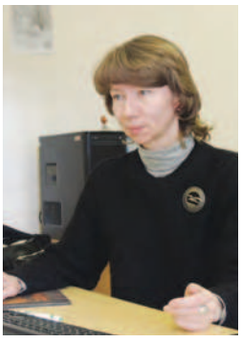
Стандартизация электронных документов



О. В. Барышева, ведущий программист РНБ

О. Н. Шорин, заместитель генерального директора РНБ

Мы предлагаем ознакомиться с позицией специалистов Российской национальной библиотеки по вопросу стандартизации электронных документов, включая постановку проблемы, подход к проблемной ситуации и предложения по отбору понятий и формулировке определений. Это предварительные материалы для дальнейшей совместной работы всех заинтересованных организаций, осуществляющих библиотечно-информационную и научно-информационную деятельность, официально выпускающих в публичное обращение электронные документы с целью их дальнейшего использования в научно-исследовательских, культурно-просветительных и иных целях.

Современная жизнь перенасыщена электронными документами самого разного профиля. В том числе и библиотеки генерируют растущее в геометрической прогрессии множество различных электронных документов, которые являются несомненным интеллектуальным активом и широко используются, в частности при обслуживании читателей/пользователей. Библиотечные процессы немыслимы сегодня без обращения к электронным документам, хранящимся в библиотеке и вне ее, в режиме локального и удаленного доступа.

Ценность электронных документов определяется не только их совокупным содержанием, но в значительной степени и тем, каким образом они обрабатываются, хранятся, предоставляются в пользование и обращаются. Именно бесперебойная циркуляция электронных документов сегодня определяет в значительной мере и социальную значимость библиотеки как культурно-информационного центра. Поэтому электронные документы, как и любые другие и даже в еще большей степени, нуждаются в стандартизации.

Актуальность создания национального библиотечного стандарта, отражающего основные характеристики и виды электронного документа, несомненна. В первую очередь это касается терминологии: необходимо дать определение самому электронному документу и процессам его создания, а также унифицировать, по возможности, типовые процедуры. При этом терминология не должна противоречить ни понятийному аппарату смежных областей деятельности, ни системе международных терминов и стандартов. Должна быть принята во внимание и «Концепция развития национальной системы стандартизации Российской Федерации на период до 2020 года», одобренная распоряжением Правительства РФ от 24 сентября 2012 г. № 1762-р.

Первоочередная задача для библиотек и органов НТИ сводится к разработке национального определения электронного документа и видов его проявлений. Дальнейшая перспектива – объединение усилий всех крупных поставщиков и владельцев электронных документов для разработки оптимальных представлений о форме и формате электронных документов, используемых в библиотеках. Разработка стандарта должна стать впоследствии базой как для выработки общих рекомендаций, так и для дальнейшей стандартизации отдельных процессов и процедур. Также важно правильно обозначить пределы использования электронных документов в библиотеках, определить типо-видовую характеристику документов, с которыми необходимо работать.

Начать следует с терминов и определений основных понятий в области электронных документов, набора характеристик, позволяющих проводить их идентификацию, базовые требования к формату представления электронных данных. При всем многообразии существующих стандартов СИБИД и смежных областей, оказывается, что нормативная база, на которую можно реально опираться при создании новых стандартов, весьма невелика:

- ГОСТ Р 7.0.83–2012. Система стандартов по информации, библиотечному и издательскому делу. Электронные издания. Основные виды и выходные сведения;
- ГОСТ Р 52292–2004. Информационная технология. Электронный обмен информацией. Термины и определения;
- ГОСТ Р ИСО/МЭК 26300–2010. Информационная технология. Формат Open Document для офисных приложений (OpenDocument) v 1.0.;
- ГОСТ 8.417–2002. Государственная система обеспечения единства измерений.

Единицы величин.

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Нами предлагаются следующие термины с соответствующими определениями:

- *ЭЛЕКТРОННЫЙ ДОКУМЕНТ:* созданный программными средствами, наделенный самостоятельным контентом и оформлением нетиражный электронный объект, анализ содержания которого может быть представлен в формализованном виде, предназначенный для передачи во времени и пространстве в целях хранения и использования, которое может быть регламентировано административными, правовыми и другими нормами;

● *ЭЛЕКТРОННЫЙ ОБЪЕКТ:* файл (совокупность файлов), формируемый в компьютерной программе пользователя или автоматизированной системе и содержащий в зафиксированном виде данные, предназначенные для восприятия компьютером или человеком с помощью соответствующего аппаратного и программного обеспечения.

Примечание. *Понятие «электронный объект» является родовым по отношению к электронному документу.*

● *ЭЛЕКТРОННЫЙ (ИНФОРМАЦИОННЫЙ) РЕСУРС:* комплекс электронных источников информации, программного обеспечения и аппаратных средств, служащих для удовлетворения информационных потребностей.

Примечание. *Понятие «электронный документ» является видовым по отношению к электронному ресурсу.*

● *КОНТЕНТ ЭЛЕКТРОННОГО ДОКУМЕНТА:* содержимое, наполнение электронного документа в плане содержания (в отличие от формы).

● *ФОРМА/ОФОРМЛЕНИЕ ЭЛЕКТРОННОГО ДОКУМЕНТА:* выполнение формальной обработки содержимого документа в соответствии с целевым назначением и/или правилами/нормами его использования при неизменности контента.

Примечание. *Форма, как правило, соотносится с определенным шаблоном электронного документа (например, электронный документ в форме письма), оформление – с изменением его внешнего вида (например, цветовое оформление). Соотнесение формы электронного документа с родом и/или видом/жанром заключенного в нем произведения, контента (например, электронный документ в форме стихотворения, марша, сборника) не рекомендуется.*

● *ВЕРСИЯ ЭЛЕКТРОННОГО ДОКУМЕНТА:* формально идентифицированное уникальное качественное состояние контента электронного документа во временном ряду по отношению к электронным документам с тем же основным контентом.

Примечание. *Качественно новый, оригинальный электронный документ не имеет версии.*

● *ИДЕНТИФИКАТОР ВЕРСИИ:* обозначение единицы в нумерованной/именованной последовательности качественных изменений контента электронного документа.

● *ЭТАЛОННАЯ ВЕРСИЯ ЭЛЕКТРОННОГО ДОКУМЕНТА:* образец, шаблон контента электронного документа для создания последующих версий/редакций.

● *РЕДАКЦИЯ ЭЛЕКТРОННОГО ДОКУМЕНТА:* результат процесса редактирования – создание обработанного и исправленного варианта существующего электронного документа или одной из его версий (в том числе локализация; изменение формы/оформления; непринципиальные изменения контента, не ведущие к качественному преобразованию, например исправление ошибок, перестановка абзацев и т. д.).

Примечание. *Редакция не имеет обязательного формального идентификатора.*

● *КОМПИЛЯЦИЯ:* способ составления контента электронного документа на основе использования/заимствования данных из уже существующих сторонних электронных документов.

● *КОПИЯ ЭЛЕКТРОННОГО ДОКУМЕНТА:* результат процесса копирования – дублирование, повторение электронного документа способом, отличным от способа его создания.

Примечание. *При создании копии возможно изменение формы/оформления, формата, знаковой природы первичного документа, но не ее контента. Копирование может быть произведено как с аналогового, так и с электронного документа, с оригинала, версии, редакции, другой копии.*

● *СЖАТИЕ/КОДИРОВАНИЕ ЭЛЕКТРОННОГО ДОКУМЕНТА:* алгоритмическое преобразование данных, производимое с целью оптимизации использования электронного документа.

Примечание. *Сжатие/кодирование производится, как правило, для уменьшения объема или ускорения загрузки электронного документа. Различие между ними состоит в том, что сжатый документ пригоден для непосредственного использования, а для кодированных электронных документов необходимо использование декодера. Степень сжатия и качественного изменения данных электронного документа зависят от используемого коэффициента/алгоритма. Все методы сжатия и кодирования данных делятся на два основных класса: с потерями и без потерь. Электронный документ может быть сжат/закодирован как целиком, так и частично, причем для разных составных частей могут быть использованы различные алгоритмы и коэффициенты сжатия.*

● *МЕТАДАННЫЕ:* зафиксированный в определенной форме структурированный набор характеристик электронного документа, организованных в соответствии с определенной схемой, и предназначенный для идентификации, поиска, оценки и управления электронными документами.

Примечание. *Метаданные формируются на основе* схемы методом выделения общего для всех электронных документов и обязательного для использования при их обработке набора полей, правил структурирования областей и элементов, извлечения данных из электронного документа и приведения их в соответствии с предписанным синтаксисом.

● *СХЕМА МЕТАДАННЫХ:* стандартизованный набор и структура представления метаданных, предназначенный для формального описания электронных документов.

Примечание. *Схема метаданных включает в себя набор полей (атрибутов, свойств, элементов), отражающих характеристики электронного документа.*

● *ФОРМАТ ДАННЫХ:* конкретная форма представления данных, в которой установлены ограничения типа данных (по ГОСТ Р 52292–2004).

Примечание. *Формат файла является частной формой формата данных.*

● *ФОРМАТ (ФАЙЛА) ЭЛЕКТРОННОГО ДОКУМЕНТА:* определенная спецификация, описывающая структуру файла, в соответствии с которой пакеты данных могут быть сохранены как файлы, переданы по сети в виде потока данных, и интерпретированы.

Примечание. *В ряде операционных систем расширение имени файла является видимым для пользователя символьным идентификатором типа файла, недостаточным для полной идентификации формата электронного документа.*

● *РАЗМЕР (ФАЙЛА) ЭЛЕКТРОННОГО ДОКУМЕНТА:* автоматически определяемое компьютером количество информации в стандартных единицах (по ГОСТ 8.417–2002).

Примечание. *Фактический объем дискового пространства, занимаемого файлом, зависит от конкретной файловой системы.*

● *ОТКРЫТЫЙ ФОРМАТ:* свободная от лицензионных ограничений при использовании общедоступная спецификация (стандарт) хранения цифровых данных, позволяющая переносить их с одной программной платформы на другую без искажения формы, структуры, содержания.

Примечание. *Не следует смешивать понятия открытого формата и свободной лицензии на использование. Открытость заключается в доступности спецификаций и соответствии открытого формата электронного документа стандарту, понятие свободы относится к передаче прав и является одной из моделей лицензирования.*

● *ИДЕНТИФИКАЦИЯ ЭЛЕКТРОННОГО ДОКУМЕНТА:* анализ электронного документа по одному или нескольким характерным признакам с целью опознания, определения сходства/различия, отнесения к конкретному классу/виду/типу.

● *ИДЕНТИФИКАТОР:* выбранный по какому-либо основанию деления признак, фиксирующий конкретную характеристику электронного документа, а также его обозначение.

● *АРХИВАЦИЯ/АРХИВНОЕ ХРАНЕНИЕ:* помещение электронного документа в условия, оптимальные для надежного долговременного хранения с целью последующего обращения к нему в будущем.

ВИДЫ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

1. ПО ЗНАКОВОЙ ПРИРОДЕ КОНТЕНТА:

● текстовый электронный документ – электронный документ, контент-основу которого составляет человекочитаемая информация преимущественно в виде слов;

● графический электронный документ – электронный документ, контент-основу которого составляет визуальное представление объектов/сущностей;

● звуковой (аудио) – электронный документ электронный документ, контент-основу которого составляет информация в форме, предназначенной для прослушивания.

2. ПО СТЕПЕНИ ОДНОРОДНОСТИ:

● электронный документ, в котором объединены контент-элементы разной знаковой природы (например, текстово-визуальные, аудио-визуальные);

● электронный документ, в котором объединены контент-элементы разных динамических характеристик (например, текстово-звуковые, графико-звуковые).

3. ПО СОСТАВУ ЭЛЕМЕНТОВ:

● однородный (гомогенный) электронный документ состоящий из контент-объектов одной знаковой природы;

● разнородный (гетерогенный) электронный документ, имеющий в своем составе контент-объекты различной знаковой природы.

4. ПО ДИНАМИЧЕСКИМ ХАРАКТЕРИСТИКАМ:

● статический (неподвижный) электронный документ: статическое визуальное представление контент-элементов (например, фотография);

● динамический (движущийся) электронный документ: серия последовательного представления контент-элементов, которая приводит к эффекту движения и/или воспроизведению сигналов во времени (например, музыка, видео).

5. ПО КОЛИЧЕСТВУ ЭЛЕМЕНТОВ:

● простой (односоставный) электронный документ – электронный документ, состоящий из единственного контент-элемента (например, фотография);

● составной (многосоставный) электронный документ – электронный документ, состоящий из более чем одного контент-элемента (например, слайд-шоу).

6. ПО СТРУКТУРЕ КОНТЕНТА:

● плоский электронный документ – электронный документ с последовательной линейной связью контент-элементов;

● объемный электронный документ – электронный документ с пространственной нелинейной связью контент-элементов.

7. ПО ПРОЦЕССАМ ДЕРИВАЦИИ (ПОРОЖДЕНИЯ ЭЛЕКТРОННОГО ДОКУМЕНТА):

● впервые созданный электронный документ;

● электронный документ с измененным контентом;

● электронный документ с измененной формой/оформлением без качественного изменения кон-

тента, электронный документ с измененной знаковой природой;
- электронный документ с измененным форматом (файла);
- электронный документ с измененным размером (файла);
- дубликат электронного документа (полная идентичная копия, отличная по времени создания).

8. По происхождению контента:
- новый электронный документ (в том числе копия аналогового документа, ранее не представленная в электронной форме);
- редакция;
- версия;
- компиляция.

9. По производности:
- оригинальный, созданный впервые в электронной форме;
- копия электронного документа;
- конвертированный электронный документ (переведенный из одного формата в другой);
- трансформированный электронный документ (переведенный из одной знаковой системы в другую методом синтеза или анализа).

Примечание. *К наиболее распространенным способам трансформации относятся: автоматическое распознавание текста, речи, знаков; автоматический перевод; автоматический синтез речи.*

ИДЕНТИФИКАЦИЯ ЭЛЕКТРОННОГО ДОКУМЕНТА

Идентификация производится на основе анализа блока постоянных характеристик электронного документа.

Определение постоянных характеристик электронного документа осуществляется в процессе его обработки – комплекса документальных и информационных процессов, в основе которых лежит формально-содержательный анализ.

Результатом обработки является создание метаданных по определенной схеме. Метаданные могут формироваться полностью или частично автоматически при создании и/или автоматизированной обработке электронного документа.

Схема метаданных представляет собой набор элементов метаданных, предназначенных для конкретного практического применения, например описания электронного документа. Определение значений самих элементов называется семантикой схемы. Содержание, присваиваемое элементам метаданных, называется значением. Схема метаданных в целом определяет имена элементов и их семантику, а также правила приведения значений (например, правила оформления, перечень допустимых значений) и синтаксические правила, определяющие кодировку элементов и их значений. Схема метаданных, в которой не установлены правила синтаксиса, называется синтаксически независимой, т. е. метаданные могут кодироваться в любой определяемой синтаксической системе.

Выбор схемы метаданных зависит от условий, в которых осуществляется функционирование электронного документа, т. е. от пользовательской среды, информационных процессов, целевого и пользовательского назначения, объектов и субъектов информационного взаимодействия.

В зависимости от конкретного назначения схема метаданных может модифицироваться с помощью расширения и профиля.

РАСШИРЕНИЕ – добавление элементов к уже разработанной схеме для поддержки метаданных конкретного вида электронных документов или создание метаданных для конкретной группы пользователей. Профиль используется для ограничения числа используемых элементов метаданных, для уточнения определения элементов при описании конкретного вида электронных документов, для определения значений, который может принимать тот или иной элемент.

Модификация схемы выполняет функцию разделения метаданных на универсальные (для всех электронных документов) и специальные (отдельные виды/ориентация на группы пользователей).

Универсальный набор метаданных для электронных документов содержит следующие блоки:
- данные об электронном документе как интеллектуальном объекте (сведения о создателе, заглавии, ответственности, содержании и языковой принадлежности);
- сведения об электронном документе как о физическом объекте (формат, размер, компоненты, адресная информация);
- характеристики жизненного цикла электронного документа (даты и иные параметры времени);
- данные о связи электронного документа с другими (сведения о версии, взаимном цитировании, отношениях «род–вид» и «часть–целое»);
- сведения о доступе к электронному документу (условия, права и правила использования).

СВЯЗЬ МЕЖДУ МЕТАДАННЫМИ И ЭЛЕКТРОННЫМ ДОКУМЕНТОМ, КОТОРЫЙ ОНИ ОПИСЫВАЮТ, МОЖЕТ ОСУЩЕСТВЛЯТЬСЯ ДВУМЯ СПОСОБАМИ:

- метаданные могут содержаться в записи, хра-

нящейся отдельно от описываемого электронного документа;

- метаданные могут храниться непосредственно в теле электронного документа и извлекаться по мере необходимости (например, для построения поискового индекса).

Типы метаданных:

- описательные метаданные (данные для поиска и идентификации контента электронного документа);
- структурные метаданные (данные о том, каким образом расположены и соединены элементы контента);
- административные метаданные (данные для управления и обеспечения сохранности электронного документа, включая технические и правовые аспекты).

БАЗОВЫЕ ТРЕБОВАНИЯ К ФОРМАТАМ ПРЕДСТАВЛЕНИЯ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

Цели определения базовых требований к форматам электронных документов:

- поддержка стратегического планирования в отношении цифрового контента электронных документов;
- обеспечение долгосрочного хранения и инвентаризации электронных документов, включая выявление инструментов и документации, необходимых для управления их контентом;
- разработка стратегии для поддержки стабильных форматов электронных документов в устойчивых технологических и пользовательских средах;
- разработка стратегии для конвертирования электронных документов нестабильных форматов в целях сохранения их контента в неустойчивых технологических и пользовательских средах;
- определение политики сбора электронных документов;
- разработка политики развития и использования медиа-независимых форматов электронных документов;
- обеспечение экономической эффективности создания, хранения и работы с электронными документами;
- выявление форматов электронных документов, оптимальных для использования с конкретными видами контента;
- разработка механизмов технической защиты и миграции файлов электронных документов, а также стратегии поддержки парка аппаратного обеспечения (для аппаратно-зависимых форматов электронных документов при невозможности их конвертации) или стратегии портирования программного обеспечения.

КЛАССИФИКАЦИЯ ФОРМАТОВ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

Общепринятой единой классификации форматов в настоящее время не существует. Наиболее распространенными основаниями для деления форматов на классы являются:

- расширение имени файла электронного документа (например, *.doc или *.docx, *.txt);
- тип информации: Интернет медиа-типов типа MIME (например, текст/HTML);
- цель использования (например, форматы электронных книг);
- служебное назначение или область применения (например, коммуникационные форматы ГИС);
- конкретные устройства (например, *.raw цифровых камер);
- операционные системы и носители (например,*.iso образ диска);
- алгоритм сжатия (например, *.jpg – формат сжатия графических файлов);
- степени защиты контента (например, pdf-файлы с AdobeDRM).

Примечание. *Различные форматы файлов различаются степенью детализации, один формат может накладываться на другой или использовать элементы других форматов.*

ТРЕБОВАНИЯ К ФОРМАТАМ:

Требования, перечисленные ниже, являются универсальными, т. е. относятся к цифровым форматам для всех видов электронных документов.

Для электронных документов отдельных видов могут дополнительно применяться специфические требования.

ОТКРЫТОСТЬ

Разработка открытых форматов электронных документов вызвана необходимостью создания условий для эффективного обмена электронными документами между всеми участниками процессов коммуникации на основе гармонизации способов и средств взаимодействия между информационными системами различных производителей и унификации существующих форматов электронных документов.

Соблюдение требования открытости формата представления электронного документа позволяет обеспечивать взаимодействие различных информационных систем; поддерживать возможность коллективной работы с электронными документами; предоставлять электронный документ в государственные органы, физическим и юридическим лицам; обеспечивать унифицированную обработку по стандартной схеме метаданных и надежность долговременного хранения электронных документов.

Использование открытых форматов файлов необходимо для организации публичных сервисов, создания электронных документов, которые вводятся в публичное обращение, при проведении государственных тендеров на разработку или закупку программного обеспечения и т. п.

Основные разработчики стандартов открытых форматов:

- ISO – International Organization for Standardization;
- ECMA – European Computer Manufacturers Association; Ecma International – European association for standardizing information and communication systems;
- NISO – National Information Standards

Organization, a non-profit association accredited by the American National Standards Institute (ANSI);
- OASIS – Organization for the Advancement of Structured Information Standards;
- W3C – The World Wide Web Consortium;
- ITU-T – Telecommunication Standardization Sector of the International Telecommunication Union.

РАСПРОСТРАНЕННОСТЬ

Распространенность предполагает максимально широкий круг пользователей электронных документов, представленных в данном формате, включая первичных создателей, распространителей, пользователей электронных документов.

Распространенность включает в себя использование формата:
- в качестве мастер-формата электронного документа;
- для доставки электронного документа конечным пользователям;
- как средство обмена между системами.

Распространенность формата замедляет его устаревание и предполагает при разработке новых форматов электронных документов параллельное развитие инструментов их конвертации без дополнительных экономических затрат на миграцию, портирование и эмуляцию систем для работы с электронными документами устаревших распространенных форматов.

Свидетельством распространенности формата электронного документа является: 1) поддержка его максимально большим количеством конкурирующих программных средств для создания, просмотра, поиска, воспроизведения (вне зависимости от производителя и его лицензионной политики) и 2) выбор электронных документов в данном формате максимальным количеством конечных пользователей при наличии альтернативных форматов представления тех же электронных документов.

Распространенность формата электронных документов является основой для совместной работы организаций, осуществляющих библиотечно-информационную деятельность, органов научно-технической информации, организаций, официально выпускающих в публичное обращение электронные документы с целью их массового использования.

ПРОЗРАЧНОСТЬ

Прозрачность предполагает минимальное количество аппаратно-программных средств, которое требуется задействовать для того, чтобы контент электронного документа стал доступен конечному пользователю. Соответственно, форматы, которые предполагают различные виды пост-обработки электронных документов с целью оптимизации (особенно шифрование и сжатие), будут обладать меньшей прозрачностью, чем электронные документы в формате, где сжатие не использовалось.

Прозрачность усиливается, если текстовое содержание (в том числе текст метаданных, внедренных в файлы электронных документов с графическим (нетекстовым) контентом) кодируется в стандартных кодировках (например, UNICODE в кодировке UTF-8) и хранится в естественном порядке чтения.

Существуют некоторые виды контента, которые даже при создании электронного документа не могут быть сохранены в несжатом виде. Для унификации требований прозрачности необходимо определение коэффициента прозрачности форматов, используемых для разных видов электронных документов, для разных целей использования и разных видов деятельности.

ВНЕДРЕНИЕ МЕТАДАННЫХ

Форматы, позволяющие внедрение метаданных, т. е. создание и хранение их непосредственно в теле документа, являются наиболее предпочтительными для создания/перевода в них электронных документов. При этом оптимальным является использование форматов с автометаданными, т. е. тех, в которых часть значений полей автоматически формируется программными средствами. Максимальная полнота автометаданных, документирующих жизненный цикл электронного документа с момента его создания, облегчает его поддержку в долгосрочной перспективе и обеспечивает наименьшую уязвимость во всех неинформационных процессах, связанных с обеспечением сохранности.

В современном динамично развивающемся обществе электронные документы становятся таким же стратегическим ресурсом, как традиционные материальные и энергетические ресурсы. Эволюция информационного общества немыслима без использования информационных ресурсов в электронном виде, а использование электронных информационных ресурсов стало бы крайне осложнено в будущем без стандартизации процессов перевода аналоговых документов в электронную форму и создания электронных информационно-насыщенных систем. Уже достигнутый качественно иной уровень производства, хранения и распространения информации обеспечивает все более широкое и эффективное использование электронных документов и поэтому настойчиво требует придать им соответствующий статус.