УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2021 Том 21 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2021 Vol. 21 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2021-21-2-256-266
УДК 004.822

# Построение графов знаний нормативной документации на основе семантического моделирования и автоматического извлечения терминов

**Дмитрий Ильич Муромцев[1]✉, Иван Андреевич Шилин[2], Дмитрий Алексеевич Плюхин[3], Ильдар Раисович Баймуратов[4], Резеда Раитовна Хайдарова[5], Юлия Юрьевна Дементьева[6], Денис Александрович Ожигин[7], Татьяна Алексеевна Малышева[8]**

[1,2,3,4,5,8] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[6,7] ООО «Нанософт разработка», Москва, 108811, Российская Федерация

[1] mouromtsev@itmo.ru✉, http://orcid.org/0000-0002-0644-9242
[2] shilinivan@itmo.ru, http://orcid.org/0000-0002-7846-1398
[3] zeionara@gmail.com, https://orcid.org/0000-0001-5790-7883
[4] baimuratov.i@gmail.com, http://orcid.org/0000-0002-6573-131X
[5] mignolowa@gmail.com, http://orcid.org/0000-0001-8270-9192
[6] yulia.dementeva@gmail.com, https://orcid.org/0000-0002-6829-6245
[7] denis@nanocad.ru, https://orcid.org/0000-0002-9581-241X
[8] tamalysheva@itmo.ru, https://orcid.org/0000-0002-1171-379X

**Аннотация**

**Предмет исследования.** Предложено новое комплексное решение для автоматического анализа и идентификации терминов в нормативно-технической документации. Идентификация терминов в документации является актуальной задачей в цифровизации отрасли проектирования и строительства зданий и сооружений. В настоящий момент поиск и проверка требований нормативно-технической документации выполняется вручную, что влечет существенное количество ошибок. Автоматизация подобных задач позволит существенно повысить качество автоматизированного проектирования. **Метод.** Разработанный алгоритм основан на таких методах анализа естественного языка как токенизация, поиск лемм и основ слов, анализ стоп-слов, подсчет векторных представлений токенов и словосочетаний, частеречная и синтаксическая разметка и др. **Основные результаты.** Эксперименты по автоматическому извлечению терминов в нормативной документации показали большие возможности предложенного алгоритма для построения графов знаний в предметной области проектирования. Точность распознавания на примере 202 отобранных экспертами документов составила 79 % по совпадению наименований и 37 % по совпадению идентификаторов терминов. Это является сопоставимым результатом с известными подходами к решению данной проблемы. **Практическая значимость.** Результаты работы могут использоваться в системах автоматического проектирования на основе Building Information Modeling моделей, а также для автоматизации экспертизы проектной документации.

**Ключевые слова**

семантический анализ текста, онтологии, извлечение терминов, векторные представления, глубокие нейронные сети

**Ссылка для цитирования:** Муромцев Д.И., Шилин И.А., Плюхин Д.А., Баймуратов И.Р., Хайдарова Р.Р., Дементьева Ю.Ю., Ожигин Д.А., Малышева Т.А. Построение графов знаний нормативной документации на основе семантического моделирования и автоматического извлечения терминов // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 2. С. 256–266. doi: 10.17586/2226-1494-2021-21-2-256-266

# Building knowledge graphs of regulatory documentation based on semantic modeling and automatic term extraction

**Dmitry I. Mouromtsev[1]✉, Ivan A. Shilin[2], Dmitrii A. Pliukhin[3], Ildar R. Baimuratov[4], Rezeda R. Khaydarova[5], Yulia Yu. Dementyeva[6], Denis A. Ozhigin[7], Tatiana A. Malysheva[8]**

[1,2,3,4,5,8] ITMO University, Saint Petersburg, 197101, Russian Federation
[6,7] "Nanosoft razrabotka LLC", Moscow, 108811, Russian Federation

[1] mouromtsev@itmo.ru✉, http://orcid.org/0000-0002-0644-9242
[2] shilinivan@itmo.ru, http://orcid.org/0000-0002-7846-1398
[3] zeionara@gmail.com, https://orcid.org/0000-0001-5790-7883
[4] baimuratov.i@gmail.com, http://orcid.org/0000-0002-6573-131X
[5] mignolowa@gmail.com, http://orcid.org/0000-0001-8270-9192
[6] yulia.dementeva@gmail.com, https://orcid.org/0000-0002-6829-6245
[7] denis@nanocad.ru, https://orcid.org/0000-0002-9581-241X
[8] tamalysheva@itmo.ru, https://orcid.org/0000-0002-1171-379X

**Abstract**

The paper proposes a new complex solution for automatic analysis and terms identification in regulatory and technical documentation (RTD). The task of terms identification in the documentation is one of the key issues in the digitalization dealing with the design and construction of buildings and structures. At the moment, the search and verification of RTD requirements is performed manually, which entails a significant number of errors. Automation of such tasks will significantly improve the quality of computer-aided design. The developed algorithm is based on such methods of natural language analysis as tokenization, search for lemmas and stems, analysis of stop words and word embeddings applied to tokens and phrases, part-of-speech tagging, syntactic annotation, etc. The experiments on the automatic extraction of terms from regulatory documents have shown great prospects of the proposed algorithm and its application for building knowledge graphs in the design domain. The recognition accuracy for 202 documents selected by experts was 79 % for the coincidence of names and 37 % for the coincidence of term identifiers. This is a comparable result with the known approaches to solving this problem. The results of the work can be used in computer-aided design systems based on Building information modeling (BIM) models, as well as to automate the examination of design documentation.

**Keywords**

semantic text analysis, ontologies, term extraction, word embeddings, deep neural networks

**For citation:** Mouromtsev D.I., Shilin I.A., Pliukhin D.A., Baimuratov I.R., Khaydarova R.R., Dementyeva Yu.Yu., Ozhigin D.A., Malysheva T.A. Building knowledge graphs of regulatory documentation based on semantic modeling and automatic term extraction. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 2, pp. 256–266 (in Russian). doi: 10.17586/2226-1494-2021-21-2-256-266

## Введение

Эффективность цифровизации экономических и производственных процессов зависит в первую очередь от того, в какой степени она охватывает этапы жизненного цикла новых услуг или продукции. Важнейшим из этапов жизненного цикла является проектирование, которое, в свою очередь, должно учитывать множество нормативных документов. Современные системы автоматизированного проектирования позволяют создавать качественные цифровые модели. Однако проверка данных моделей на соответствие нормативной документации во многом остается ручным трудом.

В настоящее время существуют вполне развитые проекты и решения по управлению и визуализации информации, основанные на концепции управления данными о продукте (Product Data Management, PDM) и информационном моделировании (Information Modeling, IM). Для строительной отрасли стандартом де-факто по управлению информацией стало информационное моделирование зданий (Building Information Modeling, BIM) [1]. Для обеспечения доступности смысловых элементов BIM-моделей для приложений за пределами среды моделирования создана онтологическая модель данных Industry Foundation Classes (IFC) [2]. В частности, в работе [3] предложено семантическое моделирование на основе онтологий для осуществления проверки семантических правил при проектировании и строительстве зданий. В работе [4] рассмотрена прототипная реализация инструмента проверки представления BIM на основе IFC с использованием открытых стандартов и нормативных требований.

Одно из преимуществ онтологического представления данных — возможность комбинирования формальной семантики и пользовательских терминологических словарей, а также связывания гетерогенной информации о структуре документов с их метаданными, с учетом определений и синонимов терминов для заданной предметной области [5]. Все это позволяет автоматизировать процесс построения баз знаний на основе семантического анализа нормативных документов.

Для эффективной работы с терминологическими словарями необходимы алгоритмы, использующие методы автоматического анализа естественного языка (Natural Language Processing, NLP) и учитывающие специфику предметной области. В работе [6] предложен подход автоматизированного извлечения информации из строительных нормативных документов на английском языке, основанный на правилах, подходах и методах NLP, а также учитывающий отраслевую модель IFC.

В настоящей работе рассмотрены первые результаты исследований по созданию семантической модели (базы знаний), содержащие технические требования норма-

тивных документов на русском языке, представленные в виде множества терминов и ссылок на соответствующие разделы этих документов. Такая модель может быть интеллектуальным ядром системы автоматической выдачи рекомендаций и поиска требований для систем автоматизированного проектирования. В процессе решения поставленной задачи авторами разработан алгоритм идентификации терминов в нормативных документах, позволяющий автоматически формировать семантическую модель в процессе парсинга этих документов. В основе семантической модели лежит разработанная онтология нормативного документа. Отличительные особенности онтологии — ее гибкость при описании структурных компонентов документов, а также выделенных терминов, входящих в состав представленного заказчиком классификатора. Для описания базовых абстракций онтологии использованы как существующие стандартные онтологии, так и специфические, разработанные для данного проекта сущности. Такая комбинация позволяет добиться более легкой расширяемости семантической модели при добавлении новых словарей и классификаторов, а также расширения спектра предметных областей нормативных документов.

Разработанный алгоритм автоматической идентификации терминов реализован в виде программного обеспечения, основными функциями которого являются: парсинг документов, идентификация терминов, экспорт полученных результатов и автоматическое построение RDF (Resource Description Framework)-графа знаний.

### Онтологии для описания нормативно-технической документации

Разработанная семантическая модель представлена в виде онтологии на языке OWL[1]. Пайплайн разработки семантической модели состоял из четырех этапов.

Первый этап — анализ исходных документов с целью получения первичного представления о предметной области, содержащихся в ней классах и отношениях между ними. На основе произведенного анализа были выявлены контексты данных, и отобраны существующие онтологии, которые могли быть использованы при разработке новой семантической модели.

Второй этап — разработка формализованной модели предметной области в виде онтологии. Осуществлялся поиск необходимых сущностей в существующих онтологиях, если подходящих сущностей не находилось, то добавлялись новые.

Третий этап — формирование примеров использования семантической модели для описания предметной области и разработки базы знаний. Сформированы примеры двух типов: индивиды в составе самой онтологии и размеченные с ее помощью документы.

Четвертый этап — тестирование результатов автоматического парсинга документов, основанного на разработанной онтологии, с целью оценки полноты и корректности знаний, извлеченных из документов. Схема пайплайна разработки семантической модели представлена на рис. 1.

При разработке семантической модели выполнен анализ исходных документов с целью получения первичного представления о предметной области, контекстах содержащихся в документах данных и основных необходимых сущностях. В результате произведенного анализа выделено четыре контекста данных, извлекаемых при автоматическом парсинге:

1) метаданные документа;
2) структура документа;
3) термины, встречающиеся в документе;
4) лингвистические свойства текста документа.

В табл. 1 использованные онтологии сопоставлены с контекстами, выделенными в результате анализа документов.

В качестве отправной точки при разработке онтологии было решено импортировать сущности из существующих онтологий, подобранных в соответствии с выявленными на этапе анализа контекстами данных. В табл. 2 представлена информация об импортированных сущностях.

Разработанная онтология состоит из 1049 аксиом, которые описывают 83 класса, 31 объектное свойство и 34 свойства данных. Примерами импортированных классов являются такие сущности, как *Документ*, *Тема*, *Параграф*, *Содержание*, *Фраза* и др. (всего 22 класса). В онтологию были добавлены такие специфические классы, как *Приказ*, *Свод правил*, *Стандарт*, *Физическое лицо*, *Логическая операция*, *Ключевые слова*, *Нормативные ссылки*, *Область применения* и др. Все объектные свойства в онтологии являются оригинальными. Например, это такие свойства, как *Имеет синоним*, *Предметное свойство документа*, *Издан в* и др. Импортированными свойствами данных являются *Идентификатор*, *Название*, *Дата изменения*, *Дата создания*, а примерами оригинальных свойств — *Код страны*, *Свойство данных слова*, *Индекс первого/последнего символа* и пр.

В качестве примеров применения разработанной онтологии для описания данных документов сформирован набор индивидов[2], демонстрирующих использование основных классов и свойств. На рис. 2 представлен

*Таблица 1*. Сопоставление онтологий с выделенными контекстами

*Table 1.* Mapping of ontologies with selected contexts

| Контекст | Онтологии |
|---|---|
| Метаданные документа | dc [7], document[1] |
| Структура документа | doco [8], document |
| Термины | — |
| Лингвистические свойства | lemon [9] |

[1] Документация OWL [Электронный ресурс]. Режим доступа: https://www.w3.org/OWL/, свободный. Яз. англ. (дата обращения: 23.03.2021).

[1] Документация Knora [Электронный ресурс]. Режим доступа: https://docs.knora.org/paradox/02-knora-ontologies/index.html, свободный. Яз. англ. (дата обращения: 12.02.2021).

[2] Индивид (NamedIndividual) — это тип сущностей в онтологии, конкретные объекты, экземпляры классов.

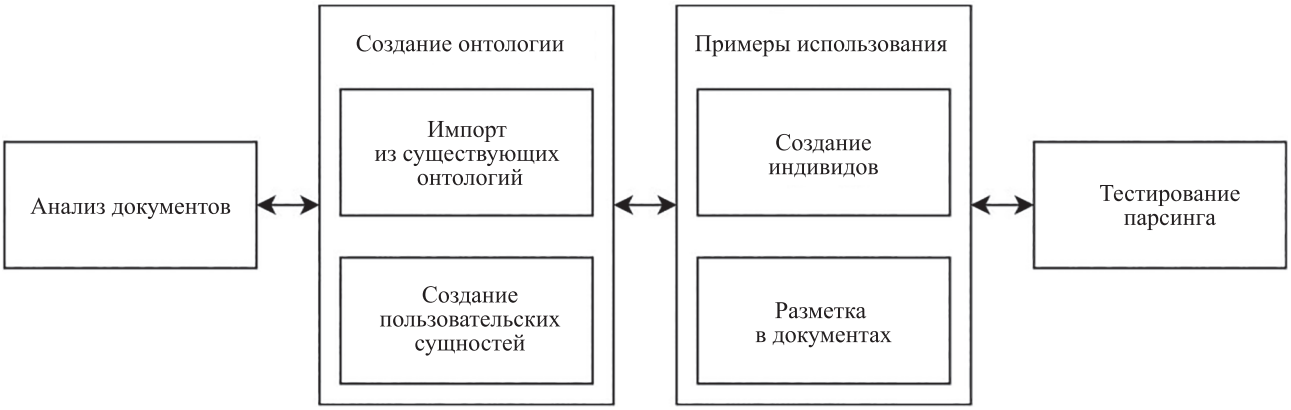


*Рис. 1*. Пайплайн разработки семантической модели

*Fig. 1*. Semantic model development pipeline

*Таблица 2*. Импортированные сущности

*Table 2.* Imported entities

| Онтология | Тип | Список импортированных сущностей |
|---|---|---|
| dc | Свойство данных | Альтернативное название, Дата изменения, Дата принятия, Дата создания, Идентификатор, Название |
| doco | Класс | Введение, Параграф, Предисловие, Приложение, Раздел, Рисунок, Сноска, Содержание, Список, Список литературы, Ссылка, Таблица, Термины и определения, Формула |
| knora | Класс | Документ, Элемент |
| lemon | Класс | Определение, Слово, Тема, Фраза |

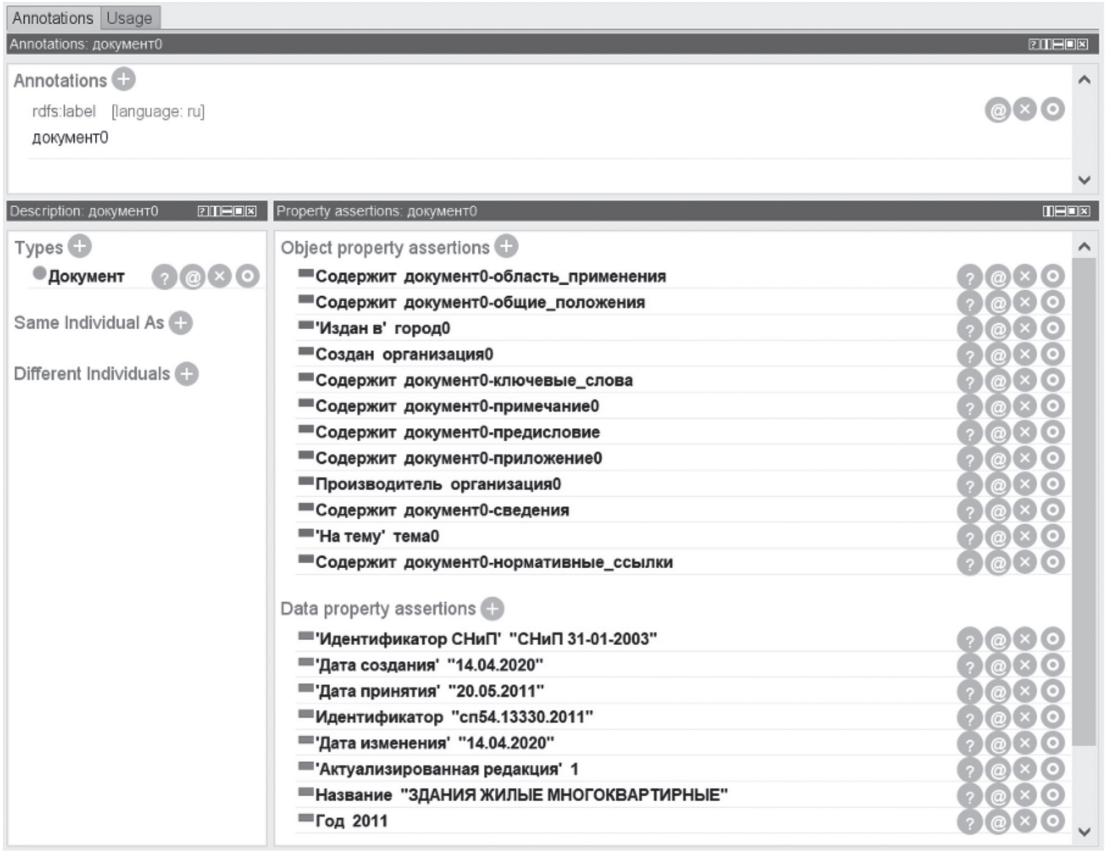


*Рис. 2*. Пример онтологического описания документа в Protege

*Fig. 2*. An example of an ontological description of a document in Protégé

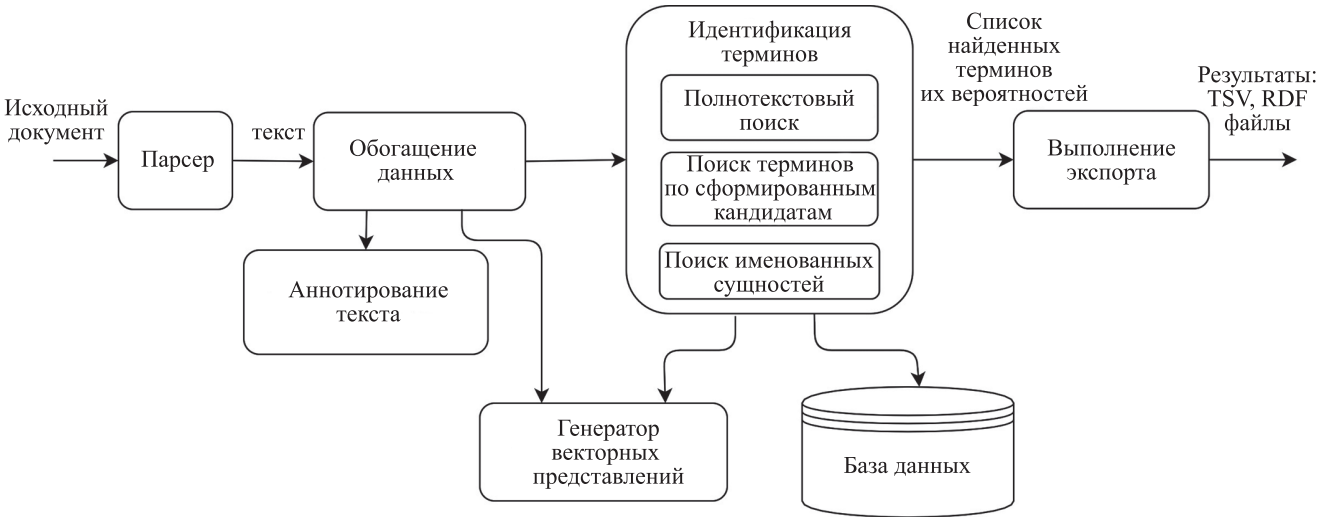


*Рис. 3*. Структура компонентов системы обработки нормативных документов

*Fig. 3*. The structure of the components of the regulatory documents processing system

пример описания документа в редакторе онтологий Protege. Документ содержит разделы: приложение, область применения, предисловие, общие положения, сведения о документе, ключевые слова, нормативные ссылки и примечание. Указана организация, создавшая документ, его тема, город, в котором он издан, дата создания, дата принятия, идентификатор, дата изменения, название, идентификатор СНиП. Также указано, что это актуализированное издание.

### Алгоритм извлечения терминов

Структура компонентов, реализующих алгоритм автоматического извлечения терминов для задачи семантического моделирования нормативных документов представлена на рис. 3.

В алгоритме на первом этапе выполняется парсинг исходного нормативного документа, в результате которого исходный документ преобразуется в представление в виде объектной модели. После парсинга текст на естественном языке обогащается данными за счет его аннотирования и формирования векторных представлений. Далее выполняется идентификация терминов, которая включает три алгоритма:

— полнотекстовый поиск;
— поиск терминов по сформированным кандидатам;
— поиск именованных сущностей.

При идентификации терминов используется база данных, содержащая исходный словарь терминов и их метаданные, а также генератор векторных представлений. В качестве базы данных применен Elasticsearch[1], в модуле генерации векторных представлений — модель BERT [10], обученная на 8899 нормативных документах.

Далее выполняется экспорт полученных результатов в форматах RDF и TSV (Tab-Separated Values). Рассмотрим подробно основные этапы алгоритма.

**Парсинг документа.** Парсинг документа заключается в сегментации документа на элементы его метаданных и текстовой информации с использованием структуры формата и паттернов, характерных для файлов использованного набора. Приведем основные признаки, используемые для генерации значений полей внутреннего представления документа и составления на ее основе семантической модели.

— Регулярные выражения и подстроки — наиболее общий способ обработки текста модулем парсинга, независящий от строгости соблюдения формата документа и заключающийся в выделении из текста обобщенных фрагментов, удовлетворяющих требованиям, зафиксированным в заданном шаблоне.
— Размер и расположение элементов документа — данный способ более прост в реализации, однако и более существенно зависит от строгости соблюдения формата при составлении документа. Суть подхода заключается в проверке количества объектов того или иного типа в процессе поиска элементов документа, содержащих необходимые данные.
— Xml-теги и атрибуты — наиболее гибкий с точки зрения использования и трудоемкий с точки зрения реализации способ, который использован для извлечения данных, не извлекаемых сторонними библиотеками либо извлекаемых ими частично или полностью некорректно.
— Метаданные файлов — способ извлечения информации о файлах из метаданных, полученных сторонними библиотеками. Наиболее прост с точки зрения реализации.
— Метаданные элементов документа — способ поиска значений необходимых полей, заключающийся в анализе метаданных элементов документа; зависит от строгости соблюдения формата документа при его составлении, и по данной причине применен для обработки ограниченного набора типов элементов документов.

[1] Документация elasticsearch [Электронный ресурс]. Режим доступа: https://www.elastic.co/guide/index.html, свободный. Яз. англ. (дата обращения: 12.02.2021).

**Обогащение данных.** На этапе обогащения данных применяются следующие подходы к обработке текста на естественном языке:

1) токенизация;
2) поиск лемм (приведение словоформ к нормальной (словарной) форме);
3) поиск стемм (основ словоформ);
4) разбиение токенов на классы в соответствии с регистрами символов;
5) установление принадлежности токена к категории стоп-слов и последовательностей символов пунктуации;
6) подсчет векторных представлений токенов и словосочетаний;
7) частеречная разметка;
8) синтаксическая разметка.

Применение частеречной и синтаксической разметок позволило сократить область и повысить точность поиска упоминаний терминов в тексте. Результаты применения подходов 1–5 используются для оценки схожести словосочетаний текста и описания того или иного термина на ранних этапах работы пайплайна, на которых осуществляется формирование предварительного набора релевантных терминов. Векторные представления токенов и словосочетаний применяются на более поздних этапах работы алгоритма.

**Поиск терминов.** Поиск терминов в тексте документа осуществлялся следующими способами:

— полнотекстовый поиск, основанный на идее использования встроенных алгоритмов базы данных для поиска записей, соответствующих заданному тексту, без предварительных этапов генерации системой словосочетаний, которые могут соответствовать упоминаниям терминов;
— итеративный поиск по базе данных, состоящий из повторения двух шагов: генерация словосочетаний-кандидатов, предположительно являющихся упоминаниями терминов; поиск соответствий в базе данных;
— использование моделей машинного обучения, основанных на глубоких нейронных сетях, решающих задачу извлечения именованных сущностей.

**Экспорт результатов.** В результате выполнения предыдущих этапов формируется семантическое представление документа в формате RDF, позволяющее сформировать граф знаний и программный интерфейс, обеспечивающий доступ к результатам анализа нормативного документа. Также создается представление документа в формате TSV, ориентированное на использование разметчиками для уточнения результатов работы системы и оценки их качества.

### Методика оценки точности работы алгоритма

Для оценки качества работы алгоритма извлечения терминов разработан модуль формирования и подсчета метрик. Рассмотрим основные метрики.

Выполнялась оценка как полного совпадения результатов ручной разметки с данными, сгенерированными автоматически (например, полное совпадение идентификаторов и названий терминов), так и частичного, в ходе которой допускалось различие морфологических характеристик слов и семантическая омонимия названий терминов.

Рассмотрим две основные конфигурации используемых метрик:

— оценка полного совпадения идентификаторов терминов, что подразумевает совпадение и значений других полей, составляющих описание термина (условно обозначим «id»);
— оценка совпадения названий выделенных терминов при том, что их идентификаторы могут отличаться (обозначим «name»).

Например, в случае использования следующего словаря терминов (указаны только значения полей, содержащие идентификатор и название термина соответственно):

1942 Высота потолка
1234 Высота потолка
2414 Высота этажа

При оценке качества работы алгоритма, если в ручной разметке некоторому упоминанию термина в тексте присвоен концепт «1942 Высота потолка», то использование описанных конфигураций метрик приведет к результатам, представленным в табл. 3.

**Результаты построения семантической модели на основе разработанного программного обеспечения и алгоритмов.** Тестовые запуски выполнялись на сервере со следующими характеристиками: CPU: Intel core i9-9900K, GPU: Nvidia RTX 2080 Ti @ 11 GB, RAM: 64 GB.

Выполнение тестовых запусков системы на сервере для сбора данных о времени обработки датасета, состоящего из 8 документов, представлено в табл. 4.

В табл. 4 использованы следующие обозначения:

n-pipelines — количество одновременно поддерживаемых экземпляров модели частеречной и синтаксической разметки, используемых в процессе работы системы;

id-sorting-order — наименование используемого алгоритма сортировки результатов обращения к базе данных по числовому идентификатору термина;

null — алгоритм не задан (т. е. использование сортировки по умолчанию);

asc — сортировка по возрастанию;

top-n-common — количество наиболее релевантных записей, запрашиваемых при обращении к базе данных в процессе выполнения алгоритма итеративного поиска;

n-workers — наибольшее количество файлов, обрабатываемых одновременно (количество процессов в основном пуле системы);

pre-computed-embeddings — бинарное поле, показывающее, были ли векторные представления подсчитаны заранее (до начала процесса обработки документов, продолжительность которого измерялась).

После получения оценок влияния параметров запуска алгоритма на скорость его работы, для построения семантической модели совместно с экспертами предметной области создана более крупная выборка из 202 нормативных документов. На основе этой выборки с помощью алгоритма была сформирована семантиче-

*Таблица 3*. Результаты сравнения термина, размеченного вручную, с результатом работы алгоритма путем использования двух конфигураций метрик

*Table 3.* The results of comparing a manually labeled term with the result of the algorithm execution using two configurations of metrics

| Метрика | Найденный термин | Совпадает ли найденный термин с исходным |
|---|---|---|
| id | 1942 Высота потолка | Да |
| id | 1234 Высота потолка | Нет |
| id | 2414 Высота этажа | Нет |
| name | 1942 Высота потолка | Да |
| name | 1234 Высота потолка | Нет |
| name | 2414 Высота этажа | Нет |

ская модель, содержащая 46 431 378 триплетов. Для формирования этой модели выполнено множество запусков с разными конфигурационными параметрами, каждый из которых в среднем выполнялся от 20 до 35 ч. Итоговый запуск длился 28,734 ч.

**Оценка работы алгоритма по извлечению терминов.** Для оценки точности работы алгоритма осуществлено сравнение результатов автоматического извлечения терминов с ручной разметкой на выборке из трех документов. Для ручной разметки сформирована группа аннотаторов (неспециалистов в предметной области) и экспертов.

При подсчете оценки использовались стандартные метрики (значения приведены в табл. 5):

— precision — отношение количества верно распознанных объектов к общему количеству распознанных объектов;
— recall — отношение количества верно распознанных объектов к общему количеству объектов в исходной выборке;
— f1-score — гармоническое среднее значений precision и recall (обеспечивает возможность комплексной оценки качества работы системы).

При подсчете метрик могут использоваться следующие стратегии усреднения:
— макро-среднее (условно обозначено macro-average) — заключается в подсчете среднего арифметического набора значений метрики;
— взвешенное среднее (условно обозначено weighted-average) — аналогична макро-среднему macro-average; единственное отличие заключается во взвешивании отдельных значений пропорционально длинам последовательностей элементов документа, на основе которых данные значения были подсчитаны;
— микро-среднее (условно обозначено micro-average) — основана на идее агрегации промежуточных результатов и дальнейшем выполнении алгоритма подсчета метрики (в противоположность выполнению данного алгоритма сначала для подсчета отдельных значений метрик и последующем усреднении полученных результатов, что реализовано в альтернативных подходах, описанных выше).

При подсчете оценки использована стратегия усреднения weighted-average, значения и фрагмент результатов оценки которой приведены в табл. 5.

*Таблица 4*. Результаты измерения времени обработки системой тестового набора данных с использованием различных конфигураций параметров

*Table 4.* The results of measuring the processing time of the test dataset by the system using various configurations of parameters

| Конфигурация (использованные значения параметров) | | | | | Время выполнения, мин |
|---|---|---|---|---|---|
| n-pipelines | id-sorting-order | top-n-common | n-workers | pre-computed embeddings | |
| 2 | null | 100 | 2 | Нет | 29,149 |
| | | | | Да | 22,236 |
| | | | 4 | Нет | 27,618 |
| | | | | Да | 20,656 |
| | | 1000 | 2 | Нет | 30,396 |
| | | | | Да | 23,459 |
| | | | 4 | Нет | 29,243 |
| | | | | Да | 22,247 |
| | asc | 100 | 2 | Нет | 31,578 |
| | | | | Да | 24,570 |
| | | | 4 | Нет | 30,211 |
| | | | | Да | 24,302 |
| | | 1000 | 2 | Нет | 33,378 |
| | | | | Да | 26,519 |
| | | | 4 | Нет | 38,245 |
| | | | | Да | 25,256 |

*Таблица 5*. Значения метрик по результатам тестирования конфигураций системы

*Table 5*. The values of metrics based on testing results of system configurations

| Конфигурации системы | | name | | | id | | |
|---|---|---|---|---|---|---|---|
| Номер | Название | f1-score | precision | recall | f1-score | precision | recall |
| 1 | basic | 0,6611 | 0,6200 | 0,7078 | 0,3562 | 0,2975 | 0,4438 |
| 2 | full_text_search | 0,6619 | 0,6205 | 0,7093 | 0,3566 | 0,2977 | 0,4445 |
| 3 | without_synonyms | 0,6650 | 0,6246 | 0,7111 | 0,3586 | 0,2998 | 0,4460 |
| 4 | ner_basic | 0,7855 | 0,7728 | 0,7984 | 0,3493 | 0,2913 | 0,4362 |
| 5 | all_algorithms | 0,7997 | 0,8012 | 0,7982 | 0,3715 | 0,3144 | 0,4535 |

В табл. 5 использованы следующие обозначения:

basic — базовая конфигурация системы, в которой используется только алгоритм итеративного поиска;

full_text_search — конфигурация системы, в которой в дополнение к алгоритму итеративного поиска используется модуль полнотекстового поиска;

without_synonyms — упрощенная конфигурация системы, в которой используется только алгоритм итеративного поиска без учета синонимов;

ner_basic — конфигурация системы, в которой используются алгоритм итеративного поиска и модель машинного обучения, решающая задачу распознавания именованных сущностей;

all_algorithms — полная конфигурация системы, в которой используются все алгоритмы.

Визуализация результатов оценки из табл. 5 представлена на рис. 4, на котором отражены зависимости результатов тестирования системы с использованием

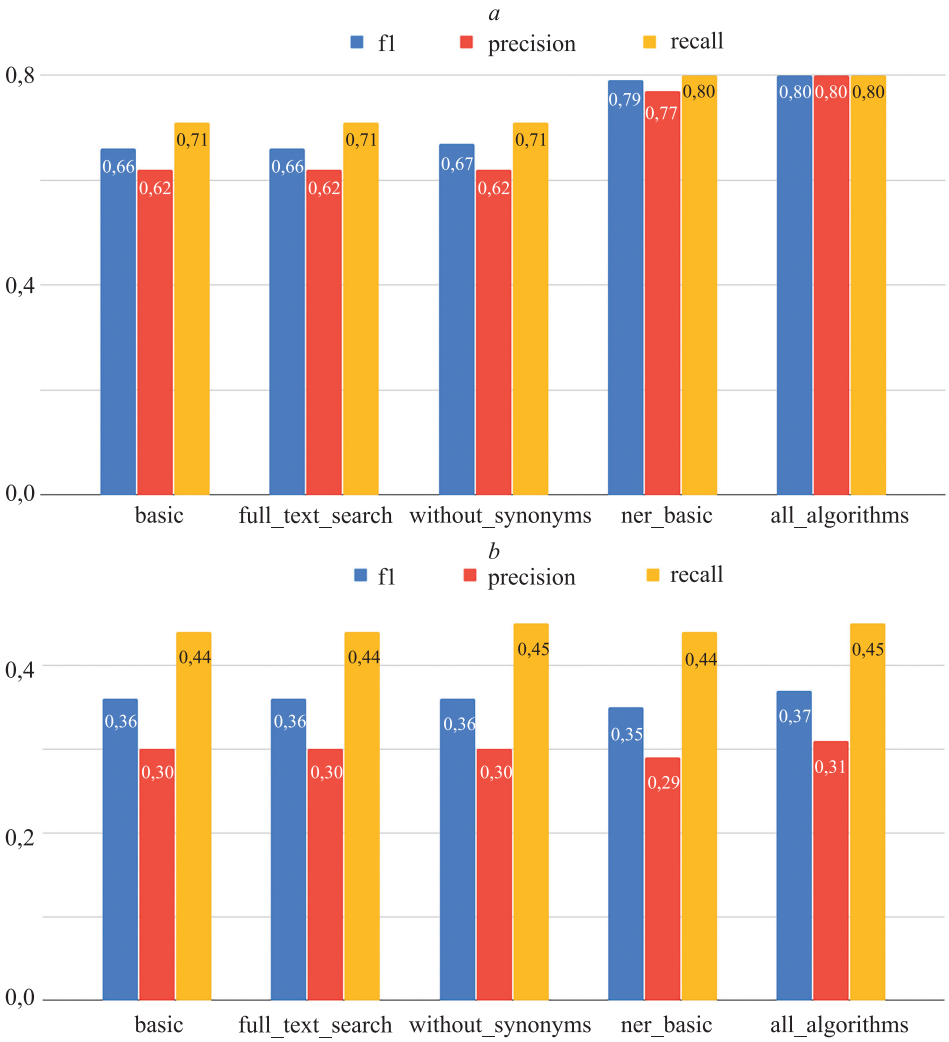


*Рис. 4*. Результаты тестирования конфигураций системы с использованием метрик, основанных на оценке совпадений названий «name» (*a*) и идентификаторов «id» (*b*) терминов

*Fig. 4*. The testing results of system configurations using metrics based on similarity of matches between the term names “name” (*a*) and term identifiers “id” (*b*)

метрик, основанных на оценке совпадения названий и идентификаторов терминов, соответственно, от используемой конфигурации системы.

В результате сформированной оценки можно сделать вывод о наибольшей точности запуска, состоящего из композиции всех алгоритмов.

### Примеры размеченных документов и запросов

В результате работы системы сформированы два представления в форматах RDF и TSV. Визуальное представление фрагмента TSV-файла в приложении WebAnno показано на рис. 5.

Фрагмент семантической модели в формате RDF представлен на рис. 6.

В качестве примера использования семантической модели приведен один из запросов к документу. Рассмотрим пример поиска используемых терминов из словаря для запроса на языке SPARQL:

```
select DISTINCT ?entity_elastic_id ?entity_elastic_text
where{
    {?entity a ncs:TermOccurrence. } UNION
    {?entity a ncs:OrganisationOccurrence}
    ?entity ncs:text ?text;
        ncs:occurrenceOf ?entity_elastic.
    ?entity_elastic terms:id ?entity_elastic_id ;
        ncs:text ?entity_elastic_text.
    FILTER (?entity_elastic_id > 2000)
}
```

В результате выполнения запроса для рассматриваемого документа сформирован список идентификаторов (entity_elastic_id) и названий терминов (entity_term_text), размеченных в документе. Фрагмент списка представлен в табл. 6.

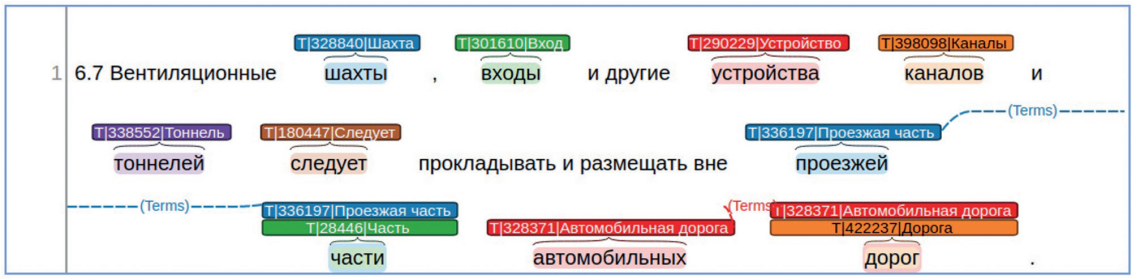


*Рис. 5*. Пример визуализации разметки текста в формате TSV WebAnno

*Fig. 5*. An example of visualisation for the text markup in TSV WebAnno format

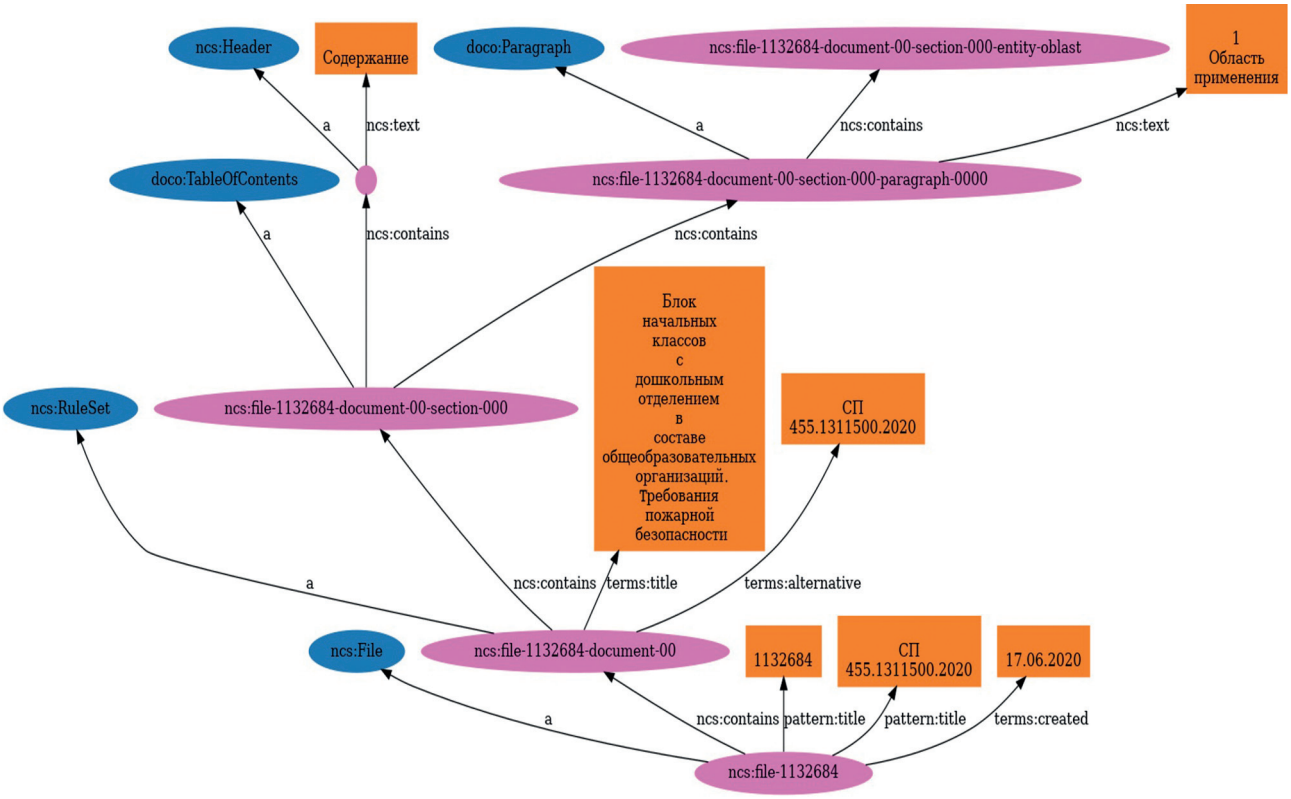


*Рис. 6*. Фрагмент семантической модели в формате RDF

*Fig. 6*. A fragment of the semantic model in RDF format

*Таблица 6*. Пример фрагмента результатов выполнения SPARQL-запроса для поиска наиболее часто встречающихся терминов

*Table 6.* An example of the SPARQL query result for searching the most common terms

| entity_elastic_id | entity_elastic_text |
|---|---|
| 14505 | Федеральное агентство по техническому регулированию и метрологии |
| 25507 | ФГБУ Всероссийский научно-исследовательский институт труда Министерства труда и социальной защиты Российской Федерации |
| 5219 | Министерство Российской Федерации по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий |
| 5635 | Главное управление Министерства Российской Федерации по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий |
| 421860 | Комбинация |
| 422805 | Вентиляция |
| 424253 | Отделение |

## Заключение

В работе создан принципиально новый подход к обработке нормативной документации на русском языке на основе онтологий и методов обработки естественного языка. До настоящего времени известные решения ограничены структурным моделированием научно-технической документации и ручной рубрикацией разделов документов. Встречающиеся в научных исследованиях описания прототипов систем для автоматизированной обработки англоязычной документации, главным образом, используют классический подход на основе правил. Спецификой предлагаемого авторами решения является обучение и использование новейших моделей векторных представлений текстовой информации в комбинации с семантическим моделированием и различными эвристиками для автоматического извлечения терминов из текстов нормативной документации.

В ходе разработки семантической модели выполнен анализ существующих стандартных онтологий, а также добавлены специфические для данного проекта сущности. Такая комбинация позволила добиться упрощения расширяемости семантической модели при добавлении новых словарей и классификаторов, а также обеспечивает в перспективе возможность расширения спектра предметных областей нормативных документов. Разработанный алгоритм автоматической идентификации терминов реализован в виде программного обеспечения, имеющего следующие основные функции: парсинг документов, идентификация терминов, экспорт полученных результатов и автоматическое построение RDF-графа знаний. Для экспериментальной проверки сформирована выборка из 202 документов, на основе которой построен граф знаний, включающий множество найденных из документов терминов и ссылок на соответствующие разделы документов. Для данного графа знаний также разработан программный интерфейс (API), и показана возможность получения информации о терминах с помощью семантических запросов на языке SPARQL. Примеры выполнения этих запросов показали высокую гибкость и простоту использования семантической модели и графа знаний в сторонних приложениях.

## Литература


1. Eastman C.M., Teicholz P., Sacks R., Liston K. BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors. John Wiley & Sons, 2011. 640 p.
2. Liebich T. et al. Industry foundation classes IFC2x edition 3 technical corrigendum 1 / International Alliance for Interoperability (Model Support Group). 2012.
3. Pauwels P., Van Deursen D., Verstraeten R., De Roo J., De Meyer R., Van De Walle R., Van Campenhout J. A semantic rule checking environment for building performance checking // Automation in Construction. 2011. V. 20. N 5. P. 506–518. doi: 10.1016/j.autcon.2010.11.017
4. Zhang C., Beetz J., Weise M. Model view checking: automated validation for IFC building models // eWork and eBusiness in Architecture, Engineering and Construction: Proc. 10th European Conference on Product and Process Modelling, ECPPM. 2014. P. 123–128. doi: 10.1201/b17396-24
5. Pauwels P., Terkaj W. EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology // Automation in Construction. 2016. V. 63. P. 100–133. doi: 10.1016/j.autcon.2015.12.003


## References


1. Eastman C.M., Teicholz P., Sacks R., Liston K. *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors*. John Wiley & Sons, 2011, 640 p.
2. Liebich T. et al. *Industry foundation classes IFC2x edition 3 technical corrigendum 1*. International Alliance for Interoperability (Model Support Group), 2012.
3. Pauwels P., Van Deursen D., Verstraeten R., De Roo J., De Meyer R., Van De Walle R., Van Campenhout J. A semantic rule checking environment for building performance checking. *Automation in Construction*, 2011, vol. 20, no. 5, pp. 506–518. doi: 10.1016/j.autcon.2010.11.017
4. Zhang C., Beetz J., Weise M. Model view checking: automated validation for IFC building models. *eWork and eBusiness in Architecture, Engineering and Construction: Proc. 10th European Conference on Product and Process Modelling, ECPPM*, 2014, pp. 123–128. doi: 10.1201/b17396-24
5. Pauwels P., Terkaj W. EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology. *Automation in Construction*, 2016, vol. 63, pp. 100–133. doi: 10.1016/j.autcon.2015.12.003

6. Dawood H., Siddle J., Dawood N. Integrating IFC and NLP for automating change request validations // Journal of Information Technology in Construction. 2019. V. 24. P. 540–552. doi: 10.36680/J.ITCON.2019.030
7. Hernández E.G., Piulachs J.M. Application of the Dublin Core format for automatic metadata generation and extraction // Proc. 5th International Conference on Dublin Core and Metadata Applications (DC-2005). 2005. P. 213–216.
8. Constantin A., Peroni S., Pettifer S., Shotton D., Vitali F. The document components ontology (DoCO) // Semantic Web. 2016. V. 7. N 2. P. 167–181. doi: 10.3233/SW-150177
9. Villegas M., Bel N. PAROLE/SIMPLE 'lemon'ontology and lexicons // Semantic Web. 2015. V. 6. N 4. P. 363–369. doi: 10.3233/SW-140148
10. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT. V 1. 2019. P. 4171–4186.

6. Dawood H., Siddle J., Dawood N. Integrating IFC and NLP for automating change request validations. *Journal of Information Technology in Construction*, 2019, vol. 24, pp. 540–552. doi: 10.36680/J.ITCON.2019.030
7. Hernández E.G., Piulachs J.M. Application of the Dublin Core format for automatic metadata generation and extraction. *Proc. 5th International Conference on Dublin Core and Metadata Applications (DC-2005)*, 2005, pp. 213–216.
8. Constantin A., Peroni S., Pettifer S., Shotton D., Vitali F. The document components ontology (DoCO). *Semantic Web*, 2016, vol. 7, no. 2, pp. 167–181. doi: 10.3233/SW-150177
9. Villegas M., Bel N. PAROLE/SIMPLE 'lemon'ontology and lexicons. *Semantic Web*, 2015, vol. 6, no. 4, pp. 363–369. doi: 10.3233/SW-140148
10. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT*. V 1. 2019, pp. 4171–4186.

## Авторы

**Муромцев Дмитрий Ильич** — кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 55575780100, http://orcid.org/0000-0002-0644-9242, mouromtsev@itmo.ru
**Шилин Иван Андреевич** — кандидат технических наук, ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57021593400, http://orcid.org/0000-0002-7846-1398, shilinivan@itmo.ru
**Плюхин Дмитрий Алексеевич** — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0001-5790-7883, zeionara@gmail.com
**Баймуратов Ильдар Раисович** — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57204417976, http://orcid.org/0000-0002-6573-131X, baimuratov.i@gmail.com
**Хайдарова Резеда Раитовна** — кандидат технических наук, преподаватель практики, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57031874500, http://orcid.org/0000-0001-8270-9192, mignolowa@gmail.com
**Дементьева Юлия Юрьевна** — магистр бизнес-информатики, руководитель отдела разработки инновационных систем, ООО «Нанософт разработка», Москва, 108811, Российская Федерация, https://orcid.org/0000-0002-6829-6245, yulia.dementeva@gmail.com
**Ожигин Денис Александрович** — магистр информационных технологий, технический директор, ООО «Нанософт разработка», Москва, 108811, Российская Федерация, https://orcid.org/0000-0002-9581-241X, denis@nanocad.ru
**Малышева Татьяна Алексеевна** — кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57191042146, https://orcid.org/0000-0002-1171-379X, tamalysheva@itmo.ru

## Authors

**Dmitry I. Mouromtsev** — PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 55575780100, http://orcid.org/0000-0002-0644-9242, mouromtsev@itmo.ru
**Ivan A. Shilin** — PhD, Assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57021593400, http://orcid.org/0000-0002-7846-1398, shilinivan@itmo.ru

**Dmitrii A. Pliukhin** — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0001-5790-7883, zeionara@gmail.com
**Ildar R. Baimuratov** — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57204417976, http://orcid.org/0000-0002-6573-131X, baimuratov.i@gmail.com

**Rezeda R. Khaydarova** — PhD, Lecturer, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57031874500, http://orcid.org/0000-0001-8270-9192, mignolowa@gmail.com

**Yulia Yu. Dementyeva** — MBI, Head of Innovation Development Department, "Nanosoft razrabotka LLC", Moscow, 108811, Russian Federation, https://orcid.org/0000-0002-6829-6245, yulia.dementeva@gmail.com
**Denis A. Ozhigin** — MIT, Technical Director, "Nanosoft razrabotka LLC", Moscow, 108811, Russian Federation, https://orcid.org/0000-0002-9581-241X, denis@nanocad.ru

**Tatiana A. Malysheva** — PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57191042146, https://orcid.org/0000-0002-1171-379X, tamalysheva@itmo.ru

*Статья поступила в редакцию 18.02.2021*
*Одобрена после рецензирования 10.03.2021*
*Принята к печати 30.03.2021*

*Received 18.02.2021*
*Approved after reviewing 10.03.2021*
*Accepted 30.03.2021*