DOI 10.37882/2223–2966.2021.08.14

ПОСТРОЕНИЕ ОБЪЕКТНОЙ МОДЕЛИ ЭТАЛОННОГО ТЕКСТОВОГО ДОКУМЕНТА ДЛЯ СЕРВИСА АВТОМАТИЗИРОВАННОГО НОРМОКОНТРОЛЯ

CREATION THE OBJECT MODEL OF THE TEXT BENCHMARK DOCUMENT FOR AUTOMATION THE NORM RULE CHECKING SERVICE

E. Kobets
N. Nasyrov
P. Tartynskikh
N. Gorlushkina

Summary. The scientific paper describes the development of the object model of the benchmark document for the norm rule checking service. An approach is presented of feature highlighting of execution of text documents elements from standards. In the research results have been highlighted and classified basic structural elements and their constituent sub-elements. These elements formed the basis for the creation of the object model of the text benchmark document.

The results of the implementation of the object model were used in the process of forming a dataset for machine learning algorithms for the classifying of structural elements. Based on the achieved classification results the further verification in progress of elements is carried out according to the requirements of a particular class.

The model allows to prepare rules for automation search of errors in the execution of text documents. The model helps to implement the software realization of this search.

Keywords: object model, automation, norm rule checking, benchmark document, formalization errors, execution of text documents, standards, unification.

Кобец Елизавета Александровна
Ведущий инженер, аспирант, Университет ИТМО (Санкт-Петербург)
www.kobets@yandex.com
Насыров Наиль Фаизович
Ассистент, аспирант, Университет ИТМО (Санкт-Петербург)
pasdel@mail.ru
Тартынских Петр Сергеевич
Инженер, Университет ИТМО (Санкт-Петербург)
tartynskikh.ps@yandex.ru
Горлушкина Наталия Николаевна
К.т.н., с.н.с., доцент, Университет ИТМО (Санкт-Петербург)
nagor.spb@mail.ru

Аннотация. В статье описывается разработка объектной модели эталонного документа для сервиса автоматизированного нормоконтроля. Представлен подход выделения признаков оформления элементов текстовых документов из стандартов. В результате анализа были выделены и классифицированы базовые структурные элементы (классы) и составляющие их подэлементы. Эти элементы легли в основу создания объектной модели эталонного текстового документа. Результаты реализации объектной модели были использованы в процессе формирования датасета для алгоритмов машинного обучения классификации структурных элементов. На основании полученных результатов классификации выполняется дальнейшая проверка элементов по требованиям отдельного класса.

Модель позволяет подготовить правила для автоматизированного поиска ошибок оформления текстовых документов. Модель помогает осуществить программною реализацию этого поиска.

Ключевые слова: объектная модель, автоматизация, нормоконтроль, эталонный документ, ошибки оформления, оформление текстовых документов, стандарты, унификация.

Введение

Разработка объектной модели представляет собой творческий процесс, который слабо формализуется. Однако, без разработки такой модели сложно создать оптимальную структуру будущей системы. Поэтому объектной модели, описывающей совокупность объектов и их взаимодействие в системе, уделяется большое внимание [1,2]. Целью представляемого исследования является разработка объектной модели эталонного документа для сервиса автоматизированного нормоконтроля, которая позволит расширять функции сервиса, унифицировать объекты, повышать точность проверки требований оформления.

Под эталонным текстовым документом понимается используемая в рамках сервиса абстрактная модель, которая инкапсулирует эталонные значения свойств (или диапазоны этих значений) документа, согласно тем или иным требованиям.

Таблица 1. Формализация требований по тексту ГОСТов

Элемент	Правила для разработчика		Вывод
	ГОСТ 7.32–2017	ГОСТ Р 7.0.11–2011	Сходства / Отличия
1. Поля, колонтитулы	п.6.1.1 <…> Текст отчета следует печатать, соблюдая следующие размеры полей: левое — 30 мм, правое — 15 мм, верхнее и нижнее — 20 мм. <…>	п.5.3.7 <…> Страницы диссертации должны иметь следующие поля: левое — 25 мм, правое — 10 мм, верхнее — 20 мм, нижнее — 20 мм. <…>	Сходства: в оформлении нижнего поля, в оформлении верхнего поля, Отличия: в оформлении левого поля, в оформлении правого поля. в размере абзацного отступа.
2. Абзацный отступ	п.6.1.1 <…> Абзацный отступ должен быть одинаковым по всему тексту отчета и равен 1,25 см. <…>	п.5.3.7 <…> Абзацный отступ должен быть одинаковым по всему тексту и равен пяти знакам. <…>	
…	…	…	…

В статье описывается подход выделения признаков оформления элементов текстовых документов. Объектная модель, рассматриваемая в статье, создается для инструмента "Сервис автоматизированного нормоконтроля документов и обучения оформлению документации" [3,4] (в дальнейшем Сервис). Проектирование такого Сервиса — задача нетривиальная и требует креативного ассоциативного мышления. При решении этой задачи использовались методы анализа, классификации, машинного обучения, что позволило создать инструмент для корректного оформления текстов, таких как выпускные квалификационные работы. Но разнообразие текстовых документов и требований к их оформлению имеет широкий спектр, поэтому необходимо создание инструмента с универсальным алгоритмом, который позволит добавлять/расширять функциональные возможности при увеличении количества нормативных документов, по которым проверяется оформление. Алгоритм формализации требований к оформлению документов для Сервиса был описан в работе [5].

Разработка объектной модели эталонного текстового документа предполагает формализацию требований оформления текстов по ГОСТам и стандартам. Именно для изучения возможностей использования требований, зафиксированных в различных стандартах и нормативных документах, проводится анализ нескольких ГОСТов.

Объектная модель эталонного текстового документа будет служить основой подготовки правил разработчикам для последующей реализации автоматизированного поиска ошибок оформления в текстах работ.

Основная часть

Изучение ГОСТов

Для изучения и сравнения были выбраны ГОСТ 7.32–2017 [6] и ГОСТ Р 7.0.11–2011 [7], потому что именно в них фиксируются основные требования к оформлению научных текстов (отчетов, выпускных квалификационных работ, диссертаций и авторефератов диссертаций). В процессе исследования было выявлено:

1. В ГОСТ 7.32–2017:
1. Описываются требования к документу, который содержит систематизированные данные о научно-исследовательской работе, состояние научно-технической проблемы, процесс, результаты научно-технического исследования.
2. Дается описание по структуре и правилам оформления соответствующих текстовых документов.
3. Приводятся требования (разделы 5 и 6) к оформлению отдельных элементов документа, а также общие требования к оформлению документов.
4. В ГОСТ Р 7.0.11–2011:

Полный состав документа может различаться по способу представления работы:

a) диссертация в виде рукописи (раздел 4, 5),
b) диссертация в виде научного доклада (раздел 6, 7),
c) автореферат диссертации (раздел 8, 9).

В ходе анализа было выявлено, что ряд типичных элементов текстов отчетов и диссертаций имеют как сходства, так и различия в требованиях к оформлению.

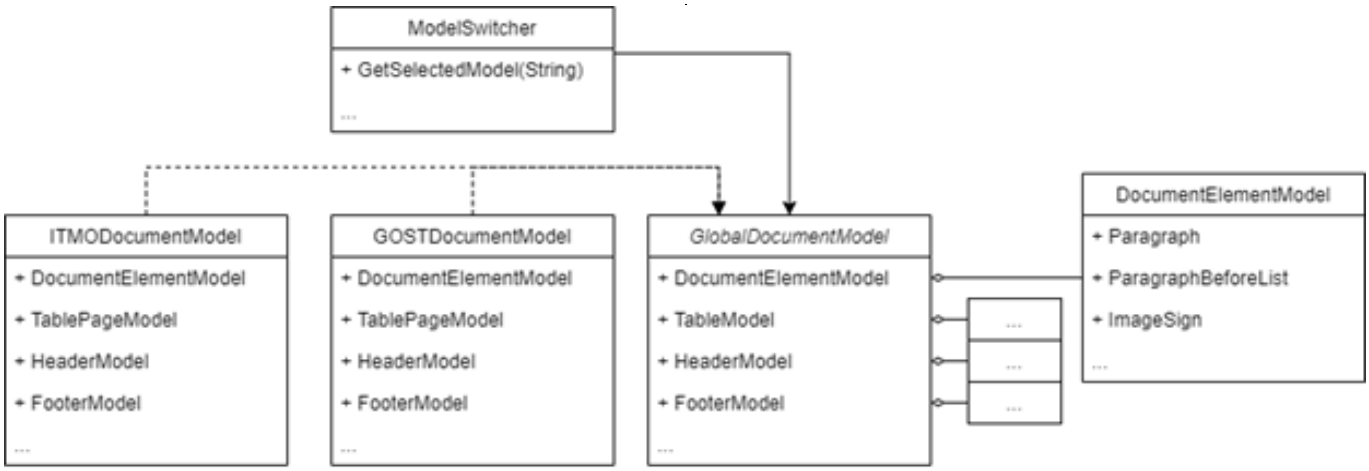

Рис. 1. Диаграмма классов (модель документа)

Например, для заголовков в обоих ГОСТах идентичными являются требования к заголовкам относительно:

- расположения на странице: посередине страницы и без точки на конце (ГОСТ 7.32–2017 п. 6.2, ГОСТ 7.0.11–2011 п. 5.3.5),
- переносов слов, которые — не допускаются (ГОСТ 7.32–2017 п. 6.2.4, ГОСТ 7.0.11–2011 п. 5.3.5),
- отточия, которым соединяется последнее слово заголовка и соответствующие ему номера страниц — в оглавлении (ГОСТ 7.32–2017 п. 5.4.1, ГОСТ 7.0.11–2011 п. 5.2.3),
- записи приложений и их перечисления: с прописной буквы, полужирным шрифтом, отдельной строкой по центру без точки в конце (ГОСТ 7.32–2017 п. 6.17.3, ГОСТ 7.0.11–2011 п. 5.7.4).

В то же время существуют отличия в требованиях оформления заголовков. В частности:

- заголовки разделов и подразделов основной части отчета следует начинать с абзацного отступа и размещать после порядкового номера, печатать с прописной буквы, полужирным шрифтом, не подчеркивая, без точки в конце (ГОСТ 7.32–2017 п. 6.2.3).

В результате анализа были выделены и классифицированы базовые структурные элементы (классы) и составляющие их подэлементы в количестве 57 единиц для последующей автоматизированной проверки. Отобранные классы с одной стороны стали основой для алгоритмов машинного обучения (классификация каждого из отдельных структурных элементов внутри документа), а с другой стороны — идентификаторами для проверки оформления.

Таким образом, была изучена структура ГОСТов, разделы и пункты, связанные с правилами оформления текстовых документов и выделены базовые структурные элементы (классы) без составляющих их подэлементов: поля, колонтитулы, отступы, нумерация страниц, нумерация разделов, оглавление, заголовок, перечисления (список), таблица, подпись таблицы, рисунок, подпись рисунка, формула, подпись формулы, приложения, список литературы, разделитель “SPACE”, элемент листинга.

Архитектура сервиса

В рамках проектирования Сервиса, его работа была распределена на несколько модулей. Модуль для взаимодействия с документами в этой системе занимает одно из ключевых мест. Более подробное описание архитектуры Сервиса и процесс его работы представлено в статье [8].

После анализа ошибок оформления и инструментов для работы с docx файлами, стало очевидно, что функцию проверки оформления документа можно разделить на функции проверки отдельных его аспектов. Так, например, были выделены: форматирование абзацев, оформление титульного листа, колонтитулов, содержания, списка литературы, таблиц, рисунков и т.д. Эти процессы являются независимыми друг от друга, и, следовательно, могут выполняться параллельно. Каждый из аспектов содержит набор свойств, имеющих эталонные значения для соответствующих требований оформления. Таким образом, необходимо было перейти к проектированию объектной модели эталонного документа на уровне работы модуля.

Объектная модель

По требованиям к сервису функции проверки должны поддерживать выбор различных требований оформления. Для этого была предложена модель эта-

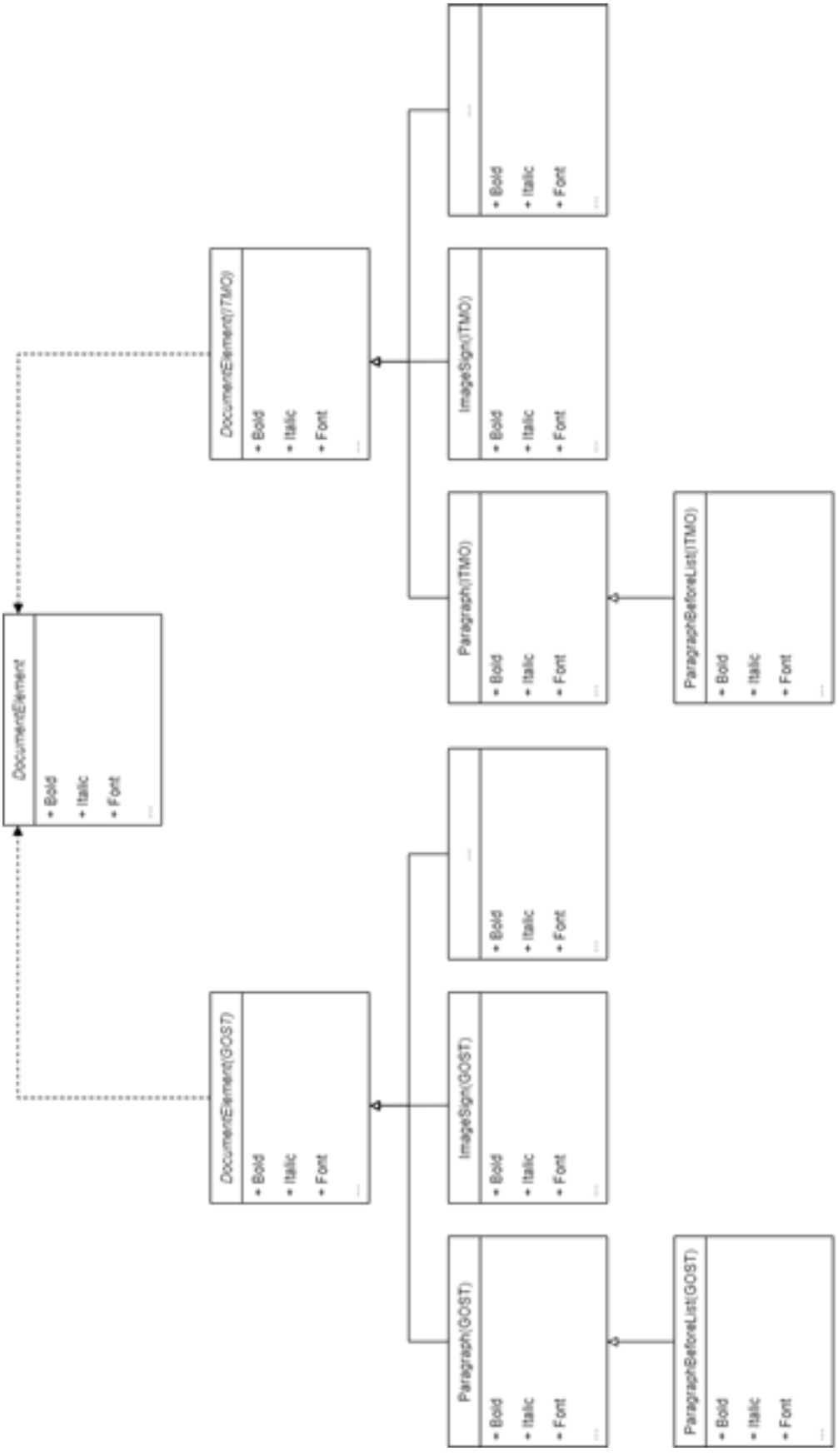

Рис. 2. Диаграмма классов (элементы документа)

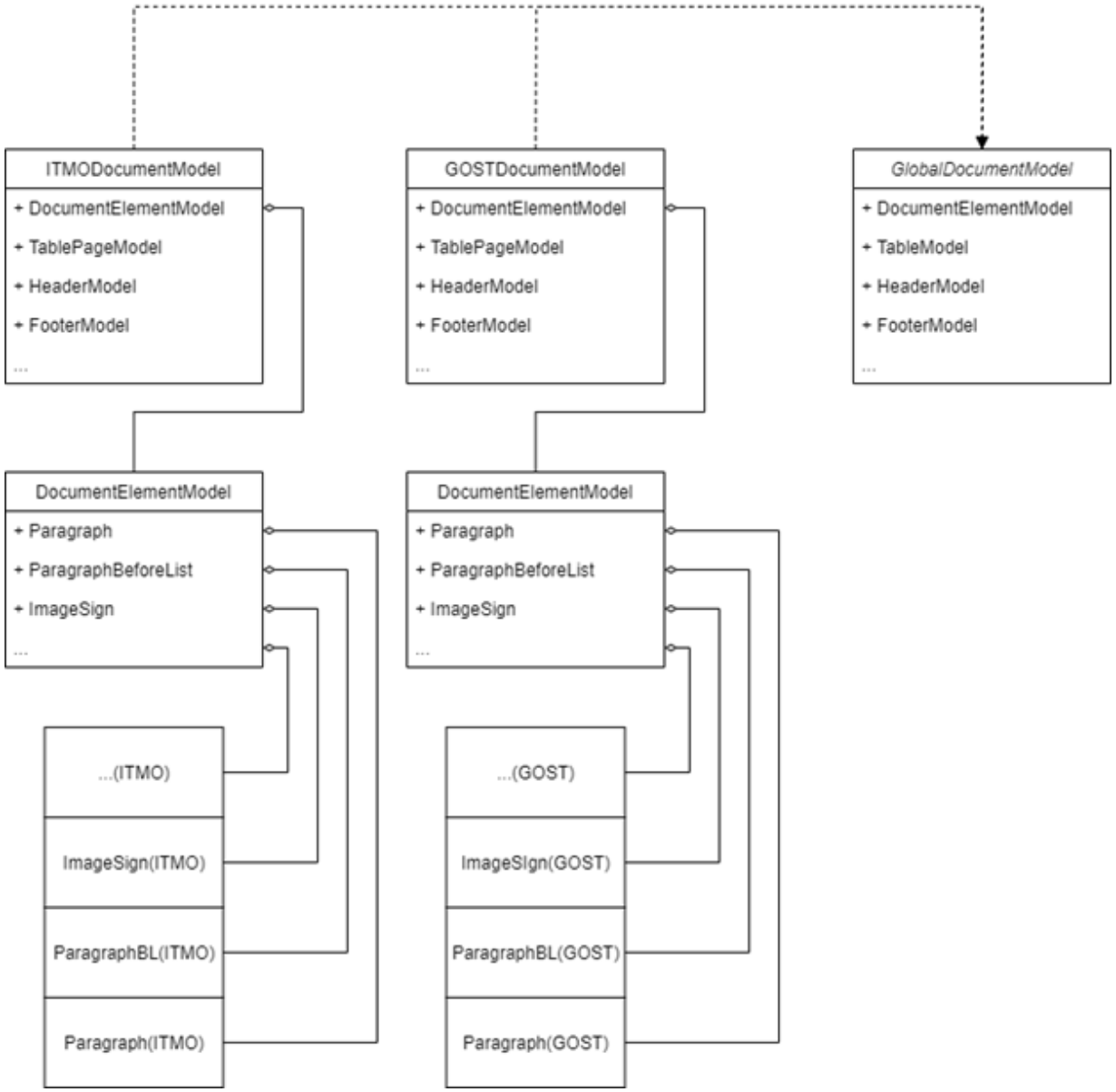

Рис. 3. Диаграмма классов (модели для требований)

лонного документа, которая должна использоваться при разработке модуля.

Для определения корректных значений (или диапазонов значений) свойств для классов необходимо было проанализировать, как было показано выше, требования оформления по ГОСТ 7.32–2017 и ГОСТ Р 7.0.11–2011, и возможности разработчиков при работе с документом на уровне кода.

В качестве основного инструмента для взаимодействия с docx файлами был взят API, предоставляемой библиотекой GemBox.Document. Более подробное обоснование этого выбора можно прочитать в исследовании [9].

На рисунке 1 изображена диаграмма классов, которая определяет общую структуру модели документа. Рассмотрим ее более подробно.

Абстрактный класс GlobalDocumentModel определяет несколько полей, олицетворяющих отдельные аспекты документа. Так, можно выделить модель форматирования абзацев, модель форматирования таблиц, модель форматирования списка литературы и др. В свою очередь существуют не абстрактные классы, которые являются конкретными реализациями модели документа для соответствующих требований. На схеме (рисунок 1) они обозначены GOSTDocumentModel и ITMODocumentModel.

Таблица 2. Значения свойств классов объектной модели документа по ГОСТ Р 7.0.11–2011

	Базовый элемент	Обычный абзац	Абзац перед списком	Абзац перед формулой	Заголовок первого уровня
Класс		**c1**	**c2**	**c3**	**b1**
Свойства ParagraphFormat					
Alignment	HorizontalAlignment.Justify				Horizontal Alignment. Center
BackgroundColor	**Color.Empty, Color.White**				
Border	**BorderStyle.None**				
KeepLinesTogether	**false**				
KeepWithNext	false	false	true	true	false
LeftIndentation	**0**				
LineSpacing	**1,5**				
LineSpacingRule	**LineSpacingRule.Multiple**				
MirrorIndents	**false**				
NoSpaceBetweenParagraphsOfSameStyle	**false**				
OutlineLevel	OutlineLevel.BodyText				OutlineLevel.Level1
PageBreakBefore	false				true
RightIndentation	**0**				
RightToLeft	**false**				
SpaceAfter	**0**				
SpaceBefore	**0**				
SpecialIndentationLeftBorder	-36,85				0
SpecialIndentationRightBorder	-35,45				0
WidowControl	**true**				
Свойства CharacterFormat для всего абзаца					
AllCaps	false				true
BackgroundColor	**Color.Empty**				
AlternativeBackgroundColor	**Color.White**				
Bold	false				true
Border	**SingleBorder.None**				

Объект ModelSwitcher находится на самом верхнем уровне и позволяет получить соответствующую реализацию наследника класса GlobalDocumentModel по текстовому обозначению правил в программе (например, «GOST» для выбора модели для ГОСТ 7.32–2017).

Модель форматирования абзацев определяет оформление для классов элементов документа таких как абзац, абзац после списка, подпись к рисунку и др. Все эти классы элементов являются дочерними классами от базовых классов для соответствующих стандартов (на схеме обозначены DocumentElement(GOST) и DocumentElement(ITMO)), которые в свою очередь являются дочерними классами от DocumentElement (рисунок 2). В нем определены общие для всех классов элементов свойства (шрифт, начертание, выравнивание и др.). Такая иерархия позволяет создавать независимые объектные модели для различных стандартов с отдельными функциями проверки специфичных для них свойств.

В свою очередь наследники класса GlobalDocumentModel ссылаются на соответствующие для требований реализации этих классов, тем самым инкапсулируя все условия правильности оформления в одном объекте (рисунок 3).

Для реализации классов, изображенных на рисунке 2 в коде необходимо определить эталонные значения соответствующих свойств. Как уже было сказано ранее, этот процесс требовал анализа соответствующих ГОСТов и документации библиотеки GemBox.Document. Результат этого анализа был воплощен в виде электронных таблиц. Часть такой таблицы, представлена в таблице 2.

В первом столбце расположены свойства документа, с которыми может взаимодействовать библиотека GemBox.Document. После этого идут столбцы классов элементов. Значения свойств, которые не отличаются от базовых записаны в широких ячейках. В таблице представлены значения свойств для базового элемента

и классов c1, c2, c3, b1. Всего было составлено две таблицы для ГОСТ 7.32–2017 и ГОСТ Р 7.0.11–2011. Каждая таблица имеет более 50 свойств.

На текущем этапе Сервис ориентирован и реализован для проверки форматирования абзацев текстового документа, это составляет 17 элементов из выделенных ранее 57, а остальные 40 элементом учтены в разработанной объектной модели, и могут быть добавлены и активированы в любое время по необходимости с целью проверки текстовых документов в рамках нормоконтроля, таким образом объектная модель может расширяться. Сервис разработан с учетом поддержки проверок других аспектов документа.

Заключение

В ходе исследования была изучена структура ГОСТов, их разделов и пунктов, связанных с правилами оформления текстовых документов и выделены базовые структурные элементы (классы). Проведенная работа по подготовке правил для разработчиков позволяет осуществлять последующую реализацию автоматизированного поиска ошибок в тексте документов. При реализации объектной модели были выделены предикторы, необходимые для реализации алгоритмов машинного обучения для классификации структурных элементов, на основании которых выполняется дальнейшая проверка элементов по требованиям отдельного класса.

Таким образом, разработанная объектная модель позволяет:

- расширять функции сервиса, что было продемонстрировано на использовании нескольких ГОСТов,
- унифицировать объекты и общие требования в разных ГОСТах,
- повышать точность проверки требований оформления за счет возможности внедрения корпуса алгоритмов для автоматизированной проверки различных аспектов документа.

ЛИТЕРАТУРА

1. Ивановский А.А. Объектная модель системы избирательного распространения информации. Научные и технические библиотеки. 2019;(4):61–75. https://doi.org/10.33186/1027–3689–2019–4–61–75
2. Ульянов В.Н., Каюров Н.К., Лукьянов Э.Е. Проблемы построения систем автоматизации. Объектная модель в системах управления // Автоматизация, телемеханизация и связь в нефтяной промышленности 2020 № 12 (569) Стр. 27–32
3. Berezhkov A.V., Nasyrov N.F., Valitova Y.O., Ivanov S.E., Gorlushkina N.N., Kobets E.A. Organizational models of student peer assessment // 1st International Conference on Computer Technology Innovations dedicated to the 100th anniversary of the Gorky House of Scientists of Russian Academy of Science (ICCTI-2020) — 2020, pp. 40–45
4. Насыров Н.Ф., Кобец Е.А., Горлушкина Н.Н. Автоматизированная генерация учебных подзадач на основе методики тегов и критериев // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки — 2020. — № 3. — С. 102–107
5. Кобец Е.А., Насыров Н.Ф., Комаров М.С., Горлушкина Н.Н. Алгоритм формализации требований к оформлению документов для сервиса автоматизированного нормоконтроля // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки — 2021.
6. ГОСТ 7.32–2017 // URL: https://docs.cntd.ru/document/1200157208 (даты обращения 10.10.2019–01.06.2021)
7. ГОСТ Р 7.0.11–2011 // URL: https://docs.cntd.ru/document/1200093432 (даты обращения 10.10.2019–01.06.2021)
8. Nasyrov N.F., Komarov M.S., Tartynskikh P.S., Gorlushkina N.N. Automated formatting verification technique of paperwork based on the gradient boosting on decision trees.//Procedia Computer Science Volume 178, 2020, Pages 365–374 9th International Young Scientists Conference in Computational Science, YSC2020; Virtual, Online; Greece; 7 September 2020 до 12 September 2020; Код 165573, 2020, pp. 365–374
9. Тартынских П.С., Комаров М.С., Насыров Н.Ф. Подходы к автоматизированному анализу оформления электронных документов формата docx//Сборник тезисов докладов конгресса молодых ученых. Электронное издание. — СПб: Университет ИТМО, 2020. Электронное издание. — [2020, электронный ресурс]. — Режим доступа: https://kmu.itmo.ru/digests/article/4592, своб. — 2020