О создании и перспективах использования корпуса текстов российских правовых актов как набора открытых данных[1]

Д.А. Савельев

научный сотрудник Института проблем правоприменения при Европейском университете в Санкт-Петербурге, кандидат юридических наук. Адрес: 191187, Санкт-Петербург, Гагаринская ул., 6/1. E-mail: dsaveliev@eu.spb.ru

Аннотация

Развивающиеся в настоящее время методы компьютерного анализа текстов могут быть полезны для исследований в юридической науке и практике. Очевидным требованием для такого анализа является наличие открытого и структурированного корпуса текстов. Статья представляет такой корпус текстов правовых актов федерального и регионального законодательства в машиночитаемой форме (набор данных) RusLawOD. Он опубликован в открытом доступе на Интернет-портале Github. Созданный набор данных основан на открытых источниках правовых актов, прежде всего на данных официального Интернет-портала правовой информации (pravo.gov.ru), полученных в результате работы по интеграции открытых данных об официальном опубликовании правовой информации и данных ИПС «Законодательство России». Основным исследовательским вопросом в сфере права при разработке данного ресурса стал вопрос, каким образом осуществлять публикацию текстов правовых актов и метаданных о них. Необходимо прийти в общегосударственном масштабе к общему стандарту описания правовых актов в машиночитаемой форме для возможностей обмена данными между разными информационными системами. Для этого нужно определиться с единообразным наименованием атрибутов, идентифицирующих документ, а также его внутреннюю структуру. В статье предлагаются решения, которые можно взять за основу для этого. Помимо описания данных приводятся примеры, как указанные данные могут помочь в решении научных юридических задач. Такими примерами служат классификация правовых актов и определение частоты коллокаций определенных терминов. На основе анализа опубликованных на указанном портале карточек документов составлен классификатор используемых на практике тематик и произведен подсчет частоты использования каждой из тематик. Автор сравнивает существующую классификацию правовых актов, которая производится при создании ИПС «Законодательство России», и результаты использования методов компьютерной лингвистики для определения наиболее часто используемых в законодательстве тематик, приходя к выводу о том, что современные методы машинного анализа текстов позволяют получать достоверные и значимые результаты.

[1] Автор выражает признательность за методическую помощь научному руководителю Института Вадиму Волкову и ведущему научному сотруднику Института Дмитрию Скугаревскому. Публикация подготовлена в рамках научного проекта № 17-18-01618, поддержанного Российским научным фондом.

Ключевые слова

правовая информация, законодательство, открытые данные, набор данных, XML, правовой акт, машиночитаемый корпус, компьютерная лингвистика, анализ текста.

Библиографическое описание: Савельев Д. А. О создании и перспективах использования корпуса текстов российских правовых актов как набора открытых данных // Право. Журнал Высшей школы экономики. 2018. № 1. С. 26–44.

УДК: 340 DOI: 10.17323/2072-8166.2018.1.26.44

1. Новые возможности информационных технологий и проблемы правовой информатизации

С появлением компьютерных технологий перед правовой наукой открылись новые возможности, связанные со сбором, хранением и обработкой правовой информации в электронной форме. В настоящее время появились мощности для обработки огромного количества текстов с высокой скоростью. Не только поиск документов по определенным словам может быть улучшен с помощью этих информационных технологий. Такие задачи, как автоматизированная аннотация и классификация текстов, анализ использования в текстах тех или иных терминов на уровне каждого отдельного слова, а не на уровне документа (анализ правового тезауруса), автоматизированное извлечение новой информации и построение выводов вплоть до подготовки документов и ответов на вопросы на естественном языке, стали возможными благодаря новым технологиям: статистической обработке массивов текстов как «больших данных», машинному обучению, онтологическому моделированию, компьютерной лингвистике.

Для реализации таких технологий необходимы исходные данные в виде текстов в «машиночитаемой» форме. Остановимся на основных достижениях и проблемах в этой сфере. Начало правовой информатизации мы отсчитываем с 1982 года — со времени появления формата «АИПС-законодательство», созданного в Научном центре правовой информации при Министерстве юстиции СССР[2]. В дальнейшем государство и научное сообщество продолжали работу в этом направлении[3]. К настоящему времени основными в данном направлении стали три достижения. Во-первых, закреплено правило, что неопубликованные акты, затрагивающие права граждан,

[2] О НЦПИ при Министерстве юстиции Российской Федерации см.: [Электронный ресурс]: // URL: http://www.scli.ru:8080/about/ (дата обращения: 15.11.2017)

[3] См.: История развития правовой информатизации России [Электронный ресурс]: // URL: http://pravo.gov.ru/Inform/pravinfarticles/articles/pravinfarticles_7.html (дата обращения: 15.11.2017); Официальное электронное опубликование: История, подходы, перспективы. М., 2012.

не применяются, а обязанность органов государственной власти направлять ответы на индивидуальные обращения граждан в законодательстве была дополнена институтом обязательной публикации информации для всеобщего сведения в Интернете, включая тексты принятых правовых актов и судебных решений. Во-вторых, это официальное электронное опубликование правовых актов на портале, который ведет ФСО России[4]. Здесь можно отметить и другие источники электронной правовой информации — Информационная правовая система «Законодательство России» (далее — ИПС), которую ведет НТЦ «Система»[5], и реестр нормативных правовых актов Министерства юстиции[6]. В третьих, на государственном уровне была закреплена политика публикации наборов открытых данных. Однако проблемой, на наш взгляд, пока является то, что эти три направления не совместились в одно целое.

Как правило, единицей поисковой выдачи при поиске в справочной правовой системе или на любом государственном портале, предоставляющем доступ к правовым актам, является документ. Это снимает базовую потребность профессионалов в ознакомлении с информацией, но ни один из этих источников не позволяет выполнить научную работу по анализу массивов правовых актов в появившемся в последние годы подходе «текст как данные» (англ. — text as data). Основной информационный ресурс здесь — официальное электронное опубликование правовых актов осуществляется в форме графических копий (сканов) бумажного документа, который не содержит текста в формате, пригодном для работы компьютерной программы непосредственно с текстом.

Все указанные источники правовой информации, равно как и коммерческие справочные правовые системы (Гарант, Кодекс, КонсультантПлюс, Право.ру и др.) предназначены для использования непосредственно человеком для ознакомления с размещенными в них документами. На данный момент задача опубликования текстов правовых актов как машиночитаемых открытых данных, которые могут подвергаться массовой автоматизированной обработке, не решена.

2. Источники корпуса текстов правовых актов и его полнота

Прежде всего необходим корпус текстов в формате, который легко поддается обработке компьютерными программами («машиночитаемом», в тер-

[4] Официальное опубликование правовых актов [Электронный ресурс]: // URL: http://publication.pravo.gov.ru/ (дата обращения: 05.12.2017)

[5] Информационно-правовая система «Законодательство России» [Электронный ресурс]: // URL: http://pravo.gov.ru/ips.html (дата обращения: 05.12.2017)

[6] Научный центр правовой информации при Минюсте. Базы данных [Электронный ресурс]: // URL: http://www.scli.ru:8080/bd/ (дата обращения: 05.12.2017)

минах 1990-х годов). Для этого нужно преобразовать опубликованные для всеобщего сведения сканы страниц, а также страницы на государственных сайтах в подходящий текстовый формат. Такая работа была проведена, а также к текстам привязана информация о них (метаданные). Полученный в результате набор данных опубликован на платформе Github[7] для возможного использования другими исследователями. Общее количество документов опубликованной версии набора данных — 409 тыс. (состояние на август 2017 г.[8]), объем архивных файлов — 1.8 Гб, в распакованном виде — 10.7 Гб.

Указанный корпус текстов предназначен прежде всего независимым исследователям, которые не обладают прямым (позволяющим создать и использовать свои программы обработки данных) доступом к данным государственных информационных ресурсов и коммерческих справочных правовых систем. Содержание корпуса текстов в виде открытых данных также означает, что результаты тех или иных исследований могут быть повторены другими учеными.

В оценке набора данных важна его полнота. Мы можем утверждать, что единого, непротиворечивого и абсолютно актуального перечня даже действующих, не говоря уже о когда-либо принятых правовых актах в Российской Федерации не существует. Различные источники собирают от 25 тыс. до 50 млн. различных правовых документов.

Важной характеристикой в правовых информационных ресурсах является законченность наполнения ресурса, которая позволяла бы не только выполнить подборку наиболее значимых документов, но и оставляла бы возможность сказать, что, если документ не найден, значит, в действующих актах его не существует. На данный момент таких электронных ресурсов нет, и все лишь в той или иной степени приближаются к полноте представления данных.

В описываемый набор данных включены правовые акты федерального уровня, уровня субъектов федерации и муниципального уровня. К моменту написания настоящей статьи подготовлены для использования данные Официального интернет-портала правовой информации, состоящие из двух массивов. Первый массив — это тексты, прошедшие на портале официальное опубликование в форме графических образов страниц (сканов), затем распознанные нами при помощи свободного программного обеспечения Tesseract[9]. При этом объем обработанных сканов составил около 200 Гб.

---

[7] Russian Law as Open Data [Электронный ресурс]: // URL: https://github.com/irlcode/RusLawOD (дата обращения: 07.12.2017)

[8] К моменту публикации настоящей статьи продолжается работа пополнение всеми документами 2017 г., в связи с чем не все составляющие набора данных полностью содержат документы 2017 г.

[9] Tesseract Open Source OCR Engine (main repository) [Электронный ресурс] // URL: https://github.com/tesseract-ocr (дата обращения: 05.12.2017)

Они включают акты всех уровней с декабря 2011 года, однако не все органы власти и субъекты федерации подключились к официальному электронному опубликованию одномоментно. Некоторые субъекты федерации на этом портале отсутствуют (например, Москва), а муниципальные акты пока включаются только в отдельных случаях. Поэтому очевиден существенный рост по годам официально опубликованных в электронной форме документов (см. табл. 1).

Таблица 1

Количество документов, найденных в разделе официального опубликования правовых актов портала pravo.gov.ru

Год опубликования	2011	2012	2013	2014	2015	2016	2017
Количество документов	405	2012	4498	14382	78231	90802	101395

Второй массив — это документы, опубликованные на том же портале вне рамок официального электронного опубликования в Информационно-правовой системе «Законодательство России» (далее — ИПС «Законодательство»). Она включает только федеральное законодательство, преимущественно с 1991 года. При сравнении количества новых правовых актов за год в ИПС Законодательство (см. табл. 2, рис. 1) можно констатировать постепенное его увеличение и резкое увеличение в 2016 году. Однако установить, насколько действительно увеличивается законодательная масса, по этим данным трудно, так как полнота базы не известна достоверно. Например, при поиске на Интернет-портале системы КонсультантПлюс за период с 1.01 по 31.12 2016 только по видам документов «Закон, Постановление, Распоряжение, Указ, Кодекс» находится более 25 тыс. документов, тогда как в ИПС за этот период — менее 12 тыс. Точное сравнение по сайтам провести невозможно, по нашей оценке, ввиду различных подходов к формированию списков документов.

Таблица 2

Количество документов, найденных в разделе ИПС «Законодательство России» портала pravo.gov.ru (выборочно, за указанный год)

Год	Документов
До 1991	364
1991	3702
1992	7043
2002	6426
2012	8584

Год	Документов
2013	8020
2014	8669
2015	9024
2016	11729
2017	11855

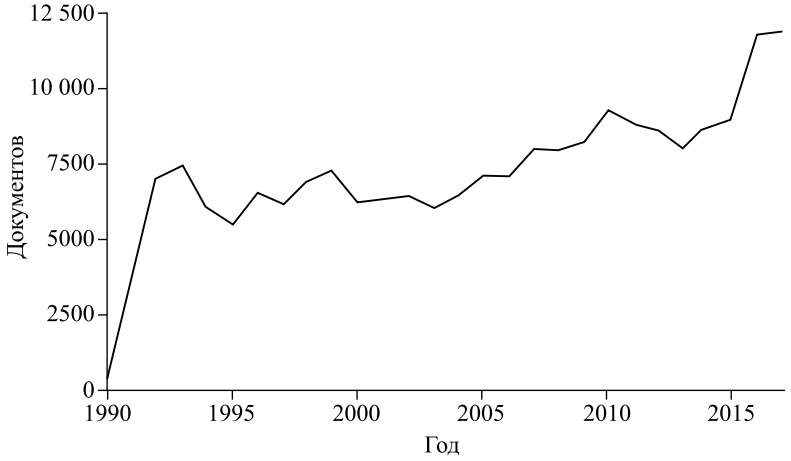

Рис.1. Количество документов в ИПС «Законодательство» по году принятия акта

Эти два массива частично пересекаются, поэтому в некоторых документах набора есть два текста, и дублирование оставлено для исследовательских целей. В дальнейшем планируется дополнить информацию из других источников.

При составлении описываемого корпуса использовались тексты только первоначальных редакций документов. Консолидированные редакции (с внесенными изменениями и дополнениями) не включаются во избежание дублирования текстов и в целях правильного подсчета статистики. Поскольку среди опубликованных есть документы, объем которых превышает 1000 страниц, а отдельные документы превышают и 20 000 страниц. Чаще всего объемные документы — это программы капитального ремонта, описания географических карт и бюджетно-финансовые данные, которые в целом не дают много текстовой информации. Распознавание документов не осуществлялось после 500-й страницы. Распознавание (англ. — OCR) текста осуществлялось автоматически, без корректуры. Поэтому в текстах возможны опечатки. Прежде всего они встречаются там, где поверх текста поставлена печать, вписаны от руки реквизиты документа и форматированные таблицы. Тем не менее для

целей формирования корпуса главным было охватить существенную часть текста. Если имеется два текста — официальное опубликование и ИПС «Законодательство России», то в настоящий корпус включаются оба текста, что позволяет оценить качество распознавания остальных текстов.

3. Особенности «машиночитаемого» формата корпуса текстов правовых актов

Подготовка набора данных начинается с выбора формата данных. Наиболее простым для использования компьютерными программами является «чистый текст» (англ. — plain text), где отсутствует форматирование, кроме разбивки на абзацы. Однако такой текст невозможно разметить, а также хранить в нем метаданные (сведения о документе). Поэтому был выбран формат XML, который предусматривает такие возможности. XML позволяет внутри текстового файла специальными метками (тегами) разметить текстовую информацию на ее структурные части, а также включить внутрь такого текста сведения о документе (метаданные) таким образом, что впоследствии различные компьютерные программы могут воспроизвести и использовать эти структурированные данные. Но XML — лишь стандарт записи структурных единиц в тексте, который не говорит, какие разделы должны быть внутри файла, как они должны называться. Чтобы можно было осуществлять обмен открытой правовой информацией между системами, необходимо разработать общий стандарт наименования элементов файла.

При этом логично использовать мировой опыт в этой сфере. Структура документа была исполнена, ориентируясь на формат Akoma Ntoso[10], который впоследствии публикуется как OASIS LegalDocML[11]. Однако на данный момент формат не полностью соответствует этому стандарту, так как он предполагает разбивку внутренней структуры документа, которая к настоящему моменту еще не разработана в нашем корпусе текстов.

При разработке набора данных пришлось преодолевать не только проблему графического формата текстов. В официальных источниках поля данных недостаточно структурированы для автоматизированной обработки. Например, открытые данные об официальном опубликовании содержат строку <title>, в которой одновременно содержатся и вид (тип) документа, и наименование принявшего документ органа, и дата подписания, и номер.

[10] Akoma Ntoso. Официальный сайт [Электронный ресурс]: // URL: http://akomantoso.org (дата обращения: 15.11.2017)

[11] Akoma Ntoso Version 1.0. Part 2: Specifications [Электронный ресурс]: // URL: http://docs.oasis-open.org/legaldocml/akn-core/v1.0/cs01/part2-specs/akn-core-v1.0-cs01-part2-specs.html (дата обращения: 05.12.2017)

Поскольку это разные свойства документа, для исследований данных их необходимо различать. Эти данные разбиваются на разные элементы разметки при помощи «основанного на правилах» (англ. — rule-based) подхода: то, что может быть определено по словарю известных видов документов и органов, определяется в соответствии со словарями, а также идентифицируются дата и номер по типовым признакам. Поскольку мы собираем данные более чем из одного источника, встает еще и проблема дублирования данных. При этом для максимальной точности сохранения данных мы параллельно сохраняем значения атрибутов документа из всех источников.

В итоге схема структуры документа, из которого состоит набор данных, состоит из следующих разделов:

<act> («акт») — общее указание на тип документа — правовой акт в отличие от других видов правовых документов — законопроекта и т.п. В раздел входят подразделы идентификации, сведений об опубликовании и текста;

<identification> («идентификация») — раздел идентификационных данных, в котором содержатся атрибуты (реквизиты, свойства) документа, позволяющие его уникально идентифицировать (выделить единственный документ среди других);

<publication> («опубликование») — раздел данных о публикации документа, в котором приводятся сведения об источниках. Однако в отечественном документообороте данные опубликования тоже носят идентифицирующий характер;

<classification> («классификация») — раздел данных о классификации документа — отнесения его к рубрикам классификатора или назначении ключевых слов;

<body> («тело») — раздел текста документа, который в настоящий момент состоит из двух вариантов текста — официального опубликования и текста ИПС «Законодательство России» (одного из них или двух вместе).

Особенности заполнения отдельных метаданных требуют пояснения. В практике создания правовых баз данных имеются разночтения относительно названий и состава атрибутов, идентифицирующих документ. Так, термины «постановление», «закон» и т.п. называют и видом, и типом документа, а в понятие вида документа может входить или не входить принимающий орган. Название у многих документов часто отсутствует, но при этом создается при создании информационного ресурса операторами.

Полное описание всех элементов (схема) дано в приложении, а также его можно найти по адресу набора данных в Интернете, приведенному выше, и в этом источнике также будут публиковаться обновления схемы и набора данных. Все элементы имеют комментарии, которые опубликованы вместе с набором данных.

Наряду с набором данных важным техническим моментом является составление единых перечней (словарей) видов документов и принимающих документы органов власти, которые также опубликованы вместе с набором данных. Данная инициатива предполагает привлечение независимых разработчиков и специалистов органов власти для дальнейшего совершенствования формата публикации документов.

4. Правовые вопросы доступа

Вопросы доступности правовых актов в современной России отражены в законодательстве. В частности, Заключение Комитета конституционного надзора СССР впервые установило следующее правило: «Опубликование законов и других нормативных актов, касающихся прав, свобод и обязанностей граждан, то есть доведение их тем или иным способом до всеобщего сведения, является обязательным условием применения этих актов»[12]. Это правило впоследствии было закреплено и в п. 3 ст. 15 Конституции России. При этом в Заключении справедливо отмечается: «Предоставление должностным лицам права определять, имеют ли правовые акты «общее значение» или не имеют, а также засекречивать информацию и в зависимости от своего решения публиковать или не публиковать нормативные акты открывает возможность произвольного, неправомерного и не контролируемого обществом ограничения прав и свобод, для возложения на граждан дополнительных, нередко обременительных обязанностей, а также предоставления неоправданных льгот отдельным категориям граждан»[13].

В дальнейшем, при развитии электронных технологий и увеличении массы законодательных актов, опубликование только на бумаге стало недостаточным. В 2003 году впервые стало обязательным размещение в сети Интернет правовых актов, принимаемых Правительством Российской Федерации и федеральными органами исполнительной власти[14]. Данное положение было затем развито в законах об обеспечении доступа к информации о деятельности государственных органов, органов местного самоуправления и судов[15].

---

[12] Заключение Комитета конституционного надзора СССР от 29.11.1990 № 12 (2-12) «О правилах, допускающих применение неопубликованных нормативных актов о правах, свободах и обязанностях граждан» // СПС КонсультантПлюс.

[13] Там же.

[14] См.: Постановление Правительства Российской Федерации от 12.02.2003 № 98 «Об обеспечении доступа к информации о деятельности Правительства Российской Федерации и федеральных органов исполнительной власти» (с последующими изменениями).

[15] Федеральный закон от 09.02.2009 № 8-ФЗ «Об обеспечении обращения к информации о деятельности государственных органов и органов местного самоуправления» (с последующими

Наконец, современный этап развития Интернет-технологий привел к пониманию, что размещение текстов и простых форм поиска для ознакомления граждан недостаточно. Появилось понятие «открытые связанные данные», суть которого заключается в размещении в сети Интернет больших массивов данных («наборов данных»), которые могут быть свободно использованы разработчиками различных программ и информационных систем в своих продуктах[16]. В таких продуктах могут применяться современные средства обработки информации, и новое знание может быть получено на основе интеграции («связывания») различных видов информационных ресурсов. В связи с этим на Официальном интернет-портале правовой информации[17] публикуются наборы открытых данных об официальном опубликовании правовых актов. Однако эти данные включают только сведения о правовом акте, но не его машиночитаемый текст.

Пп. 1 п. 6 ст. 1259 Гражданского кодекса Российской Федерации исключает тексты правовых актов из сферы действия авторских прав. В соответствии с этим такую информацию можно признать общедоступной. В части открытых данных об официальном опубликовании Официальный интернет-портал не ограничивает коммерческое или некоммерческое использование таких данных при условии атрибуции (указания ссылки на первоначальный источник) и неискажения данных[18]. Что касается данных ИПС «Законодательство России», то в описании на портале указано «Система предназначена для: <...> автоматизированной поддержки юридической обработки правовой информации»[19].

Информация, публикуемая для использования исследователями в сети Интернет, должна сопровождаться понятными во всем мире условиями использования, в связи с чем публикация описываемого набора данных осуществляется на условиях лицензии Creative Commons Attribution-NonCommercial 4.0 International[20].

изменениями); Федеральный закон от 22.12.2008 № 262-ФЗ «Об обеспечении обращения к информации о деятельности судов в Российской Федерации» (с последующими изменениями).

[16] Подробнее о правовом регулировании в сфере открытых данных см.: Открытое правительство / Открытые данные [Электронный ресурс]: // URL: http://opendata.open.gov.ru/event/5598184/ (дата обращения: 05.12.2017)

[17] На основании пп. «г» п. 2 Указа Президента Российской Федерации от 7.05.2012 № 601 «Об основных направлениях совершенствования системы государственного управления» // СПС Гарант.

[18] Официальное опубликование правовых актов. Открытые данные [Электронный ресурс]: // URL: http://publication.pravo.gov.ru/od/ (дата обращения: 05.12.2017)

[19] Информационно-правовая система «Законодательство России» [Электронный ресурс]: // URL: http://pravo.gov.ru/ips.html (дата обращения: 05.12.2017)

[20] Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) [Электронный ресурс]: // URL: https://creativecommons.org/licenses/by-nc/4.0/ (дата обращения: 05.12.2017)

5. Примеры использования представляемых данных

Публикация набора данных дает новые возможности использования текстов правовых актов, которые не доступны исследователям без прямого доступа к базам данных органов власти или коммерческих структур, позволяющего написать собственную компьютерную программу для анализа. Ниже мы приводим примеры такого использования, которые стали результатом работы специально созданного программного обеспечения.

5.1. Изучение классификации корпуса текстов по общеправовому классификатору отраслей законодательства и перспектив его использования

Классификатор, утвержденный почти 25 лет назад[21], последний раз изменялся 12 лет назад. В ходе исследования нами установлено следующее. В данных ИПС «Законодательство России» содержатся сведения о классификации правовых актов в соответствии с упомянутым выше классификатором. Причем, если в официально опубликованном классификаторе содержится три уровня, то в ИПС используется пять уровней, и нам не удалось найти опубликованными тематики всех уровней в сводном виде.

Использованный нами инструментарий позволил вывести актуальный классификатор с рубриками, которые используются в карточках документов составителями ИПС и частотой их использования. При этом мы видим, что из общего числа в примерно 190 тыс. документов, принятых по 2016 год включительно, примерно 165 тыс. документов имеют ссылку хотя бы на одну рубрику классификатора. В итоге по этим документам определено, что в классификаторе создано 4112 рубрик, хотя в официально утвержденном классификаторе 1151 рубрика; в самой часто встречающейся рубрике более 39 тыс. документов (см. табл. 3). Среднее количество тематик в документе в целом по массиву 3.74, при этом подавляющее большинство документов отнесены не более чем к пяти тематикам, но встречаются документы (около 40) которые связаны с более чем 100 тематиками.

При этом по содержательному наполнению многих тематик классификатора невозможно анализировать или осуществлять поиск документов о нормативно-правовом регулировании или, например, сделать вывод о количестве нормативно-правовых актов, регулирующих те или иные вопросы. Большинство документов в тематиках — не нормативно-правое регулирова-

[21] См.: Указ Президента Российской Федерации от 16.12.1993 № 2171 «Об общеправовом классификаторе отраслей законодательства»; Указ Президента Российской Федерации от 15.03.2000 № 511 «О классификаторе правовых актов» (с последующими изменениями) // СПС Гарант.

ние, а индивидуально-правовые акты по частным вопросам. Но даже если не брать это в расчет, то такие тематики, как «Правила, инструкции, указания, порядки и иные решения», не несут практически никакой функции по поиску или идентификации документов.

Таблица 3

10 тематик классификатора, к которым отнесено наибольшее количество документов

Тематика	Количество документов в тематике
Стадии прохождения законопроектов	39203
Отмена, изменение и дополнение нормативных правовых актов	32800
Награждения государственными наградами Российской Федерации	7329
Правила, инструкции, указания, порядки и иные решения	7179
Присвоение (лишение) почетных званий	5458
Награждения грамотами, объявление благодарности Президента Российской Федерации, Правительства Российской Федерации	4781
Представители при рассмотрении законопроектов	4611
Акционерное общество	4452
Образование, реорганизация и ликвидация юридических лиц (см. также 010.170.010.020, 190.020.070)	4188
Формы финансовой помощи (субвенции, дотации, ссуды, субсидии)	4138

Очень грубую попытку оценить распределение законодательства по наиболее общим тематикам можно сделать, определив число документов, отнесенных к самому верхнему уровню классификатора (см. табл. 4)

Однако если внимательно посмотреть на лидирующую с большим отрывом тематику «Конституционный строй», то мы отметим, что самые часто встречающиеся подтематики (табл. 3) «Стадии прохождения законопроектов» и «Отмена, изменение и дополнение нормативных правовых актов» относятся именно к ней. Но логически трудно согласиться с тем, что большинство правовых актов в общей массе посвящено именно конституционному строю. То же касается и международных отношений, где статистически много назначений и ратификации международных актов.

Приведенные выше данные являются наглядной иллюстрацией, почему сейчас классификаторы в информационном поиске не используются — их

ведение в адекватном виде вручную крайне трудоемко[22] и вследствие этого, по нашему мнению, на больших объемах бессмысленно. В настоящее время практическое использование специалистами классификатора для поиска информации сводится к минимуму, поскольку существуют возможности полнотекстового поиска. Примерно то же самое можно сказать и о заранее выделяемых ключевых словах, которые также назначаются документам в ИПС.

Таблица 4

Распределение документов по тематикам верхнего уровня классификатора

010 Конституционный строй	161804
020 Основы государственного управления	78273
200 Международные отношения. Международное право	61530
030 Гражданское право	52353
080 Финансы	50885
210 Индивидуальные правовые акты по кадровым вопросам, вопросам награждения, помилования, гражданства, присвоения почетных и иных званий	46176
090 Хозяйственная деятельность	40107
060 Труд и занятость населения	15588
130 Образование. Наука. Культура	15264
110 Природные ресурсы и охрана окружающей природной среды	14805
070 Социальное обеспечение и социальное страхование	11840
150 Оборона	10435
100 Внешнеэкономическая деятельность. Таможенное дело	9918
160 Безопасность и охрана правопорядка	9386
180 Правосудие	8753
140 Здравоохранение. Физическая культура и спорт. Туризм	8732
120 Информация и информатизация	6542
050 Жилище	6403
170 Уголовное право. Исполнение наказаний	3617
190 Прокуратура. Органы юстиции. Адвокатура. Нотариат	1945
040 Семья	1706

5.2. Статистический анализ употребления терминов в названиях и текстах документов

Чтобы представить себе общую картину тематики законодательства, «популярности» той или иной темы в тот или иной период, на какие группы распределяется принимаемое законодательство без использования ручной

[22] *Шаршун В.А.* О едином правовом классификаторе Республики Беларусь // Информационное право. № 3. 2015. С. 7–11.

классификации, следует проанализировать коллокации, или n-граммы слов, как минимум, в названиях документов. В термине «n-грамма» переменная n означает целое число — количество слов в словосочетании, т.е. n-граммы — это словосочетания из двух или более слов. В компьютерной лингвистике[23] под коллокациями понимают устойчивые словосочетания[24]. При подсчете сочетаний слов, которые наиболее часто встречаются в названиях документов, мы можем составить мнение о том, какие темы чаще всего рассматриваются в документах.

При программной обработке названий того же корпуса правовых актов, который был рассмотрен выше применительно к классификации, получены данные о статистике использования словосочетаний, и, как наиболее показательные, выбраны сочетания из двух слов (биграммы). Чтобы статистика использования словосочетаний не зависела от форм слова, необходимо каждое слово привести к начальной форме (стемминг или лемматизация[25]) Наиболее часто встречающиеся из них показаны в табл. 5.

Таблица 5

Наиболее часто встречающиеся в названиях документов биграммы

Словосочетание	Встречается, раз
Российский Федерация	136065
Федеральный закон	60721
Внесение изменение	54004
закон внесение	29581
Правительство Российской	24311
проект федеральный	22719
изменение статья	13312
Кодекс российской	12076
Федеральный собрание	11349
статья федеральный	11346

Как мы видим из таблицы, вполне ожидаемое первое место занимает упоминание Российской Федерации, а наиболее часто употребимые словосоче-

[23] См. напр.: [Электронный ресурс]: // URL: https://dic.academic.ru/dic.nsf/ruwiki/977041 (дата обращения: 05.12.2017)

[24] Подробнее см., напр.: *Кочеткова Н.А.* Статистические языковые методы. Коллокации и коллигации // Новые информационные технологии в автоматизированных системах. 2013. № 16 [Электронный ресурс]: // URL: http://cyberleninka.ru/article/n/statisticheskie-yazykovye-metody-kollokatsii-i-kolligatsii (дата обращения: 07.12.2017)

[25] С использованием свободного программного обеспечения Pymorphy2. См.: Морфологический анализатор pymorphy2. [Электронный ресурс]: // URL: http://pymorphy2.readthedocs.io/en/latest/ (дата обращения: 05.12.2017). Подробнее об этом: *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. Basel, 2015. P. 320–332.

тания в названиях связаны с внесением изменений в ранее принятые акты. Но внесение изменений не составляет смысловой темы правового акта. Чтобы увидеть значимые словосочетания, нужно провести экспертную оценку их значимости. Например, на наш взгляд, значимыми являются приведенные в табл. 6 словосочетания.

Таблица 6

Частота упоминаний биграмм в названиях документов (экспертный отбор значимых коллокаций)

Словосочетание	Встречается, раз
государственный награда	3676
награда российский	3607
награждение государственный	3582
официальный представитель	3508
акционерный общество	3414
почетный звание	3014
награждение орден	2962
административный правонарушение	2953
присвоение почетный	2919
почетный грамота	2526
подписание соглашение	2519
ратификация соглашение	2468
награждение почетный	2405
профессиональный образование	2368
федеральный собственность	2276
звание заслужить	2106
грамота правительство	1988
государственный образовательный	1809
военный служба	1788
чрезвычайный ситуация	1787
федеральный агентство	1774
территория российский	1769
чрезвычайный полномочный	1726

Таким образом, по массиву названий документов мы быстро можем определить, что среди правовых актов наиболее часто выходят документы о внесении изменений и дополнений в ранее принятые акты. Содержательно в законодательстве наиболее часты документы о награждениях, назначениях, а также вопросы отдельных акционерных обществ, федеральной собственности, образования, военной службы и ЧС. Эти результаты в целом похожи на рассмотренное в предыдущем разделе распределение тематик классификатора, но могут быть и более точными. Однако результаты статистического

исследования получены программным способом за считанные минуты, тогда как работа по назначению операторами тематик конкретным документам (это более 600 тыс. поставленных вручную тематик) заняла, по всей видимости, несравнимо больше времени и человеческих ресурсов.

Показанный подход к анализу названий — далеко не единственный. Существуют и значительно более сложные технологии, основанные на обработке больших массивов текста, которые можно реализовать с помощью описываемого в настоящей статье корпуса текстов. Упомянутая здесь технология приводится в качестве примера начального уровня, для ознакомления юридической общественности с возможностями, которые открываются при использовании корпуса текстов.

6. Заключение и выводы

К настоящему моменту впервые сформирован и открыт для свободного использования корпус текстов и метаданных законодательства Российской Федерации, субъектов федерации в машиночитаемом формате (как «набор данных»), который позволяет разработчикам и исследователям использовать его для углубленного анализа законодательства путем использования различного программного обеспечения. В ходе создания такого корпуса возник вопрос о стандартизации подходов к разметке текстов правовых актов, и в качестве вывода сформулировано предложение о формировании такого стандарта на основе международного опыта и интернациональных стандартов. В настоящее время существенно возрастает и объем публикуемых в электронной форме правовых актов, и возможности современных информационных технологий, в частности, компьютерной лингвистики, при обработке текстов. Современные технологии анализа текстов, которые можно использовать на рассматриваемом корпусе, быстрее и точнее приводят к результатам правовых исследований. Приведенные в качестве примеров исследования на опубликованном наборе данных — далеко не единственные, и публикация набора данных открывает возможности для использования других, самых современных методик.

Библиография

Баранов В.М., Кузнецов А.П., Маршакова Н.Н. Классификация в российском законодательстве (теоретико-прикладное исследование): М.: Юрлитинформ, 2014. 160 с.

Боярский К.К. Введение в компьютерную лингвистику. СПБ: НИУ ИТМО, 2014. 72 с.

Будаков А.С. Вопросы официального опубликования правовых актов в электронном виде / Получение, хранение и использование информации в электронной среде:

публично-правовое и частноправовое регулирование: матер. международной конференции. СПБ.:Президентская библиотека, 2013. С. 25–30.

Вершинин А.П. Электронный Свод законов и правовая информатизация в России // Известия высших учебных заведений. Правоведение. 2010. № 4. С. 98–108.

Вершинин А.П. От свода законов Российской империи к автоматизированной систематизации российского законодательства // Государство и право. 2016. № 10. С. 90–91.

Захаров Г.Н. Классификатор правовых актов // Вестник Тверского государственного университета. Серия «Право». 2015. № 3. С. 20–25.

Звягинцев М.Н. Классификация муниципальных правовых актов // Экономика и управление. № 4. 2007. С. 54–56.

Исаков В.Б. Формирование правовой основы системы официального электронного опубликования / Получение, хранение и использование информации в электронной среде. СПБ.: Президентская библиотека, 2013. С.18–24.

История развития правовой информатизации России. URL: http://pravo.gov.ru/Inform/pravinfarticles/articles/pravinfarticles_7.html (дата обращения: 15.11.2017)

Николаев И.С., Митренина О.В., Ландо Т.М. Прикладная и компьютерная лингвистика. М.: URSS, 2017. 320 с.

Официальное электронное опубликование: история, подходы, перспективы. М.: Формула права, 2012. 320 с.

Ткаченко Н.В. Статистический анализ федерального законодательства. URL: https://csr.ru/wp-content/uploads/2017/02/Issledovanie_TSSR_statistika-po-zakonoproektam.pdf (дата обращения: 15.11.2017)

Шаршун В.А. О едином правовом классификаторе Республики Беларусь // Информационное право. № 3. 2015. С. 7–11.

Lodder A., Oskamp A. Information Technology and Lawyers. Advanced Technology in the Legal Domain from Challenges to Daily Routine. Berlin: Springer, 2006. 198 p.

Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages / Analysis of Images, Social Networks and Texts. Basel: Springer International, 2015. P. 320–332.

On Creating and Using Text of the Russian Federation Corpus of Legal Acts as an Open Dataset

Denis Saveliev

Researcher, Institute for Implementing Law, European University in Saint Petersburg, Candidate of Juridical Sciences. Address: 6/1 Gagarinskaya Str., Saint Petersburg 191887, Russian Federation. E-mail: dsaveliev@eu.spb.ru

Abstract

Methods of computer-aided text analysis that are currently being developed can be useful for research in legal science and in practice. An obvious requirement for such an analysis is the availability of an open and structured corpus of texts. The article presents such a corpus of texts of legal acts of federal and regional legislation in a machine-readable form (of a dataset) RusLawOD. It is publicly available on the Github Internet portal. The created

data set is based on open sources of legal acts, primarily on the data of the Official Internet Portal of Legal Information (pravo.gov.ru) as a result of integration of open data about published officially legal acts and the *Zakonodatelstvo Rossii* legal information system. The main research issue in the field of law in the development of this resource was the question how to publish the texts of legal acts and metadata about them. It is necessary to come on a nationwide scale to the general standard for the description of legal acts in machine-readable form for the possibilities of data exchange between different information systems. To do this, we need to determine the uniform name of the attributes that identify the document, as well as its internal structure. The article suggests solutions that can be taken as a basis for this. In addition to describing the data, examples are given how the data presented can help in solving research legal problems. Such examples are the classification of legal acts and the definition of the frequency of collocations of certain terms. On the basis of analysis of metadata about documents published in the official site, the classifier of really used themes was reconstructed, and theme usage was counted. The author compares existing classification of legal acts and the use of methods of computer linguistics to determine the most frequently used subjects in legislation, coming to the conclusion that modern methods of computer-based text analysis make it possible to get valuable and proven results.

Keywords

legal information, legislation, open data, dataset, XML, legal act, machine-readable corpus, computer linguistics, text as data.

Citation: Saveliev D.A. (2018) On Creating and Using Text of the Russian Federation Corpus of Legal Acts Acts as Open Dataset. *Pravo. Zhurnal Vysshey shkoly ekonomiki*, no 1, pp. 26–44 (in Russian)

DOI: 10.17323/2072-8166.2018.1.26.44

References

Baranov V.M., Kuznetsov A.P., Marshakova N.N. (2014) *Klassifikatsiya v rossiyskom zakonodatel'stve (teoretiko-prikladnoe issledovanie)* [Classification in Russian legislation (theoretical and applied research]. Moscow: Yurlitinform, 160 p. (in Russian)

Budakov A.S. (2013) Voprosy ofitsial'nogo opublikovaniya pravovykh aktov v elektronnom vide [Issues of formal publishing legal acts in electronic form]. *Poluchenie, khranenie i ispol'zovanie informatsii v elektronnoy srede: publichno-pravovoe i chastnopravovoe regulirovanie* [Retrieving, keeping and applying information in the electronic environment. N.A. Shevelev (ed.)]. Saint Petersburg: Presidential Library, p. 25–30.

Boyarskiy K. K. (2014) *Vvedenie v komp'yuternuyu lingvistiku* [Introduction into computer linguistics]. Saint Petersburg: NIU ITMO Press, 72 p.

Isakov V.B. (2013) Formirovanie pravovoy osnovy sistemy ofitsial'nogo elektronnogo opublikovaniya [Forming legal basis of official electronic publication]. *Poluchenie, khranenie i ispol'zovanie informatsii v elektronnoy srede: publichno-pravovoe i chastnopravovoe regulirovanie...* [Retrieving, keeping and applying information in the electronic environment...]. Saint Petersburg: Presidential Library, p.18–24.

Istoriya razvitiya pravovoy informatizatsii Rossii (2014) [History of legal information system in Russia]. Available at: URL: http://pravo.gov.ru/Inform/pravinfarticles/articles/pravinfarticles_7.html (accessed: 15.11. 2017)

Korobov M.V. (2015) Morphological analyzer and generator for Russian and Ukrainian languages. *Analysis of images, social networks and texts*. Basel: Springer International, p. 320–332.

Lodder A., Oskamp A. (2006) *Information technology and lawyers. Advanced technology in the legal domain, from challenges to daily routine*. Berlin: Springer, 198 p.

Nikolaev I.S., Mitrenina O.V., Lando T.M. (2017) *Prikladnaya i komp'yuternaya lingvistika* [Applied and computer linguistics]. Moscow: URSS, 320 p. (in Russian)

Officialnoye electronnoye opublikovamie: isrotia, podhody, perspectivy (2012) [Official electronic publishing: history, approaches, prospects]. V.B Isakov, ed. Moscow: Formula prava, 320 p. (in Russian)

Sharshun V.A. (2015) O edinom pravovom klassifikatore Respubliki Belarus' [On the unified nomenclature of the Republic of Belarus]. *Informatsionnoe pravo*, no 3, p. 7–11.

Tkachenko N.V. (2016) *Statisticheskiy analiz federal'nogo zakonodatel'stva* Available at: URL: https://csr.ru/wp-content/uploads/2017/02/Issledovanie_TSSR_statistika-po-zakonoproektam.pdf (accessed: 15.11.2017)

Vershinin A.P. (2010) Elektronnyy Svod zakonov i pravovaya informatizatsiya v Rossii [Electronic digest of laws and legal information system in Russia]. *Izvestiya vysshikh uchebnykh zavedeniy. Pravovedenie*, no 4, p. 98–108.

Vershinin A.P. (2016) Ot svoda zakonov Rossiyskoy imperii k avtomatizirovannoy sistematizatsii rossiyskogo zakonodatel'stva [From The Digest of Laws of the Russian Empire to automatic system of Russian law]. *Gosudarstvo i pravo*, no 10, p. 90–91.

Zakharov G.N. (2015) Klassifikator pravovykh aktov [Nomenclature of legal acts]. *Vestnik Tverskogo universiteta*, no 3, p. 20–25.

Zvyagintsev M.N. (2007) Klassifikatsiya munitsipal'nykh pravovykh aktov [Nomenclature of municipal legal acts]. *Ekonomika i upravlenie*, no 4, p. 54–56.