УДК 004.912

ИНТЕГРАЦИЯ АЛГОРИТМОВ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ О СТРУКТУРНЫХ ЭЛЕМЕНТАХ ДОКУМЕНТОВ ФОРМАТА ODT В СЕРВИС АВТОМАТИЗИРОВАННОГО НОРМОКОНТРОЛЯ

Терещенко В. В. (Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»)

Научный руководитель - доцент, Горлушкина Н. Н. (Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»)

Введение. ODT (OpenDocument Format) – открытый формат документов, созданных при помощи программного обеспечения для обработки текстов, которое поддерживает работу с документами формата OpenDocument [1]. Документы формата ODT представляют собой архив, содержащий папки и XML-файлы. XML-файлы [2] хранят как текст документа, так и метаданные, такие как например информация о дате и времени создании, последнем изменении документа, количестве слов и символов, информация об авторе документа и текстовом редакторе, в котором рассматриваемый документ был создан. Также в отдельных файлах хранятся данные об используемых стилях, структурных элементах, шрифтах, различные параметры печати и вывода на экран [3].

Автоматизация нормоконтроля связана со сложностью парсинга ODT документов, которые зависят от XML разметки и организации иерархии стилей внутри самих документов. В связи с описанными особенностями работы с форматом необходимо создать объекты, хранящие свойства и характеристики, получаемые из структурных элементов документа с помощью алгоритмов парсинга.

Основная часть. В электронных документах формата ODT основная информация о структурных элементах хранится в файле content.xml в виде атрибутов стилей. Стили для таблиц описывают концепцию сеток строк, сеток столбцов и ячеек. Строки и столбцы отображаются в соответствующих группах, которые определяют следует ли продолжать строку или столбец на следующей странице. Для адаптации существующей концепции был реализован ряд объектов - таблица, столбец, строка и ячейка.

В свою очередь списки состоят из заголовка, за которым следует любое количество элементов. Каждый список содержит счетчик для своих элементов и любого вложенного списка, который он может содержать. Также каждый список содержит стиль, который применяется ко всем его элементам и вложенным спискам. На основе формата списков и их свойств, были созданы объекты стиля, маркированного и нумерованного списков.

Информация об изображениях хранится в двух объектах - frame, который является контейнером для расширенного содержимого и хранит его характеристики, и image, хранящий метаданные изображения. Поскольку описанные графические объекты хранят связанную информацию, было решено агрегировать необходимые атрибуты в одном общем объекте.

Алгоритмы для работы с объектами имеют общую концепцию. В них выполняется обход всех элементов в автоматических стилях. Для каждого элемента, который содержит любой из рассматриваемых структурных объектов, извлекаются все характеристики главного и дочерних элементов. Затем определяется к какому типу объект относится, создается его экземпляр и добавляется в результирующую коллекцию объектов. Следует отметить, что для получения информации об изображениях необходимо было получить данные из структуры element_dict, хранящей записи обо всех типах используемых объектов и самих объектов. Для этого были получены все ключи словаря элементов и преобразованы в список кортежей, и затем проходя циклам по всем элементам списка находился нужный токен для извлечения.

Затем, используя полученный токен, извлекались записи обо всех интересующих объектах, хранящихся в документе.

Выводы. В результате были созданы объекты, хранящие характеристики таблиц, списков и изображений электронных документов формата ODT. Созданные объекты и алгоритмы для извлечения их характеристик добавлены к другим созданным унифицированным сущностям сервиса нормоконтроля и лягут в основу подготовки данных для последующего создания классификатора и автоматизации процесса нормоконтроля.

Список использованных источников:

1. Open Document Format for Office Applications (OpenDocument) Version 1.2. Part 1: OpenDocument Schema. 2010-06-10.
2. P. Walmsley, Definitive XML Schema - A.: Prentice Hall, 2013 - 768 с.
3. Марцинкевич В. И., Ларионова Г. С., Терещенко В. В., Ситникова К. А. Анализ возможностей парсинга электронных текстовых документов для автоматизации нормоконтроля // Экономика. Право. Инновации. - 2022. - Т. 1, №.3. - С. 39-49.

Терещенко В. В. (автор) Подпись

Горлушкина Н. Н. (научный руководитель) Подпись