Разработка математической модели метаданных электронного документа долговременного хранения

А.В. Соловьев

Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия

Аннотация. В данной статье представлена разработка математической модели метаданных электронного документа долговременного хранения. К метаданным отнесены не только собственно данные о документе, например, такие как автор (авторы), время и место создания, время последнего изменения, сведения о формате, название документа и другие. Но также и данные о специфических операциях над документом, таких как заверение документа электронной подписью, проверка интерпретируемости и аутентичности документа. Кроме того, к метаданным отнесены индексные данные, сведения о нормативно-справочной информации, связанной с документом, данные о связях с другими документами. В ходе исследования выполнено достаточно подробное моделирование этих метаданных электронного документа. Результаты моделирования могут использоваться для организации долговременного хранения электронных документов в плане решения проблемы их интерпретируемости, проектирования программных средств хранения. В ходе дальнейших исследований автор планирует разработать математическую модель оценки интерпретируемости электронных документов долговременного хранения.

Ключевые слова: *долговременное хранение, сохранность документа, электронный документ, метаданные, модель содержания.*

DOI: 10.14357/20790279220311

Введение

В работах автора [1, 2] выполнено математическое моделирование электронного документа (ЭлД) долговременного хранения. Было показано, что математическое моделирование крайне важно для решения проблемы интерпретируемости ЭлД на длительном сроке хранения [3]. Решение проблемы интерпретируемости делает возможным решение более общей задачи обеспечения долговременной сохранности электронных документов [4].

В работе [2] предложена математическая модель ЭлД долговременного хранения, представляющая собой объединение выделенных по смыслу частей ЭлД – семантических блоков. Определен состав семантических блоков верхнего уровня (макроблоков), необходимый и достаточный для организации долговременного хранения. В ходе исследования выполнено достаточно подробное моделирование макроблоков верхнего уровня оригинала ЭлД, а также его нормализованной копии. Под нормализацией понимается приведение оригинала электронного документа к единому формату (набору форматов) данных, предназначенного для долговременного хранения. Показано, что состав семантических блоков ЭлД зависит от создаваемой его информационной системы. Приведен пример ЭлД системы электронного документооборота (СЭД) и определен состав его семантических блоков.

В работе [2] были определены важные составные части ЭлД долговременного хранения. К таким частям относятся метаданные ЭлД, сведения о связанной (зависимой) нормативно-справочной информации (НСИ), данные о других связанных (зависимых) документах. Приведем разработку математической модели метаданных ЭлД долговременного хранения.

1. Разработка математической модели метаданных

Как было показано в работе [2], в общем виде математическая модель ЭлД долговременного хра-

нения с точки зрения состава макроблоков представляется в следующем виде:

$$D = OrD \cup OdfD \cup DMD \cup CLI \cup LDI , \quad (1)$$

где D – ЭлД долговременного хранения;
OrD – математическая модель оригинала ЭлД;
$OdfD$ – математическая модель нормализованной копии оригинала ЭлД;
DMD – математическая модель метаданых ЭлД;
CLI – математическая модель НСИ и классификаторов ЭлД;
LDI – математическая модель связанных с ЭлД других документов.

Математические модели OrD и $OdfD$ приведены также в работе [2].

Перейдем к разработке математической модели метаданных ЭлД DMD. В принципе НСИ также можно отнести в широком смысле к метаданным ЭлД. Однако, с другой стороны, НСИ – это отдельные ЭлД. Поэтому в случае долговременного хранения общая модель хранения данных скорее представляет собой семантическую сеть, состоящую из различных ЭлД D, которые могут быть связаны, и собственно связей между ЭлД. Причем к ЭлД также должны быть отнесены НСИ и связанные (или же зависимые) документы.

Метаданные необходимо хранить для решения проблемы интерпретации ЭлД спустя годы или десятилетия [3].

Как было показано в [2] к метаданным DMD относятся собственно данные о ЭлД, например, установленные стандартом «дублинского ядра» (Dublin Core) DCD [5].

Для того, чтобы не потерять важную информацию о форме и структуре ЭлД, к метаданным следует отнести модель содержания документа CMD, форму представления (отображения) ЭлД пользователю VMD. К метаданным также могут быть отнесены данные о действиях с ЭлД, т.е. различные выписки из журналов инвентаризаций (аутентичности, интерпретируемости), операций, безопасности (журнал доступа к ЭлД) и т.д. $OperD$, данные индексов, в том числе и полнотекстовых, $DIdx$. Данные индексов и операций с документами служат для удобства работы с ЭлД. В принципе эти семантические блоки могут быть необязательными, но играют важную роль в дополнительном контроле аутентичности, интерпретируемости и безопасности хранения ЭлД, а также для работы с поиском по содержимому ЭлД. Разумеется, такие данные необходимы, если ЭлД имеет высокую ценность.

Тогда математическая модель метаданных может быть представлена в следующем виде:

$$DMD = DCD \cup CMD \cup VMD \cup OperD \cup DIdx \quad (2)$$

1.1. Математическая модель данных о документе

Математические модели семантических блоков DCD в соответствии со стандартом [5] могут быть представлены в следующем виде:

$$DCD = \cup_{(i=1,N)}(DCD_i) , \quad (3)$$

где DCD_i – элементы метаданных, такие как:
- Название ЭлД ($Dtitle$);
- Авторы ЭлД (массив $Dcreators = \cup_{(i1=1,N1)}(Dcreator_{i1})$);
- Ключевые слова (массив $Dkeyws = \cup_{(i2=1,N2)}(Dkeyw_{i2})$);
- Текстовое содержимое ЭлД ($Dtext$);
- Редакторы ЭлД (массив $Deditors = \cup_{(i3=1,N3)}(Deditor_{i3})$);
- Издатели ЭлД (массив $Dpublishers = \cup_{(i4=1,N4)}(Dpublisher_{i4})$);
- Дата/время создания ЭлД ($Dcrdate$);
- Дата/время приема ЭлД в долговременное хранение ($DLTdate$)
- Даты/время изменения ЭлД (массив $Deddates = \cup_{(i5=1,N5)}(Deddate_{i5})$);
- Форматы ЭлД (массив $Dformats = \cup_{(i6=1,N6)}(Dformat_{i6})$);
- Права доступа к ЭлД (массив $Drights = \cup_{(i7=1,N7)}(Dright_{i7})$);
- Языки создания ЭлД (массив $Dlangs = \cup_{(i8=1,N8)}(Dlang_{i8})$);
- География создания ЭлД (массив $Dgpss = \cup_{(i9=1,N9)}(Dgps_{i9})$);
- Тип документа ($Dtype$).

Безусловно, права доступа к ЭлД ($Drights$) должны соответствовать модели оригинала документа (OrD и $OdfD$), приведенных в [2]. География создания ЭлД ($Dgpss$) может быть представлена массивом GPS-координат тех географических локаций, где ЭлД создавался или редактировался.

Тогда математическая модель DCD (3) представляется в следующем виде:

$$DCD = Dtitle \cup (\cup_{(i1=1,N1)}(Dcreator_{i1})) \cup (\cup_{(i2=1,N2)}(Dkeyw_{i2})) \cup Dtext \cup (\cup_{(i3=1,N3)}(Deditor_{i3})) \cup (\cup_{(i4=1,N4)}(Dpublisher_{i4})) \cup Dcrdate \cup DLTdate \cup (\cup_{(i5=1,N5)}(Deddate_{i5})) \cup (\cup_{(i6=1,N6)}(Dformat_{i6})) \cup (\cup_{(i7=1,N7)}(Dright_{i7})) \cup (\cup_{(i8=1,N8)}(Dlang_{i8})) \cup (\cup_{(i9=1,N9)}(Dgps_{i9})) \cup Dtype \quad (4)$$

Математическая модель CMD может быть представлена в виде графи семантических блоков оригинала ЭлД $G(OrD)$, как это отображено в [2].

Математическая модель VMD – это описание модели визуализации, она может быть дана в виде

«пустографки», в которой представлены в виде графических элементов или окон графического интерфейса те данные, которые являются постоянными и не зависят от экземпляра ЭлД. А также постоянную текстовую информацию. Вместо переменных данных *Dtext*, зависящих от экземпляра ЭлД остаются пустые места.

Важным замечанием является то, что пересечение множеств (*CMD* U *VMD*) и *DD* дает пустое множество, т.е.:

$$(CMD \cup VMD) \cap Dtext = \varnothing.$$

Действительно, форма и содержание ЭлД определяются согласно [6, 7] следующим образом. Все, что извлекается из текста ЭлД и имеет переменное значение в зависимости от экземпляра ЭлД, мы будем называть содержанием ЭлД. 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 ЭлД.

Следовательно, форма и содержание ЭлД содержат непересекающуюся информацию.

Т.к. *CMD* и *VMD* не изменяются в зависимости от экземпляра ЭлД, то они относятся к форме ЭлД. *Dtext* уникально для каждого экземпляра, следовательно это содержание ЭлД. Следовательно, (*CMD* U *VMD*) ∩ *Dtext* = ∅.

При долговременном хранении разумно хранить *CMD* и *VMD* отдельно от *DCD,* т.к. таким образом можно существенно экономить место не сохраняя эти семантические блоки с каждым ЭлД. Тогда необходимо для *DCD* тип документа (*Dtype*) разработать единый словарь типов, который бы по существу определял, какие *CMD* и *VMD* относятся к данному экземпляру ЭлД.

Тогда в составе метаданных ЭлД можно хранить не сами *CMD* и *VMD*, а ссылки на их ЭлД *CMDLink* и *VMDLink*. В этом случае математическая модель метаданных (2) представляется следующим образом:

$$DMD = DCD \cup CMDLink \cup VMDLink \cup OperD \cup DIdx. \quad (5)$$

Метаданные ЭлД могут также содержать электронную подпись (ЭП) или подписи как один из семантических блоков ЭлД. ЭП в данном случае может выполнять ту же функцию, как и в математической модели оригинала ЭлД [2] для обеспечения аутентичности ЭлД [8].

1.2. Математическая модель данных журналов операций

Для того, чтобы разработать математическую модель выписки из журналов действий с ЭлД (*OperD*), необходимо представить, что собой представляют журналы действий. Очевидно, что журнал действий состоит из отдельных записей, в которых помещаются сведения о проведенных инвентаризациях аутентичности и интерпретируемости, доступе к ЭлД и др. Таких журналов может быть много, но для удобства долговременного хранения выписка из этих журналов при конкретном документе должна быть одна. Для решения проблемы аутентичности [4,8] выписки из журнала, она либо заверяется при приеме на длительное хранение целиком, либо каждая запись журнала в отдельности.

Т.к. при долговременном хранении будут производиться инвентаризации ЭлД в части интерпретируемости и аутентичности, то выписка о проведенных действиях может дополняться новыми сведениями. Следовательно, второй вариант с заверением каждой записи журнала ЭП предпочтительней.

Тогда математическая модель *OperD* представляет собой множество записей о выполненных операциях:

$$OperD = \cup_{(i=1,N)}(OperD_i), \quad (6)$$

где $OperD_i$ – одна запись о выполненной операции, $OperD_i = OperType_i \cup OperRes_i \cup OperDate_i \cup OperInfo_i \cup DSign_i$.

Элемент множества $OperD_i$ представляет собой информацию о типе операции ($OperType_i$), о результате операции ($OperRes_i$), дату и время выполнения операции ($OperDate_i$), дополнительную информацию о содержании операции ($OperInfo_i$), о заверенной ЭП ($DSign_i$). Информация о типе операции для удобства должна строиться на основе типового словаря типов: миграция, проверка ЭП, перезаверение новой ЭП с сохранением авторства старых ЭП (см. подробно [8]), инвентаризация интерпретируемости и т.д. *DSign* включает в себя сведения о сертификатах, сведения о новых ЭП и др, предусмотренное моделью ЭП [2]. *OperRes* содержит сведения о результате операции: например, результат проверки ЭП, результат интерпретации ЭлД и т.д.

Дополнительная информация ($OperInfo_i$) может содержать любые дополнительные сведения об операции с ЭлД, а также о целях и причинах ее проведения.

Безусловно, *OperD* также относится к метаданным документа.

1.3. Математическая модель индексов документа

Любая индексная информация ЭлД необходима для ускорения его поиска. В первую очередь, интерес представляет полнотекстовый индекс ЭлД.

Чтобы не переиндексировать весь текст ЭлД при приеме в долговременное хранение, что может

занимать определенное время, можно при наличии такой возможности, перенести уже существующий полнотекстовый индекс ЭлД в систему, предназначенную для долговременного хранения. Безусловно, если такая возможность принципиально реализуема.

Разумеется, наличие индексной информации при ЭлД не обязательно для долговременного хранения. Наличие такой информации всего лишь служит для выполнения сервисных функций и повышения удобства пользования.

При полнотекстовой индексации выполняется нормализация текста документа, т.е. приведение всех слов оригинала документа к единственному числу, именительному падежу (для существительных), неопределенной форме (глаголов), мужскому роду (прилагательные).

Модель полнотекстового индекса *FTIdx* ЭлД можно представить следующим образом:

$$FTIdx = \mathrm{U}_{(i=1,N)}(FTWrd_i)\,, \qquad (7)$$

где $FTWrd_i$ – элемент (слово) полнотекстового индекса. Набор нормализованных слов содержимого документа представляет собой множество (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.

Точно также может быть представлен и неполнотекстовый индекс данных ЭлД *RIdx*, например, реквизитный индекс. Тогда математическая модель индексов ЭлД может быть представлена в следующем виде:

$$DIdx = FTIdx \cup RIdx \cup DSign. \qquad (8)$$

Индекс *DIdx* не должен изменяться в процессе хранения ЭлД, т.к. оригинал ЭлД при долговременном хранении не подлежит изменению. Для этого индексная информация может быть заверена ЭП или множеством ЭП (*DSign*).

2. Разработка математических моделей макроблоков зависимых данных

Теперь перейдем к разработке математических моделей связанных (или иначе зависимых) данных. В работе [2] были определены важные макроблоки связанных (зависимых) данных ЭлД:

- НСИ (классификаторы, словари, нормативные документы) *CLI* на которые ссылается основной ЭлД. Если ЭлД, подлежащий долговременному хранению, ссылается на определенную НСИ, то НСИ должна сохраняться вместе с ЭлД. Причем НСИ должна сохраняться именно в той версии, которая была актуальна на момент создания ЭлД. Дело в том, что НСИ может изменяться со временем и может привести к тому, что версии НСИ не совпадут с актуальными версиями НСИ на момент создания ЭлД. Это, в свою очередь, приведет к проблеме правильной интерпретации ЭлД.
- Данные о документах *LDI* (или же сами документы), связанных с ЭлД, подлежащим долговременному хранению. Наличие этого макроблока также связано с необходимостью решения проблемы интерпретации ЭлД. Проблемы здесь те же, что и в случае с НСИ – наличие связанных ЭлД, наличие актуальных на момент создания ЭлД версий связанных документов.

Вместо НСИ и данных о связанных документах целесообразнее использовать связки с ЭлД НСИ и связанных (зависимых) ЭлД. В этом случае хранилище ЭлД будет представлять собой семантическую сеть, состоящую из ЭлД и связей между ними. Такое хранение сложнее организовать, но зато в результате организации будет существенная экономия места для хранения.

Тогда математическую модель ЭлД (1) можно представить в виде:

$$D = OrD \cup OdfD \cup DMD \cup CLILink \cup LDILink. \qquad (9)$$

Перейдем к разработке математических моделей макроблоков *CLI*, *LDI* (*CLILink*, *LDILink*).

2.1. Математическая модель классификаторов и связанной нормативно-справочной информации

Правила архивного хранения ЭлД прямо предполагают наличие классификации ЭлД при хранении [9–13]. Однако хранение при каждом ЭлД всех связанных с ним классификаторов приведет к непомерному увеличению размера хранилища ЭлД и вряд ли оправдано. В то же время, потеря ценной информации о классификации ЭлД может его обесценить.

Из данного противоречия вытекает необходимость создания хранилища классификаторов и НСИ при организации долговременного хранения ЭлД. Как минимум необходимо хранить классификаторы и НСИ, связанные с подмножеством ЭлД, в неизменном состоянии на момент поступления ЭлД в долговременное хранение.

Наличие классификаторов – это одна из ключевых особенностей, которая характеризует электронные архивы. Можно выделить следующие классификаторы документов в архивах и корпоративных хранилищах:

1. Классификатор «дело-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. 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. Документы (ЭлД) НСИ, связанные с подлежащим долговременному хранению ЭлД. Важно, чтобы ЭлД НСИ хранились в хранилище классификаторов НСИ в версии актуальной на момент создания ЭлД.

Хотя, строго говоря, классификаторы и НСИ – это разные сущности, тем не менее, удобно рассматривать их в модели совместно, т.к. часто системы классификации определены в нормативно-справочных документах. Также НСИ часто содержат в себе элементы классификации в виде справочников и словарей.

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 ЭлД. При этом не так редки ситуации, когда вычисленные функции расстояния позволяют отнести ЭлД, как к одной, так и к другой вершине классификатора.

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

Тогда математическую модель связок с классификаторами и НСИ можно представить в следующем виде:

$$CLILink = \mathrm{U}_{(i=1,N)}(CLILink_i)\,, \qquad (10)$$

где $CLILink_i$ – связка с классификатором или ЭлД НСИ. Каждая связка представляет собой идентифицирующую информацию классификатора или ЭлД НСИ ($IdInfoL_i$), а также может включать в себя ЭП или множество ЭП ($DSign_i$) для обеспечения аутентичности $CLILink_i$. Каждая связка может быть одного из следующих типов:

- *MDLink* – связка с ЭлД НСИ *MD*. Каждый ЭлД НСИ, по сути, представляет такой же ЭлД, описанный в модели ЭлД в [2];
- *DTLink* – связка с классификатором «дело-том». *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 ЭлД допускается только в томе дела. Наличие данного классификатора обязательно для любого электронного архива;
- *OrgSLink* – связка с классификатором структуры организации. *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;
- *HCLink* – связка с иными классификаторами. *HCL* – как правило, иерархический классификатор, предназначенный для создания альтернативной *OrgS-DT* классификации документов.

Тогда идентифицирующая информация связки $IdInfoL_i$ включает в себя тип классификатора, и другую информацию, позволяющую однозначно определить связанный ЭлД классификации.

Основная проблема использования классификаторов состоит в необходимости автоматизации классификации ЭлД. 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 ЭлД 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 см. [14, 15].

2.2. Математическая модель связанных документов

Как правило, ЭлД связан с другими документами. В этом случае при организации долговременного хранения речь может идти о создании хранилища связанных ЭлД [16] или организации семантической сети, узлами которой являются ЭлД, ребрами – связи между ЭлД.

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

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

Тогда математическая модель связок ЭлД с другими ЭлД:

$$LDILink = \mathrm{U}_{(i=1,N)}(LDILink_i)\,, \qquad (11)$$

где $LDILink_i$ – элемент множества связок документа с другими ЭлД. Каждая связка может включать в себя ЭП или множество ЭП (*DSign*) для обеспечения аутентичности $LDILink_i$. Тогда математическую модель каждой связки можно представить как:

$$LDILink_i = IdInfoLDoc_i \,\mathrm{U}\, DSign_i$$

Элемент вектора связок представляет собой идентифицирующую информацию связанного документа $IdInfoLDoc_i$, которая заверена ЭП ($DSign_i$), установившего связь либо же архивной ЭП, если связи устанавливаются автоматически по каким-то признакам.

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

Заключение

В результате проведенного исследования были разработаны математические модели метаданных ЭлД. К метаданным относятся не только собственно данные о ЭлД, определенные, например, стандартом [5], но и данные о модели содержания, модели представления в форме ЭлД, данные индексов, в том числе, полнотекстовых.

К метаданным относятся также данные о связанных с ЭлД долговременного хранения классификаторах, нормативно-справочной информации и других ЭлД.

По мнению автора исследования без подробного моделирования состава семантических блоков ЭлД невозможно полностью обеспечить реализацию технологии организации долговременного хранения ЭлД [4]. Без моделирования состава информации ЭлД обеспечить стабильность таких важных характеристик ЭлД долговременного хранения, как аутентичность [8] и интерпретируемость [3].

В ходе дальнейших исследований автор планирует разработать математическую модель оценки интерпретируемости электронных документов долговременного хранения.

Литература

1. *Соловьев А.В.* Проблема определения электронного документа долговременного хранения // Информационные технологии и вычислительные системы. 2022. №1. С.47-54.
2. *Соловьев А.В.* Математическая модель электронного документа долговременного хранения // Информационные технологии и вычислительные системы. 2022. №2. С.30-36.
3. *Соловьев А.В.* Решение проблемы интерпретации цифровых данных долговременного хранения // Труды ИСА РАН. 2021. Том 71. Вып. 2. С. 43-49.
4. *Solovyev A.V.* Long-Term Digital Documents Storage Technology // Lecture Notes in Electrical Engineering. 2020. Vol. 641. P. 901–911.
5. ГОСТ Р 7.0.10-2019 (ИСО 15836-1:2017) Система стандартов по информации, библиотечному и издательскому делу. Набор элементов метаданных «Дублинское ядро». Основные (ядерные) элементы. (System of standards for information, librarianship and publishing. The Dublin Core

metadata element set. Basic (core) elements).
6. *Емельянов Н.Е.* Виды представления структурированных данных // Теоретические основы информационной технологии. Сборник трудов ВНИИСИ. 1988. № 22. С. 42–46.
7. *Емельянов Н.Е.* Теоретический анализ документного интерфейса: Препринт / Н.Е. Емельянов. М.: Всесоюзный научно-исследовательский институт системных исследований. 1987. 40 с.
8. *Solovyev A.V.* Authentication control algorithm for long-term keeping of digital data // IOP Conference Series: Materials Science and Engineering (MSE). 2020. Vol. 862(5). 052080.
9. ГОСТ Р 51141-98 Делопроизводство и архивное дело. Термины и определения.
10. 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.
11. Приказ Министерства культуры и массовых коммуникаций Российской Федерации № 536 от 8 ноября 2005 г. «О Типовой инструкции по делопроизводству в федеральных органах исполнительной власти».
12. Федеральный закон от 22.10.2004 № 125-ФЗ «Об архивном деле в Российской Федерации».
13. Typical requirements for automated electronic document management systems. Specification MoReq // Office for Official Publications of the European Communities as INSAR Supplement VI.
14. *Акимова Г.П., Пашкин М.А.* Аналитический подход к решению задачи мониторинга информационного пространства // Системы высокой доступности. 2006. №3-4. Т. 2. С. 44-50.
15. *Акимова Г.П. Богданов Д.С. и др.* Современные автоматизированные технологии обработки разнородных информационных потоков // «Организационное управление и искусственный интеллект» Сборник трудов Института системного анализа РАН. 2003. С. 290-304.
16. *Белова А.Н., Соловьев А.В.* Построение баз данных взаимосвязанных документов // Труды ИСА РАН. 2012. Т. 62. Вып. 3. С. 25-30.

Соловьев Александр Владимирович, Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» г. Москва, Россия. Главный научный сотрудник. Доктор технических наук. Количество печатных работ: 135. Область научных интересов: системный анализ, системы управления базами данных, теория надежности, математическое моделирование, долговременное хранение электронных документов. E-mail: soloviev@isa.ru

Development of a mathematical model of metadata of an electronic document for long-term storage

A.V. Solovyev

Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia

Abstract. This article presents the development of a mathematical model for the metadata of an electronic document for long-term storage. Metadata includes not only the actual data about the document, for example, such as the author (authors), time and place of creation, time of the last modification, format information, document name, and others. But also data on specific operations on a document, such as certification of a document with an electronic signature, verification of the interpretability and authenticity of a document. In addition, metadata includes index data, data on regulatory and reference information associated with a document, data on links with other documents. In the course of the study, a fairly detailed modeling of these metadata of the electronic document was performed. The simulation results can be used to organize long-term storage of electronic documents in terms of solving the problem of interpretability of electronic documents, designing storage software. In the course of further research, the author plans to develop a mathematical model for assessing the interpretability of electronic documents for long-term storage.

Keywords: *long-term storage, document preservation, electronic document, metadata, digital data*
DOI: 10.14357/20790279220311

References

1. *Solovyev A.V.* 2022. Problema opredeleniya elektronnogo dokumenta dolgovremennogo khraneniya [The problem of defining an electronic document for long-term storage] // Informatsionnyye tekhnologii i vychislitel'nyye sistemy [Information Technology and Computing Systems] 1: 47-54.
2. *Solovyev A.V.* 2022. Matematicheskaya model elektronnogo dokumenta dolgovremennogo khraneniya [Mathematical model of an electronic document of long-term storage] // Informatsionnyye tekhnologii i vychislitel'nyye sistemy [Information Technology and Computing Systems] 2: 30-36.
3. *Solovyev A.V.* 2021. Resheniye problemy interpretatsii tsifrovykh dannykh dolgovremennogo khraneniya [Solving the problem of interpreting digital data for long-term keeping]. Trudy ISA RAN [Proceedings of the ISA RAS]. 71(2): 43-49.
4. *Solovyev A.V.* 2020. Long-Term Digital Documents Storage Technology. Lecture Notes in Electrical Engineering. 641: 901-911.
5. GOST R 7.0.10-2019 (ISO 15836-1:2017) System of standards for information, librarianship and publishing. The Dublin Core metadata element set. Basic (core) elements.
6. *Emelyanov N.E.* 1988. Vidy predstavleniya strukturirovannykh dannykh [Types of representation of structured data]. Teoreticheskiye osnovy informatsionnoy tekhnologii. Sbornik trudov VNIISI. [Theoretical foundations of information technology. Collection of works of VNIISI]. 22: 42–46.
7. *Emelyanov N.E.* 1987. Teoreticheskiy analiz dokumentnogo interfeysa [Theoretical analysis of the document interface]. Vsesoyuznyy nauchno-issledovatel'skiy institut sistemnykh issledovaniy [All-Union Research Institute for System Research]. 40 p.
8. *Solovyev A.V.* 2020. Authentication control algorithm for long-term keeping of digital data // IOP Conference Series: Materials Science and Engineering (MSE). 862(5): 052080. doi: 10.1088/1757-899X/862/5/052080.
9. GOST R 51141-98 Records management and archiving. Terms and definitions.
10. 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.
11. 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».
12. Federal law of the Russian Federation 22.10.2004 № 125-FZ «About the archival affair in Russian Federation».
13. 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.
14. *Akimova G.P., Pashkin M.A.* 2006. Analytical approach to solving the problem of monitoring of information space // High availability systems. 3-4(2): 44-50.
15. *Akimova G.P., Bogdanov D.S.* 2003. Modern automated processing technology of heterogeneous information flows // Organizational control and artificial intelligence. Proceedings of the ISA RAS: 290-304.
16. *Belova A.N., Solovyev A.V.* 2012. Postroyeniye baz dannykh vzaimosvyazannykh dokumentov [Building databases of related documents]. Trudy ISA RAN [Proceedings of the ISA RAS]. 62(3): 25-30.

Solovyev A.V. Chief Researcher, Doctor of Technical Sciences. Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia. E-mail: soloviev@isa.ru