МЕТОДЫ И СРЕДСТВА РАЗРАБОТКИ ПРОГРАММНЫХ СИСТЕМ

В.В. Ланин

Национальный исследовательский университет
«Высшая школа экономики»
(Пермский филиал)

vlanin@live.com

КЛАССИФИКАЦИЯ ФОРМАТОВ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

Введение

Вопрос выбора представления (формата) электронных документов является крайне важным. От представления документа зависят способы его обработки, спектр решаемых задач, для которых применим данный формат, и, в конечном итоге, удобство работы с ним пользователей. Особые возможности для обработки документов обеспечивает их семантическая индексация.

Для выбора нужного формата представления документа необходимо оценить все возможности, обеспечиваемые использованием этого формата. Для решения данной задачи полезно выполнить классификацию существующих форматов. При этом особое внимание следует уделить возможности включения метаданных как основного механизма поддержки семантического индексирования.

Цель данной работы – построить классификацию электронных документов, их форматов, которая может быть использована для выбора оптимального формата при решении задач представления и хранения, обработки электронных документов в информационных системах различного назначения.

Существующие классификации электронных документов

Рассмотрим понятие формата документа и связанное с ним понятие формата файла. Заметим, что хотя эти понятия используются повсеместно, но четкая их формулировка до сих пор не устоялась, исследователи продолжают дискуссию по формулировке данных определений [1-5].

Формат файла – набор семантических, синтаксических схем и правил сериализации для преобразования абстрактной информации в определенные наборы байт [4]. В свою очередь, *формат документа* – это формат файла для хранения документов на носителе информации.

В настоящее время используется большое количество различных форматов, существенно отличающихся по концепции и возможностям представления содержания. Общепринятой классификации форматов на данный момент нет. Чаще всего форматы документов классифицируют *в зависимости от задач*, на решение которых они ориентированы, и *подходов к обработке документов* в этих форматах. Например, в [6] вводится классификация, учитывающая особенность обработки документов при решении задач поиска. Форматы делятся на три класса:

1) устаревшие и закрытые коммерческие форматы (Microsoft Word, Lexicon, WordStar);
2) форматы, ориентированные на графическое представление документа (PostScript, PDF);
3) форматы, содержащие динамически выполняющиеся элементы (современные Интернет-сайты).

В [7] описана классификация электронных документов (ЭД) *по метаданным*, которые исследователь определяет как характеристики технологических процессов, необходимых для визуализации. Автор отмечает, что современные ЭД – это комбинированные сложноформатные документы, а тенденции, наблюдаемые в компьютерной отрасли, говорят об их дальнейшем усложнении. В основе классификационной модели – *ранжирование ЭД* в зависимости от комбинации метаданных. Выделяется пять классов документов:

- *Одноранговые* документы состоят из одного типа данных, объединенных одной структурой, записанных в один файл определенного формата (простые текстовые файлы, графические изображения, оцифрованные аудио- и видеодокументы).
- *Двухранговые* документы состоят из нескольких типов данных, объединенных одной структурой, записанных в один файл определенного формата (файлы электронных таблиц, документы векторных графических форматов).

– *Трехранговые ЭД* состоят из нескольких типов данных, объединенных в несколько структур, записанных в один файл определенного формата.
– *Четырехранговые ЭД* состоят из нескольких типов данных, объединенных в несколько структур, записанных в нескольких файлах общего формата.
– *Пятиранговые ЭД* состоят из нескольких типов данных, объединенных в несколько структур и записанных в нескольких файлах разных форматов.

Внутри ранга электронные документы можно разделить *по конкретным технологиям, в рамках которых они были созданы*: текстовым или графическим процессорам, типам СУБД, гипертекстовым системам и т.п.

Ранжирование ЭД позволяет подходить к ним с общих методологических позиций. Рассмотрение происходит в зависимости от объема материальных и интеллектуальных затрат, необходимых для проведения последующей миграции. Чем выше ранг ЭД, тем сложнее ее осуществить. Для обеспечения долговременной сохранности четырех- и пятиранговых ЭД требуются большие машинные ресурсы, труд высококвалифицированных программистов и ясное представление о их строении.

Следует упомянуть также *стандарт MIME* (Multipurpose Internet Mail Extensions – многоцелевые расширения Интернет-почты) – стандарт, описывающий передачу различных типов данных по электронной почте, а также, шире – спецификация для кодирования информации и форматирования сообщений таким образом, чтобы их можно было пересылать по сети Интернет.

Многоаспектная классификация форматов документов

Для анализа возможностей различных форматов ЭД введем *многоаспектную классификацию*.

В зависимости от *типа лицензии* на него формат может быть проприетарным или открытым. *Проприетарные* (*proprietary*) форматы создаются и контролируются частными организациями и являются собственностью авторов или правообладателей. *Открытый* (*non-proprietary*) формат обычно разрабатываемая некоммерческой организацией по стандартизации, спецификация формата доступна, использование формата свободно от лицензионных ограничений.

Спецификация формата может быть *закрытой* (*closed specifications*) или *открытой* (*open specifications*). Доступность специ-

фикации часто связывают с проприетарностью формата, что в большинстве случаев действительно так (формат Microsoft Word), но встречаются и исключения (формат PDF). Для поддержки форматов с закрытой спецификаций в продуктах третьих лиц применяется обратный инжиниринг, что зачастую негативно влияет на стабильность работы и нарушает условия использования формата.

В зависимости от *способа представления данных в файле* можно выделить *бинарные* и *текстовые форматы*. Первым форматом, используемым для представления текстовой информации, является *простой текстовый формат* (*plain text*), в котором в явном виде представлены символы текста, а из служебных символов используются только команды управления кареткой. Практически без каких-либо изменений он повсеместно применяется и по сей день. Одно из основных преимуществ текстового формата заключается в возможности его обработки без каких-либо специализированных программных средств. На базе текстового представления файла построено множество других форматов (например, RTF (*Rich text format*), XML и др.). Основными недостатками текстового представления являются значительный объем занимаемой памяти и последовательный доступ к содержимому. *Бинарные форматы*, напротив, достаточно компакты и поддерживают произвольный доступ к содержимому. Однако для доступа к их содержимому требуются знания об их структуре, что часто используется как «защищающий» фактор для проприетарных форматов.

Если классифицировать форматы электронных документов *по возможностям форматирования* (под форматированием понимается изменение «внешнего вида» текста, при котором не изменяется его содержание), то можно выделить следующие классы:

- «простые» текстовые форматы, ориентированные на представление только текстовой информации и не допускающие форматирования без изменения содержания;
- текстовые документы с поддержкой форматирования.

Также можно ввести разделение форматов электронных документов в зависимости *от способа использования*:

- форматы, ориентированные на редактирование;
- форматы, ориентированные на представление информации.

По *возможности расширения формата* (включения в файл дополнительных метаданных):

- расширение формата невозможно;
- расширение формата возможно неспецифичными средствами;
- расширение формата поддерживается спецификацией.

Результаты проведённой классификации представлены в табл. 1.

Таблица 1. Классификация наиболее популярных форматов ЭД

Формат	Тип лицензии	Спецификация	Способ представления	Форматирование	Назначение формата	Расширение формата
eXtensible Markup Language, XML	откр.	откр.	текст.	нет	ред.	есть
HyperText Markup Language, HTML	откр.	откр.	текст.	есть	просм.	нет
Rich Text Format, RTF	пропр.	откр.	текст.	есть	ред.	нет
Open XML	откр.	откр.	текст.	есть	ред.	есть
OpenDocument Format, ODF	откр.	откр.	текст.	есть	ред.	есть
Portable Document Format, PDF	пропр.	откр.	бинарн.	есть	просм.	нет
XML Paper Specification, XPS	откр.	откр.	текст.	есть	просм.	есть
Microsoft Word, Doc	пропр.	закр.	бинарн.	есть	ред.	есть
TeX	откр.	откр.	текст.	есть	просм.	нет
WordPerfect	пропр.	откр.	бинарн.	есть	ред.	нет
DjVu	пропр.	откр.	бинарн.	нет	просм.	нет

Заключение

Представленная классификация отражает все существенные характеристики форматов ЭД и может быть использована для выбора оптимального формата при решении задач представления и хранения, обработки документов в информационных системах различного назначения.

Библиографический список

1. *Brown A.* Selecting file formats for long-term preservation The National Archives (UK) Digital preservation guidance note 1. 2008. URL: http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf.
2. *Авдеев А.Н., Ларин М.В.* Об организации электронного документооборота // Документация в информационном обществе: проблемы оптимизации документооборота: докл. и сообщ. на XVIII Междунар. науч.-практ. конф., 26-27 окт. 2011 г. / Федер. арх. агентство (Росархив), Всерос. науч.-исслед. ин-т документоведения и арх. дела (ВНИИДАД). М.: ВНИИДАД, 2012. С. 389-404.
3. *McLellan E.P.* General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation. URL: http://www.interpares.org/display_file.cfm?doc=ip2_gs11_final_report_english.pdf.
4. *Todd M.* Technology Watch Report: File formats for preservation. URL: http://www.dpconline.org/component/docman/doc_download/375-file-formats-for-preservation.
5. *Abrams S.* Instalment on “File Formats”. URL: https://www.era.lib.ed.ac.uk/bitstream/handle/1842/3351/Abrams%20file-formats.pdf?sequence=1.
6. *Губин М.В.* Модели и методы представления текстового документа в системах информационного поиска: автореф. дис. канд. физ.-мат. наук, СПб., 2005.
7. *Тихонов В.И.* Сущностные характеристики, состав и классификация электронных документов // Документация в информационном обществе: электронное делопроизводство и электронный архив. М.: Росархив. ВНИИДАД. РОИА. 2000. С. 204-218.