Информационные технологии

Архивные хранилища и электронные архивы документов, основные постулаты и проблемы разработки

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

Аннотация. В статье изложены основные понятия и определения, характерные для систем электронного архива. Представленная работа является обобщением опыта создания электронных архивных систем и содержит описание проблем и подводных камней, с которыми столкнулись авторы и которые могут возникнуть при проведении аналогичных исследований. Приведенные основы методологии количественной оценки качества функционирования электронных архивных систем, в том числе и территориально-распределенных, могут быть полезны при проектировании систем электронного архива.

Ключевые слова: *электронный документооборот, системы управления содержимым, электронный архив, система управления электронными документами, электронный документ, электронный архивный документ, автоматизация архивного дела.*

Введение

В современном производстве программного обеспечения выделились классы систем, обеспечивающие оперативную работу с электронными документами в организации, такие как системы электронного документооборота (СЭД), электронные системы хранения финансовых документов, информационно-аналитические системы (ИАС) и др. Системы такого рода обычно предназначены для автоматизации различных бизнес-процессов делопроизводства, хранения и анализа электронных документов в организации. Несомненно, что неотъемлемой частью таких процессов является архивное дело. Сравнительно недавно начавшийся бум внедрения СЭД, ИАС и др. 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 является вполне объяснимым, поскольку указанный срок — это среднее время хранения документов в «оперативных» архивах или в БД СЭД до их массовой передачи в вышестоящие архивы. Не менее важным фактором является отсутствие насыщения потока поступления документов в архив, определяющих массовость. В результате сложилась ситуация, что для электронных архивов до сих пор не определены базовые термины, не описаны основные характеристики, слабо развита теоретическая составляющая, порой отсутствует нормативная база.

В предлагаемой статье обобщены имеющиеся в настоящее время мнения о том, что представляет собой полноценный электронный архив, что является

единицей хранения в архиве, каким образом можно классифицировать хранящиеся в нем документы, в т. ч. для облегчения поиска и аналитической обработки данных. Выделены проблемы и «подводные камни», которые могут возникнуть при разработке электронных архивов, предложены критерии, позволяющие оценить их эффективность. Авторами реализовано достаточно большое количество электронных архивных систем, и в статье подведен итог многолетнего опыта разработки и внедрения таких систем.

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).

Особенности электронных архивов. Документы, составляющие основу электронного архива, 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 ведется в соответствии с правилами, оговоренными нормативными документами [1, 2, 6, 7].

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 разноформатные документы, которые могут быть структурированы. Как правило, такие хранилища позволяют включать и удалять документы (а также прочие информационные ресурсы и файлы), находящиеся в доступе в конкретной организации, в т. ч. в различных ее информационных системах. Единая классификация 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. На рис. 1 показано место Электронного архива и архивного хранилища среди информационных систем.

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

Рис. 1. Место электронного архива и архивного хранилища среди классов информационных систем

ECMS подразделяется на несколько классов информационных систем, таких как СЭД, кадровые системы, CRM (системы взаимодействия с клиентами) 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 т. ч. книг), снабженных средствами навигации и поиска.

Система автоматизации документооборота, **система электронного документооборота** — 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 — 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.

2. Документ в электронном архиве и корпоративном хранилище данных

Для представления документа в электронном архиве и корпоративном хранилище данных введем определение: документ в электронном архиве представляет собой граф (дерево), состоящий из взаимосвязанных семантических блоков B_i. Блоки представляют собой подграфы (поддеревья), также состоящие из семантических блоков следующего уровня.

Действительно, в любом документе всегда можно выделить заголовок, подзаголовки, повторяющиеся части, агрегаты (массивы, структуры данных), атомарные данные (листы дерева). Между документами могут существовать отношения, т. е. лес документов может быть связан в единый граф. При этом в вершинах деревьев можно указывать неявные связи с другими документами.

Так же для документа в ЭА может быть выделена иерархия семантических блоков, каждый из которых несет вполне определенную смысловую нагрузку с точки зрения ЭА. Документ в ЭА может быть представлен следующими семантическими блоками верхнего уровня:

$$\mathrm{DAr} = \sum_{i=1}^{N} (B_i) = \mathrm{ArCard} + \mathrm{OrD} + \mathrm{FTIdx} + \mathrm{CLIdx}, \qquad (1)$$

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 из них, 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.;

OrD — оригиналы документов (электронные оригиналы документов или оцифрованные изображения оригинальных бумажных документов, которые далее также будем обозначать как оригиналы) — неизменяемая часть электронного документа;

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;

CLIdx — вектор связок между электронным документом и классификаторами $<\mathrm{CLIdx}_1, \ldots, \mathrm{CLIdx}_k, \ldots, \mathrm{CLIdx}_K>$ ($k = 1, K$) — изменяемая часть электронного документа, т. к. набор связок может изменяться или дополняться. Является необязательной частью документа.

2.1. Архивная карточка документа

$$\mathrm{ArCard} = \mathrm{ArCardCM} + \mathrm{ArCardF} + \mathrm{ArCardC}, \qquad (2)$$

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

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

ArCardC — содержимое архивной карточки (значения реквизитов или дерево описания данных, заполненное значениями данных).

2.2. Оригиналы документов

$$\mathrm{OrD} = \sum_{i=1}^{N} (\mathrm{OrDoc}_i + \sum_{j=1}^{M_1} \mathrm{Sign}_{ij}), \qquad (3)$$

где OrD — оригинал документа (файл, оцифрованная копия). Каждый документ имеет свой граф (обычно дерево) описания данных, называемый моделью содержания (в иностранной литературе — ContentModel). 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 представле-

на отдельной оцифрованной копией), каждый из которых заверен ЭП[1];

$Sign_{ij}$ — 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) — неизменяемая часть электронного документа.

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 формула для оригиналов документов может быть записана в следующем виде:

$$\mathrm{OrD} = \sum_{i=1}^{N_1} (\mathrm{OrDoc}_i + \sum_{j=1}^{M_1} \mathrm{DSign}_{ij}) + \sum_{i=1}^{N_2} (\mathrm{OrRes}_i + \sum_{j=1}^{M_2} \mathrm{RSign}_{ij}) + \sum_{i=1}^{N_3} (\mathrm{OrAgr}_i + \sum_{j=1}^{M_3} \mathrm{ASign}_{ij}) + \sum_{i=1}^{N_4} (\mathrm{OrExe}_i + \sum_{j=1}^{M_4} \mathrm{ESign}_{ij}) + \sum_{i=1}^{N_5} (\mathrm{OrMet}_i + \sum_{j=1}^{M_5} \mathrm{MSign}_{ij}),$$

где DSign_{ij}, RSign_{ij}, ASign_{ij}, ESign_{ij}, MSign_{ij} — j-я электронная подпись i-й части оригинала документа.

2.3. Полнотекстовый индекс

$$\mathrm{FTIdx} = <\mathrm{FTWrd}_1, \ldots, \mathrm{FTWrd}_p, \ldots, \mathrm{FTWrd}_P>, \qquad (4)$$

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.

2.4. Связи между электронным документом и классификаторами

$$\mathrm{CLIdx} = <\mathrm{CLIdx}_1, \ldots, \mathrm{CLIdx}_k, \ldots, \mathrm{CLIdx}_K>,$$

где CLIdx_k — элемент вектора связок документа и классификаторов. Существует проблема создания такого вектора для каждого документа, особенно для архивов большого объема. Частично проблема решается с помощью создания словарей ключевых слов на этапе проектирования электронного архива.

3. Классификаторы в электронном архиве и архивном хранилище

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 одно из ключевых особенностей, которая характеризует ЭА. Более того, отсутствие классификации в ЭА противоречит документам, регламентирующим архивную деятельность, о которых говорилось в начале статьи.

Итак, какие классификаторы документов можно выделить в ЭА и корпоративных хранилищах.

1. Иерархическая структура хранения данных в соответствии с правилами делопроизводства (дело, том или фонды, пачки или др.).
2. Над единицей хранения (дела, пачки), как правило, существует еще один классификатор, определяющий иерархию структуры организации. Тем самым, единицы хранения связываются с отдельными подразделениями организации, при этом каждое дело имеет только одно «родительское» подразделение. Удобнее всего объединить эти два классификатора в единый 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).

[1] ЭП (электронная подпись) — информация в электронной форме, присоединенная к другой информации в электронной форме (электронный документ) или иным образом связанная с такой информацией. Используется для определения лица, подписавшего электронный документ [15]. 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.

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).

Набор классификаторов документов в архиве можно представить следующим набором:

$$CLSF = <DT, OrgS, HCL, MDS>, \quad (5)$$

где $MDS = <MDS_1, \ldots, MDS_N>$, при этом 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 дела тома, 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, это затратная операция по времени. 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 [13, 14].

4. Проблемы, сдерживающие развитие ЭА

При разработке и эксплуатации ЭА эксплуатирующие организации и разработчики ЭА сталкиваются с целым рядом фундаментальных проблем, наличие которых тормозит развитие ЭА. Разрешение данных проблем может стать одним из факторов продвижения систем архивных хранилищ и ЭА на рынке программного обеспечения (ПО).

4.1. Проблема потокового ввода

Одним из главных препятствий, сдерживающих развитие электронных архивов, является проблема организации потокового ввода огромного количества документов для наполнения БД архива.

Дело в том, что к началу внедрения электронного архива в организации, как правило, существует архив бумажных документов, занимающий подчас достаточно большие площади. При использовании системы электронного архива встает проблема оцифровки бумажных документов и размещения их в БД ЭА, т. к. технические и программные средства, позволяющие провести данную работу, дорогостоящие и под силу только крупным организациям. Огромный объем трудозатрат или дополнительных расходов (некоторые разработчики электронных архивов предлагают услуги оцифровки) часто останавливает внедрение электронных архивных систем в организациях.

Несколько проще обстоит дело с электронными документами, однако, как правило, на момент внедрения электронного архива в организации обычно

уже есть несколько разнообразных информационных систем, в которые введены документы различных форматов (бухгалтерия, кадры, СЭД, офисные приложения и т. д.). Вводить каждый электронный документ по отдельности в ЭА тоже дело затратное.

Однако, в отличие от оцифровки бумажных оригиналов, в этом случае имеется решение, которое устроило бы большинство организаций, внедряющих ЭА. Речь идет об использовании функций потокового автоматического ввода. Такие функции позволили бы максимально автоматизировать не только процесс первичного наполнения БД ЭА, но и ввод новых документов. При наличии такого решения можно организовать автоматическую загрузку документов в БД ЭА.

Имеющиеся сейчас программные средства позволяют удовлетворительно решать большинство задач по организации потокового ввода и продолжают свое развитие. Однако, как правило, настройка этих средств производится только высококвалифицированным персоналом, которого в организации, внедряющей ЭА, нет, что влечет за собой значительные расходы на привлечение сторонних специалистов.

Основное развитие исследований по данной проблеме, как мы считаем, должно лежать в области упрощения процедуры настройки программных средств автоматического ввода на подлежащие переводу в ЭА данные, что позволит сильно сократить расходы на первоначальное наполнение ЭА.

4.2. Проблема долгосрочного хранения

Еще одной проблемой, тормозящей развитие ЭА, является проблема долгосрочного хранения.

Срок хранения документа в архиве зависит от типа документа, некоторые документы хранят бессрочно, другие — десятилетиями или годами. Для поддержания работоспособности архивов электронных документов длительного хранения необходима не только надежная техника, но и гарантия неизменности документа в течение срока хранения.

Для гарантии неизменности должны применяться как организационные меры, так и программные средства. К организационным мерам обычно относят защиту оборудования от несанкционированного доступа непосредственно к серверным стойкам, коммуникационному оборудованию и пр. К программным средствам относят разграничение прав доступа на электронные документы и организацию контроля целостности, который реализуется с помощью хранения хешей электронных документов или использования ЭП. В таком решении ЭП автоматически устанавливается программными средствами ЭА при вводе электронного документа в БД. При необходимости, например, при истечении срока действия ключа, все электронные документы в БД переподписываются новым ключом ЭП. Надо понимать, что такая схема не исключает подмены документов административным персоналом, эксплуатирующим ЭА.

Следующей проблемой ЭА является проблема обеспечения юридической значимости электронных документов. На настоящий момент ключевым решением данной проблемы является использование ЭП. Однако сертификаты и открытые ключи ЭП обладают ограниченным сроком действия, поэтому спустя год или 5 лет ЭА может выдать сообщение о некорректной ЭП, что поставит под сомнение подлинность документа.

Кроме того, 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 (в 2009 г. была взломана ЭП с 768-битным ключом, но пока это возможно 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.

Кроме того, 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 прогнозировать размер БД и возможный ее рост, чтобы предоставить адекватные средства защиты информации.

Основной проблемой при использовании ЭП является возникновение недействительных ЭП по истечении срока действия сертификатов подписи, ключей подписи, отзыва сертификата, отбраковки сертификата (признание удостоверяющим центром (УЦ) сертификата недействительным или утратившим доверие) и др.

Проработан данный вопрос достаточно слабо. Существуют так называемые усовершенствованные (усиленные) ЭП, содержащие подтвержденную метку времени. Это дает возможность безошибочной проверки ЭП даже после истечения срока действия сертификата (максимально разрешенный по закону срок действия сертификата — 6 лет), в частности, таким способом решается коллизия второго рода. Однако остается проблема отозванных сертифика-

тов, и здесь вопрос достаточно сложный: считать ли сертификат недействительным вообще или только с момента отзыва? Кроме того, через 30 лет станет недействительным и ключ ЭП (максимальный срок действия). В этом случае со временем достаточно сложно будет доказать юридическую значимость документа. Данная причина также тормозит как развитие долгосрочных ЭА, так и применение ЭП в принципе.

Следующая проблема появляется из-за сложности взаимодействия ЭА с УЦ. Она возникает, если ЭА хранит электронные документы, подписанные ЭП, которые выданы разными УЦ, в т. ч. в различных регионах РФ. В таком случае часто ЭА не может проверить ЭП, кроме того, нет никаких гарантий хранения сертификатов ЭА самими УЦ. На данный момент решения данной проблемы нет. В качестве одного из промежуточных решений предлагают в ЭА организовать хранение всех сертификатов, списки отзыва сертификатов (СОС) и много дополнительной информации, на основании которой может быть проведено расследование и установлена подлинность документа. По сути функции УЦ переносятся в ЭА. Юридическая значимость документов при таком подходе сомнительна.

4.3. Проблема обратной связи

На данном этапе развития большинство организаций стремится автоматизировать деятельность, связанную с делопроизводством, управлением документами, архивным делом и т. д. Однако на этапе проектирования и создания таких систем мало кто строит модели эффективности данной автоматизации. Особенно это касается больших и, возможно, территориально-распределенных, информационных систем (ИС), и ЭА в частности. Их отсутствие может привести к расхождению планируемых целей, поставленных перед началом создания ИС (ЭА), с достигнутыми. Поэтому перед началом проектирования ЭА необходимо определить показатели качества, характеризующие качество работы ЭА в процессе эксплуатации. Данные показатели необходимо рассчитывать (или по крайней мере оценивать) на этапе проектирования ЭА и точно подсчитывать на этапе эксплуатации.

Поскольку перед созданием абстрактного ЭА часто невозможно определить точные показатели качества ввиду отсутствия необходимой информации, будем использовать критерии, характеризующие эффективность работы системы, т. е. степень достижения целей, которые можно поставить при создании ЭА.

К таким показателям относятся:

- своевременность выполнения функций ЭА;
- достоверность функционирования ЭА;
- надежность хранения данных и функционирования ЭА.

5. Показатели качества ЭА

Каждый из перечисленных выше показателей качества ЭА может быть выражен отдельной математической моделью. Модели показателей качества могут быть связаны друг с другом или выступать ограничениями в математических моделях других показателей. Для каждого показателя качества должны быть определены допустимые и предельные значения.

Расчеты (оценки) показателей качества позволят скорректировать проект ЭА, спрогнозировать возникновение проблем при проектировании и эксплуатации, оптимизировать расходы на разработку и эксплуатацию ЭА.

Ниже приведены математические модели для перечисленных показателей качества ЭА, которые могут быть рекомендованы для разработчиков и проектировщиков ЭА. Формулы приведены без вывода, из-за ограничений объема данной статьи, основные выводы приведены в [9-12].

5.1. Своевременность выполнения функций ЭА

Своевременность — обеспечение в заданные временные рамки выполнения операций различных типов по обработке информации в ЭА: создание иерархических классификаторов, ввод документов, индексация ключевых реквизитов архивной карточки, классификация, полнотекстовая индексация, поиск, извлечение документов, проверка сохранности (ЭП, и др.) документов, анализ содержимого документа, распознавание и оцифровка бумажных образов.

Основной интегральный показатель своевременности может быть определен по следующей формуле [11]:

$$P_И(t \le T_Д) = P\,(t \le T_Д)\,K_{о.г.},$$

где $P\,(t \le T_Д)$ — вероятность обработки информации в процессе выполнения различных операций (типы операций см. выше) в ЭА за время t не больше заданного (допустимого) $T_Д$ при условии надежного функционирования средств обработки;

$K_{о.г.}$ — коэффициент оперативной готовности ЭА (см. ниже).

Для точного расчета всех временных интервалов T необходимо получить статистические данные и определить допустимое время $T_Д$, позволяющее получить оценку вероятности своевременного выполнения операции.

Оценка вероятности выполнения операции i-го типа за время, не превышающее допустимое, рассчитывается по формуле:

$$P(T_i \le T_{Дi}) = n_i/N(i), \qquad (6)$$

где n_i — число значений времени выполнения операции i-го типа T_i, которые удовлетворяют неравенству $T_i \le T_{Дi}$; $N(i)$ — количество выполненных операций i-го типа (см. типы операций в определении своевременности); $l = [1, N(i)]$.

При необходимости получения вероятностной оценки времени выполнения операции i-го типа с учетом надежности ЭА формула (6) принимает вид:

$$P^*(T_i \le T_{Дi}) = P(T_i \le T_{Дi})\, K_{о.г.i},$$

где $K_{о.г.i}$ — коэффициент оперативной готовности ЭА при выполнении операции i-го типа (см. в определении своевременности типы операций).

5.2. Достоверность функционирования ЭА

Достоверность функционирования ЭА — свойство, обуславливающее безошибочность производимых в ЭА преобразований информации [8] при выполнении операций создания иерархических классификаторов, ввода документов, индексации ключевых реквизитов архивной карточки, классификации документов, полнотекстовой индексации, поиска, извлечения документов, проверки сохранности (ЭП и др.) документов, анализа содержимого документа, распознавания и оцифровки бумажных образов.

Для описания достоверности обработки информации в ЭА целесообразно применить систему показателей [11, 12].

1. Доверительная вероятность необходимой точности (достоверность) — $P = 1 - Q_{ош}$ — вероятность того, что в пределах заданного массива данных отсутствуют грубые погрешности, приводящие к нарушению необходимой точности обработки информации (это касается процессов ввода реквизитов архивной карточки, хранения, индексации, анализа содержимого документа и т. д.). $Q_{ош}$ — вероятность появления ошибки при вводе документов, индексации, классификации, анализе содержимого документа и т. д.
2. Средняя наработка информации на ошибку — отношение объема информации, преобразуемой в системе, к математическому ожиданию количества ошибок, возникающих в информации (например, при ошибках классификации или поиска документов в ЭА).
3. Вероятность коррекции ошибки в заданное время — $P_{корр}(\tau)$ — вероятность того, что время, затрачиваемое на идентификацию и исправление ошибки в данных архивной карточки, полнотекстовой индексации, распознавания, поиска или классификации документов, не превысит заданное τ .
4. Среднее время коррекции информации — $T_и$ — математическое ожидание времени, затрачиваемого на идентификацию и исправление ошибки.
5. Коэффициент информационного-технического использования ЭА:

$$K_{ти} = \frac{T_{раб} - (T_к + T_и)}{T_{раб}},$$

где $T_{раб}$ — математическое ожидание планируемого времени работы системы на преобразование информации (см. типы операций в определении достоверности),

$T_к$ — математическое ожидание времени контроля,

$T_и$ — математическое ожидание времени идентификации и исправления ошибок.

5.3. Надежность функционирования ЭА

Надежность функционирования ЭА — свойство сохранять во времени в установленных пределах значения всех параметров, характеризующих способность выполнять основное назначение при воздействии неисправностей (отказов и сбоев) технических средств, ошибок в программах и данных, ошибок персонала и пользователей в заданных режимах и условиях эксплуатации при известных характеристиках системы технического обслуживания и ремонта [3].

5.3.1. Надежность технического обеспечения ЭА

Надежность ЭА в целом не отличается от надежности функционирования информационных систем. Функционирование ЭА может быть определено согласно ГОСТ 27.003–90 (Надежность в технике [5]) как циклическая работа по назначению. Тогда применительно к выбранным показателям надежности основными можно считать следующие.

1. Коэффициент оперативной готовности

$$K_{о.г.} = K_г P(t),$$

где $K_г$ — коэффициент готовности — вероятность того, что ЭА будет работоспособен в произвольный момент времени; $P(t)$ — вероятность безотказной работы при наработке $t = N$ часов.
2. Среднее время восстановления $T_в$ (по всем видам отказов).

Под безотказной работой понимается способность технического (программного) обеспечения выполнять свои функции даже в условиях отказа (полно-

го или с восстановлением) отдельных частей ЭА. Коэффициент готовности определяется как

$$K_г = \frac{T_0}{T_0 + T_в},$$

где T_0 — средняя наработка на отказ (по всем видам отказов), $T_в$ — среднее время восстановления.

3. Вероятность безотказной работы системы $P(t)$

$$P_{nmc}^{(j)}(t) = e^{-\lambda_j t},$$

где $P_{nmc}^{(j)}(t)$ — вероятность безотказной работы элемента надежности, $\lambda = 1/T_o$ — интенсивность потока отказов для одного элемента надежности.

5.3.2. Надежность программного обеспечения ЭА

Модель надежности ПО ЭА основана на допущении, что интенсивность обнаружения ошибок пропорциональна количеству ошибок, остающихся по истечении (i-1) интервала времени, суммарному времени, уже затраченному на отладку к началу текущего интервала, средней длительности поиска ошибки в текущем i-м интервале времени отладки t_i; что позволяет оценить вероятность безотказной работы системы и коэффициент готовности. Считается, что ошибки постоянно исправляются, выходят обновления, т. е. ПО ЭА постоянно модифицируется, а в процессе модификации возникают новые ошибки. Таким образом, общий характер зависимости исправления ошибок носит «пилообразный» характер. Нужно отметить, что разработанная модель, основанная на известной модели Шика—Волвертона, дает заниженные (пессимистические) оценки надежности [9, 10].

Вероятность безотказной работы определяется как:

$$P(t) = \exp(-(K_{JM}(E_0 - M))\ t/2),$$

где K_{JM} — коэффициент пропорциональности; E_0 — количество ошибок в начале отладки; M — полное количество временных интервалов, на каждом из которых обнаружена хотя бы одна ошибка. В нашем случае оно равно количеству обнаруженных ошибок, т. к. принимается допущение, что на каждом временном интервале произошла одна ошибка.

Средняя наработка между обнаруженными ошибками:

$$T = \sqrt{\left(\pi / \left(2\left(K_{JM}\left(E_0 - M\right)\right)\right)\right)}.$$

Для оценки параметров модели K_{JM} и E_0 необходимо решить систему уравнений (решается итерационно, затем округляется E'_0 в большую сторону и получается E_0):

$$\begin{cases} K_{JM} = M / \left(\left(\sum_{i=1}^{M}(t_i)\right)\left(E'_0 + 1 - \dfrac{\sum_{i=1}^{M}(i \cdot t_i)}{\sum_{i=1}^{M}(t_i)}\right)\right), \\ E'_0 = M/\left(\sum_{i=1}^{M}\left(1/\left(E'_0 - i + 1\right)\right)\right) + \dfrac{\sum_{i=1}^{M}(i \cdot t_i)}{\sum_{i=1}^{M} t_i} - 1. \end{cases}$$

Оценка остаточного количества ошибок ПО: $E_{ост} = E_0 - M$.

5.3.3. Модель надежности ЭА в целом

Вышеприведенные модели могут объединяться в сложные схемы надежности в зависимости от конфигурации, в которых учитывается надежность серверов ЭА, автоматизированных рабочих мест (АРМ), оборудования КС, ПО отдельных компонентов ЭА и связанных с ним программ.

В общем случае ЭА может быть многоуровневым (центр, филиалы, отделения) и распределенным (обязательный учет надежности КС), поэтому общая схема для построения модели надежности трехуровневого иерархического распределенного ЭА (расчет для коэффициента готовности) может выглядеть следующим образом [9, 10]:

$$К_{г\ ЭА} = К_{г\ в.у.} \sum_{i=1}^{N_{с.у.}} (B_i\ К_{г\ кс\ с.у.\ i}\ К_{г\ с.у.\ i}\ (\sum_{j=1}^{N_{н.у.i}} A_{ij} \times \times К_{г\ кс\ н.у.\ ij}\ К_{г\ н.у.\ ij})), \quad (7)$$

где $К_{г\ в.у.}$ — показатель надежности для объектов ЭА верхнего уровня (произведение показателей надежности объектов типа центры обработки данных (ЦОД), центральные серверы, ПО центральной БД и т. д.);

$B_i = Об_i / \sum_{i=1}^{N_{с.у.}} Об_i$ ($Об_i$ — количество объектов (отделений), обслуживаемых i-м звеном ЭА среднего уровня (филиал), причем $\sum_{i=1}^{N_{с.у.}} B_i = 1$) — процент объектов, обслуживаемых i-м звеном ЭА среднего уровня;

$N_{с.у.}$ — количество звеньев ЭА среднего уровня;

$К_{г\ кс\ с.у.\ i}$ — коэффициент готовности оборудования и каналов связи от верхнего до среднего уровней;

$К_{гс.у.\ i}$ — коэффициент готовности i-го звена ЭА (группы объектов ЭА) среднего уровня;

$A_{ij} = Об_{ij} / Об_i$ — процент объектов (АРМ), обслуживаемых j-м нижним звеном ЭА в среднем звене i $\left(\sum_{j=1}^{N_{н.у.i}} A_{ij} = 1\right)$;

$N_{н.у.i}$ — количество элементов нижнего уровня (АРМ ЭА) в i-ом районном звене;

$К_{г\ кс\ н.у.\ ij}$ — коэффициент готовности оборудования и КС уровня i-й объект среднего уровня — j-й объект нижнего уровня;

$К_{гн.у.\ ij}$ — коэффициент готовности объекта (АРМ) нижнего уровня в j-м элементе нижнего уровня.

Формулу (7) можно упрощенно представить в виде:

$$К_{г\ ЭА} = К_{г\ в.у.} \sum_{i=1}^{N_{с.у.}} (B_i\, К_{г\ кс\ с.у.\ i}\, К_{г\ с.у.\ i} \times$$

$$\times \Big(\sum_{j=1}^{N_{н.у.i}} К_{г\ кс\ н.у.\ ij}\, К_{г\ н.у.\ ij}) / N_{н.у.i}). \qquad (8)$$

Формулы (7), (8) позволяют учесть взаимное влияние среднего и нижнего уровней системы. Аналогично коэффициенту готовности ($K_{гЭА}$) рассчитывается и вероятность безотказной работы $P(t)_{ЭА}$ путем произведения показателей $P(t)$ соответствующих элементов схемы надежности. Например, модель вероятности безотказной работы системы, согласно формуле (7), примет вид

$$P(t)_{ЭА} = P(t)_{в.у.} \sum_{i=1}^{N_{с.у.}} (B_i P(t)_{кс\ с.у.\ i} P(t)_{с.у.\ i} \times$$

$$\times \Big(\sum_{j=1}^{N_{н.у.i}} A_{ij} P(t)_{кс\ н.у.\ ij} P(t)_{н.у.\ ij})).$$

С учетом приведенной модели надежности можно сформулировать состояние полного отказа системы: отказ системы в целом возможен при отказе объекта верхнего уровня, и/или отказе всех элементов среднего уровня ЭА, и/или всех элементов нижнего уровня ЭА.

Частичный отказ системы: отказ одного элемента среднего уровня, что равносильно отказу всех элементов нижнего уровня, обслуживаемых данным звеном среднего уровня ЭА.

Данная модель надежности прошла серьезную проверку практикой, в частности использовалась при построении методики и расчета показателей надежности ГАС «Выборы» [9, 12].

Заключение

Данная статья явилась попыткой систематизировать знания и опыт, полученные при разработке архивных систем, в частности электронных архивов для Пенсионного фонда РФ, Газпромбанка, коммерческих и государственных предприятий. Авторы предложили описание проблематики и теории построения электронных архивов, выделили круг проблем, с которыми неизбежно столкнется разработчик.

Тема является весьма актуальной, поскольку через несколько лет после внедрения СЭД, эксплуатирующие их организации неизбежно столкнутся с проблемой архивного хранения документов, прошедших полный цикл делопроизводства. Отставание во внедрении электронных архивов будет составлять 5–7 лет, как раз срок, после которого документы требуется либо уничтожить, либо отправить на архивное хранение согласно действующим правилам в РФ.

Иерархическая структура, которая, как правило, присуща электронным архивам, в т. ч. и территориально-распределенным, помогает создать единую методику оценки надежности и эффективности их работы, в т. ч. и проектную, на основе уже опробованного аппарата оценки. Более того, данная методика применима и к «сетевым» конфигурациям ЭА, поскольку сетевой граф можно разбить на критичные сечения-деревья, или же выделить остовные деревья, соответствующие иерархическим связям в ЭА и уже для них многократно применить описанные модели оценки.

Список разработанных электронных архивных систем

Авторами статьи разработан ряд крупных систем электронного архива и архивных хранилищ.

1. Архив документов персонифицированного учета Пенсионного фонда РФ (в эксплуатации с 2004 г. во всех субъектах РФ, более 1500 пользователей на апрель 2012 г., в 2010 г. загружено более 300 млн документов).
2. Архив финансово-распорядительных документов ОАО «Газпромбанк» (в эксплуатации с 1997 г., прием до 60 000 документов в день).
3. Архив документов ООО «Юридическая фирма Городисский и Партнеры» (в эксплуатации с 2003 г., прием до 2000 документов в день).
4. Архив документов НПФ «Благосостояние» (2003 г.).
5. Стандартное коробочное решение «Архив документов подразделения» (2010 г.).
6. Система ведения реестра для ЗАО СР «ДРАГа» (в эксплуатации с 1998 г.).
7. Архивная информационная система фондов Центрального музея древнерусской культуры и искусства им. Андрея Рублева (2004 г.).
8. Архивная информационная система «Фонды и выставки Музея истории города Москвы» (2005 г.).
9. Подсистема учета материально-технических ресурсов для ГАС «Правосудие» (2007 г.).
10. Информационно-аналитическая система «Астарта» (2002 г.).

Кроме перечисленных, авторами разработан также ряд небольших электронных архивов для коммерческих и государственных организаций.

Литература

1. Федеральный закон от 22.10.2004 № 125-ФЗ «Об архивном деле в Российской Федерации».
2. 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.
3. ГОСТ 27.002–89. Надежность в технике. Основные понятия. Термины и определения. М.: Изд-во стандартов, 1989. 36 с.
4. *Дружинин Г. В.* Надежность автоматизированных производственных систем. М.: Энергоатомиздат, 1986. 480 с. ил.
5. ГОСТ 27.003–90. Надежность в технике. Состав и общие правила задания требований по надежности. М.: Изд-во стандартов, 1990, 27 с.
6. Приказ Министерства культуры и массовых коммуникаций Российской Федерации № 536 от 8 ноября 2005 г. «О Типовой инструкции по делопроизводству в федеральных органах исполнительной власти».
7. Государственный стандарт РФ ГОСТ Р 51141–98 «Делопроизводство и архивное дело. Термины и определения» (утвержден Постановлением Госстандарта РФ № 28 от 27 февраля 1998 г.).
8. *Кульба В. В., Ковалевский С. С., Шелков А. Б.* Достоверность и сохранность информации в АСУ. М.: СИНТЕГ, 2003.
9. *Акимова Г. П., Соловьев А. В.* Анализ оценки надежности иерархической территориально-распределенной информационной системы на примере ГАС «Выборы» // Труды Института системного анализа РАН (ИСА РАН). Т. 45. «Технология программирования и хранения данных» М.: Ленанд/URSS, 2009. С. 355–372.
10. *Акимова Г. П., Соловьев А. В.* Методология оценки надежности иерархических информационных систем // Системный подход к управлению информацией / Труды ИСА РАН. Т. 23. М.: КомКнига/URSS, 2006. С. 18–47.
11. *Акимова Г. П., Соловьев А. В., Янишевский И. М.* Методология оценки эффективности иерархических информационных систем // Системный подход к управлению информацией / Труды ИСА РАН. Т. 23. М.: КомКнига/URSS, 2006. С. 48–66.
12. *Акимова Г. П., Пашкина Е. В., Соловьев А. В.* Анализ оценки эффективности иерархической территориально-распределенной системы на примере ГАС «Выборы» / Труды Института системного анализа РАН (ИСА РАН) «Обработка изображений и анализ данных» М.: Книжный дом «Либроком», 2010. Т. 58. С. 27–42.
13. *Акимова Г. П., Пашкин М. А.* Аналитический подход к решению задачи мониторинга информационного пространства, Системы высокой доступности. № 3–4. Т. 2, с. 44–50, 2006
14. *Акимова Г. П., Богданов Д. С., Мусатов И. В., Пашкин М. А., Солдатов Д. В., Сомин Н. В. С*овременные автоматизированные технологии обработки разнородных информационных потоков. // «Организационное управление и искусственный интеллект» Сборник трудов Института системного анализа РАН, 2003. С. 290–304.
15. Федеральный закон Российской Федерации от 6 апреля 2011 г. № 63-ФЗ.

Акимова Галина Павловна. В. н. с. ИСА РАН. К. т. н. Окончила МФТИ в 1978 г. Количество печатных работ: 43. 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. Н. с. ИСА РАН. Окончил МГТУ «Станкин» в 2001 г. Количество печатных работ: 11. Область научных интересов: системное программирование, информационные технологии, информационно-аналитические системы, электронный архив. E-mail: matveyp@cs.isa.ru

Пашкина Елена Владимировна. Н. с. ИСА РАН. Окончила МГУ в 2003 г. Количество печатных работ: 8. Область научных интересов: системное программирование, информационные технологии, электронный документооборот, электронный архив. E-mail: alena@cs.isa.ru

Соловьев Александр Владимирович. В. н. с. ИСА РАН. К. т. н. Окончил МГТУ им. Н. Э. Баумана в 1994 г. Количество печатных работ: 28. Область научных интересов: системный анализ, системы управления базами данных, теория надежности, влияние человеческого фактора, математическое моделирование, электронный документооборот. E-mail: alexsol@cs.isa.ru