Электронные архивы: разработка математической модели электронного документа при долговременном хранении

А.В. Соловьев

Аннотация. В статье рассматривается разработка математической модели электронного документа при долговременном хранении. В статье определяется необходимый состав и структура информации для долговременного хранения электронных документов. Приводится декомпозиция разработанные модели электронного документа для выделения составных частей общей информации. Статья предназначена для создания теоретической основы долговременного хранения электронных документов.

Ключевые слова: электронный документооборот, электронный архив, система управления электронными документами, электронный документ, долговременное хранение.

Введение

Как было показано в предыдущих статьях [12-16], важной задачей при организации долговременного хранения (определение долговременного хранения дано в [12, 16]) является необходимость определить, что должно храниться. Разнообразие определений понятия электронный документ, отсутствие единого понимания, что именно нужно хранить, порождает необходимость решения данной задачи путем создания математической модели электронного документа (ЭлД) при долговременном хранении.

Деловые документы, составляющие основу электронного архива (ЭА), как правило, связаны с делопроизводственными процессами в организации. Структурирование документов производится на основе размещения документов в более крупной единице хранения, названной делом. Разбивка по делам в РФ ведется в соответствии с правилами, оговоренными нормативными документами [3, 8, 9, 17].

Кроме ЭА есть еще ряд систем, которые не подчиняются строгим правилам ведения архива, однако могут хранить документы в течение длительного срока. При этом документы также могут быть связаны друг с другом некоторой системой классификации, информацией о владельце и др.

Приведем определения систем хранения электронных документов, распространенных в настоящее время в информационных технологиях.

Корпоративное хранилище данных – структурированное хранилище разнородных электронных документов, позволяющее управлять этими документами на основе единых правил, разработанных для нужд конкретного предприятия (организации).

В архивное хранилище обычно помещают разнородные и разноформатные документы, которые могут быть определенным образом классифицированы или упорядочены. Как правило, такие хранилища позволяют включать и удалять документы (а также прочие информа-

ционные ресурсы и файлы), находящиеся в доступе в конкретной организации, в том числе в различных ее информационных системах.

Единая классификация документов в таких системах может осуществляться как путем автоматического индексирования по заранее определенным ключевым реквизитам, позволяющим осуществлять поиск в архивном хранилище, так и путем автоматической классификации документов на основе обучаемого классификатора и полнотекстового индексирования документов.

Система управления данными предприятия (*Enterprise Content management system*, ECMS) – информационная система, используемая для обеспечения и организации совместного процесса создания, редактирования и управления документами.

ECMS подразделяется на несколько классов информационных систем, таких как системы электронного документооборота (СЭД), кадровые системы, системы взаимодействия с клиентами (CRM – Customer relationship manager) и др.

Главной задачей такой системы является возможность собирать в единое целое и объединять на основе ролей и задач все разнотипные электронные документы, доступные как внутри организации, так и за ее пределами, а также возможность обеспечения взаимодействия сотрудников, рабочих групп и проектов с созданными ими базами знаний, информацией и данными так, чтобы их легко можно было найти, извлечь и повторно использовать привычным для пользователя образом.

Информационно-поисковая система (электронная библиотека) – упорядоченная коллекция разнородных электронных документов (в том числе книг), снабженных средствами навигации и поиска.

Система автоматизации документооборота, система электронного документооборота (СЭД) – автоматизированная многопользовательская система, сопровождающая процесс управления работой организации с целью обеспечения выполнения ею своих функций. При этом предполагается, что процесс управления опирается на человеко-читаемые документы, содержащие инструкции, обязательные к исполнению сотрудниками организации.

Информационно-аналитическая система (ИАС) – информационная система, которая помимо задач хранения и поиска информации способна решать аналитические задачи, например, помощь в принятии решения и построение прогнозов.

Для создания эффективного ЭА подобная система должна обладать возможностями хранилища данных, классификации документов на основе правил архивного хранения, а также автоматической тематической классификацией. Одной из необходимых функций ЭА является полнотекстовая индексация документов архивного хранилища, которая является «базовой» для многих поисковых и аналитических функций.

Электронные архивы должны быть связаны с оперативными информационными системами в единую промышленную цепочку, позволяющую быстро загружать документы в архив и, наоборот, осуществлять поиск архивных документов из оперативной системы.

1. Краткий обзор методологий моделирования электронных документов при долговременном хранении

В настоящее время сложилась следующая противоречивая ситуация. Общая тенденция развития компьютерных средств работы с электронными документами говорит о том, что в ближайшее время вытеснение бумажных документов станет массовым явлением, и подходы к их хранению должны быть выработаны уже сейчас.

Учитывая изменчивость программно-аппаратной информационной среды, задача обеспечения сохранности электронных документов становится вовсе нетривиальной. Современная программно-техническая среда, т.е. набор технических средств (компьютеры, устройства хранения, записи, воспроизведения) и программных средств (операционные системы, средства создания и просмотра электронных документов) – это не нечто неизменное, статичное, а наоборот, крайне динамично меняющаяся во времени структура, работая с которой важно учитывать такие факторы как технологическое старение, износ, обновление программной среды.

Если бумажный документ может быть один раз создан и далее он хранится в неизменном, если пренебречь старением бумаги, виде в течение десятилетий, не нуждаясь в средствах отображения, то электронный документ без таких средств существовать не может.

Данная противоречивая ситуация определяет необходимость решения важной научно-технической проблемы обеспечения сохранности ЭлД, включая доступность, аутентичность (неизменность), интерпретируемость (читаемость) ЭлД в динамически меняющейся программно-аппаратной среде в течение всего длительного (годы, десятилетия) срока хранения.

Попытки разрешения данной проблемы активно предпринимаются, в частности, как в РФ, так и за рубежом делаются попытки систематизировать проблемы долговременной сохранности, предпринимаются попытки создания электронных архивов (ЭА). Однако, судя по отсутствию универсального решения проблемы, необходимо продолжать попытки ее решения, т.к. рост количества ЭлД идет лавинообразно.

Отсутствие универсального решения проблемы связано с несколькими факторами: отсутствует четкое понимание, что такое ЭлД с точки зрения состава информации, которую нужно сохранять, отсутствует методология измерения параметров среды хранения ЭлД, не все проблемы сохранности ЭлД до конца изучены и систематизированы.

Если кратко рассмотреть современные исследования (подробный анализ невозможен из-за ограничений объема журнальных статей), суженные до проблемы создания математической модели документа при долговременном хранении, можно отметить, что имеют место несколько основных тенденций при моделировании документов при долговременном хранении:

1. нормализация, то есть конвертация исходных документов в формат долговременного хранения при приеме в архив;
2. ограничение форматов документов для приема в архив;
3. включение в модель документа метаданных (сведений о документе кроме собственно самого документа);
4. создание распределенных аппаратных сред хранения с многократным дублированием, как носителей информации, так и копий документов;
5. использование средств криптозащиты информации для обеспечения неизменности документов.

Следование или не следование этим тенденциям определяет и модели документов, разработанные в разных странах.

Так, например, разработчики NARA (National Archives and Records Administration, USA) считают документом любую информацию в цифровом виде, в том числе включающую и программно-аппаратную реализацию [18]. Это приводит к тому, что в архив помещаются любые такие документы от файла до сервера (включая всю аппаратную часть, а также средства раскодирования и отображения документа). При этом упор с самого начала сделан на создание распределенных средств хранения с многократным дублированием [19]. Данный подход приводит к тому, что информация надежно сохраняется, однако возникают проблемы с чтением документов (раскодированием форматов, интерпретацией) спустя многие годы [20]. В связи с чем архив работает в настоящее время в ограниченном режиме [18]. Понимая данные проблемы, идет тенденция к сокращению набора принимаемых форматов данных [22]. Кроме того, можно отметить, что NARA не использует средства криптозащиты, делая упор на многократное резервирование данных. Модель документа включает полнотекстовый индекс, используемый для быстрого поиска [21], который можно отнести к метаданным документа. Многие наработки по моделированию документов при долговременном хранении стандартизируются [23].

Аналогичным путем с NARA следуют разработчики электронных архивов Австралии и Новой Зеландии. Однако в дополнение к моделям NARA, в документ добавляются метаданные документа (сведения о создании, об авторстве, о системе классификации документа) [24]. Исключается из модели аппаратная среда. Впрочем, метаданные четко не описаны. Моделирование электронного документа в архивах Новой Зеландии включает в себя также необходимость миграции данных [27].

Разработчики архивов Великобритании определяют состав метаданных электронного документа, которые включаются в модель документа [25, 26]. Тем самым в модель документа включаются: трехуровневая схема классификации документов (класс-папка(дело)-том), сведения о пользователях (авторе и менявших документ), сведения о датах создания и изменения документов.

Разработчики архивов Германии делают упор при моделировании документа на обеспечение неизменности и подтверждения авторства документа с использованием средств криптозащиты информации. Досконально прописываются нормы использования электронной подписи (ЭП), в том числе и перезаверения документа более новой ЭП с сохранением сведений об авторе старой ЭП [28]. Тем самым в модель документа включается ЭП, однако речи о хранении метаданных не идет. Не включаются также и аппаратная составляющая.

Аналогичным путем идут разработчики архивов Республики Корея, делая упор на ЭП при моделировании документов долговременного хранения [29]. Похожие модели, включающие ЭП, создавались для архивов Украины [30].

Архивы Дании, Швейцарии делают упор на нормализацию, тем самым моделирование документа включает трансформацию из оригинала в нормализованную копию в форматах PDF/A, TIFF, JPEG2000 (Швейцария) [31], PDF (нормализованная копия) + XML (метаданные документа) – Дания [32, 33]. Дания использует метаданные типа «дублинское ядро» при описании модели документов.

Тем самым, видно, что проблема создания модели электронного документа крайне актуальна. Тем не менее, в литературе модели документов описаны «крупноблочно», зачастую даже без применения математических формул, и не описывается методология создания подобных моделей применительно к архивам долгосрочного хранения. Очень важно также понять какой подход выбрать, и каким тенденциям следовать при разработке собственной модели.

2. Разработка модели документа

Опираясь на разработанную в ИСА РАН теорию документного интерфейса [4, 5, 7], разработанную д.т.н., проф. Емельяновым Н.Е. и ее развитие [6, 11] под документом будем понимать структурированную информацию, как совокупность взаимосвязанных семантических блоков. Документ (деловой документ), безусловно, имеет четкую структуру, форму и содержание. Электронный документ – документ, семантические блоки которого и взаимосвязи между ними представлены в электронно-цифровой форме.

Семантические блоки — некоторые фрагменты документа, выделенные по смысловому содержанию. Всякий реальный документ разбивается на взаимосвязанные части (разделы, подразделы, пункты и т.д.), которые мы будем называть семантическими блоками.

Графически модель документа в электронном архиве можно представить в виде графа (или дерева, если до корня из любой листовой вершины имеется единственный путь), 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) [10], т.е. лес документов может быть связан в единый граф. При этом в вершинах деревьев можно указывать неявные связи с другими документами. Если эти связи сделать явными, то лес превратится в сеть, разработка модели документа станет более сложной.

Учитывая рекомендации, необходимые для решения проблем, приведенных в [12, 13, 16], необходимо отметить, что документ в ЭА должен содержать дополнительную информацию: метаданные документа, связь с классификаторами, индексы, ЭП (электронная подпись), сертификаты ЭП и СОС (списки отзыва сертификатов). Кроме этого должны быть также заверенные выписки из журналов аудита ЭП.

При длительном хранении документа кроме классификаторов и индексов, являющихся неотъемлемой частью электронного документа и проходящих вместе с ним возможных миграций данных, документ дополняется нормализованной копией документа.

Нормализованная копия представляет собой преобразование документа в один из форматов долгосрочного хранения (открытых, документированных форматов) XML, ODF, PDF/A. Она может быть также представлена сочетанием форматов, например XML для хранения содержимого (текста) документа, метаданных, индексов, информацией о связи с другими документами. Если необходимо сохранить внешний вид документа, как можно точно повторяющий внешний вид оригинала, то лучше использовать TIFF (для черно-белых документов) или PNG (для цветных).

Хранение же программных компонент, а также аппаратных представляется в такой модели излишним, т.к. сильно перегружает архив данными, разобраться в которых будет крайне сложно. Кроме того, автор рассматривает электронный документ, как объект управления, максимально оторванный от программно-аппаратной среды хранения (полный отрыв невозможен), что делает возможным его относительно безболезненный перенос (миграцию) из одной программно-аппаратной среды хранения в другую (проблеме отчуждаемости будет посвящена отдельная статья автора, т.к. ее рассмотрение сильно увеличит объем статьи).

Ниже опишем методологический подход к созданию математической модели документа применительно к долговременному хранению.

Математическая постановка задачи создания модели ЭлД при долговременном хранении.

Дано:

1) Множество ЭлД $D = \{ DAr_i \}$

Найти:

1) Модель ЭлД $DAr_i \in D$, описывающую ЭлД с точки зрения состава информации, необходимой для долговременного хранения

Решение:

Можно утверждать, что математическая модель делового документа в ЭА при долговременном хранении представляет собой объединение (оператор U) семантических блоков документа (1.1).

Каждый семантический блок имеет отдельное назначение в разработанной модели.

В общем виде модель документа в ЭА в смысле представления необходимого состава информации определяется следующим образом:

$$DAr = \mathrm{U}_{(i=1,N)}(B_i)$$

Документ разбивается на семантические блоки:

- если $B_i \cap B_j \neq 0$, то $B_i \subseteq B_j$ или $B_j \subseteq B_i$;
- если $\exists B_j \subset B_i$, то $B_i = \cup B_j$ для всех $B_j \subset B_i$.

Каждому документу D поставим в соответствие некоторый граф $\Gamma(D)=(V,E)$, где V — множество вершин графа, E — множество дуг. $V=\{B_i\}$. $E=\{(B_i,B_j): B_i \rightarrow B_j\}$ (Рис. 1).

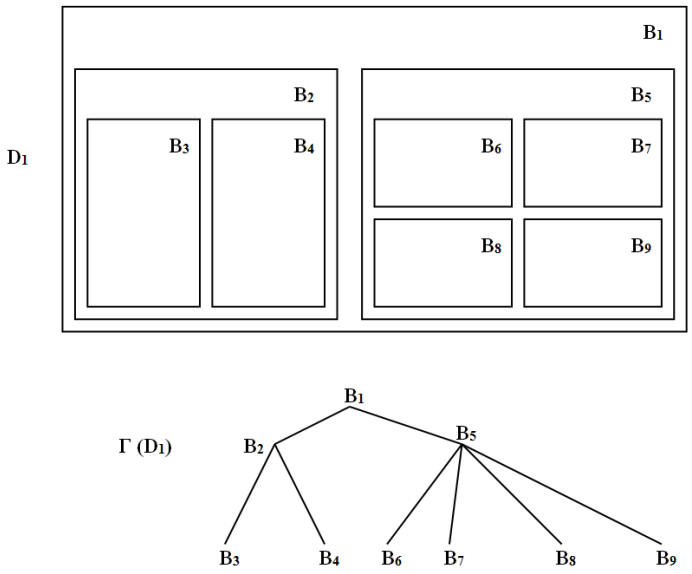

Рис. 1. Пример разбиения документа на семантические блоки и соответствующий граф документа

Каждый семантический блок B_i – представляет собой определенную часть документа, фрагмент части (подчасть), фрагмент подчасти и т.д.:

$$DAr = ArCard \cup OrD \cup OdfD \cup FTIdx \cup CLIdx \cup LDIdx \cup OperD, \quad (1.1)$$

где *ArCard* – архивная карточка документа (состоит из набора реквизитов, которые могут задаваться древовидной схемой) – изменяемая часть электронного документа, т.к. может меняться форма карточки, а также состав ее реквизитов;

OrD – оригинал документа (электронный документ (оригинал) или оцифрованное изображение оригинального бумажного документа, которое далее также будем обозначать как оригинал) – неизменяемая часть электронного документа (может включать ЭП, проставленные, например, в системе электронного документооборота – [12, 14]);

OdfD – нормализованная копия (под нормализацией здесь понимается приведение документов к единому формату (набору форматов) долговременного хранения) оригинала документа – неизменяемая часть электронного документа, которая создается при приеме документов в ЭА;

FTIdx – нормализованный текст (под нормализацией здесь понимается приведение всех слов в тексте к единственному числу именительного падежа и инфинитиву для глаголов) оригинала документа, представляет собой набор всех слов оригинала документа, приведенных к единственному числу, именительному падежу (для существительных), неопределенной форме (глаголов) и т.д. Является необязательной частью документа, ссылки на элементы *FTIdx* должны также содержаться в *OdfD*;

CLIdx – модель системы классификации ЭлД $<CLIdx_1, \dots, CLIdx_k, \dots, CLIdx_K>$ (k=1,K) – изменяемая часть электронного документа, т.к. набор классификаторов 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;

LDIdx – вектор связей с другими документами. Эта часть может изменяться, например, если в процессе исследования архивных документов выяснится, что документ связан еще с какими-то документами, массив связей может быть дополнен. Вектор связей позволяет связать (при необходимости) все множество документов в сеть;

OperD – выписка из журнала инвентаризации ЭП, сама выписка и/или ее отдельные записи также заверяются ЭП. Сама выписка – изменяющаяся часть документа. Как минимум она состоит из единственной записи – проверке ЭП при приеме в ЭА. Каждая инвентаризация ЭП фиксируется в журнале инвентаризации и заверяется ЭП.

Аналогично (как и *OperD*) к документу могут присоединяться выписки из журналов инвентаризации носителей, инвентаризации интерпретации и др. Однако, выписка из журнала инвентаризации ЭП должна быть обязательно (2.9). Это необходимый минимум дополнительных данных для удостоверения целостности и неизменности ЭлД.

Ограничения:

Модель документа представляет собой лес деревьев семантических блоков.

Свойства ЭлД:

1. Являться связанным объектом. ЭлД в общем случае – это не файл, отдельно хранящийся на электронном носителе (компакт-диске, например), а существующий в рамках информационной системы (ЭА) объект, который связан с классификаторами, рубрикаторами, другими документами.

2. Легкость миграции. Должна быть максимально облегчена будущая миграция документа, поэтому информация о нем (метаданные) должны также мигрировать вместе с документом, чтобы не было потери данных, а для этого метаданные также должны быть всегда с ЭлД, следовательно включены в модель ЭлД.

3. Быть структурированным. Документ должен быть разделен логически и физически на семантические блоки, т.к. выдача из ЭА документа «по запросу» должна выполняться с учетом разных уровней доступа пользователей к частям документа.

Каждый семантический блок документа (назовем его макроблоком), представленный в формуле (3.1) также делится на семантические блоки следующего уровня.

Оценка сложности модели ЭлД при составлении модели для конкретного ЭА:

Назовем рангом (или порядком) семантического блока B_i^k и обозначим $r(B_i^k)$ — длину максимальной цепочки вложенных блоков с началом в B_i^k. Иначе $r(B_i^k) = \max\{\text{m: } \exists \text{ блоки } B_{i1}, \ldots, B_{im} \text{ такие, что } B_{im} \subset \ldots \subset B_{i1} \subset B_i^k\}$. Если блок B_i^k прост, то $r(B_i^k) = 0$.

Модель ЭлД считаем сложной, если $r(B_i^k) \geq 10$, простой, если $r(B_i^k) \leq 4$.

Все макроблоки, кроме *OdfD* и *OrD*, назовем метаданными документа, т.к. они так или иначе содержат информацию о документе.

Ниже подробно описаны модели каждого макроблока ЭлД в ЭА долгосрочного хранения.

3. Разработка моделей макроблоков документа

3.1. Модель архивной карточки

Дано:

1) Множество архивных карточек $ArCard_i$ ЭлД $D = \{ DAr_i \}$

2) Модель состава семантических блоков верхнего уровня (1.1) ЭлД *DAr*

Найти:

1) Модель архивной карточки $ArCard_i$ ЭлД $DAr_i \in D$, описывающую архивную карточку ЭлД с точки зрения состава информации, необходимой для долговременного хранения

Решение:

Пусть *ArCard* – архивная карточка документа (состоит из набора реквизитов, которые могут задаваться древовидной схемой) – изменяемая часть электронного документа, т.к. может меняться форма карточки, а также состав ее реквизитов.

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.

Тогда модель архивной карточки ЭлД:

$$ArCard = ArCardCM \cup ArCardF \cup ArCardC, \quad (2.1)$$

где *ArCardCM* – модель содержания архивной карточки (дерево описания данных (реквизитов) архивной карточки), подразумевается, что архивная карточка одна на один архивный ЭлД;

ArCardF – экранная форма для заполнения и показа пользователю реквизитов карточки;

ArCardC – содержимое архивной карточки, т.е. значения реквизитов или, что тоже самое, вершин дерева описания данных.

Конечно, и форма и содержание также разбиваются на части – семантические блоки.

Можно утверждать, что пересечение формы и содержания архивной карточки дает пустое множество, т.е.:

$$(ArCardCM \cup ArCardF) \cap ArCardC = \varnothing.$$

Действительно, форма и содержание определяются согласно [11] следующим образом. Все, что извлекается из документа и его метаданных и имеет переменное значение в зависимости от экземпляра документа, мы будем называть содержанием архивной карточки. Состав и структуру семантических блоков, все постоянные тексты, опорные линии, шрифты (гарнитура, кегль, тип выделения и т.п.) и иную информацию, не зависящую от экземпляра ЭлД назовем формой архивной карточки.

Следовательно, форма и содержание содержат непересекающуюся информацию.

Так как *ArCardCM* и *ArCardF* не изменяются в зависимости от экземпляра ЭлД, то они относятся к форме документа. *ArCardC* уникально для каждого экземпляра, следовательно это содержание документа. Следовательно, $(ArCardCM \cup ArCardF) \cap ArCardC = \varnothing$.

Можно утверждать, что *ArCard* относится к метаданным документа.

Действительно, *ArCard* содержит информацию о документе, т.к. по сути это описание (форма) и набор (содержание) поисковых реквизитов и/или иной служебной информации, необходимой для идентификации и поиска документа в ЭА.

3.2. Модель оригинала документа

Дано:

1) Множество оригиналов OrD_i ЭлД $D = \{ DAr_i \}$

2) Модель состава семантических блоков верхнего уровня (1.1) ЭлД *DAr*

Найти:

1) Модель оригинала OrD_i ЭлД $DAr_i \in D$, описывающую оригинал с точки зрения состава информации, необходимой для долговременного хранения

Решение:

Пусть оригинал документа – *OrD* – электронный оригинал документа или оцифрованное изображение оригинального бумажного документа – неизменяемая часть электронного документа.

Тогда модель оригинала документа представляется в общем виде так:

$$OrD = U_{(i=1,N)}(OrDoc_i \, U \, (U_{(j=1,M1)} Sign_{ij})), \quad (2.2)$$

где *OrD* – оригинал документа (электронный документ (в общем случае файл), оцифрованная копия).

Каждый документ имеет свой граф (обычно дерево) описания данных, называемый моделью содержания (в иностранной литературе Content Model).

Оригинал – неизменяемая часть архивного документа. Исключение могут составлять случаи, когда добавляется более точная копия файла (например, более четкий оцифрованный образ), в этом случае допустимо удалить старый образ и добавить новый, но делается это только уполномоченным лицом, например, администратором безопасности, а в протоколе безопасности обязательно ставится отметка о произведенной замене.

Предпочтительнее добавлять новые образы без удаления старых;

$OrDoc_i$ – часть оригинала документа, может состоять из набора файлов (например, если каждая страница многостраничного документа представлена отдельной оцифрованной копией), каждый из которых заверен ЭП;

$Sign_{ij}$ – *j*-я электронная подпись (ЭП) *i*-го оригинала документа.

Может содержать в себе сертификат подписавшего, а также цепочку сертификатов, или же ссылки на сертификаты подписи, сертификаты удостоверяющих центров (УЦ), списки отзыва сертификатов. Это неизменяемая часть электронного документа.

Для правильного хранения важно составить точную модель ЭлД, чтобы не потерять необходимые семантические блоки при миграции данных.

Например, частями оригинала документа для СЭД могут быть:

OrRes – оригиналы листов резолюций документа (эта часть документа и последующие могут возникнуть, например, при передаче документа из СЭД в архив), в общем случае также набор файлов, заверенный множеством ЭП;

OrAgr – оригиналы листов согласований,

OrExe – оригиналы листов исполнения (например, исполнения поручений по документу в СЭД),

OrMet – оригиналы листов ознакомлений.

В этом случае модель оригинала документа может быть записана в следующем виде:

$$\begin{aligned} OrD = ((&U_{(i=1,N1)}(OrDoc_i \, U \, (U_{(j=1,M1)} DSign_{ij}))) \, U \\ &(U_{(i=1,N2)}(OrRes_i \, U \, (U_{(j=1,M2)} RSign_{ij}))) \, U \\ &(U_{(i=1,N3)}(OrAgr_i \, U \, (U_{(j=1,M3)} ASign_{ij}))) \, U \\ &(U_{(i=1,N4)}(OrExe_i \, U \, (U_{(j=1,M4)} ESign_{ij}))) \, U \\ &(U_{(i=1,N5)}(OrMet_i \, U \, (U_{(j=1,M5)} MSign_{ij})))) \\ &U \, SignOrD, \end{aligned} \quad (2.3)$$

где $DSign_{ij}$, $RSign_{ij}$, $ASign_{ij}$, $ESign_{ij}$, $MSign_{ij}$ – *j*-я электронная подпись *i*-й части оригинала документа. *SignOrD* – ЭП контролирующая целостность оригинала документа в целом. Эта подпись, так же как и остальные подписи частей может дополняться «новыми» подписями по мере прохождения процедур инвентаризации ЭП для обеспечения целостности документа.

Можно утверждать, что разбиение документа на подобные семантические блоки оправдано, учитывая разделение прав доступа при выдаче документа из ЭА.

Одной из важнейших операций в ЭА является выдача документов по запросу. При этом необходимо учитывать тот факт, что разные части документа могут иметь разные грифы секретности. Следовательно, выдавать такой документ можно только частями.

В этом случае вопрос о том, как документ хранить и как выдавать пользователю по запросу, должен базироваться на возможности раз-

ного доступа к частям документа. Декомпозиция документа на части при хранении в таком случае неизбежна и оправдана.

Недостатком данной схемы является необходимость контроля целостности всего документа, а не только отдельных частей, иначе части могут быть потеряны.

3.3. Модель нормализованной копии документа

Дано:

1) Множество нормализованных копий $OdfD_i$ ЭлД $D = \{ DAr_i \}$

2) Модель состава семантических блоков верхнего уровня (1.1) ЭлД *DAr*

Найти:

1) Модель нормализованной копии $OdfD_i$ ЭлД $DAr_i \in D$, описывающую состав информации, необходимой для долговременного хранения нормализованной копии ЭлД

Решение:

Пусть нормализованная копия оригинала документа – *OdfD* – неизменяемая часть электронного документа долговременного хранения в ЭлД.

Копия создается в результате процедуры нормализации при приеме оригинала документа в ЭА; *OdfD* заверяется ЭП (в общем случае несколькими) при приеме в ЭА.

Тогда модель нормализованной копии документа:

$$OdfD = (U_{(i=1,N1)} OdfDoc_i) \cup (U_{(j=1,N2)} OdfPic_j) \cup (U_{(k=1,N3)} Sign_k), \quad (2.4)$$

где *OdfDoc* – преобразованное к формату долгосрочного хранения (нормализованное) содержимое частей [1–*N*1] оригинала ЭлД (XML, ODF, PDF/A),

OdfPic – множество [1–*N*2] нормализованной графической информации (растровые и векторные изображения, элементы презентаций и др.), подлежащей преобразованию из сдаваемых документов в графические форматы долгосрочного хранения (TIFF, JPEG, PNG). *OdfDoc* содержит ссылки на графические материалы,

Sign – множество ЭП [1–*N*3], заверяющих нормализованный документ (содержит в себе сертификаты подписавших, цепочку сертификатов, сертификаты удостоверяющих центров (УЦ), возможно СОС);

Можно утверждать, что если оригинал документа логически можно разбить на семантические блоки, как в формуле (2.3), то и нормализации должна подвергаться каждая часть *OrD*.

Следует учесть, что если формат долгосрочного хранения (например, PDF/A) спустя десятилетия (сейчас гарантия поддержки формата со стороны производителя заявлена на 50 лет), то потребуется повторная нормализация оригинала документа.

При этом появится проблема, если формат оригинала документа более не интерпретируется, потребуется провести нормализация формата долговременного хранения. В этом случае, например, юридическая значимость документа может быть поставлена под сомнение.

Чтобы постараться обойти эту проблему, структура семантических блоков и обязательность заверения их по отдельности, и всей нормализованной копии в целом должны повторять структуру деления для оригинала (см. формулу (2.3)).

ЭП нормализованной копии в этом случае должны содержать сведения об авторах документа из ЭП оригинала.

3.4. Модель полнотекстового индекса документа

Дано:

1) Множество ЭлД $D = \{ DAr_i \}$

2) Модель состава семантических блоков верхнего уровня (1.1) ЭлД *DAr*

Найти:

1) Модель полнотекстового индекса ЭлД $DAr_i \in D$

Решение:

Нормализуется текст оригинала документа *OrD*. Нормализованный текст оригинала документа не следует путать с нормализацией ЭлД, т.е. приведением оригинала документа к формату долгосрочного хранения.

Нормализация текста документа – это приведение всех слов оригинала документа к единственному числу, именительному падежу (для существительных), неопределенной форме (глаголов), мужскому роду (прилагательные).

Эта часть документа является необязательной. Однако элементы *FTIdx* предназначены для улучшения качества поисковых процедур. При долговременном хранении важно не поте-

рять полнотекстовый индекс, т.к. переиндексация при переносе ЭлД из одной среды хранения в другую может занимать значительное время.

Модель полнотекстового индекса *FTIdx* ЭлД:

$$FTIdx = <FTWrd_1, \dots, FTWrd_p, \dots, FTWrd_P>, \quad (2.5)$$

где $FTWrd_p$ – элемент (слово) полнотекстового индекса, p=[1,P]. Набор нормализованных слов содержимого документа представляет собой вектор, в общем случае достаточно большой размерности.

Многие промышленные информационные системы и платформы (СУБД) позволяют автоматически построить полнотекстовый индекс, что значительно упрощает индексацию документа в архиве, но требует дополнительного места для хранения индекса.

Чтобы не выполнять полнотекстовую индексацию после миграции данных, что для большого количества документов может занять недопустимо много времени, желательно по возможности переносить ранее сформированный полнотекстовый индекс. Сам индекс не должен изменяться, т.к. оригинал документа при долгосрочном хранении в ЭА не подлежит изменению.

3.5. Модель классификаторов ЭлД

Дано:

1) Множество ЭлД $D = \{ DAr_i \}$

2) Модель состава семантических блоков верхнего уровня (1.1) ЭлД *DAr*

Найти:

1) Модель классификаторов ЭлД $DAr_i \in D$

Решение:

ЭлД не существует в архиве в отрыве от систем классификации, но хранение при каждом документе всех связанных с ним классификаторов приведет к непомерному увеличению роста БД ЭлД. Однако, потеря ценной информации о классификации документа может его обесценить.

Из данного противоречия вытекает необходимость создания архива классификаторов. Как минимум необходимо хранить классификаторы, связанные с подмножеством ЭлД в ЭА, в неизменном состоянии на момент сдачи документа в ЭА. Впрочем, данная задача в виду ее сложности выходит за рамки настоящего исследования.

В корпоративных хранилищах данных, как правило, реализуется только реквизитный и/или полнотекстовый (поиск по тексту электронных документов). Использование классификаторов для упорядочивания корпоративного хранилища данных менее распространено, несмотря на то, что разноплановая классификация важна при хранении больших объемов электронных документов.

Наличие классификаторов – это одна из ключевых особенностей, которая характеризует ЭА. Более того, отсутствие классификации в ЭА противоречит документам, регламентирующим архивную деятельность [3, 8, 9, 17].

Итак, можно выделить следующие классификаторы документов в ЭА и корпоративных хранилищах:

1. Классификатор «дало-том». Это иерархическая структура классификации ЭлД в соответствии с правилами делопроизводства (дело, том или фонды, пачки или др.).

2. Структура организации. Над иерархической структурой хранения (дела, пачки), как правило, существует еще один классификатор, определяющий структуру организации. Тем самым, дела связываются с отдельными подразделениями организации, при этом каждое дело имеет только одно «родительское» подразделение. При реализации в рамках *PAR* удобнее всего объединить эти два классификатора в один древовидный классификатор для обеспечения простоты представления данных в приложениях, распределении прав доступа и т.д.

3. Иерархический классификатор (классификация ручная или автоматизированная согласно заранее выбранной классификации, например, на основе реквизитной информации) или классификация (авторубрикация документов на основе анализа содержимого документа).

4. Метаданные реквизитов поиска. По сути, данный классификатор складывается либо из вектора реквизитов, либо, в более сложном случае, из реквизитов дерева описания данных архивной карточки документа. Путем организации запросов данных с последующей группировкой по различным реквизитам можно динамически получить упорядоченный набор документов.

Каждый из этих классификаторов представляет собой дерево (граф в общем случае). Циклы при такой классификации, как правило, исключены, чтобы не сводить задачу обработки и классификации к задаче существенно более высокой сложности.

Иерархические классификаторы позволяют представлять данные в архиве таким образом, что каждый документ может иметь более одного «родителя» - вершины дерева классификации. Это происходит потому, что, как правило, классификация производится на основе выделенного (или полученного в результате обучения) набора ключевых слов каждого документа. При этом не так редки ситуации, когда вычисленные функции расстояния позволяют отнести документ, как к одной, так и к другой вершине классификатора.

Наличие классификаторов делает корпоративное хранилище полноценным электронным архивом при условии сохранения требования неизменяемости оригиналов документов (оцифрованных копий).

Модель классификаторов документа:

$$CLIdx = < CLIdx_1, \dots , CLIdx_k, \dots , CLIdx_K>, \quad (2.6)$$

где $CLIdx_k$ – элемент классификации ЭлД (k=[1,K]). Существует проблема создания вектора $CLIdx$ для каждого документа, особенно для архивов большого объема.

Частично проблема может быть решена с помощью создания словарей ключевых слов на этапе проектирования электронного архива.

Набор классификаторов документов в архиве можно представить следующим набором:

$$CLSF = <DT, OrgS, HCL, MDS>, \quad (2.7)$$

где $MDS = <MDS_1, \dots , MDS_N>$, где MDS_i – метаданные реквизитов поиска i-го типа документа (i=[1,N]), определяются реквизитами архивной карточки i-го типа документа (в вырожденном случае – одна карточка на все типы документов, если классификацию по типам произвести невозможно);

DT – классификатор «дело-том», представляющий собой лес деревьев как правило высотой 2, верхний уровень – дело, нижний – том (искусственное деление для электронного архива, однако имеющее место в обычном архивном деле для удобства хранения), размещение документов допускается только в томе дела. Наличие данного классификатора обязательно для электронного архива;

OrgS – иерархическая структура организации (как правило объединяется с *DT* для удобства представления в приложениях, работающих с электронным архивом). Наличие данного классификатора, как правило, предполагается в электронном архиве, т.к. дела не «висят в воздухе», а ведутся в определенных подразделениях организации;

HCL – иерархический классификатор (может отсутствовать в электронном архиве), предназначенный для создания альтернативной *OrgS-DT* классификации документов.

Основная проблема использования классификаторов состоит в необходимости автоматизации привязки электронных документов к классам. Для решения данной проблемы существует несколько подходов: первый заключается в написании правил отнесения документов к классам, второй – в использовании машинного обучения.

В случае первого подхода результаты классификации сильно зависят от компетентности специалиста, описывающего правила. Кроме того, временная емкость этой операции классификации достаточно высокая.

В случае использования машинного обучения, требования к квалификации специалиста значительно меньше, временные затраты при этом сильно зависят от наличия множества электронных документов для составления обучающей выборки.

На практике наиболее разумным оказывается комбинирование обоих подходов к решению проблемы автоматизации отнесения электронных документов к классам. Подробнее о принципах построения и обучения классификаторов на примере разработанной информационно-аналитической системы «Астарта» показано в [1, 2]. В разработке ИАС «Астарта» авторы данного исследования принимали непосредственное участие.

3.6. Модель связок документа с другими документами

Дано:

1) Множество ЭлД $D = \{ DAr_i \}$

2) Модель состава семантических блоков верхнего уровня (1.1) ЭлД *DAr*

Найти:

1) Модель связок ЭлД DAr_i с другими ЭлД из множества D

Решение:

Как правило, документ связан с другими документами, однако хранить всю семантическую сеть документов вряд ли целесообразно. В этом случае будет сложно организовать выдачу документа по запросу, т.к. придется корректно «откреплять» его от семантической сети.

Чтобы документ оставался отдельной единицей хранения, и в то же время был связан с другими документами, необходимо предусмотреть семантические блоки в метаданных документа, в которых хранилась бы идентифицирующая информация связанного документа. В этом случае для поиска связанного документа необходимо использовать данную идентифицирующую информацию.

Данная часть является необязательной частью документа (связей может и не быть). Вектор связок может меняться, если в процессе работы с архивными документами выяснится их связанность с другими документами.

Модель связок документа с другими ЭлД:

$$LDIdx = <LDIdx_1, \dots, LDIdx_q, \dots, LDIdx_Q>, \quad (2.8)$$

где $LDIdx_q$ – элемент вектора связок документа с другими документами (q=[1,Q]).

$$LDIdx_q = IdInfoLDoc \cup DSign$$

Элемент вектора связок представляет собой идентифицирующую информацию связанного документа *IdInfoLDoc*, которую, скорее всего, необходимо заверить ЭП пользователя ЭА (*DSign*), установившего связь либо же архивной ЭП, если связи устанавливаются автоматически по каким-то признакам.

Идентифицирующая информация связанного документа должна включать ключевые реквизиты документа, извлеченные из его оригинала и не должна включать специальные идентификаторы (ключи) базы данных, искусственно создаваемые при хранении документа. В последнем случае при миграции, ключи могут быть автоматически изменены, что приведет к потере информации о связях документа.

3.7. Модель выписки из журнала инвентаризации ЭП

Дано:

1) Множество ЭлД $D = \{ DAr_i \}$

2) Модель состава семантических блоков верхнего уровня (1.1) ЭлД *DAr*

Найти:

1) Модель выписки из журнала инвентаризации ЭлД $DAr_i \in D$

Решение:

Пусть выписка из журнала инвентаризации ЭП – *OperD* – состоит из отдельных записей, заверенных архивной ЭП (вычисляется автоматически или устанавливается оператором ЭА).

Это обязательная, изменяющаяся часть документа. Как минимум состоит из единственной записи – проверке ЭП при приеме в ЭА.

Каждая инвентаризация ЭП фиксируется в журнале инвентаризации и заверяется ЭП. При миграции выписка целиком также должна быть заверена ЭП для контроля целостности журнала инвентаризации.

Выписка с течением времени может изменяться, т.к. очередная инвентаризация будет добавлять запись в журнал. Заверение ЭП записи журнала производится автоматически или оператором, выполняющим инвентаризацию.

Математическая модель представляет собой вектор записей о проведенных инвентаризациях:

$$OperD = <OperD_1, \dots, OperD_s, \dots, OperD_S>, \quad (2.9)$$

где $OperD_s$ – элемент вектора – одна запись о проведенной инвентаризации (s=[1,S]), а $OperD_s = OperInfo \cup DSign$.

Элемент вектора представляет собой информацию об инвентаризации (*OperInfo*), которая заверена ЭП (набором ЭП) оператора ЭА и членов инвентаризационной комиссии (*DSign*), проводивших инвентаризацию. ЭП может вычисляется автоматически. Информация об инвентаризации включает в себя: сведения о процедуре, в рамках которой выполняется инвентаризация (миграция, проверка ЭП, перезаверение новой ЭП с сохранением авторства старых ЭП (подробно [15]), инвентаризация носителей), сведения о сертификатах и ЭП, в том числе об авторах ЭП. Сведения о новых ЭП и причина замены ЭП, результата проверки ЭП.

OperInfo содержит время проведения инвентаризации, состав комиссии, результат инвентаризации.

OperD относится к метаданным документа. Подлежит рассмотрению вне ЭА только при отрицательном результате инвентаризации.

Аналогично (как и *OperD*) описываются модели выписки из журналов инвентаризации носителей, инвентаризации интерпретации и др.

Заключение

В данной статье представлена разработанная автором исследования математическая модель документа в ЭА долговременного хранения, а также методологический подход к созданию математических моделей документа при долговременном хранении.

Можно утверждать, что эта модель документа является необходимой для обеспечения моделирования долговременного хранения электронного документа.

Действительно, как показано в [12, 13, 16] при долговременном хранении существует целый набор проблем, например сохранения аутентичности, интерпретируемости, «старения» носителей информации, изменения программно-аппаратной среды хранения, надежности и устойчивости среды хранения (рассмотрение данных проблем выходит за рамки данной статьи). Кроме того, существует проблема использования документа, при решении которой может потребоваться разный уровень доступа к частям документа.

Для решения проблемы аутентичности должны быть применены методы шифрования, гарантирующие неизменность. Таким методом в РФ, безусловно, является использование ЭП с помощью сертифицированных средств криптозащиты в качестве «незаинтересованной», независимой компоненты защиты данных. Подробнее решение данной частной проблемы показано в статье автора [15]. Чтобы решить проблему использования, необходима декомпозиция единого документа на семантические блоки – части, которые могут быть использованы по отдельности.

Учитывая, что срок действия ЭП ограничен по времени из-за причин указанных выше, необходима периодическая инвентаризация ЭП. Следовательно, данные об инвентаризации тоже должны сохраняться и мигрировать вместе с документом.

Проблема интерпретируемости должна быть решена с помощью нормализации документа и хранения вместе с оригиналом нормализованной копии, также заверенной ЭП. Причем в ЭП включаются сведения об авторах оригинала документа.

Таким образом, будет достигнуто практическое решение проблемы интерпретируемости документа.

Для решения проблем «старения» носителей и обновления программно-аппаратной среды ЭА необходимо производить миграцию документов ЭА. Для этого документ должен быть максимально откреплен от среды хранения. Для решения проблем надежности и устойчивости необходим отдельный комплекс технологий, предложенный автором исследования, не рассматриваемый в данной статье.

Т.е. все необходимые сведения о документе должны сохраняться вместе с документом, не быть привязанными к среде хранения ЭА.

Таким образом, достигается решение проблемы отчуждаемости от конкретной программно-аппаратной реализации, как степень соответствия представленной модели электронного документа в ЭА (модели (1.1), (2.1) – (2.9)) текущему представлению документов в ЭА.

Для решения этих проблем в модель документа введены метаданные, позволяющие документировать операции с документом, а также связанность документа с системой классификации, другими документами, индексами (пп. 3.4-3.6).

Следовательно, выполняя требования представленной выше математической модели документа, можно с большой вероятностью утверждать, что поставленные в начале исследования проблемы долговременного хранения будут решены.

Литература

1. Акимова, Г.П. Аналитический подход к решению задачи мониторинга информационного пространства / Г.П. Акимова, М.А. Пашкин // Журнал «Системы высокой доступности». – 2006 – №3-4, т.2 – С.44-50.
2. Акимова, Г.П. Современные автоматизированные технологии обработки разнородных информационных потоков / Г.П. Акимова, Д.С. Богданов, И.В. Мусатов,

М.А. Пашкин, Д.В. Солдатов, Н.В. Сомин // «Организационное управление и искусственный интеллект» Сборник трудов Института системного анализа РАН – 2003 – С.290-304.
3. ГОСТ Р 51141-98 Делопроизводство и архивное дело. Термины и определения (утвержден Постановлением Госстандарта РФ № 28 от 27 февраля 1998 г.).
4. Емельянов, Н.Е. Виды представления структурированных данных / Н.Е. Емельянов // Теоретические основы информационной технологии / Сб. тр. ВНИИСИ – 1988 – Вып. 22. — М.:ВНИИСИ, 1988 – С.42–46.
5. Емельянов, Н.Е. Генерация информационных систем по формам входных и выходных документов / А.Н. Богачева, Н.Е. Емельянов, А.П. Романов // PC Magazine. – 1993 – №1 – С.85-89.
6. Емельянов, Н.Е. Применение новых информационных технологий в делопроизводстве / Г.П. Акимова, А.С. Богданов, Н.Е. Емельянов, А.В. Соловьев, В.А. Тищенко //Развитие безбумажной технологии в информационных системах. Сборник трудов Института системного анализа РАН/Под ред. д.т.н., проф. Арлазарова В. Л. и д.т.н., проф. Емельянова Н. Е. — М.: Эдиториал УРСС, — 1999. — С. 17—27.
7. Емельянов, Н.Е. Теоретический анализ документного интерфейса: Препринт / Н.Е. Емельянов — М.: Всесоюзный научно-исследовательский институт системных исследований. — 1987 – 40с.
8. Правила организации хранения, комплектования, учета и использования документов Архивного фонда РФ и других архивных документов в государственных и муниципальных архивах, музеях и библиотеках, организациях Российской академии наук. Утверждены приказом Министерства культуры и массовых коммуникаций Российской Федерации № 19 от 18.01.2007.
9. Приказ Министерства культуры и массовых коммуникаций Российской Федерации № 536 от 8 ноября 2005 г. «О Типовой инструкции по делопроизводству в федеральных органах исполнительной власти».
10. Соловьев, А.В. Построение баз данных взаимосвязанных документов / А.Н. Белова, А.В. Соловьев // Труды Института системного анализа РАН (ИСА РАН) – 2012 – т.62, вып.3 – С.25-30.
11. Соловьев, А.В. Разработка методов и средств взаимодействия объектно-ориентированных систем управления базами данных с электронными издательскими комплексами: диссертация на соискание ученой степени кандидата технических наук: 05.13.10 / А.В. Соловьев ИСА РАН – 2000 – 130 с.
12. Соловьев, А.В. Проблемы долгосрочного хранения электронных деловых документов / Г.П. Акимова, Е.В. Пашкина, М.А. Пашкин, А.В. Соловьев // Журнал «Делопроизводство». – 2014 – №1 – С.96-104.
13. Соловьев, А.В. Проблемы долгосрочного хранения электронных документов / Г.П. Акимова, Е.В. Пашкина, М.А. Пашкин, А.В. Соловьев // Труды XX Международной научно-практической конференции «Документация в информационном обществе: эффективное управление электронными документами» (Москва, РГАСПИ, 20-21 ноября 2013 г.) – 2014 – С.396-400.
14. Соловьев, А.В. Электронные архивы: возможные решения проблем долгосрочного хранения данных / Г.П. Акимова, Е.В. Пашкина, М.А. Пашкин, А.В. Соловьев // Труды Института системного анализа РАН (ИСА РАН) – 2013 – т.63, вып.4 – С.39-49.
15. Соловьев, А.В. Решение проблем оценки и сохранения аутентичности электронных документов при долговременном хранении / А.В. Соловьев // Журнал «Системы высокой доступности». – 2014 – №4, т.10 – С.99-106.
16. Соловьев, А.В. Электронные архивы: о постановке задачи долговременного хранения электронных документов / А.В. Соловьев // Информационные технологии и вычислительные системы. – 2014 – №4 – С.74-78.
17. Федеральный закон от 22.10.2004 № 125-ФЗ «Об архивном деле в Российской Федерации».
18. Miller, J. NARA to suspend development of ERA starting in 2012 [Электронный ресурс] / J. Miller – 2012 – Режим доступа: FederalNewsRadio.com http://www.federalnewsradio.com/?sid=2204570&nid=35
19. Рысков, О.И. Основные направления деятельности национальных архивов США и Соединенного Королевства Великобритании и Северной Ирландии в области управления электронными документами правительственных учреждений / О.И. Рысков // Отечественные архивы. – 2004. - № 3.
20. Carlstrom, G. Is DoD's new pay system fair? [Электронный ресурс] / G. Carlstrom // FederalTimes.com – 2008 – Режим доступа: http://federaltimes.com/index.php?S=3502888
21. Lipowicz, A. NARA officials defend searchability of electronic archive [Электронный ресурс] / A. Lipowicz // Federal Computer Week. – 2011 – Режим доступа: http://fcw.com/articles/2011/11/01/nara-officials-defending-searchability-of-electronic-archive.aspx.
22. Блог Национальных Архивов США [Электронный ресурс] – 2013 – Режим доступа: http://blogs.archives.gov/records-express/2013/11/01/opportunity-for-comment-transfer-guidance-bulletin/
23. Рысков, О.И. Об основных направлениях деятельности зарубежных архивных органов в области исследования и нормативного регулирования работы с электронной документацией / О.И. Рысков // Секретарское дело. – 2005. - № 3. – С.76.
24. Храмцовская, Н.А. Сравнение подходов к обеспечению долговременной сохранности электронных материалов, часть III. [Электронный ресурс] / Ангеваape, И., перевод Храмцовская Н.А. – 2012 – Режим доступа: http://rusrim.blogspot.com/2012/09/iii.html.
25. Стандарт MoReq 2008 года. [Электронный ресурс] – 2008 – Режим доступа: MoReq2 Collateral Website (http://www.moreq2.eu/).
26. Типовые требования к автоматизированным системам электронного документооборота. Спецификация MoReq // Office for Official Publications of the European Communities as INSAR Supplement VI, ISBN 92-894-1290-9.
27. Hutar, J. Rendering Matters [Электронный ресурс] / J. Hutar – 2012 – Режим доступа: http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering.

28. Preservation of Evidence of Cryptographically Signed Documents // BSI Technical Guideline TR-03125 – Version 1.1 – Federal Office for Information Security – 2011 – 111 p.
29. Храмцовская, Н.А. Южная Корея: Стандарт функциональных требований к системе управления архивными документами – и не только [Электронный ресурс] – 2010 – Режим доступа: http://rusrim.blogspot.ru/2010/11/blog-post_12.html.
30. Про особливості работи з електронними документами в Україні // «Секретарь-референт» - 2013 - №1 (121) - С.24-29.
31. Das Archiv muss zu den Leuten gehen [Электронный ресурс] // Das Zürcher Staatsarchiv zügelt ins Internet. Neue Zürcher Zeitung Online. – 2014 – Режим доступа: (http://www.nzz.ch/nachrichten/medien/das_archiv_muss_zu_den_leuten_gehen_1.740545.html?printview=true)
32. Statens Arkiver. Format- og Strukturkonverteringsprojektet. [Электронный ресурс] – 2013 – Режим доступа: http://www.sa.dk/content/dk/forskning_og_udvikling/udviklingsprojekter/format-_og_strukturkonverteringsprojektet.
33. Виладсен, К.К. Конверсия коллекции электронных документов датских Национальных Архивов. [Электронный ресурс] – 2008 – Режим доступа: http://www.dlm2008.com/img/pdf/villadesn_ab_gb.pdf.

Соловьев Александр Владимирович, Заместитель директора по научной работе ИСА ФИЦ ИУ РАН. Окончил МГТУ им. Н.Э. Баумана в 1994 году. Доктор технических наук. Количество печатных трудов: 51. Область научных интересов: системный анализ, системы управления базами данных, теория надежности, математическое моделирование, электронный документооборот, электронный архив, долговременное хранение электронных документов. E-mail: soloviev@isa.ru

Electronic archives: development of mathematical models of electronic documents for long-term storage

A.V. Solovyev

Abstract. The article discusses the development of a mathematical model of an electronic document for long-term storage. The article defines the desired composition and structure of information for long-term storage of electronic documents. We present decomposition model developed by an electronic document to highlight components of general information. This article is intended to create a theoretical basis for long-term storage of electronic documents.

Keywords: Electronic document management, electronic archives, electronic document management system, electronic document, long-term storage.

References

1. Akimova, G.P. Analytical approach to solving the problem of monitoring of information space / G.P. Akimova, M.A. Pashkin // High availability systems. – 2006 – №3-4, Part.2 – P.44-50.
2. Akimova, G.P. Modern automated processing technology of heterogeneous information flows / G.P. Akimova, D.S. Bogdanov, I.V. Musatov, M.A. Pashkin, D.V. Soldatov, N.V. Somin // Organizational control and artificial intelligence – 2003 – P.290-304.
3. GOST R 51141-98 Records management and archiving. Terms and definitions.
4. Emelyanov, N.E. Types of representation of structured data / N.E. Emelyanov // The theoretical foundations of information technology – 1988 – Part.22. — M.:ВНИИСИ, 1988 – P.42–46.
5. Emelyanov, N.E. Generation of information systems according to the forms of input and output documents / A.N. Bogacheva, N.E. Emelyanov, A.P. Romanov // PC Magazine. – 1993 – №1 – P.85-89.
6. Emelyanov, N.E. The application of new information technologies in administration / G.P. Akimova, A.S. Bogdanov, N.E. Emelyanov, A.V. Solovyev, V.A. Tischenko // The development of paperless technology in information systems. The collection of works of Institute of system analysis RAS — M.: Editorial URSS, — 1999. — P.17—27.
7. Emelyanov, N.E. Theoretical analysis of the document interface: Preprint / N.E. Emelyanov — M.: All-Union scientific research Institute for system studies. — 1987 – 40p.
8. The rules of the organization of storage, acquisition, accounting and use of documents of Archival Fund of the Russian Federation and other archival documents in state and municipal archives, museums and libraries, institutions of the Russian Academy of Sciences. Approved by order of the Ministry of culture and mass communications of the Russian Federation № 19 18.01.2007.
9. Order of the Ministry of culture and mass communications of the Russian Federation № 536 8 November 2005. «About the standard instruction on records management in the Federal bodies of Executive power».
10. Solovyev, A.V. The construction of a database of interlinked documents / A.N. Belova, A.V. Solovyev // Proceedings of Institute of system analysis RAS (ISA RAS) – 2012 – T.62, Part.3 – P.25-30.

11. Solovyev, A.V. Development of methods and means of interaction of object-oriented database management systems with electronic publishing complexes: the dissertation on competition of a scientific degree of candidate of technical Sciences: 05.13.10 / A.V. Solovyev ISA RAS – 2000 – 130p.
12. Solovyev, A.V. Problems long term storage of electronic business documents / G.P. Akimova, E.V. Pashkina, M.A. Pashkin, A.V. Solovyev // The Journal "Records Management". – 2014 – №1 – P.96-104.
13. Solovyev, A.V. Problems long term storage of electronic documents / G.P. Akimova, E.V. Pashkina, M.A. Pashkin, A.V. Solovyev // Proceedings of the XX International scientific and practical conference "Documentation in information society effective management of electronic documents" (Moscow, 20-21 November 2013) – 2014 – P.396-400.
14. Solovyev, A.V. Electronic archives: possible solutions to the problems of long-term data storage / G.P. Akimova, E.V. Pashkina, M.A. Pashkin, A.V. Solovyev // Proceedings of Institute of system analysis RAS (ISA RAS) – 2013 – T.63, Part.4 – P.39-49.
15. Solovyev, A.V. The problems of assessment and conservation of authenticity of electronic documents for long term storage / A.V. Solovyev // High availability systems. – 2014 – №4, Part.10 – P.99-106.
16. Solovyev, A.V. Electronic archives: the formulation of long-term storage of electronic documents / A.V. Solovyev // Information technology and computer systems. – 2014 – №4 – P.74-78.
17. Federal law of the Russian Federation 22.10.2004 № 125-FZ «About the archival affair in Russian Federation».
18. Miller, J. NARA to suspend development of ERA starting in 2012 [Electronic resource] / J. Miller – 2012 – Access mode: FederalNewsRadio.com. http://www.federalnewsradio.com/?sid=2204570&nid=35
19. Ryskov, O.I. The main activities of the national archives of the United States and the United Kingdom of great Britain and Northern Ireland to the field of electronic document management of government agencies / O.I. Ryskov // Domestic archives. – 2004. - № 3.
20. Carlstrom, G. Is DoD's new pay system fair? [Electronic resource] / G. Carlstrom // FederalTimes.com – 2008 – Access mode: http://federaltimes.com/index.php?S=3502888
21. Lipowicz, A. NARA officials defend searchability of electronic archive [Electronic resource] / A. Lipowicz // Federal Computer Week. – 2011 – Access mode: http://fcw.com/articles/2011/11/01/nara-officials-defending-searchability-of-electronic-archive.aspx.
22. Blog the National Archives of the United States [Electronic resource] – 2013 – Access mode: http://blogs.archives.gov/records-express/2013/11/01/opportunity-for-comment-transfer-guidance-bulletin/
23. Ryskov, O.I. About the main directions of activities of foreign archival authorities in the field of research and regulatory work with electronic documentation / O.I. Ryskov // Secretarial work. – 2005. - № 3. – P.76.
24. Chramzovskaya, N.A. Comparison of approaches to ensure long-term preservation of electronic materials, part III. [Electronic resource] / Angevaare, I., translation by Chramzovskaya N.A. – 2012 – Access mode: http://rusrim.blogspot.com/2012/09/iii.html.
25. Standard MoReq 2008. [Electronic resource] – 2008 – Access mode: MoReq2 Collateral Website (http://www.moreq2.eu/).
26. Typical requirements for automated electronic document management systems. Specification MoReq // Office for Official Publications of the European Communities as INSAR Supplement VI, ISBN 92-894-1290-9.
27. Hutar, J. Rendering Matters [Electronic resource] / J. Hutar – 2012 – Access mode: http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering.
28. Preservation of Evidence of Cryptographically Signed Documents // BSI Technical Guideline TR-03125 – Version 1.1 – Federal Office for Information Security – 2011 – 111 p.
29. Chramzovskaya, N.A. South Korea: Standard functional requirements to the control system archival documents – and not only [Electronic resource] – 2010 – Access mode: http://rusrim.blogspot.ru/2010/11/blog-post_12.html.
30. Of features of work with electronic documents in Ukraine // «Secretary-referent» - 2013 - №1 (121) - P.24-29.
31. The archive must go to the people [Electronic resource] // The Zurich state archives curbs on the Internet. Neue Zürcher Zeitung Online. – 2014 – Access mode: (http://www.nzz.ch/nachrichten/medien/das_archiv_muss_zu_den_leuten_gehen_1.740545.html?printview=true)
32. The State Archives. Format and Strukturkonverteringsprojektet [Electronic resource] – 2013 – Access mode: http://www.sa.dk/content/dk/forskning_og_udvikling/udviklingsprojekter/format-_og_strukturkonverteringsprojektet.
33. Viladsen, K.K. The conversion of the collection of electronic documents to the Danish National Archives. [Electronic resource] – 2008 – Access mode: http://www.dlm2008.com/img/pdf/villadesn_ab_gb.pdf.

Solovyev Alexandr Vladimirovich Deputy Director ISA FRC CSC RAS. BMSTU 1994. Number of publications: 54. Area of scientific interests: system analysis, database management system, reliability theory, mathematical modeling, electronic document management, electronic archive, long-term storage of electronic documents. E-mail: soloviev@isa.ru