичный архив ведется в ИППИ РАН и поддерживается на развернутой в институте локальной сети IBM-совместимых компьютеров, работающих в среде Windows. В качестве СУБД в настоящее время используется Access 97, в перспективе рассматривается возможность конвертирования созданной базы данных в более производительную и надежную среду. Необходимое для работы с первичным цифровым архивом прикладное программное обеспечение разработано с использованием встроенных программных средств, в том числе языка VBA.

Помимо своего основного назначения – выступать в роли первичного тексто-графического и фотоархива – разработанная база данных несет ряд служебных функций, главная из которых состоит в управлении вводом и обработкой исходной информации. Эта деятельность ведется параллельно на нескольких рабочих местах и в многосменном режиме значительным коллективом сотрудников, что предъявляет высокие требования к учету и контролю целостности вводимой информации. Благодаря упомянутым служебным функциям базы данных, фиксирующей в режиме реального времени все этапы ввода и обработки изображений и текстовой информации, этот ответственный процесс удалось держать под эффективным контролем.

В настоящее время в описываемую уникальную базу данных уже внесены краткие биографические данные всех членов Академии с 1724 г. (около 5 тыс. человек), их портреты (иногда – несколько штук), сведения о современном персональном составе и организационной структуре РАН, о научных организациях Академии и издаваемых ими научных журналах, обо всех присуждаемых Академией наградах и ученых, их удостоенных, очерк истории Академии и вклада российских ученых в мировую науку, картографический материал и многие другие сведения. В ходе выполнения работы были проведены компьютерный ввод и цифровая реставрация свыше 6,5 тысяч черно-белых и цветных негативов, фотоснимков, гравюр, живописных портретов и других графических материалов.

Описанная база данных внесена в Государственный реестр баз данных, частичный доступ к ней уже сейчас предоставлен через специально разработанный сетевой узел Интернет в ИППИ РАН. Тем не менее, несмотря на большую заинтересованность в получении этой впервые собранной воедино уникальной информации, неограниченный доступ к базе данных предоставлять не планируется, поскольку хранимые в первичном архиве высококачественные цифровые изображения уникальны и имеют большую коммерческую ценность, а часть текстовой информации носит приватный характер. Указанное противоречие как раз и разрешается путем создания широкодоступных вторичных архивов, о чем выше уже шла речь. Применительно к описываемому проекту, вторичный архив был реализован в обеих упоминавшихся формах – сетевого узла Интернет и CD-ROM.

Выпущенный ограниченным тиражом в ИППИ РАН в рамках проекта CD-ROM "Российская Академия Наук. 1724-1999 г.г." демонстрировался на выставке во время юбилейной сессии РАН 2-4 июня 1999 г., а также на выставке EVA'99, проходившей одновременно с конференцией EVA'99 в Государственной Третьяковской галерее в октябре 1999 г. Данный компакт-диск фактически содержит два продукта, построенных на единой вторичной информационной базе: Web-CD для просмотра с помощью Интернет-браузера (на любой платформе, позволяющей читать диски в формате ISO 9660), и специализированная оболочка для работы с базой данных в среде Windows.

Достоинством первого варианта является программно-аппаратная независимость и низкая ресурсоемкость, позволяющая эффективно использовать диск на недорогих персональных компьютерах. Такая ориентация и работа в автономном режиме (без сервера) не позволила в полной мере применить современные стандарты представления страниц в Интернет, включая активные страницы и средства языка JAVA. Это, разумеется, ограничило возможности активного поиска информации, но в целом удалось организовать достаточно эффективный и удобный интерфейс (рис. 1). Данный Web-CD содержит свыше 30 тыс. Web-страниц, причем подавляющее большинство их было сформировано автоматически на основании первичного архива, для чего в рамках СУБД Access на языке VBA были разработаны особые программы, генерирующие HTML-файлы. В необходимых случаях эти программы с учетом правил языка формируют связный текст из информации, разнесенной по полям базы данных. Отметим здесь, что вся текстовая информация в базе данных, на диске и сетевом узле Интернет, включая и язык интерфейсов пользователя, представлены на двух языках – русском и английском, причем для русского языка поддерживаются все основные кодировки.

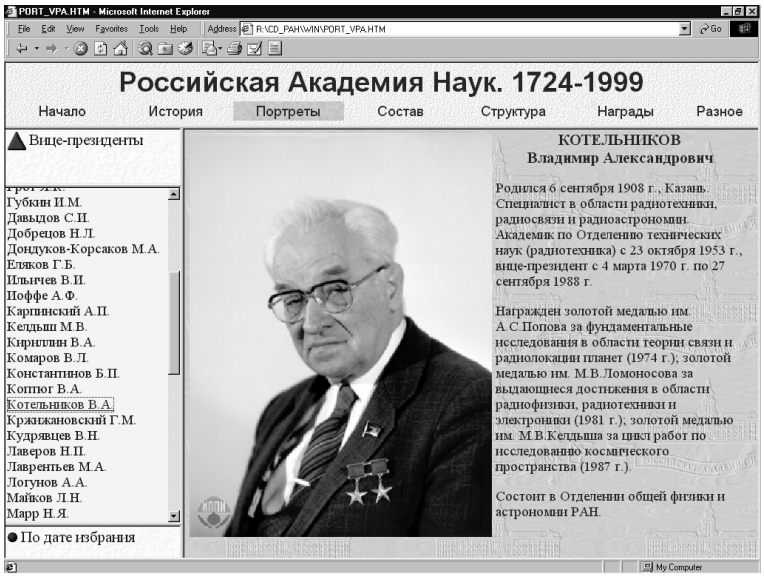

Рис. 1. Пример экранной формы в режиме Web-CD

Второй продукт на компакт-диске представляет собой специально разработанное прикладное программное обеспечение, призванное дать пользователю простой и удобный в работе инструмент непосредственного доступа к содержимому вторичной информационной базы – портретам и сопровождающим их текстовым данным. Это программное обеспечение реализовано для платформы PC и мо-

жет работать только под управлением операционных систем Windows 95/98/NT-4.0 или выше, но зато в нем более широко использованы средства мультимедиа и имеются более развитые возможности интерактивного поиска и фильтрации данных и даже возможность выбора музыкального сопровождения от классического до джазового. Таким образом, графический интерфейс пользователя программы построен в традиционном стиле Windows с главным меню, представленным на рис. 2.

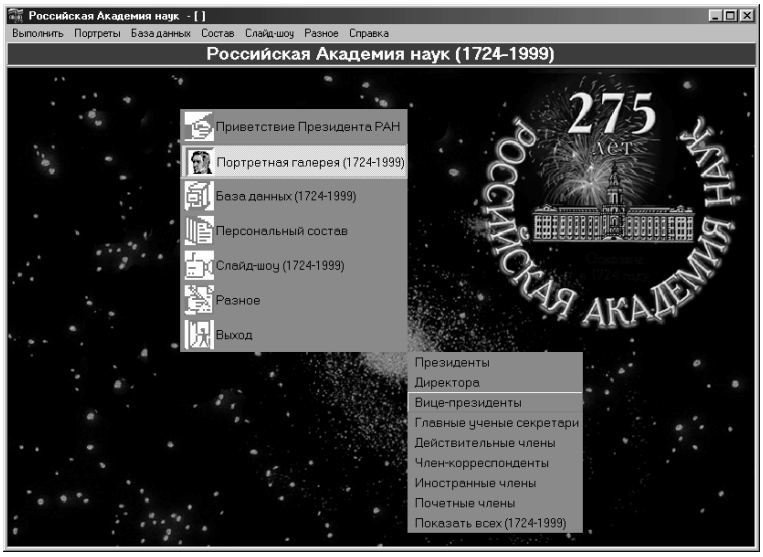

Рис. 2: Графический интерфейс пользователя с главным меню в работе

Выбранные на рис. 2 пункты главного меню запускают показ портретов и кратких биографических данных всех президентов РАН в режиме слайд-шоу с регулируемой периодичностью смены кадров и музыкальным сопровождением в виде отрывков классической, джазовой и популярной музыки в зависимости от вкусов и интеллектуального уровня пользователя. Что касается способов представления архивной информации, то в целом они типичны для прикладного программирования баз данных. Так, на рис. 3 изображены графические формы, используемые при работе пользователя с цифровым портретным архивом на диске.

Рис. 3. Графические формы для работы с цифровым архивом портретов

Каждая строка этой формы содержит фамилию, полное имя и уменьшенный портрет члена Академии, к которому относится текущая запись выбранной таблицы базы данных. Если нажать кнопку "Биография" ("Biography"), то выводится новая графическая форма с краткими биографическими сведениями и увеличенным изображением (рис. 4).

Оба варианта пользовательского интерфейса работают с одной и той же вторичной архивной базой данных, в которой архив собственно портретов занимает основной объем CD-ROM, несмотря на то, что отреставрированные цифровые изображения подверглись значительному понижению разрешения и хранятся во вторичном архиве в формате JPEG с высокой степенью сжатия информации. В итоге получаемые из архивной базы портреты при просмотре имеют довольно высокое визуальное качество, хотя и недостаточное для полиграфических целей (для этого необходимы изображения, которые хранятся в первичном цифровом архиве).

Рис. 4. Пример отображения увеличенного портрета

Другая реализация вторичного тексто-графического и фотоархива построена в виде сетевого узла Интернет (http://hp.iitp.ru). За некоторыми несущественными отличиями, этот сайт повторяет описанный выше Web-CD.

Несмотря на то, что работа по созданию тексто-графического и фотоархива РАН еще продолжается, построенные к настоящему времени разделы цифрового фотоархива уже активно используются и встречают положительную оценку пользователей, что подтверждает правильность и продуктивность выбранной методики и позволяет рекомендовать ее для применения в других отраслях архивного дела.

Работа поддержана Министерством науки и технической политики РФ в рамках проекта № 037.03.319.17/1-99.

Литература

1. Бокштейн И.М., Кузнецов Н.А., Мерзляков Н.С., Рубанов Л.И. “Возможности и средства цифровой реставрации архивных рукописных текстов. "Информационные технологии и вычислительные системы”, М.: ИВВС РАН, № 1, 1997, с. 1-15.
2. Бокштейн И.М., Карнаухов В.Н., Кузнецов Н.А., Мерзляков Н.С., Рубанов Л.И. Разработка баз данных архивных изображений на основе современных технологий их обработки и хранения. // Компьютерная оптика, 1998, Вып.15, 116-124.
3. Bockstein I.M., Karnaukhov V.N., Kuznetsov N.A., Merzlyakov N.S., Rubanov L.I. Digital restoration, enhancement, and archiving of photo-documents. In: Digital Image Processing and Computer Graphics (DIP-97). Wenger E., Dimitrov L.I. (editors), Proceedings of SPIE, 1998, Vol. 3346, 350-356.
4. Bockstein I.M., Karnaukhov V.N., Kuznetsov N.A., Merzlyakov N.S., Rubanov L.I. Automation of archival image database population. Pattern Recognition and Image Analysis, 1998, Vol. 8, No.4, 582-600.
5. Karnaukhov V., Wenger E., Merzlyakov N., Haidinger A., Lackner F. Thematic processing and retrieving of watermarks. Image Processing and Computer Optics, SPIE, 1996, Vol. 2363, 32-39.
6. Karnaukhov V.N., Merzlyakov N.S., Rubanov L. Image Processing and Storage in Digital Archives of Manuscripts and Photo-Documents. In: Proceedings of 5th Open German-Russian Workshop on Pattern Recognition and Image Understanding, Herrsching (Germany), 1998, September 21-25, pp. 30-34.