УДК 681.3

Оцифровка, каталогизация, хранение и поиск архивной документации

С. В. Смирнов

СПб ГУП «Санкт-Петербургский информационно-аналитический центр». E-mail: serge.smir@gmail.com

М. В. Белозерова

СПб ГУП «Санкт-Петербургский информационно-аналитический центр». E-mail: belozerova@iac.spb.ru

Рассмотрены основы построения электронного архива – информационной системы, предназначенной для долговременного хранения информации. Описаны внутренняя функциональная структура системы. Рассмотрены способы помещения, структурирования и поиска информации в электронном архиве. Показаны варианты взаимодействия системы с внешней средой.

***Ключевые слова**: электронный архив, оцифровка, сканирование, каталогизация, поиск документов, автоматизация деятельности сотрудников архивов, долговременное хранение информации, управление данными, планирование процессов хранения документов, проектирование архивной информационной системы.*

Basis for building an electronic archive, designed for long-term storage of information is considered. The internal functional structure of the designed system is described. Ways of ingest, structuring and retrieval of information in the electronic archive are considered. Variants of the interaction of the archive with the external environment are shown.

***Keywords**: electronic archive, digitization, cataloging, preservation, scanning, retrieval of archives, automation of archive, long-term storage of information, data management, preservation planning processes, archival information system design.*

Введение

Архивы, работающие с бумажными каталогами и документами, уже перестали обеспечивать удовлетворительную оперативность, полноту и достоверность выполнения запросов к непомерно разрастающимся фондам документов. Более того, бумажные ценности, как известно, со временем приходят в негодность и безвозвратно исчезают. О масштабах проблемы говорит тот факт, что только сеть государственных архивов России насчитывает более 2 тыс. бумажных хранилищ, содержащих свыше 140 млн. дел. Огромный поток документов и информационных материалов, обращающихся внутри крупных коммерческих структур, придает новый импульс построению архивов электронных документов. Дело касается уже не только компактного, безопасного хранения и быстрого поиска документов, но и вопросов оперативного анализа.

В наше время доступной остается только та информация, которой регулярно пользуются, постоянно адаптируя ее к новым техническим и программным средствам. Если же такую информацию оставить без внимания (скажем, поместить в архив) всего на несколько лет, никто не сможет гарантировать, что ее удастся прочесть с устаревших носителей и воспроизвести с помощью новых версий даже той программы, в которой она была записана. Другими словами, электронным данным на всем протяжении срока хранения требуется бдительный контроль. Причем активное участие в поддержании доступности к данным и их понятности должны принимать сами источники информации, т. е. роль источника и архивиста зачастую сливаются воедино.

Определение электронного архива

Формально можно дать следующее определение электронному архиву (ЭА) – автоматизированная информационная система, представляющая собой интегрированный комплекс программных и аппаратных средств, предназначенных для долговременного хранения и управления данными и информацией, в том числе для обработки, поиска и предоставления доступа к информации в соответствие с нуждами ключевой аудитории. Под ключевой аудиторией понимается круг людей, обладающих необходимым багажом знаний для адекватного восприятия содержания документов. Зачастую ЭА воспринимают как традиционный архив, где вместо бумажных дел содержатся соответствующие документы на машинных носителях (дискеты, компакт-диски). При таком подходе разница между бумажным и электронным

вариантами архива невелика, а эффективность использования такого архива ограничена скоростью работы человека, наличием свободного места для хранения машинных носителей и т.п.

Основные функции электронного архива

Цель создания ЭА состоит в обеспечении оперативного и полноценного доступа ко всем хранящимся и поступающим документам. Для этого требуется решить три основные задачи: ввести массив имеющихся в архиве документов, систематизировать их и обеспечить возможность оперативного полнотекстового доступа к электронным документам.

Общую идею можно обрисовать следующим образом. Организуется развертывание высокопроизводительной сети, включающей графические рабочие станции и мощные серверы ввода и обработки информации. Для ввода документов с бумажных носителей низкого качества, используются промышленные сканеры потокового ввода и соответствующие программные средства. Система обеспечивает эффективное индексирование и полнотекстовый поиск неструктурированной информации большого объема.

Кроме обеспечения долговременного хранения данных к информационной системе, реализующей функции ЭА, предъявляется ряд следующих основных требований.

1. Устанавливать соглашения и принимать соответствующую информацию от источников (так называемое электронное комплектование).

В том числе необходимо утверждать критерии отбора материалов для помещения в архивное хранилище. Эти критерии могут быть основаны на таких факторах, как тематика, происхождение или формат.

2. Производить контроль над данными с целью обеспечения долговременного хранения.

Данное требование акцентирует ЭА на обладании достаточными правами собственности над документами, чтобы производить процедуры необходимые для долговременного хранения. Например, если ЭА требуется конвертировать документ в новый формат, в связи с развитием технического прогресса, он обязан иметь соответствующие права.

3. Определять границы ключевой аудитории и поддерживать понятность хранимых документов для этой аудитории без обращения к помощи источника информации.

Зарождение информации всегда происходит в каком-то контексте, и зачастую без знания этого контекста невозможно полностью понять ее смысл. Исходя из этого, ЭА должен хранить не только саму информацию, но и существенную часть связанного контекста для того, чтобы убедиться в ее ясности и, главное, возможности использования будущими поколениями. «Контекстуальная информация» может состоять из описания структуры или формата, в котором хранятся данные, объяснения того как и почему информация была создана и даже методов ее интерпретации. Установка границ ключевой аудитории играет важную роль в определении объема хранимого контекста и формировании требований к набору метаданных хранимой информации.

4. Следовать принятым в системе политикам и процедурам для поддержания хранимой информации в стабильном состоянии в любых ситуациях, а также для распространения подлинных копий в оригинальной или приближенной к ней форме.

5. Делать хранимую информацию доступной для ключевой аудитории путем предоставления различных механизмов и сервисов, которые покрывают пользовательские нужды и требования.

Концепция построения электронного архива

Описание концепции построения ЭА можно разделить на две взаимосвязанные части. Первая описывает внешнее окружение ЭА, вторая – функциональные компоненты и внутренние механизмы, которые обеспечивают выполнение ранее приведенных требований. Рассмотрим каждую из них по очереди.

Окружающая среда. Процессы ЭА не работают в замкнутой среде, напротив, они кооперируют с широким кругом участников, реализуя их потребности в хранении и получении информации. Схема взаимодействия приведена на рис. 1.

За пределами ЭА существуют источники информации, пользователи и управляющие.

Источники – это те лица, организации или системы, которые осуществляют передачу информации для долговременного хранения. Они представляют информацию и сопутствующие метаданные для последующей обработки и подготовки к помещению в хранилище. Взаимодействие между архивом и источником должно регулироваться законодательством. В нормативных

документах, регулирующих это взаимодействие, должны быть описаны такие детали, как типы, форматы данных, обязательный набор метаданных и пути безопасной передачи.

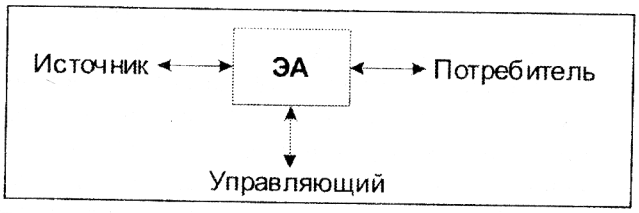

Рис. 1. Окружающая среда ЭА

В обязанности управляющих входит формулировка, модификация и контроль исполнения высокоуровневых принципов работы ЭА. Примерами могут служить: стратегическое планирование, определение тематических областей архивных коллекций, предоставление гарантий сохранности вверенных архиву документов, контроль свободного пространства и периодические проверки производительности подсистем ЭА. Следует отметить, что управляющие не обязаны быть вовлечены в ежедневные операции архива.

Пользователи – это те лица, организации или системы, которые запрашивают, получают и используют архивную информацию. Среди них можно выделить специальный класс, именуемый ключевой аудиторией. Определение границ этого класса является важным аспектом в процессе долговременного хранения. Чем шире мы раздвигаем границы, тем большее количество необходимых метаданных мы должны хранить.

Концепция взаимодействия управляющих, источников и пользователей информации, а также ключевой аудитории играет функциональную, а не организационную роль в построении системы. Все эти элементы могут представлять собой одну организационную единицу или распределяться по множеству организаций.

Функциональная модель электронного архива

Можно выделить шесть ключевых высокоуровневых подсистем, или функциональных областей, которые обеспечивают выполнение всех требований к ЭА, в том числе долговременное хранение и доступ к информации. Схема их взаимодействия приведена на рис. 2.

1. *Подсистема формирования данных* предназначена для приема информации от источника и подготовки ее к помещению в архивохранилище. Специфическими функциями являются: проверка целостности и качества поступающих данных, преобразование принятой информации в форму пригодную для хранения, составление описательных метаданных для дальнейшего поиска по ним, передача преобразованных данных в хранилище и оперативную базу данных архива.

В состав подсистемы могут быть включены следующие компоненты: а) сканирование; б)

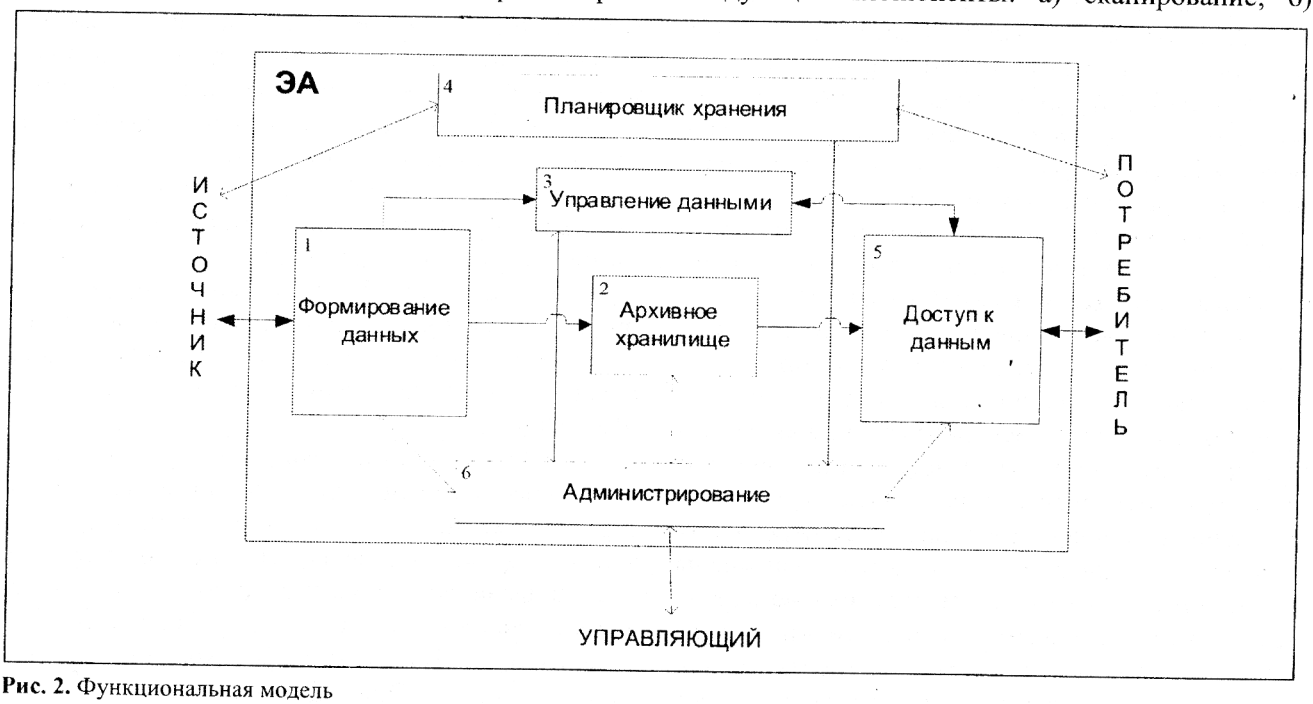

Рис. 2. Функциональная модель

распознавание; в) разбор документов; г) ручной ввод; д) конвертер формата данных из других приложений; е) миграция из других архивных систем; ж) прием данных от удаленных источников; з) импорт данных из различных СУБД; и) форматно-логический контроль; к) формирование метаданных; л) информационно-лингвистическое обеспечение.

Сканирование является серьезной задачей и требует определенного опыта и знания всех, в том числе и дополнительных возможностей оборудования и ПО. Велика вероятность того, что в процессе сканирования возникнет потребность в дополнительной обработке файлов, например в устранении перекосов, «вырезании» изображений по формату, пакетном удалении повторяющихся и ненужных частей изображений (например, изображений перфорации).

Дополнительная обработка может производиться при помощи стандартных аппаратных или программных опций. Некоторые недостатки изображений невозможно устранить в процессе сканирования при помощи дополнительных аппаратных блоков и опций программного обеспечения. Большую часть электронных копий архивных документов возможно отредактировать лишь «вручную». Пакетная обработка не приемлема для файлов, недостатки изображений в которых не имеют каких-либо закономерностей и не встречаются в других файлах.

2. *Архивное хранилище* ответственно за выбор способа хранения информации (онлайновые или оффлайновые хранилища), а также за проверку того, что набор хранимых битов остается целостным и воспроизводимым в течение всего времени хранения. Для обеспечения целостности, доступности и понятности информации подсистема должна позволять производить процедуры замены носителей и миграции форматов.

Для облегчения процессов миграции важно учитывать эти требования изначально при создании баз данных, при организации хранения электронных документов использовать наиболее распространение форматы. Иногда миграция информационных ресурсов на другие платформы по какой-то причине представляется нереальной или может существенно исказить оригиналы электронных документов. В этом случае можно использовать эмуляторы программной среды. Однако это также бывает непросто сделать, так как не для всех программных оболочек могут быть разработаны эмуляторы. Именно поэтому при создании информационных систем и электронных документов следует изначально ориентироваться не только на распространенные форматы записи, но и на распространенные ОС, СУБД и другое программное обеспечение.

Архивное хранилище также должно осуществлять контроль ошибок и обеспечивать доступность и отказоустойчивость в случаях возможных сбоев или непредвиденных обстоятельств. В конечном итоге, архивное хранилище должно возвращать запрошенные данные потребителю. Стоит отметить, что данная подсистема не должна иметь внешних интерфейсов и доступ к ней следует осуществлять только через промежуточные уровни системы.

Основными компонентами подсистемы являются: а) долговременное хранилище; б) компонент шифрования; в) компонент контроля целостности; г) компонент обеспечения доступности.

3. *Подсистема управления данными* отвечает за обслуживание описательной информации объектов хранения, а также за управление статистическими данными системы. Первичные функции включают исполнение запросов к базам данных, генерацию отчетов для других подсистем и компонент архива, проведение операций обновления и удаления оперативных данных. К компонентам данной подсистемы относятся: а) база данных для хранения информации оперативного использования; б) реестр, предназначенный для хранения метаданных, прав доступа, контрольных сумм, электронно-цифровых подписей (ЭЦП) и других административных данных; в) компонент полнотекстового поиска.

Основной задачей подсистемы является организация процесса упорядочивания и каталогизации поступающих документов.

Автоматическая категоризация документов производится путем сравнения с ранее созданными образцами, так что документ может быть обработан в соответствии с определенными ранее правилами.

4. *Подсистема планирования процесса хранения* предназначена для реализации выбранной стратегии хранения. Подсистема должна предоставлять сервисы для осуществления мониторинга ЭА на выявление произошедших изменений в области знаний ключевой аудитории, возможных инноваций в сфере технологий хранения и доступа к данным. При выявлении значительных

изменений должны быть сформированы рекомендации к обновлению политик ЭА, а также произведены сами обновления. Стратегический план хранения документа может содержать следующие пункты:

периодическое обновление сертификатов – рост вычислительных мощностей компьютеров превращает самый надежный алгоритм шифрования сегодняшнего дня в легко дешифруемый завтра;

управление жизненным циклом электронного документа – у каждого документа могут существовать такие реквизиты, как статус секретности, срок засекречивания, время хранения и т.п. Электронный архив должен периодически проверять значения реквизитов и производить необходимые обновления.

5. *Подсистема доступа к данным*, как и следует из названия, управляет процессами и сервисами, при помощи которых пользователи – в частности ключевая аудитория – запрашивают и получают информацию из архивного хранилища. Типовой процесс доступа к данным состоит из обработки пользовательского запроса и переадресации его в подсистему управления данными. Далее формируются запросы в подсистему архивного хранилища, получается хранимая информация, проводятся необходимые трансформации и содержимое отправляется пользователю. Данная подсистема обязана обеспечивать ЭА механизмами безопасности и контроля доступа к хранящейся в архиве информации.

6. *Подсистема администрирования* выполняет повседневные операции над ЭА и координирует действия других пяти подсистем. В ее обязанности входит предоставление сервисов для согласования договоров передачи данных с производителем информации, обеспечение необходимой справочной и технической поддержкой потребителей, а также реализация сервисов управления политиками и стандартами работы архива для управляющих. Подсистема является своеобразным центром всего архива: она осуществляет взаимодействие как с остальными внутренними подсистемами, так и с внешними участниками архивной деятельности.

Перечень компонент подсистемы может состоять из следующих элементов: а) компонент формирования статистики; б) аудит, создание отчетов; в) управление описаниями объектов; г) компонент создания пользовательских форм.

Заключение

Автоматизация дает нам многое, но мы, как никогда прежде, обременены миллионными массивами данных. И до тех пор, пока мы будем представлять мировое знание дискретным, разделенным не только на реальные физические, но и виртуальные документы, нам будет необходимо их находить, объединять, учитывать, а, следовательно, снабжать поисковыми и идентифицирующими признаками.

От этого зависит, что станется с накопленным информационным богатством страны, с каким интеллектуальным багажом мы окажемся среди других членов информационного общества.

ЛИТЕРАТУРА

1. *Beagrie, T., Jones, M.,* Preservation Management of Digital Materials: The Handbook. Digital Preservation Coalition. 2008.
2. ISO 14721:2003. Space data and information transfer systems. Open archival information system. Reference model.
3. *Носевич В. Л.* Архив электронных документов: белорусский опыт // Отечественные архивы. 2002. № 1. С. 44–52.

Поступила 1 апреля 2010 г.

Digitization, cataloging, preservation and retrieval of archives

S. V. Smirnov, M. V. Belozerova

The main principals of building electronic archive for long-term preservation are considered in the article.
Electronic Archive (EA) is defined as an integrated suite of software and hardware designed for long-term storage of archival documents in electronic form and providing access to them, in accordance with the needs of key audiences. Under the key audience (or designated community) understood the number of people with the necessary store of knowledge for the adequate perception of the content of documents.
Article allocates the basic obligation to be performed by the information system, which implements the function of EA.
Negotiate for and accept appropriate information from information producers.
Obtain sufficient control of the information in order to meet long-term preservation objectives.
Determine the scope of the archive's user community.
Follow documented policies and procedures to ensure the information is preserved against all reasonable contingencies, and to enable dissemination of authenticated copies of the preserved information in its original form, or in a form traceable to the original.