Интеллектуальные системы и технологии

Электронные архивы: возможные решения проблем долгосрочного хранения данных

Г. П. Акимова, М. А. Пашкин, Е. В. Пашкина, А. В. Соловьев

Аннотация. В работе систематизированы проблемы, возникающие при долгосрочном хранении электронных документов, предложены возможные варианты их решения, опробованные авторами при создании ряда систем электронных архивов. Статья является продолжением серии публикаций, посвященных проблемам создания и внедрения электронных архивов.
Ключевые слова: *электронный документооборот, системы управления содержимым, электронный архив, система управления электронными документами, электронный документ, электронный архивный документ, автоматизация архивного дела.*

Введение

Как было показано ранее [1], сравнительно недавно начавшийся бум внедрения систем электронного документооборота (СЭД) в организациях не затрагивает процесса передачи завершенных документов в полноценный делопроизводственный архив. Предположительное отставание внедрения электронных архивов от оперативных информационных систем на 3–5 лет вполне объяснимо, поскольку указанный срок — это среднее время хранения документов в «оперативных» архивах или в БД СЭД до их массовой передачи в вышестоящие архивы. Кроме того, достаточно редко можно услышать об электронных архивных системах, которые позволяют длительно хранить электронные документы (сроком не менее 5 лет).

Многие организации не желают решать задачу оцифровки большого количества документов (см. [1]) или не видят перспективы результатов такой работы, даже если и используют СЭД, и продолжают работать с архивом бумажных документов. В такой ситуации можно рекомендовать использовать ЭА РК документов, как учетную систему для ускорения и упрощения задачи поиска документов, если в архивной карточке указаны топологические данные о месте хранения документа. Решается и обратная задача — по бумажному оригиналу необходимо «поднять» всю историю работы с документом (например, выдача документа, связи данного документа с другими и т. д.). В такой ситуации электронный архив представляет собой лишь реквизитную БД, не содержащую текстовую часть документа. Такой ЭА не решает задач долговременного хранения документов, а перекладывает данную задачу на архив бумажных документов, но при этом обеспечивает эффективный способ поиска и навигации по архиву документов. Возможны варианты использования ЭА как дополнения к существующему бумажному архиву, при этом можно сохранять также и оцифрованные документы.

В предлагаемой статье рассмотрены основные способы хранения электронных документов, используемые в настоящее время, при этом особое внимание уделено документам с длительным сроком хранения. Выделены проблемы и возможные пути их решения, которые могут помочь разработчикам электронных архивов.

1. Основные понятия и определения

Электронный архив (ЭА) — структурированное хранилище неизменяемых электронных оригиналов документов (электронных изображений бумажных документов), созданное на основе законов и правил ведения архивов на конкретной территории (в конкретной стране).

Длительное хранение — хранение электронных документов не менее 5 лет.

Определение не претендует на «абсолютность», так как в конкретных архивах эти сроки могут меняться. За основу срока (5 лет) взят максимальный срок хранения документов в оперативных архивах СЭД. Электронные документы могут храниться в течение десятилетий или даже столетий или «бессрочно» в зависимости от их важности.

Долговременная сохранность — «период времени, в течение которого электронные документы поддерживаются в качестве доступного и аутентичного свидетельства (доказательства)» [2].

Аутентичный электронный документ — «электронный документ, точность, надежность и целостность которого сохраняются с течением времени» [2].

Электронная подпись — «информация в электронной форме, которая присоединена к другой информации в электронной форме (подписываемой информации) или иным образом связана с такой информацией и которая используется для определения лица, подписывающего информацию» [3].

До 2012 г. вместо ЭП использовался термин ЭЦП (электронно-цифровая подпись), определявшаяся как реквизит электронного документа, предназначенный для определения лица, подписавшего документ. В контексте данной статьи ЭП является частью хранящегося в ЭА электронного документа.

Квалифицированная электронная подпись — электронная подпись, которая соответствует следующим признакам [3]:

«1) получена в результате криптографического преобразования информации с использованием ключа электронной подписи;

2) позволяет определить лицо, подписавшее электронный документ;

3) позволяет обнаружить факт внесения изменений в электронный документ после момента его подписания;

4) создается с использованием средств электронной подписи;

5) ключ проверки электронной подписи указан в квалифицированном сертификате;

6) для создания и проверки электронной подписи используются средства электронной подписи, получившие подтверждение соответствия требованиям, установленным в соответствии с» [3].

Удостоверяющий центр — «юридическое лицо или индивидуальный предприниматель, осуществляющие функции по созданию и выдаче сертификатов ключей проверки электронных подписей, а также иные функции, предусмотренные» [3].

2. Постановка задачи долговременного хранения электронного документа и проблемы ее реализации

В простейшей постановке задача формулируется следующим образом. Требуется обеспечить длительное хранение электронных документов (см. определение в [1]) в программно-аппаратной среде, причем в течение всего срока хранения должна обеспечиваться аутентичность документа. При этом предполагается, что аутентичность документа на момент передачи его в архив подтверждена, документы не искажены, сохранность документов полная. Нет ограничений на форматы данных передаваемых в ЭА документов. ЭА сертифицирован для работы со средствами ЭП.

В рамках данной статьи предполагаем, что при длительном сроке хранения истекают сроки действия сертификатов ЭП, заканчивается оперативное хранение документов в архивах подразделений, завершается поддержка версий некоторых операционных систем (ОС), например Windows, и прикладного программного обеспечения.

На первый взгляд постановка достаточно простая, однако на практике при реализации может возникнуть множество проблем, связанных с технической сложностью реализации требований именно долговременного хранения. Выделим основные проблемы, которые всегда возникают при решении поставленной задачи:

1) аутентичность документа в течение всего срока хранения;
2) «старение» носителей информации;
3) перемещение данных и сохранность метаданных;
4) интерпретируемость и отображение электронных документов.

Указанные выше проблемы, конечно, хорошо известны и неоднократно обсуждаемы в среде разработчиков ЭА (см. например, ГОСТ Р 54989–2012 [2], являющийся переводом ISO TR 18492:2005, а также иные «переведенные» ГОСТ [4, 5] и системы требований [13]). Однако, в перечисленных документах описание проблем носит скорее рекомендательный характер и формулируется как «разработчики ЭА должны продумать» те или иные вопросы. Конкретных рекомендаций не приводится из-за

отсутствия более развитых и всеобщих (мировых) стандартов на правила хранения электронных документов, правила аутентификации, отсутствия стандартов на форматы хранения электронных архивных документов и средств интерпретации документов.

2.1. Сохранность аутентичности документа

Срок хранения в архиве зависит от вида документа, некоторые особо ценные документы должны храниться бессрочно (фактически столетиями), другие — десятилетиями, как документы по личному составу, или годами. Это накладывает определенные требования на технику, используемую в архивах длительного хранения, и требование сохранения аутентичности документа в течение всего срока хранения.

Для гарантии неизменности документа должны применяться как организационные меры, так и программные средства. К организационным мерам обычно относят защиту документов от несанкционированного доступа, например, непосредственно к хранилищам данных, носителям информации, коммуникационному оборудованию и др.

К программным средствам относят разграничение прав доступа на электронные документы и обеспечение контроля целостности, который реализуется с помощью хранения хеш-кодов электронных документов или использования ЭП. В последнем случае ЭП может автоматически устанавливаться программными средствами ЭА при вводе электронного документа в БД ЭА. Возможен вариант, когда в ЭА поступает электронный документ, подписанный ЭП. В этом случае перед ЭА стоит задача проверки корректности ЭП при вводе и, в дальнейшем, обеспечения неизменности документа в процессе хранения.

Задача оперативного хранения документов (не более 5 лет) решается достаточно просто, поскольку не приходится решать сопутствующие проблемы: большинство сертификатов ключей подписи не успевают исчерпать срок своего действия, мала вероятность, что изменения коснутся удостоверяющих центров (УЦ), например, прекращение деятельности УЦ.

При организации хранения документов свыше 5 лет разработчики и пользователи ЭА гарантированно столкнутся с проблемой просроченных сертификатов ключей ЭП, в том числе и корневых, т. е. сертификатов удостоверяющего центра, а в этом случае ЭП будет считаться недействительной, и при проверке будут зафиксированы ошибки, связанные с истечением срока действия сертификатов.

Еще одной проблемой, возникающей при использовании низкоразрядных ключей (до 256 бит) ЭП и накоплении огромных массивов электронных документов, является возможность получения так называемых коллизий первого и второго рода соответственно: подделка документов в ЭА для соответствия ЭП и появление в БД ЭА разных документов с одинаковой ЭП. Несмотря на то, что эта проблема пока маловероятна (стойкость ЭП с 256-битными ключами до 10^{30} операций), в будущем, если документы хранятся десятилетиями и массив таких документов огромен — проблема может проявиться очень остро. К тому же в связи с бурным развитием техники и технологий подделка ЭП на низкоразрядных ключах через несколько лет не составит большого труда.

И, наконец, законодательство, в частности [6], допускает наличие у одного лица (организации) нескольких ключей (сертификатов) ЭП. Также прямо не запрещено использование одного ключа (сертификатом) ЭП несколькими лицами. А это, в свою очередь, может создавать путаницу при идентификации лица, подписавшего электронный документ.

2.2. «Старение» носителей информации

Помимо возможности подмены или потери документов в процессе переноса с носителя на носитель из-за умысла или халатности персонала, существует проблема выхода из строя самих носителей информации (дисков, лент, оптических носителей и др.). Ни один производитель подобной техники не гарантирует сохранность ее в течение десятилетий (тем более столетий), а, следовательно, встает проблема своевременной диагностики носителей электронных документов и своевременной перезаписи документов на другие носители.

Гарантийные сроки хранения большинства жестких дисков — 5 лет. Производители оптических дисков однократной записи, востребованные в ЭА (носители типа WORM-write once read many), называли изначально сроки в 50–100 лет, но затем и они были существенно уменьшены (кроме того, для них нужны идеальные условия хранения) до 20–25 лет максимум, после чего данные должны быть перезаписаны. На основе опыта создания ЭА с использованием DVD-R ведущих производителей, авторы могут утверждать, что на практике срок хранения DVD-R еще ниже, проверки и перезаписи нужно осуществлять не реже 1 раза в 5 лет.

Даже для специально предназначенных для ЭА накопителей на базе технологии UDO (Ultra Density Optical, разработка компании Plasmon) на основе ультраплотной записи [7] не подтверждена возможность их работы в течение многих десятилетий. Накопители UDO служат в ЭА медицинских изображений для хранения медицинских докумен-

тов, например, медицинских карт пациентов, причем гарантированный срок хранения не превышает 5 лет. UDO представляет собой картридж 5.25 с оптическим диском внутри. Объем диска на данный момент составляет от 60 Гб до 120 Гб. Для записи может использоваться как красный лазер (650 нм), так и сине-фиолетовый (405 нм), причем во втором случае максимальный объем диска может достигать 500 Гб. Оптический диск не подвержен размагничиванию, как магнитные носители.

Магнитные ленты являются крайне неустойчивым к внешним воздействиям носителям, поскольку требуется их перемотка 1 раз в полгода и тщательная защита от размагничивания.

Использование твердотельных накопителей (SSD — Solid state disk, флэш-карт и т. д.) также пока ненадежно. Данные накопители имеют ограничение на количество циклов перезаписи (3000–10000), повышенный износ в связи с этим, высокую стоимость гигабайта информации по сравнению с жесткими дисками и оптическими дисками и невысокий объем хранения данных [8]. Проблему повышенного износа пытаются преодолеть с помощью технологии энергонезависимой памяти FRAM [1] (количество циклов перезаписи до 10^{14}) [9]. Однако, и эти носители не позволяют хранить большие объемы данных, зато отличаются высокой стоимостью. Время гарантированного хранения данных на SSD и FRAM оценивается в 10 лет.

Таким образом, промышленные средства хранения электронный информации на данный момент не могут достигнуть максимального срока хранения информации, такого как на бумаге или в виде микрофильмов (до 500 лет при идеальных условиях хранения).

Кроме того, при стремительном развитии вычислительной техники имеет место технологическое старение, поэтому с достаточно высокой вероятностью через 100 лет невозможно будет прочитать данные с современных магнитных и оптических носителей из-за отсутствия в будущем устройств их чтения, даже если информация каким-то чудом на них сохранится.

2.3. Перемещение данных и сохранность метаданных

Поскольку выше была затронута тема надежности хранения документов на внешних носителях, то с необходимостью возникает задача переноса архивных данных на новые носители информации, а, значит, встает вопрос об отсутствии потерь данных при проведении данной операции.

При этом проблема касается не только переноса самих документов, но и метаданных (см. [10]) документов, индексов (в том числе и полнотекстовых). Если сопутствующие данные (индексы, метаданные, классификаторы, рубрикаторы, связи с другими документами и др.) не могут быть корректно перенесены, то, по сути, перемещение данных (миграция) выльется в повторное создание ЭА в новой операционной среде (на новой платформе) с построением заново метаданных, индексов и т. д. К тому же, если документ является частью единицы классификации (дела, пачки) или связан с другими документами, данные связи также должны быть восстановлены, иначе целостность хранения может быть поставлена под сомнение.

2.4. Интерпретируемость и отображение данных

При долговременном хранении электронных документов возникает проблема интерпретируемости и отображения данных в новых информационных условиях, т. е. наличие возможности раскодировать хранимый формат электронного документа через десятилетия и показать документ в том или ином виде, например, отобразить на экране, распечатать и т. д.

Отсутствие стратегии в данном вопросе и превращение ЭА в склад разноформатных документов может спустя десятилетия привести к тому, что часть информации невозможно будет раскодировать из-за отсутствия (устаревания) средств интерпретации хранимых форматов данных, а также из-за утери описания хранимых форматов, в случае использования закрытых форматов представления электронных документов. Некоторым решением проблемы может быть создание конвертеров, преобразующих старые форматы в новые, но здесь следует иметь в виду, что чем позже будет поставлена задача конвертации данных, тем менее реально будет ее решение.

2.5. Прочие проблемы хранения документов в ЭА

Список рассмотренных выше проблем, естественно, не является окончательным. В частности, он не содержит таких важных задач, как скорость работы с архивом и информационная безопасность. Про подходы к решению этих задач написано достаточно много, в том числе и методики создания моделей угроз, нарушителей и т. д. В контексте этого исследования мы не будем касаться данных вопросов.

Не менее важной задачей является хранение и обработка больших объемов данных (как реквизитов, так и документов, включая задачу первоначального наполнения ЭА и потокового ввода до-

[1] Ferroelectric Random Access Memory — сегнетоэлектрическая память с произвольным доступом.

кументов). К обзору этого вопроса мы планируем вернуться в следующих статьях, посвященных электронному архиву.

Не рассматриваются в рамках данного исследования всевозможные юридические «тонкости», связанные с подписанием документов ЭП юридическими и физическими лицами, возможности непризнания электронного документа, не представляющего юридической силы. Для простоты примем допущение, что все электронные документы, заверенные ЭП и поступающие в ЭА, прошли проверку юридической значимости в момент их заверения ЭП.

Также не рассматриваем в данной статье такой важный аспект, как обеспечение катастрофоустойчивости решения ЭА. Этот вопросом мы планируем осветить в следующих статьях, посвященных созданию электронных архивов документов.

3. Предлагаемые решения проблем долговременного хранения электронных документов

В данной главе авторы делятся опытом, полученным при создании электронных архивов длительного хранения документов.

3.1. Сохранение аутентичности документа

На настоящий момент основным решением проблемы сохранения аутентичности документа является использование ЭП. Однако сертификаты и открытые ключи ЭП обладают ограниченным сроком действия, поэтому спустя год или 5 лет при обращении к документу с просроченной ЭП можно получить сообщение о некорректности ЭП, что поставит под сомнение подлинность документа. ЭП удобно использовать в системах электронного документооборота, поскольку сроки работы с документом малы, однако в системах, обеспечивающих длительное хранение, гарантированно возникнут проблемы просроченных сертификатов и ключей подписи.

При решении таких задач рекомендуется использовать для длительного хранения только усиленную квалифицированную ЭП, заверенную квалифицированным сертификатом (см. [3]), т. е. ЭП должна содержать подтвержденный штамп времени. При этом цепочка сертификатов ключей в идеале должна обязательно содержаться внутри ЭП или передаваться в ЭА вместе с ЭП. Только в этом случае есть гарантия, что спустя десятилетия подлинность документа можно будет подтвердить, если за это время, конечно, не изменятся стандарты, и будут существовать средства проверки данной ЭП. При этом нужно учесть, что при проверке ЭП может потребоваться список отзыва сертификатов (COC), актуальный на момент проставления подписи.

В качестве ключевой меры обеспечения аутентичности хранимых документов в ЭА авторами предлагается использовать архивную ЭП, которая автоматически вычисляется для всех электронных документов, помещаемых в ЭА. Процесс простановки такой ЭП должен быть возложен на операторов ввода, каждый из которых должен подписывать документ своей ЭП.

В организациях, работающих с ЭП, принято за правило периодически проводить смену ключей. Это означает, что все электронные документы, находящиеся в ЭА, следует переподписывать новым ключом ЭП (по сути новой ЭП), при этом старая ЭП сохраняется. Надо понимать, что такая схема не исключает подмены документов административным персоналом, эксплуатирующим ЭА, но гарантирует невозможность проведения данной операции операторами ввода. Кроме того, данная процедура не утверждена законодательно. Однако, переподписывание может быть включено в регламентные действия ЭА, например, стать частью процедуры инвентаризации архивных фондов. В этом случае необходимо тщательно защищать закрытые ключи электронной подписи, прописать подробно процедуру инвентаризации. Авторы считают, что процедура переподписывания документа электронной подписью оператора при вводе в архив должна быть закреплена законодательно и явиться основой для создания ЭА длительного хранения. Назовем данную процедуру инвентаризацией ЭП. В процессе инвентаризации ЭП подтверждается корректность ЭП документа, и он заверяется дополнительной ЭП (например, с ключом более высокой разрядности) в подтверждение факта инвентаризации. Новая ЭП, как более криптостойкая, исключит (или, по крайней мере, существенно снизит) риск появления в будущем документов-подделок, заверенных старыми «правильными» ЭП в БД ЭА.

Процедуру инвентаризации ЭП можно запускать в автоматическом режиме от имени оператора, работающего с архивом (особенно при огромных объемах данных), предоставляя операторам в проблемных случаях (например, нечитаемость данных, ошибка проверки ЭП и др.) принимать решения по возникшей проблеме.

Мощность компьютеров постоянно увеличивается, поэтому средства взлома ЭП, использующие полный перебор, со временем могут преодолевать все большую разрядность ключа подписи. Так, на сегодняшний момент безопасными считаются ЭП с 512-битным ключом и выше, однако в 2009 г. была взломана ЭП с 768-битным ключом, но пока это возмож-

но только за продолжительное время с использованием практически неограниченных компьютерных мощностей. Для некритичных данных можно использовать ЭП с 256-битными ключами (стойкость до 10^{30} операций).

Поэтому теоретически со временем возможна подделка документов в ЭА (коллизия первого рода), когда подбирается документ для ЭП, тем самым нарушается принцип неизменности документа в архиве. Только совместные организационные (включая политики безопасности ЭА), технические и программные способы позволят снизить вероятность взлома.

С накоплением документов при использовании низкоразрядных ключей (до 256 бит) возможна коллизия второго рода: наличие разных документов с одинаковой ЭП, что маловероятно, но теоретически возможно. Поэтому при проектировании ЭА нужно учитывать предполагаемый размер БД и возможный ее рост, чтобы предоставить адекватные средства защиты информации.

Следует обратить внимание на еще один аспект, возникающий при подтверждении аутентичности заверенных ЭП электронных документов, — сложность взаимодействия ЭА с удостоверяющим центром. Особенно часто с ним сталкиваются, когда в ЭА хранятся электронные документы, подписанные ЭП, которые выданы разными УЦ, в том числе в различных регионах РФ. В таком случае возникают ситуации, когда ЭА не может проверить ЭП поступившего документа, кроме того нет никаких гарантий хранения сертификатов ЭА самими УЦ. На данный момент решения указанной проблемы нет. В качестве одного из промежуточных решений авторы статьи предлагают непосредственно в ЭА организовать хранение всех сертификатов, списков отзыва сертификатов (СОС) и много другой дополнительной информации, на основании которой может быть проведено расследование и установлена подлинность документа. Тем самым функции УЦ переносятся в ЭА, особенно в отсутствии единой сети УЦ страны, и усложняют его.

3.2. «Старение» носителей информации

Все имеющиеся на данный момент типы носителей информации недостаточно надежны для хранения данных десятилетиями, а тем более столетиями. Более того, из-за процесса технологического старения через несколько десятилетий не останется устройств, обеспечивающих чтение актуальных на данный момент носителей информации.

Поэтому решение проблемы лежит, во-первых, в избыточности хранения информации, во-вторых, в регулярной проверке и переносе информации на новые носители данных.

Избыточность хранения данных должны быть обеспечена как хранением данных ЭА непосредственно в БД на жестком диске, так и хранением на внешних носителях копий данных ЭА. В качестве такой копии может выступать как резервная копия БД, так и копии данных, вытесненных на внешние носители. Надо отметить, что хранение данных в ЭА может быть устроено следующими способами: индексы и данные находятся в БД; индексы находятся в БД, данные на внешних носителях; индексы находятся в БД, часть данных находится в БД, часть вытеснена на внешние носители. Во всех случаях для БД ЭА необходимо организовать регулярное резервное копирование БД на внешние носители. В качестве копии данных могут выступать как внешние носители с резервной копией БД, так и совокупность внешних носителей с резервной копией БД и внешних носителей с данными. При этом должно создаваться не менее двух копий данных ЭА, причем хранить их следует в разных помещениях, а в идеальном случае в разных зданиях, удаленных друг от друга. В случае использования в качестве резервной копии компакт-дисков рекомендуется создавать не менее трех копий данных. Дополнительно может быть реализовано катастрофоустойчивое решение (зеркало, или, для особо ценных документов, — резервный центр обработки данных (ЦОД)), т. е. хранение точной копии (копий) документов. Это означает, что необходимо реализовать децентрализованное хранение копий данных с разными мандатами доступа для оперативного и административного персоналов ЭА.

Регулярная проверка и перенос информации на новые носители должны обеспечить защиту от отказов и физической деградации цифровых носителей информации. Назовем такую процедуру инвентаризацией носителей. Данная операция должна включать проверку целостности данных на носителе, оценку оставшегося времени хранения данных на носителе и, при необходимости, перенос данных на новый носитель с уничтожением старого. В случае выявления нарушения целостности данных на носителе в ходе проверки новая копия данных создается из других копий данной информации. Периоды проверки носителей данных выбираются, исходя из типа носителей информации, но в любом случае период хранения данных на неизменяемом носителе не должен превышать трех лет, т. е. раз в три года каждый носитель информации должен быть проверен и при необходимости заменен. Процесс переноса информации должен предусматривать возможность слияния данных с разных носителей,

данное условие появляется из-за постоянного увеличения объемов всех видов носителей данных.

Только в этом случае можно будет говорить о сохранности данных в ЭА.

3.3. Перемещение данных и сохранность метаданных

Миграция данных должна быть неотъемлемой частью методологии создания ЭА долговременного хранения. Другой вопрос, что должно подвергаться миграции: только ли сами документы из БД ЭА или же еще связанные с ними метаданные, классификаторы, индексы и др.

Как было показано ранее ([1]), классификаторы и индексы являются неотъемлемой частью документа, поскольку определяют контекст его использования: предметную область, структуры организаций, логику хранения и классификации, связи с другими документами и т. д. По мнению авторов статьи, потеря этих данных при миграции может оказаться критичной, документ будет вырван из контекста использования, и понять его принадлежность какой-либо тематике будет проблематично.

Поэтому решение по миграции данных должно включать не только миграцию самих электронных документов, но и метаданных документа, расширив описание формата долгосрочного хранения (см. п. 3.4) набором тегов, которые нужны для хранения метаданных (например, расширенное дублинское ядро[2] [12]) документа. Практическая реализация данного положения при разработке авторами статьи электронных архивов подтверждают его правильность.

Отдельно стоит вопрос о полнотекстовых индексах документа. Конечно, не хочется терять такую ценную информацию, однако большинство СУБД не позволяет распорядиться полнотекстовыми индексами самостоятельно, а перестройка индекса для огромного массива данных после миграции может оказаться дорогостоящей по времени процедурой. Несовместимым может оказаться и формат индексов при переносе в другую среду хранения. При решении данной проблемы рекомендуется либо переносить полнотекстовые индексы вместе с документами, либо включить процедуру перестройки индексов в процесс миграции. Во втором случае переход на новую инфраструктуру ЭА должен быть осуществлен с задержкой в эксплуатации после миграции данных, т. е. для организации постепенного ввода в строй новой версии ЭА, что, в свою очередь, означает допущение существования ЭА в двух различных средах хранения при условии полной синхронизации данных между версиями БД ЭА.

Процедуру миграции можно будет производить реже, если использовать преимущества виртуализации операционных систем — операционная система (ОС), запущенная на виртуальном компьютере, будет функционировать, даже тогда, когда она не может быть установлена на современный компьютер. Однако, рано или поздно встанет вопрос о поддержке данной старой ОС со стороны производителя. К тому же в настоящий момент существуют ограничения на использование некоторых ОС в виртуальных средах. Например, использование IBM i (старое название OS/400) возможно только в виртуальных средах на платформе Power, на платформе Intel данная ОС работать не будет даже в виртуальной среде.

3.4. Интерпретируемость и отображение электронных документов

В информационном мире существует множество различных форматов электронных документов, но со временем многие из них перестают поддерживаться, а тем самым с течением времени трудно будет найти программное обеспечение, способное проинтерпретировать документ, сохраненный десятки лет назад в некотором формате.

Согласно [2]: «Стратегия долговременной сохранности должна обеспечить, чтобы электронные документы в будущем оставались читаемыми. Для достижения этой цели составляющий электронные документы поток битов должен быть доступен на той компьютерной системе или устройстве:

- на которой(ом) он первоначально был создан, или
- на которой(ом) он в настоящее время хранится, или
- которая(ое) в настоящее время используется для доступа к нему, или
- которая(ое) будет использоваться для хранения электронной информации в будущем».

Первые два варианта относятся скорее к сохранности собственно носителей (см. п. 3.2). Рассмотрим теперь проблему читаемости собственно данных, расположенных на читаемом носителе (не столь важно новом или старом).

Для решения такой проблемы должен быть подобран формат хранения архивных документов, отвечающий требованиям: простой, открытый и документированный, которые в свою очередь снизили бы вероятность «не интерпретируемости» документов, сохраненных в ЭА в данном формате, в будущем.

[2] Стандарт метаданных. Создан см. Dublin Core Metadata Initiative (http://dublincore.org/). В России с 01.07.2011 действует ГОСТ [12].

Федеральное агентство по техническому регулированию и метрологии РФ утвердило ГОСТ Р ИСО/МЭК 26300–2010 (перевод принятого в 2006 году международного стандарта ISO/IEC 26300:2006), в котором в качестве стандартного формата для офисных приложений определен Open Document Format (ODF).

Существует поддержка ODF, начиная с MS Office 2007. Google Docs и IBM Lotus Symphony так же поддерживают ODF.

В первой версии спецификации MoReq [13] редакции 2001 г. существовал раздел, посвященный долгосрочному хранению электронных документов, в котором одним из ключевых было требование предпочтительного использования открытых, документированных форматов в противовес к проприетарным (коммерческим).

В настоящее время при использовании обычных «текстовых» форматов офисных приложений выделяют группу рисков, которые связаны с используемыми форматами файлов:

«Во-первых, это проблема скрытой информации. Потенциально любой офисный документ может содержать в себе данные о предыдущих правках, комментарии, невидимый текст, сведения о компании и авторе. Все это для окончательной редакции является лишним и не должно попадать в электронный архив.

Во-вторых, автор может использовать в документе поля, значения которых могут изменяться, что приводит к искажению всего документа. Простейший пример — поле с текущей датой. Представьте, мы распечатываем документ из архива, а он датирован сегодняшним числом. Также не следует забывать и о макросах, которые могут изменить документ.

В-третьих, документ может содержать гиперссылки на веб-страницы или на другие связанные объекты (рисунки, схемы, другие документы). Иногда это действительно необходимо для удобства пользования этим документом и для его понимания. Но при помещении такого документа в архив с этим надо что-то делать — сохранять вместе с документом копии веб-страниц, например» [14].

Для решения указанной проблемы авторы предлагают в качестве формата архивного документа использовать открытые документированные форматы XML, ODF, PDF/A, в один из них конвертировать принимаемые в архив файлы, сохраняя оригиналы файлов как приложения (в случае их заверения ЭП — сохраняя вместе с ЭП). Однако, для более строго решения необходимо законодательно утвердить правила приема документов в ЭА и их переформатирование при сдаче на длительное хранение. Тогда в процессе приема в ЭА необходимо будет перезаверить ЭП весь набор полученных файлов документов, сохраняя оригиналы документов в исходном формате и их оригинальные ЭП. Соответствующая процедура также должна быть разработана и утверждена.

Отдельно стоит вопрос работы с видео, аудио документами, презентациями, анимационными файлами, программным кодом (скрипты), исполняемыми файлами и их компонентами. Для видео и аудио документов также необходим перевод их в наиболее простые и открытые форматы (возможно, что таким стандартом станет WebM [16]), сохраняя оригиналы сдаваемых документов. Впрочем, данная тема требует отдельного исследования, авторы статьи не располагают в данном вопросе достаточным опытом.

Помимо преобразования электронных документов в форматы хранения документов, потребуется предусмотреть процедуру инвентаризации данных, в процессе выполнения которой «устаревшие» форматы электронных документов ЭА должны быть преобразованы в современные для процедуры инвентаризации данных форматы электронных документов. Например, если в ЭА хранятся видеоданные в формате программы, которая больше не развивается, то следует провести преобразование таких видеоданных в формат программы, которая будет развиваться и поддерживаться, или преобразовать в стандарт, который поддерживается многими производителями программного обеспечения.

Отметим, что при копировании данных ЭА на внешние носители информации должны сохраняться как структура описания данных, так и описание формата хранения данных. Для хранения метаданных рекомендуется использование XML-формата.

3.5. Синхронизация электронного и бумажного архивов

На данный момент задача является значимой ввиду наличия огромных бумажных архивов. Одним из вариантов решения задачи синхронизации является использование штрих-кодов (двумерные штрихкоды способны хранить до 4 КБ информации), которым надпечатывается бумажный документ при регистрации (оцифровке) в архиве. Для особо ценных документов, особенно в случае их возможной порчи при надпечатке, может быть надпечатана штрих-кодом бумажная архивная карточка документа.

При этом перед началом создания ЭА должна быть продумана топология (например, комната-стеллаж-полка) хранилища бумажных документов, информация о которой хранится в электронном архиве как реквизит документа (вплоть до координат GPS).

Без решения данной проблемы и при наличии одновременно бумажного и электронного архивов (реквизитная БД), использование архива по назначению будет затруднительно и связано с временными издержками на поиск бумажных оригиналов и, наоборот, при отслеживании связей, окружения, классификации бумажного оригинала по ЭА (обратная задача).

4. Модель документа в ЭА долгосрочного хранения

Рассмотрев проблемы долгосрочного хранения документов и возможные пути их решения, можно скорректировать модель документа в ЭА, представленную в [1].

Графически модель документа в электронном архиве можно представить в виде графа (дерева), состоящего из взаимосвязанных семантических блоков B_i. Блоки в свою очередь представляют собой подграфы (поддеревья), также состоящие из семантических блоков следующего уровня: в любом документе всегда можно выделить заголовок, подзаголовки, повторяющиеся части, агрегаты (массивы, структуры данных), атомарные данные (листы дерева). Между документами могут существовать различные отношения (связи) [15], т. е. лес документов может быть связан в единый граф. При этом в вершинах деревьев можно указывать неявные связи с другими документами.

При длительном хранении документа кроме классификаторов и индексов [1], являющихся неотъемлемой частью электронного документа и проходящих вместе с ним возможные миграции данных, документ дополняется содержимым документа, преобразованным в один из форматов долгосрочного хранения (открытых, документированных форматов) XML, ODF, PDF/A. Поэтому модель документа в ЭА преобразуется в следующую (оператор «+» в данной записи в отличие от [1] заменен на операцию «объединения», так как речь идет не о фактическом сложении, а об объединении множеств различных данных):

$$DAr = \cup_{(i=1,N)}(B_i) = ArCard \cup OdfD \cup OrD \cup FTIdx \cup CLIdx,$$

где

$ArCard$ — архивная карточка документа (состоит из набора реквизитов, которые могут задаваться древовидной схемой) — изменяемая часть электронного документа, может меняться форма карточки, а также состав ее реквизитов. Однако изменение значений реквизитов, по крайней мере тех, которые получены из оригинала документа, запрещено, либо выполняется только уполномоченными лицами. Оперативно могут изменяться только значения реквизитов, определяющих нумерацию в данном конкретном архиве, топологию (размещение физического оригинала), служебную информацию: шифры, аннотация и т. д.;

$OdfD$ — преобразованное к формату долгосрочного хранения содержимое оригиналов документов — неизменяемая часть электронного документа, создается при приеме документов в ЭА, OdfD заверяется ЭП (в общем случае несколькими) при приеме в ЭА;

$$OdfD = OdfDoc \cup (\cup_{i=1,N1} OdfPic_i) \cup (\cup_{j=1,N2} Sign_j),$$

где

$OdfDoc$ — собственно преобразованное к формату долгосрочного хранения содержимое сдаваемых документов, $OdfPic$ — набор ($1–N1$) графической информации (растровые и векторные изображения, элементы презентаций и др.), подлежащей преобразованию из сдаваемых документов в графические форматы долгосрочного хранения (TIFF, JPEG, PDF/A), при этом $OdfDoc$ содержит ссылки на графические материалы, $Sign$ — набор ЭП ($1–N2$), заверяющих преобразованный документ (содержит в себе сертификаты подписавших, цепочку сертификатов, сертификаты удостоверяющих центров (УЦ));

OrD — оригиналы документов (электронные оригиналы документов или оцифрованные изображения оригинальных бумажных документов, которые далее также будем обозначать как оригиналы) — неизменяемая часть электронного документа (может включать ЭП, проставленные, например, в системе электронного документооборота — см. [1]);

$FTIdx$ — полнотекстовый индекс, полученный на основе индексирования реквизитов и текстов документа — изменяемая часть электронного документа (строится на основе полнотекстового анализа оригиналов документов), представляет собой набор всех слов оригиналов документов, приведенных к единственному числу, именительному падежу (для существительных), неопределенной форме (глаголов) и т. д. Является необязательной частью документа, ссылки на элементы $FTIdx$ содержаться в $OdfDoc$;

$CLIdx$ — вектор связок между электронным документом и классификаторами $<CLIdx_1,\ldots, CLIdx_k,\ldots, CLIdx_K> (k = 1, K)$ — изменяемая часть электронного документа, так как набор связок может изменяться или дополняться. Является необязательной частью документа, ссылки на элементы $CLIdx$ могут содержаться в $OdfDoc$. В простейшем случае

представляет собой набор позиций классификаторов, с которыми связан архивный документ. В случае долговременного хранения данная часть документа является информацией о классифицировании и среде хранения (окружении) документа. О классификаторах и их видах было рассказано в статье [1].

Заключение

Данная статья явилась попыткой систематизировать знания и опыт, полученные при разработке архивных систем, в частности электронных архивов долговременного хранения документов для Пенсионного фонда РФ (в эксплуатации с 2001 г., сроки хранения документов до 100 лет в зависимости от типов документов), АКБ «Газпромбанк» (в эксплуатации с 1997 года, сроки хранения — десятки лет), коммерческих и государственных предприятий. Авторы систематизировали проблемы, возникающие при долговременном хранении электронных документов, с которыми неизбежно столкнется разработчик подобных систем.

Тема является весьма актуальной, поскольку электронные документы (по факту в ПФ РФ, ФНС и др.) начинают активно замещать документы бумажные, а значит, при длительных сроках хранения должна быть обеспечена их сохранность. Общая тенденция развития говорит о том, что в ближайшее время вытеснение бумажных документов станет массовым явлением, и подходы к их хранению должны быть выработаны уже сейчас.

В работе выделены проблемы, которые возникают при решении задачи долгосрочного хранения документов, приведены варианты их решения, доказана необходимость предлагаемых решений для обеспечения сохранности документов длительного хранения. Предлагаемые решения по хранению электронных документов предполагают избыточность хранения данных: хранение нескольких копий, оригиналов и переформатированных документов, наличие процедур инвентаризации носителей, ЭП и данных.

Авторы постарались показать, что современный ЭА — это не нечто неизменное, статичное, а постоянно изменяющаяся во времени структура, работая с которой необходимо регулярно обновлять носители информации, следить за действующими форматами данных, и, возможно, проводить обновление программной части ЭА.

Литература

1. *Акимова Г. П.*, *Пашкин М. А.*, *Пашкина Е. В.*, *Соловьев А. В.* Архивные хранилища и электронные архивы документов, основные постулаты и проблемы разработки / Труды Института системного анализа РАН (ИСА РАН). Т. 62. Вып. 4. М.: Красанд/URSS, 2012, С. 3–13.
2. ГОСТ Р 54989–2012 /ISO TR 18492:2005 Обеспечение долговременной сохранности электронных документов (вступает в силу с 01.05.2013).
3. Федеральный закон Российской Федерации от 6 апреля 2011 г. № 63-ФЗ «Об электронной подписи».
4. ГОСТ Р 54471–2011/ISO/TR 15801:2009 Системы электронного документооборота. Управление документацией. Информация, сохраняемая в электронном виде. Рекомендации по обеспечению достоверности и надежности.
5. ГОСТ Р ИСО 15489–1–2007 Система стандартов по информации, библиотечному и издательскому делу. Управление документами.
6. 1-ФЗ «Об электронной цифровой подписи» от 10 января 2002 г.
7. Оптические накопители Plasmon G-серии. Электронная публикация. [http://www.plasmon.ru/g-seria.shtm].
8. Наступление SSD // Журнал сетевых решений/LAN. № 11. 2010. Электронная публикация [http://www.osp.ru/lan/2010/11/13005552/].
9. *Volker Rzehak*. Особенности применения FRAM микроконтроллеров Texas Instruments // Журнал РАДИОЛОЦМАН. Апрель. 2012. Электронная публикация. [http://www.rlocman.ru/review/article.html?di=113273].
10. ГОСТ Р ИСО 23081–1–2008. Процессы управления документами. Метаданные для документов.
11. ГОСТ Р 34.10–2001. Информационная технология. Криптографическая защита информации. Процессы формирования и проверки электронной цифровой подписи.
12. ГОСТ Р 7.0.10–2010 (ИСО 15836:2003) «НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ. Система стандартов по информации, библиотечному и издательскому делу. НАБОР ЭЛЕМЕНТОВ МЕТАДАННЫХ „ДУБЛИНСКОЕ ЯДРО“».
13. Типовые требования к автоматизированным системам электронного документооборота. Спецификация MoReq. Версия 5.2.1. Март. 2001. Электронная публикация [http://www.cornwell.co.uk/moreq.html].
14. *Макаров С.* Хранение е-документов: как угнаться за ИТ? Электронная публикация. [http://www.cnews.ru/reviews/index.shtml?2011/02/08/426535].

15. *Белова А. Н.*, *Соловьев А. В.* Построение баз данных взаимосвязанных документов / Труды Института системного анализа РАН (ИСА РАН). Т. 62. Вып. 3. М.: Красанд/URSS, 2012, С. 25–30.
16. *Ходаковский К.* Google представила новый открытый видеостандарт. Электронная публикация [http://www.3dnews.ru/news/Google-predstavlyaet-noviy-otkritiy-videostandart/].
17. Обзор 10 облачных хранилищ данных. Электронная публикация [http://topobzor.com/obzor-10-oblachnyx-xranilishh-dannyx/.html].
18. Резервное копирование в «Облачное хранилище». Электронная публикация [http://habrahabr.ru/company/selectel/blog/168249/].
19. *Шамшина П. Ю.*, *Шамшина Т. А.* Риски информационной безопасности и аппаратно-программного средства защиты для облачных хранилищ данных. Рижский институт транспорта и связи. Латвия. Электронная публикация [http://mosi.ru/ru/conf/news/riski-informacionnoy-bezopasnosti-i-apparatno-programmnogo-sredstva-zashchity-dlya].

Акимова Галина Павловна. Вед. н. с. ИСА РАН. К. т. н. Окончила МФТИ в 1978 г. Количество печатных работ: 48. Область научных интересов: системное программирование, системный анализ, информационные технологии, влияние человеческого фактора, информационно-аналитические системы, электронный документооборот, электронный архив. E-mail: galina@cs.isa.ru

Пашкин Матвей Александрович. Науч. с. ИСА РАН. Окончил МГТУ «Станкин» в 2001 г. Количество печатных работ: 10. Область научных интересов: системное программирование, информационные технологии, информационно-аналитические системы, электронный архив. E-mail: matveyp@cs.isa.ru

Пашкина Елена Владимировна. Науч. с. ИСА РАН. Окончила МГУ в 2003 г. Количество печатных работ: 7. Область научных интересов: системное программирование, информационные технологии, электронный документооборот, электронный архив. E-mail: alena@cs.isa.ru

Соловьев Александр Владимирович. Вед. н. с. ИСА РАН. К. т. н. Окончил МГТУ им. Н. Э. Баумана в 1994 г. Количество печатных работ: 32. Область научных интересов: системный анализ, системы управления базами данных, теория надежности, влияние человеческого фактора, математическое моделирование, электронный документооборот. E-mail: alexsol@cs.isa.ru