СОДЕРЖАНИЕ

ВВЕДЕНИЕ

1 Аналитический раздел

1.1 Структура PDF файла

1.1.1 Представление PDF файла

PDF документ имеет иерархическую структуру (дерево), корнем которого является словарь Catalog. Визуализацию данного дерева можно рассмотреть на рисунке 1.1.

Каталог содержит ссылки на вершины описания страниц. Поддеревья страниц отсортированны, что позволяет быстро находить необходимую страницу. Словарь каждой страницы хранит ссылку на словарь ресурсов, который хранит требуемые шрифты, изображения т. д. [1].

1.1.2 Разделы PDF файла

Структура PDF файла включает 4 раздела:

1) Заголовок;
2) тело;
3) Таблица перекрестных ссылок;
4) хвост [1].

Рассмотрим каждый раздел по отдельности.

Заголовок

Заголовком называется первая строка файла. Она содержит информацию о версии PDF [1].

Тело

Все содержимое документа находится в теле файла. Информация, которая отображается пользователю представлена восемью типами данных:

1) Булевы значения. Принимают значения true или false);
2) числа. Включают два типа данных — integer (целочисленный) и real (вещественный). Дробная часть в вещественных числах отделяется точкой;

3) имена. Представляют собой последовательность ASCII символов. Они начинаются со слеша, который не входит в имя. Вместо непосредственно символов могут включать их шестнадцатеричные коды, начинающиеся с символа #;

4) строки. Ограничены длиной в 65535 байтов. Записываются в круглых либо треугольных скобках. Могут быть представлены как ASCII символами, так и шестнадцатеричными или восьмеричными кодами.

5) Массивы. Могут содержать любые PDF-объекты. Элементы разделяются пробелом и заключаются в квадратные скобки;

6) словари. Представляют коллекцию пар ключ-значение. Ключом должно быть имя, а значением может быть любой объект. Запись словаря начинается с символов «, а заканчиваются — »;

7) потоки. Потоки содержат неограниченные последовательности байтов. В них содержится основное содержимое документов. Поток начинается с ключевого слова stream и заканчивается словом endstream. Перед началом потока записывается словарь с мета-информацией. Он включает данные о количестве байтов, фильтре применимом их к обработке и так далее;

8) null-объекты. Представляются ключевым словом null [1].

PDF объектом является любой вышеперечисленный тип, содержащий информацию [1].

Хвост

Данный раздел начинается с ключевого слова trailer и содержит:

1) Словарь;

2) смещение относительно таблицы перекрестных ссылок (англ. cross-reference table);

3) маркер конца файла %%EOF.

В словарь данного раздела входят:

1) Данные о количестве объектов (ключевое слово Size);

2) ссылки на каталог документа (ключевое слово Root);

3) информационный словарь (ключевое слово Info);

4) идентификатор файла (ключевое слово ID) [1].

Таблица перекрестных ссылок

Cross-reference table позволяет получать произвольный доступ к любому объекту в файле. Данная таблица состоит из секций. Каждая секция соответствует новой версии документа, данная таблица начинается с ключевого слова xref, так что иногда ее называют xref таблицей [2].

Любой PDF-объект может быть помечен уникальным идентификатором и использоваться как ссылка. Такие объекты называются косвенными. Они начинаются с идентификатора, номера поколения и ключевого слова obj. Заканчивается косвенный объект словом endobj. На эти объекты можно ссылаться в таблице cross-reference table и любом другом объекте (для этого используется символ R) [2].

1.2 Виды PDF форматов

В данной части работы будут проанализированы существующие виды PDF документов. Существует несколько различных видов PDF документов, каждый из которых имеет свои особенности и ограничения:

1) PDF/A;

2) PDF/X;

3) PDF/E;

4) PDF/UA.

Рассмотрим каждый из них по отдельности.

1.2.1 PDF/A

Данный формат, предназначенный для долгосрочного хранения документов. Он обеспечивает сохранность и неприкосновенность содержимого даже

через длительные периоды времени. Однако, PDF/A ограничен в функциональности и не поддерживает некоторые расширенные возможности форматов PDF [3]. Данный формат также разделяется на несколько подклассов: PDF/A-1, PDF/A-2, PDF/A-3, PDF/A-4.

Также вводится новое понятие уровня соответствия, оно накладывает дополнительные требования на классы PDF/A, для предоставления дополнительных возможностей. Расссмотрим уровни соответствия.

1) Уровень b (Basic). Цель: обеспечение надёжного воспроизведения внешнего вида документа. Распространяется на файлы формата: PDF/A-1b, PDF/A-2b, PDF/A-3b;

2) уровень a (Accessible). Цель: обеспечение возможности поиска и преобразования содержимого документа. Включает все требования уровня b и дополнительно требует, чтобы была включена структура документа. Также вводит требования:

    1) Содержимое должно быть помечено деревом иерархической структуры, что означает, что такие элементы, как порядок чтения, рисунки и таблицы, явно идентифицируются с помощью метаданных.
    2) Должен быть указан естественный язык документа.
    3) Изображения и символы должны иметь альтернативный описательный текст. Файл должен включать сопоставление символов с Unicode.

    Распространяется на файлы формата: PDF/A-1a, PDF/A-2a, PDF/A-3a;

3) уровень u (Unicode). Распространяется на файлы формата: PDF/A-2u, PDF/A-3u. Требуется сопоставление символов с Unicode. Изменения: отбрасываются требования уровня a, включая встроенную логическую структуру (т. е. теги и дерево структур);

4) уровень f (Format). Распространяется на файлы формата: PDF/A-4f. Изменения: позволяет встраивать типы файлов любого другого формата;

5) уровень e (Engineering). Распространяется на файлы формата: PDF/A-4e. Изменения: поддержка аннотаций типов RichMedia и 3D [3].

PDF/A-1

PDF/A-1 - самый распространенный формат оригинального PDF/A на сегодняшний день. Он основан на PDF 1.4 и является наиболее ограниченным, так как не поддерживает JPEG 2000, вложения, слои и прозрачность. Часть 1 стандарта была опубликована 28 сентября 2005 года и определяет два уровня соответствия для файлов PDF: PDF/A-1b и PDF/A-1b [5].

PDF/A-2

PDF/A-2 предоставляет собой ряд новых функций:

1) Сжатие JPEG2000, что особенно полезно для отсканированных документов, таких как карты, книги, а также документов с цветным содержимым, таких как чеки или паспорта;

2) Вложенные файлы PDF/A через коллекции: Acrobat позволяет пользователям создавать коллекции (иногда также называемые "портфелями"), где несколько документов PDF/A объединяются в один "контейнерный"документ PDF;

3) Необязательное содержимое (слои): Необязательное содержимое, иногда также называемое слоями, полезно для приложений картографии или инженерных чертежей, где отдельные слои могут быть показаны или скрыты в соответствии с требованиями просмотра;

4) Новый уровень соответствия PDF/A-2u - "u"для Unicode. Он упрощает поиск и копирование текста Unicode для цифровых PDF-документов и PDF-документов, которые были отсканированы с последующим оптическим распознаванием символов (OCR);

5) Метаданные на уровне объекта XMP: PDF/A-2 определяет требования к настраиваемым метаданным XMP;

6) Цифровые подписи: В то время как PDF/A-1 уже позволяет использовать цифровые подписи, PDF/A-2 определяет правила, которые должны быть применены для гарантии взаимодействия [5].

PDF/A-3

PDF/A-3 полностью аналогичен PDF/A-2, однако поддерживает добавление любых файлов, а не только PDF типа A. Однако не гарантирует валидность их прочтения в будущем [5].

Также стоит отметить, что файлы данного вида возможно использовать в электронном документообороте [6].

PDF/A-4

Основное отличие данного вида, является замена уровней соответствия b и u с целью упростить стандарт. PDF/A-4 требует отображения в Юникоде для всех шрифтов в любое время [7].

1.2.2 PDF/X

Формат, разработанный специально для обмена и печати документов в издательской отрасли. Он обеспечивает точность цветов и расположения элементов страницы, что особенно важно при печати. Однако, PDF/X имеет ограниченные возможности вставки мультимедийных элементов и интерактивности [8].

1.2.3 PDF/E

Формат, предназначенный для обмена и хранения документов в инженерной отрасли. Он поддерживает вставку трехмерных моделей, векторных изображений и других инженерных элементов. Однако, PDF/E может быть ограничен в возможности обработки сложных макетов и мультимедийных элементов [8].

1.2.4 PDF/UA

Формат, предназначенный для создания доступных документов для пользователей с ограниченными возможностями. Он обеспечивает структурированное представление контента и поддержку технологий чтения вслух и управления навигацией. Однако, PDF/UA может иметь ограничения в отображении сложных макетов и интерактивных элементов [8].

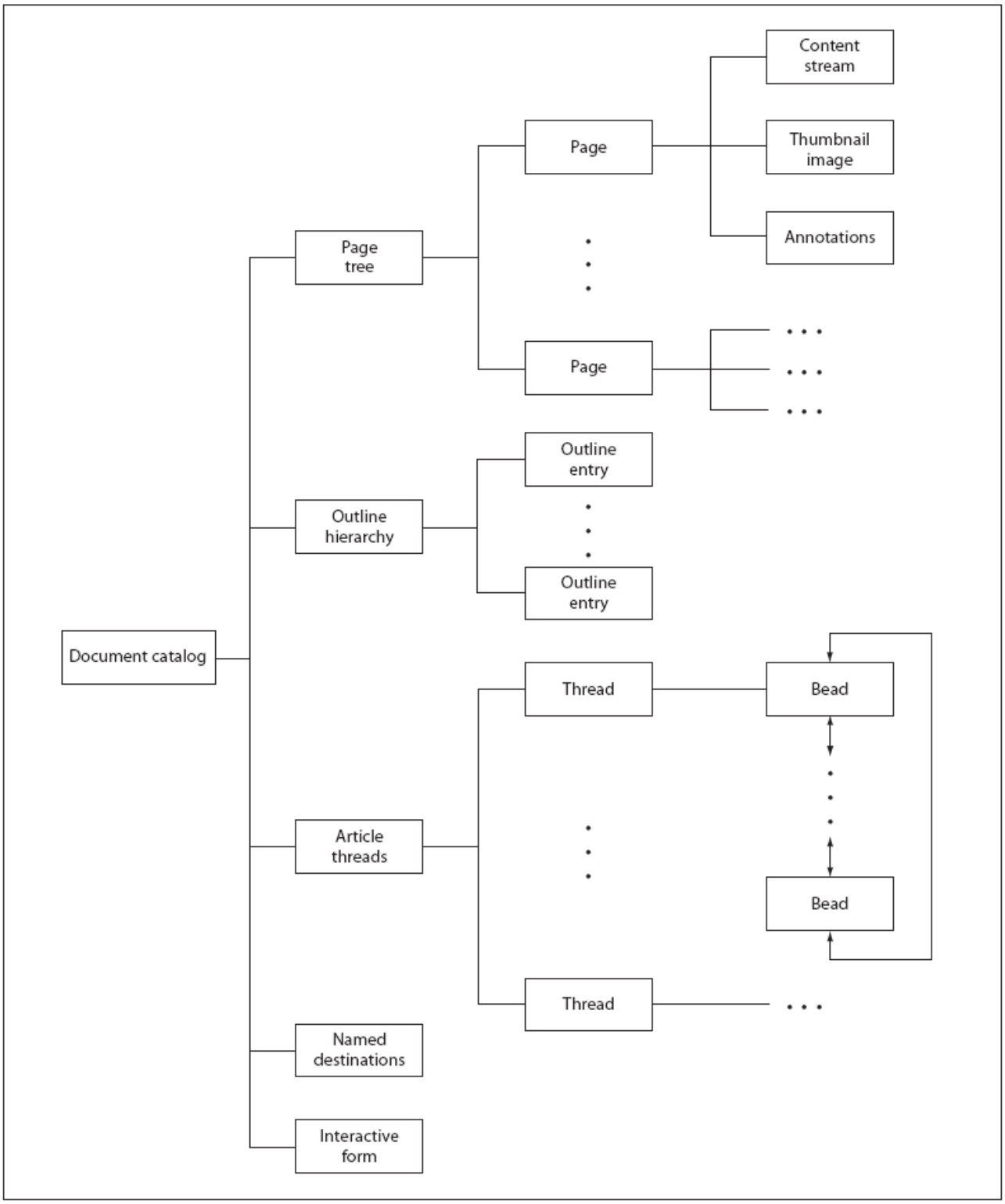

Рисунок 1.1 – Пример дерева PDF документа

2 Конструкторский раздел

2.1 Описание системы автоматической проверки отчета

С помощью использования алгоритма автоматической проверки отчета возможно существенно сократить временные ресурсы, выделяемые нормоконтролером на проверку огромного количества отчетов, однако, полностью отказаться от финального контроля результатов человеком невозможно, таким образом существует две роли при проверки отчета на соответствие ГОСТ, а именно студент и нормоконтролер.

Студент отправляет отчет на проверку, а затем получает результат со списком ошибок (если имеются). Нормоконтролер же анализирует отчет, составленный автоматической системой проверки, и при необходимости может внести необходимые поправки

3 Технологический раздел

4 Исследовательский раздел

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. What if PDF file? [Электронный ресурс]. — Режим доступа: https://docs.aspose.com/page/net/what-is-pdf-file/ (дата обращения: 26.10.2023).
2. Дружим с PDF [Электронный ресурс]. — Режим доступа: https://alexeykalina.github.io/technologies/pdf.html (дата обращения: 26.10.2023).
3. Стандарт PDF/A [Электронный ресурс]. — Режим доступа: https://yamadharma.github.io/ru/post/2021/07/30/pdf-a-standard/ (дата обращения: 26.10.2023).
4. What Are the Different Versions of PDF/A? [Электронный ресурс]. — Режим доступа: https://apryse.com/blog/pdfa-format/what-are-the-different-types-of-pdfa (дата обращения: 19.10.2023).
5. PDF/A-2 Overview [Электронный ресурс]. — Режим доступа: https://pdfa.org/wp-content/uploads/2011/10/Flyer-PDFA2-Overview-EN.pdf (дата обращения: 19.10.2023).
6. Приказ ФНС России от 24.03.2022 [Электронный ресурс]. — Режим доступа: https://www.nalog.gov.ru/rn77/about_fts/docs/12181055/ (дата обращения: 19.10.2023).
7. https://blog.avepdf.com/what-is-pdfa4/ [Электронный ресурс]. — Режим доступа: https://blog.avepdf.com/what-is-pdfa4/ (дата обращения: 19.10.2023).
8. PDF File types - specialist formats and what they're used for [Электронный ресурс]. — Режим доступа: https://www.adobe.com/uk/acrobat/resources/document-files/pdf-types.html (дата обращения: 19.10.2023).