МНОГОУРОВНЕВЫЕ ЦИФРОВЫЕ АРХИВЫ: СТРАТЕГИЯ ПОСТРОЕНИЯ И ОПЫТ СОЗДАНИЯ

Л.И. Рубанов, Н.С. Мерзляков, В.Н. Карнаухов
Институт проблем передачи информации РАН, 101447 Москва ГСП-4, Большой Каретный пер., 19
{rubanov,nick,vic}@iitp.ru

MULTILEVEL DIGITAL ARCHIVES: STRATEGY AND EXPERIENCE

L.I. Rubanov, N.S. Merzlyakov, V.N. Karnaukhov
Institute for Information Transmission Problems of the Russian Academy of Sciences 19, Bolshoj Karetnyj per., Moscow, 101447 GSP-4, Russia
{rubanov,nick,vic}@iitp.ru

The paper describes a methodology we use to translate an existing conventional archive into a digital one. The method does well for large archives comprising documents with essential graphic constituent (handwritten texts, photographs, drawings, etc.). Main structural components of our digital archive are relational database and image bank which are physically separated but logically linked together. The components make up three-level distributed structure consisting of primary archive, its regional replicas, and various secondary archives (among them subsets presented in the Web and collections of compact discs). Only authorized user are allowed to access two upper levels, and the bottom level is open for free public access. A secondary archive is created and updated automatically without special development. Such construction allows us to combine reliable storage, easy access and protection of intellectual property. The paper also presents several digital archives already implemented in the Archive of the Russian Academy of Sciences.

Культурное наследие человечества богато представлено собраниями библиотек и музеев, для которых уже имеется значительный опыт создания цифровых хранилищ, в том числе доступных массовому пользователю через Интернет. В то же время имеется огромный культурный пласт, сосредоточенный в разнообразных архивах, который пока еще не нашел достойного представления и практически недоступен мировому сообществу.

Так, Архив Российской академии наук (АРАН), образованный в 1720 г., является старейшим научным архивом России, где хранятся не только документы Академии наук за всю ее более чем 275-летнюю историю, но и материалы по истории российской и зарубежной науки. Научный потенциал АРАН содержится в более чем 2000 архивных фондах,

включающих около 1 млн. единиц хранения. Это фонды учреждений Академии наук, научных обществ, личные фонды выдающихся ученых – М.В. Ломоносова, Л. Эйлера, В.И. Вернадского, К.Э. Циолковского, С.В. Ковалевской, Н.И. Вавилова, И.И. Мечникова и др., а также различные тематические коллекции (медалей и знаков, старинных рисунков и т.п.). Сегодня основная форма работы специалистов и обычных пользователей с архивными материалами – это изучение архивных дел в читальном зале АРАН. При такой процедуре трудно обеспечить широкий доступ пользователей к архиву и при этом гарантировать физическую сохранность материалов, равно как и целостность заключенной в них интеллектуальной собственности.

С 2000 г. ИППИ РАН и АРАН ведут научно-практическую работу по созданию тексто-графической базы данных по истории российской фундаментальной науки на основе фондов Архива РАН[1]. При создании этого цифрового архива были поставлены сложные, временами противоречащие друг другу задачи – обеспечить высокую степень аутентичности цифрового представления архивных единиц, надежность их хранения, возможность цифровой реставрации полученных от времени повреждений, сочетание общедоступности архивной информации с сохранением прав коммерческой и интеллектуальной собственности, возможность использования в широком диапазоне инфраструктур и программно-аппаратных средств, и многие другие.

В ИППИ РАН накоплен более чем 30-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 [1], Фото-архив ЛАФОКИ РАН [2], Международная база данных водяных знаков в западноевропейских древних рукописях и актах [3] (в сотрудничестве с Австрийской академией наук), и др. Поэтому не случайно, что работа над тексто-графическим цифровым архивом РАН в конце 90-х гг. была поручена данному коллективу.

Ограниченность имеющихся ресурсов и колоссальный объем хранимой архивной информации потребовали выработать особую стратегию построения и наполнения цифрового архива, состоящую из следующих пяти основных компонентов: приоритетность, иерархичность, переносимость, эффективность и доступность. Подробное обсуждение этих стратегических компонентов приведено в [4], дадим поэтому лишь краткие пояснения.

---

[1] Проект поддерживает Российский фонд фундаментальных исследований (№ 00-07-90032).

Приоритетность включает в себя выбор очередности перевода в цифровую форму тех или иных архивных фондов, коллекций, единиц хранения (а в общем смысле – и самих архивов). Учитываются многие соображения: частота обращения к фонду, степень сохранности оригинала, категория архивной единицы и степень ее уникальности, коммерческая перспективность, доступные технологии и технические средства и др.

Применительно к фондам АРАН в качестве приоритетных категорий архивных единиц выбраны рукописные документы (включая рисунки, чертежи, карты и т.п.) и авторизованная машинопись, но прежде всего фотодокументы – негативы, фотографии, слайды. Приоритетными фондами признаны личные фонды выдающихся ученых, президентов Академии и возглавлявшихся ими учреждений, а также наиболее интересные тематические коллекции.

Иерархичность предполагает построение таких информационных и организационных структур, в которых было бы возможным управлять степенью полноты и детальности предоставляемой информации. Это необходимо для обеспечения постепенного наполнения цифрового архива с учетом установленных приоритетов. С другой стороны, многоуровневая структура архива создает предпосылки для его территориального распределения и репликации, а также для разграничения доступа к архивной информации. Построенная в ходе разработки иерархическая архитектура цифрового архива в настоящее время включает три уровня и подробнее описывается ниже.

Переносимость призвана учесть неизбежную за время работы смену поколений вычислительной техники, программного обеспечения и носителей информации. Чтобы результаты начатой работы не пропадали, а могли эффективно использоваться долгое время, они должны базироваться на признанных в качестве международных стандартов форматах хранения информации, типовом системном программном обеспечении, языках описания и манипулирования данными. В частности, в разрабатываемом цифровом архиве применяются переносимые стандартные форматы неподвижных изображений (TIFF, JPG, PNG), языки HTML и SQL, их расширения и усовершенствования, ставшие де факто стандартами. Основой информационного каркаса являются реляционные базы данных, слабо зависящие от конкретной программной реализации (конкретно, в нашей работе в настоящее время используется сервер баз данных Oracle). Записи базы логически связаны с банком изображений, причем динамически, в зависимости от контекста взаимодействия с архивом (рис. 1).

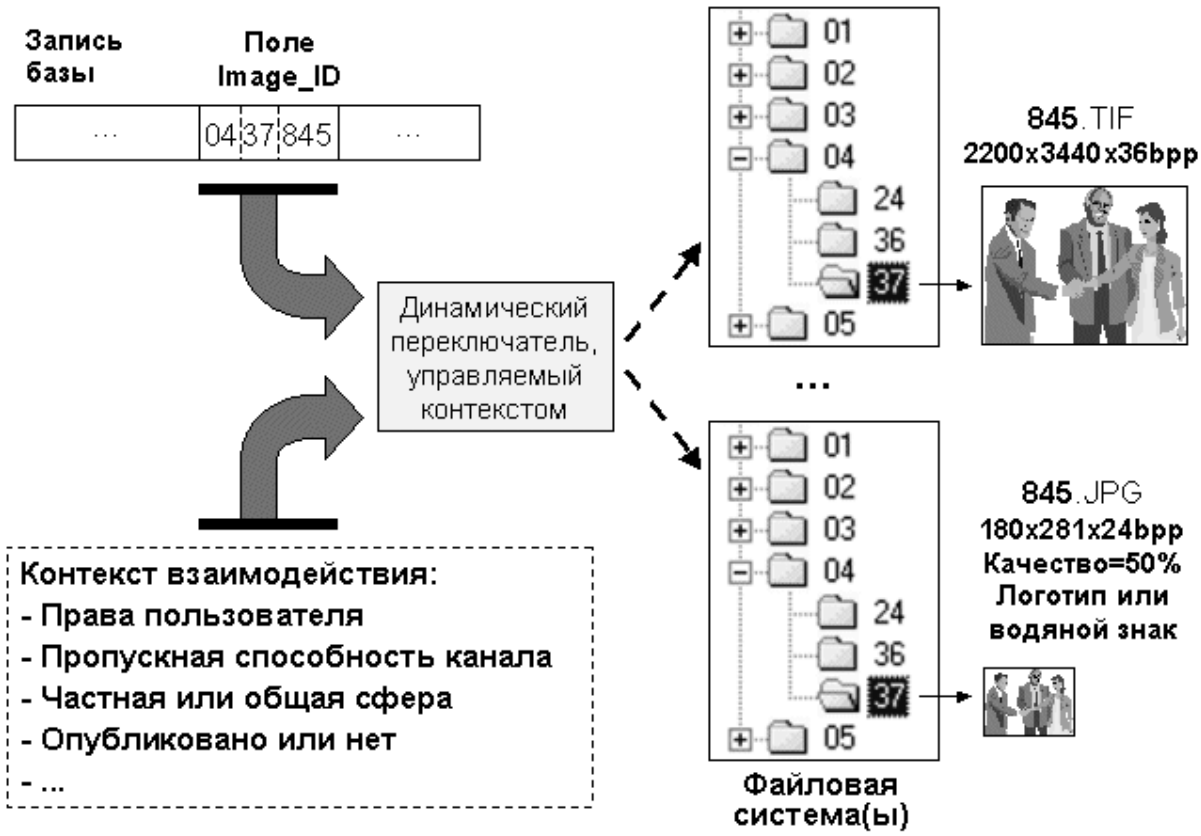

Рис. 1. Динамическая связь базы данных с банком изображений

Эффективность подразумевает нахождение оптимального баланса достигаемого качества и требуемых затрат ресурсов. Сюда входят и технические вопросы (например, выбор разрешения и глубины цвета при сканировании, определение метода и степени сжатия графической информации, выбор между графическим и текстовым представлением документа), и чисто организационные решения (например, сочетание ручных и автоматизированных методов оцифровки исходной информации, выбор объема цифровой реставрации изображений или вычитки текстовых документов после автоматического ввода).

Доступность цифрового архива рассматривается в двух аспектах. Прежде всего, это физическая возможность параллельного многоуровневого доступа к архивным базам данных в различных информационных средах в России и за рубежом, в том числе при отсутствии доступа к Интернет. Для создаваемого цифрового архива РАН принята трехуровневая структура (рис. 2).

Первый уровень (первичный цифровой архив) строится на локальной сети мощных малых машин (с перспективой перевода на большую ЭВМ). Второй уровень образуют зеркала первичного архива для повышения надежности хранения и размещения в регионах. К первичному архиву

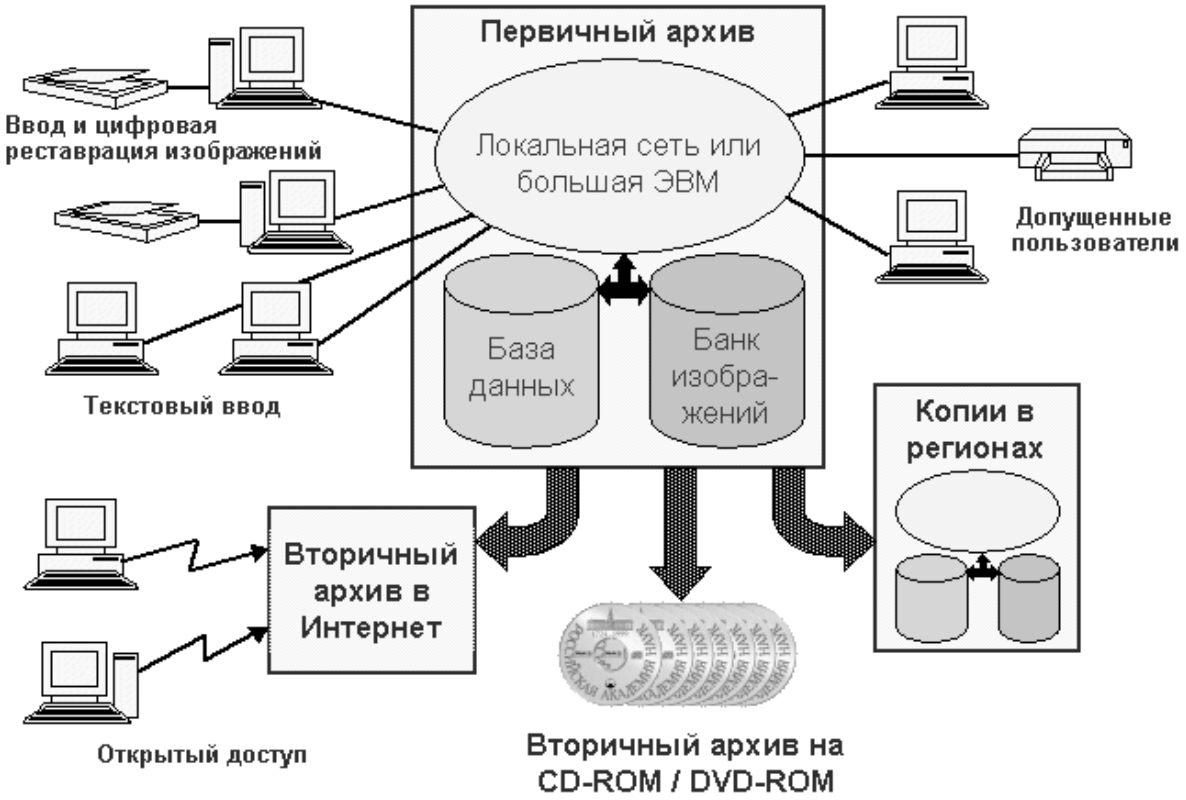

Рис. 2. Структура многоуровневого цифрового архива

предоставляется ограниченный доступ пользователей через оборудованные на первом и втором уровнях локальные рабочие места.

Третий уровень формируется на материале первичного цифрового архива в виде системы 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 передаваться в публичные библиотеки. Другой очевидной формой размещения вторичного архива может выступать сетевой узел (сайт) Интернет, что открывает глобальный доступ к содержимому архива.

Важно отметить, что создание вторичных цифровых архивов не является 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.

Другой аспект *доступности* цифрового архива – это простота и глубина поиска в нем необходимой информации. Известно, что во многих архивах возможности поиска по существу исчерпываются изучением форма-

лизованных описей по фондам, где единице хранения, занимающей многие листы, отведено лишь несколько слов. Чаще всего только хранитель соответствующего фонда имеет полное представление о его содержимом. Поэтому при переводе архива в цифровую форму в базу данных не только вносится весь существующий справочный материал (описи, обзоры и т.п.), но и по возможности он дополняется новыми дескрипторами и ассоциативными связями, упрощающими целенаправленный поиск и отбор материалов.

На основе изложенной стратегии к настоящему времени уже создано несколько разделов цифрового архива по истории науки (см. рис. 3–5):

1. База данных всех членов Академии наук с 1724 г. (4955 чел.), включающая их краткие биографии, научную специализацию, академические должности, научные награды (все на русском и английском языках), а также свыше 7000 портретов. Пример одной из форм для работы пользователей с базой данных приведен на рис. 3. На материале этого первичного цифрового архива были созданы два вторичных архива:

а) CD-ROM “Российская Академия Наук: 1724-1999” (вышло два издания);

б) Ресурс в Интернет: http://hp.iitp.ru

2. Коллекция медалей и знаков (АРАН, разряд XIII). В этой разработке была реализована возможность многомасштабной визуализации физических объектов, в том числе с большим увеличением. С этой целью нами была проведена цветная макросъемка полной коллекции с помощью цифровых фотокамер. Примеры многомасштабной визуализации приведены на рис. 4.

3. Персональные фонды президентов Академии:

а) А.П. Александров (АРАН, ф. 1916) – свыше 600 фотодокументов. Пример формы приводится на рис. 5.

б) М.В. Келдыш (АРАН, ф. 1729) – свыше 500 фотодокументов.

4. Коллекция портретов (фотографий, рисунков, гравюр) российских и зарубежных ученых, собранная Мусиным-Пушкиным (содержится внутри фонда акад. Н.А. Морозова – АРАН, ф. 543) – свыше 700 документов.

Хотя научные исследования и разработки по созданию цифрового архива РАН еще продолжаются, реализованные к настоящему времени разделы цифрового архива уже используются и встречают положительную оценку, что подтверждает правильность и продуктивность выбранной стратегии, позволяет рекомендовать ее для применения в других отраслях архивного дела. Мы надеемся, что данная технология может быть полезна при переводе в практически вечную, широкодоступную цифровую форму и других архивов – неотъемлемой составляющей мирового культурного наследия.

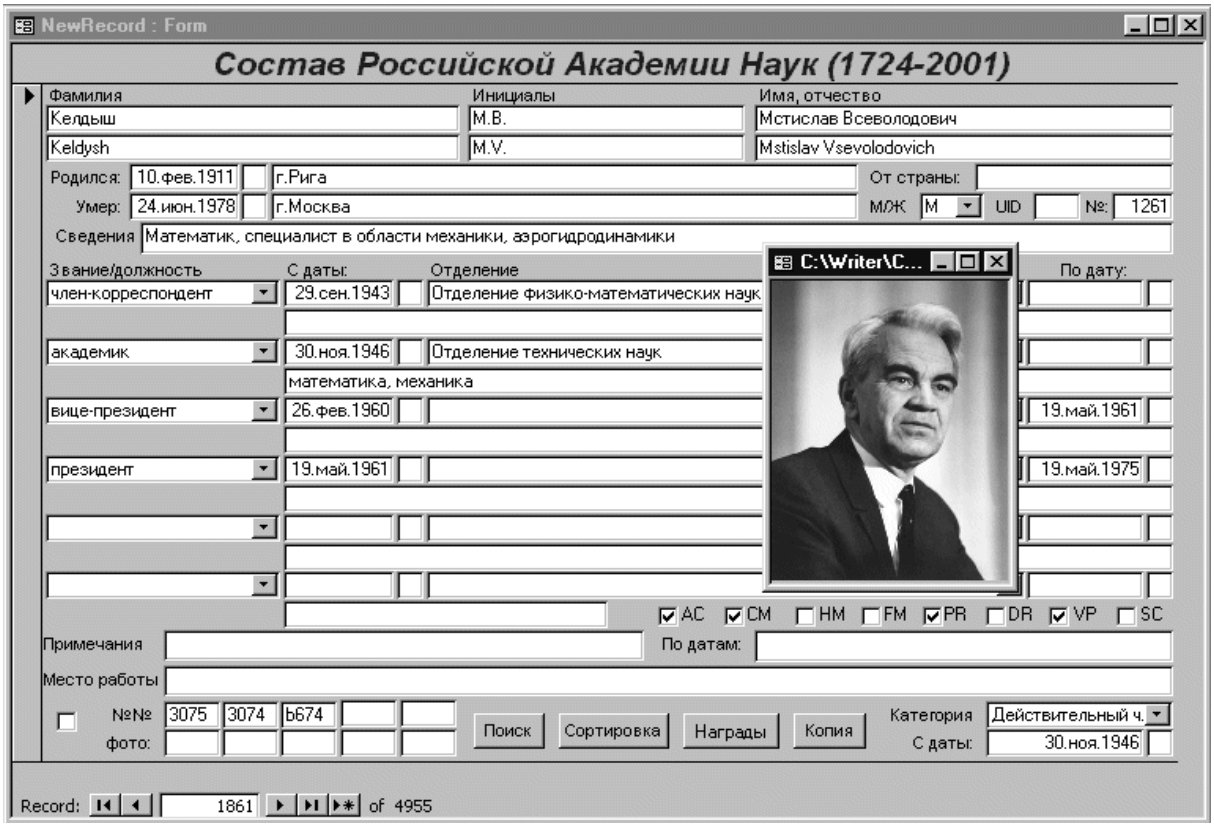

Рис. 3. Пример формы для работы с базой данных по персональному составу

а

б

в

г

Рис. 4. Примеры многомасштабной визуализации коллекции медалей и знаков

Рис. 5. Пример формы для работы с персональными фондами архива

Литература

1. И.М. Бокштейн, Н.А. Кузнецов, Н.С. Мерзляков, Л.И. Рубанов. Возможности и средства цифровой реставрации архивных рукописных текстов // Информационные технологии и вычислительные системы, № 1, 1997. М.: ИВВС РАН, 1997. С. 1-15.
2. I.M. Bockstein, V.N. Karnaukhov, N.A. Kuznetsov, N.S. Merzlyakov, and L.I. Rubanov, "Digital restoration, enhancement, and archiving of photo-documents," *Digital Image Processing and Computer Graphics (DIP-97), Proc. of SPIE*, Wenger E., Dimitrov L.I. (editors), **3346**, pp. 350-356, Vienna, 1998.
3. V. Karnaukhov, E. Wenger, N. Merzlyakov, A. Haidinger, F. Lackner, "Thematic processing and retrieving of watermarks," *Image Processing and Computer Optics (DIP-94), Proc. of SPIE*, Kuznetsov N.A., Soifer V.A. (editors), **2363**, pp. 32-39, Samara, 1996.
4. L.I. Rubanov, N.S. Merzlyakov, V.N. Karnaukhov, and N.M. Osipova, "Strategy of creation of digital archives accessible through the Internet", *Internet Imaging III, Proc. of SPIE,* G.B. Beretta, R. Schettini (editors), **4672**, pp. 181-189, San Jose, 2002.