КЛАССИФИКАЦИЯ ФОРМАТОВ ДАННЫХ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ ДЛЯ СЕРИАЛИЗАЦИИ ПРАВОВЫХ АКТОВ В МАШИНОЧИТАЕМОЙ ФОРМЕ

Лахтин С.Е.[1], Цимбал В.А.[2], Амелёнков А.А.[3]

Ключевые слова: *правовая информатизация, цифровая трансформация нормативной правовой отрасли, машиночитаемость права, нормативный правовой акт, уровни, требования, сериализация.*

Аннотация

Цель работы: создание научно-методической базы концепции развития технологий машиночитаемого права.

Методы: системный анализ, классификация, кластеризация, формальные грамматики.

Результаты исследования: обоснована классификация уровней машиночитаемости; определены технологические требования к форматам сериализации цифровых объектов, репрезентации правовых норм, документов и правовых актов; обоснованы выводы, что машиночитаемость права зависит от уровня развития информационных технологий и не может выйти за рамки их ограничений, что технологии непосредственного программирования машиночитаемых/машиноисполняемых правовых актов, вероятно, могут быть основаны на базе формально-строгих языков DSL (domain specific languages), оперирующих онтологиями, широко выходящими за рамки одного лишь права.

EDN: RDXGWO

Введение

Любая область цифровой трансформации, и *правовая информатизация* в этом смысле не исключение, опирается в *первую* очередь на свойства цифрового мира, а лишь *потом* — на предметную область, поскольку не может выйти за границы информационных технологий. Настоящая работа направлена, в частности, на уточнение такого понятия, как *машиночитаемость* в рамках форматов хранения и представления аутентичных электронных документов и правовых актов, включающих в первую очередь *нормативные правовые акты* (НПА), согласование понятия нормативно-правовой предметной области с *требованиями* к возможным реализациям в рамках форматов данных.

Необходимость точного определения термина «машиночитаемость» в области НПА вызвана потребностями устранения неоднозначного толкования в уровнях существующих и разрабатываемых спецификаций на *данные* и *программное обеспечение* (ПО) в нормативной правовой области и гармонизации в определении термина машиночитаемость в части обработки НПА в электронной форме.

Развитие машиночитаемых электронных форм НПА является важным направлением цифровой трансформации в силу целого ряда обстоятельств [1—8, 16]. Например, актуальная задача согласования (гармонизации) норм права требуется на всех уровнях законодательства Российской Федерации. Кроме этого, общий технологический базис необходим и для решения задачи поиска противоречий в документах сопредельных государств,

[1] ***Лахтин Станислав Евгеньевич,*** *руководитель проектной группы Комитета по развитию безопасной среды в сфере цифровых технологий, г. Москва, Российская Федерация.*
E-mail: e8686@yandex.ru

[2] ***Цимбал Владимир Анатольевич,*** *доктор технических наук, профессор, заслуженный деятель науки РФ, ведущий научный сотрудник Института инженерной физики, г. Серпухов, Российская Федерация.*
E-mail: tsimbal@mail.ru

[3] ***Амелёнков Андрей Алексеевич,*** *кандидат экономических наук, доцент, начальник отдела Московского государственного лингвистического университета, г. Москва, Российская Федерация.*
E-mail: amelioncov@linguanet.ru

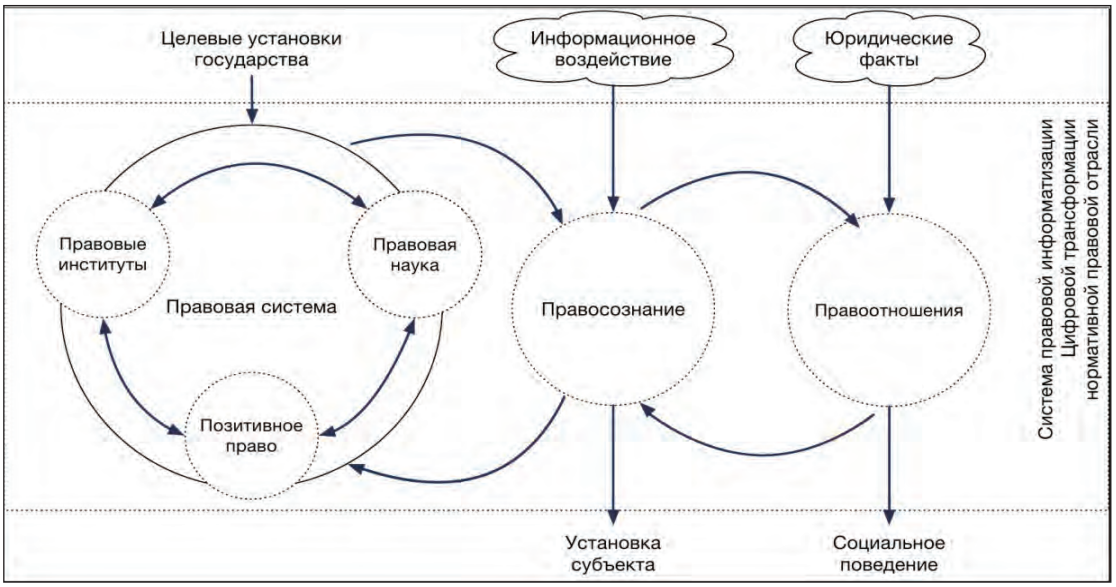

Рис. 1. Концептуальная модель системы правового регулирования

таких как Таможенный союз, ЕАЭС, организации ШОС, унаследованных документах СССР, нормативных базах сопредельных и взаимодействующих с Российской Федерацией государств. Это особенно важно для эффективной работы таможенных служб, контрольной и надзорной деятельности, антикоррупционной деятельности, противодействия незаконному обороту наркотических средств и прочим задачам трансграничного сотрудничества, где НПА должны быть гармонизированы в рамках сразу нескольких законодательств.

Решение подобного рода задачи невозможно лишь на уровне экспертного сообщества. Применение технологий машинного обучения, *Data Mining* (интеллектуальный анализ данных) [9], *Big Data* (большие данные) [13], других технологий искусственного интеллекта, машинная обработка данных, сопоставление сведений, поиск противоречий, дублирований и др. требуют формирования единого технологического и прежде всего инженерного базиса.

Постановка задачи

Информационные правоотношения — урегулированные нормами права информационные отношения, т. е. общественные отношения, возникающие по поводу информации либо юридически значимых результатов действий (бездействия) в отношении этой информации (передача, получение, преобразование, предоставление, неразглашение и др.) [3].

В связи с тем, что все устанавливаемые в процессе активного взаимодействия информационных деятелей информационные связи и отношения в зависимости от их назначения делятся на *целевые* (являющиеся средством достижения конкретных целей и одновременно результатом определенной информационной деятельности) и *обеспечивающие* (являющиеся только средством достижения конкретных целей), соответствующие информационные правоотношения также можно разделить на два больших класса [3]:

1) *целевые* (предметные, отраслевые) — информационные правоотношения;

2) *обеспечивающие* — информационные правоотношения в иных сферах: экономической, политической, социальной, т. е. в сфере социального обеспечения и страхования, образования и культуры, охраны здоровья и др.; экологической, брачно-семейной, трудовой, нравственной и др.

Система правового регулирования представляет собой единое образование множества компонентов и связей, находящихся в сложных отношениях между собой, обладающее новыми свойствами структурной и функциональной целостности и синергизма.

По целевому предназначению *система правового регулирования* является сложной открытой, неравновесной (и в силу этого незавершённой) социальной информационно-кибернетической системой, обеспечивающей правовое регулирование общественных отношений, характеризующейся высокой степенью динамичности, неустойчивости и неопределённости [10, 12].

Базовыми взаимосвязанными функциональными компонентами (подсистемами) соответствующей системы правового регулирования (международной, национальной, федеральной, территориальной) являются следующие три (*рис.* 1): правовая система (правовые институты, правовая наука, формируемая на основе «теоретического правосознания» юристов; позитивное право); правосознание (индивидуальное, групповое и общественное); правоотношения.

Поскольку правосознание граждан, наряду с другими сферами сознания, является регулятором человеческого поведения, от уровня его развития зависит состояние правопорядка в обществе и в государстве [3]. Этим обусловлена актуальность важной *государственной задачи* формирования позитивного правосознания

Рис. 2. Классификация правовой информации

населения с целью обеспечения сознательного уважения и добросовестного отношения граждан к праву и закону, практического применения ими норм права при разрешении возникающих в обществе конфликтов и споров [11]. Эта задача решается посредством актов нормативного правового характера, правового воспитания и правового информирования населения. Решение данной задачи осуществляется в рамках формирующейся в настоящее время социально-правовой системы информационно-правового воспитания [1], предназначенной для юридического всеобуча в условиях постоянного возрастания потока правовой информации (*рис*. 2).

Подходы к содержанию и составу формализованных *требований* к НПА в цифровой форме не могут возникнуть как узкая инженерная задача, никак не привязанная к общемировому состоянию уровня развития информационных технологий [1—8, 16]. Поэтому Российская Федерация принимает активное участие в работе ИСО (Международная организация по стандартизации, англ. *ISO — International Organization for Standardization*), являющейся всемирным объединением национальных органов по стандартизации (организаций — членов ИСО). При этом Российская Федерация на текущий момент обладает набором НПА, стандартизирующих понятия предметной области в рамках электронных документов, что может рассматриваться как фундамент специфических *требований* к электронным документам именно в правовой области (в области правоотношений).

Подготовка международных стандартов, как правило, осуществляется в технических комитетах ИСО. Каждая организация — член ИСО имеет право быть представленной в тех технических комитетах, тематика которых представляет для нее интерес.

Предлагается обратить внимание на основополагающий документ в рамках рассматриваемого вопроса — Технический отчет ИСО/ТО 18492[4] был подготовлен подкомитетом *SC*3 по общим вопросам технического комитета ИСО ТС 171 — «Прикладные системы создания и хранения документов» (*Document management applications*).

Однако для первичного рассмотрения и погружения в данную область необходимо изучение предшествующей нормативной базы[5].

---

[4] ГОСТ Р 54989-2012/ISO/TR 18492:2005 Обеспечение долговременной сохранности электронных документов // Электронный фонд правовой и нормативно-технической информации. URL: https://docs.cntd.ru/document/1200096286 (дата обращения: 25.11.2022).

[5] ISO 12651:1999 Electronic imaging — Vocabulary // International Organization for Standardization. URL: https://www.iso.org/ru/standard/1885.html (дата обращения: 25.11.2022).

ISO 15489-1:2016 Information and documentation — Records management — Part 1: Concepts and principles // International Organization for Standardization. URL: https://www.iso.org/standard/62542.html (дата обращения: 25.11.2022).

ГОСТ Р ИСО 15489-1-2007. СИБИД. Управление документами. Общие требования // Электронный фонд правовой и нормативно-технической информации. URL: https://docs.cntd.ru/document/1200096286 (дата обращения: 25.11.2022).

ГОСТ Р ИСО 15489-1-2019. СИБИД. Информация и документация. Управление документами. // Электронный фонд правовой и нормативно-технической информации. URL: https://docs.cntd.ru/document/1200163564 (дата обращения: 25.11.2022).

Приказ федерального агентства по техническому регулированию и метрологии «Об утверждении национального стандарта Российской Федерации» от 26.03.2019 № 101-ст // Министерство экономического развития Российской Федерации. URL: https://docs.cntd.ru/document/554239100 (дата обращения: 25.11.2022).

ИСО/ТО 15489-2 Информация и документация — Управление документами — Часть 2: Руководство. ISO/TR 15489-2:2001 Information and documentation — Records management — Part 2: Guidelines // International Organization for Standardization. URL: https://www.iso.org/standard/35845.html (дата обращения: 25.11.2022).

Как и ИСО 15489-1, многие международные стандарты утверждены в Российской Федерации в рамках серии ГОСТ Р ИСО, что позволяет осуществить агрегацию единого понятийного аппарата и сформулировать единую терминологическую базу, не отрываясь от общего информационного поля всех спецификаций ИСО, представляющих собой единое *информационное и понятийное пространство,* обладающего наборами общих в рамках отдельных предметных областей онтологиями.

В рамках указанных источников определяется модель предметной области электронных документов в виде следующего тезауруса [1—8, 16]:

- *аутентичный электронный документ* (*authentic electronic document based information*) — информационный объект, точность, надёжность и целостность которого будет сохраняться с течением времени;
- *слабоструктурированная документоориентированная информация* или *документ* (*document-based information*) — существенная информация, которую имеет смысл/следует обрабатывать как единый объект;
- *контент документа* (*document-based information content*) — содержимое документа;
- *контекст документа* (*document-based information context*) — сведения об обстоятельствах создания, контроля, использования, хранения и управления электронным документом, а также сведения о его взаимосвязях с другими подобными документами;
- *структура документа* (*document-based information structure*) — атрибуты документа, определяющие порядок (например, иерархию распознаваемых подразделов), а также определяющие способ оформления или физические параметры отображения (тип и начертание шрифта, интервалы и др.);
- *электронное архивирование* (*electronic archiving*) — обеспечение сохранения информации в определённом обособленном «пространстве» (логическом или физическом), где эта информация защищена от несанкционированного внесения изменений или деградации;
- *долговременная сохранность* (*long-term preservation*) — определение периода времени и способов поддержки электронных документов в качестве доступного и аутентичного свидетельства (доказательства);
- *метаданные* (*metadata*) — данные, описывающие содержание, структуру, контекст документа (например, ключевые слова, сведения описывающие *требования* к структуре);
- *миграция* (*migration*) — любой процесс переноса документов из одной программно-аппаратной среды или носителя информации в другую среду без изменений (с сохранением целостности) и без каких-либо изменений в контенте и контексте;
- *средство хранения информации* (*storage repository*) — организация, подразделение (или сообщество), инженерное решение, на которое возложена ответственность за хранение и поддержание аутентичных электронных документов;
- *технологическое устаревание* (*technological obsolescence*) — вытеснение с рынка ИТ существующего (признанного) технологического решения вследствие развития технологий.

Подобная классификация позволяет уточнить, какого именно понятия касается та или иная часть отдельных *требований* или рассуждений. Это имеет особенное значение при рассмотрении вопросов устаревания технологии (*рис*. 3).

Моральное (технологическое) устаревание способов записи и чтения данных, используемых технологий, носителей, соглашений о кодировании, программных реализаций и др. является неизбежным этапом «жизненного цикла» любой информационно-технологической системы. Не существует ни одной технологии, которая бы могла существовать в неизменном виде вечно. А срок жизни инженерных решений является одним из определяющих в создании долговечных и надёжных систем. В части электронных документов вопросы сохранности и возможности повторного использования данных являются наиболее важными. В связи с этим стратегия обеспечения долговременного использования (сохранности) обязана целенаправленно обосновывать и решать проблему устаревания носителей, подходов и способов доступа к данным, к которой относятся и системы управления базами данных, способы кодирования информации и др. вопросы.

Важным фактором построения информационных систем, работающих длительный период времени, является применение «технологически нейтральных» форматов, предусматривающих возможность работы с документами для будущих пользователей. Под *«технологически нейтральными»* понимаются форматы, которые не накладывают ограничений на используемый носитель, программную реализацию, технологию записи или чтения, используемые операционные системы или любые другие параметры использования формата.

Решение поставленной задачи

Единообразное и основанное на общих соглашениях использование терминов позволяет гармонизированно интерпретировать общие понятия. При этом необходимо внимательно и отдельно рассмотреть такие понятия, как *контент*, *контекст* и *метаданные*.

Очень часто даже квалифицированными специалистами допускается ошибка в дефинициях «контекст» и «метаданные». Так, распространённым заблужде-

ГОСТ Р ИСО 23081-1-2008. СИБИД. Процессы управления документами. Метаданные для документов. Часть 1. Принципы // Электронный фонд правовой и нормативно-технической информации. URL: https://docs.cntd.ru/document/1200067603 (дата обращения: 25.11.2022).

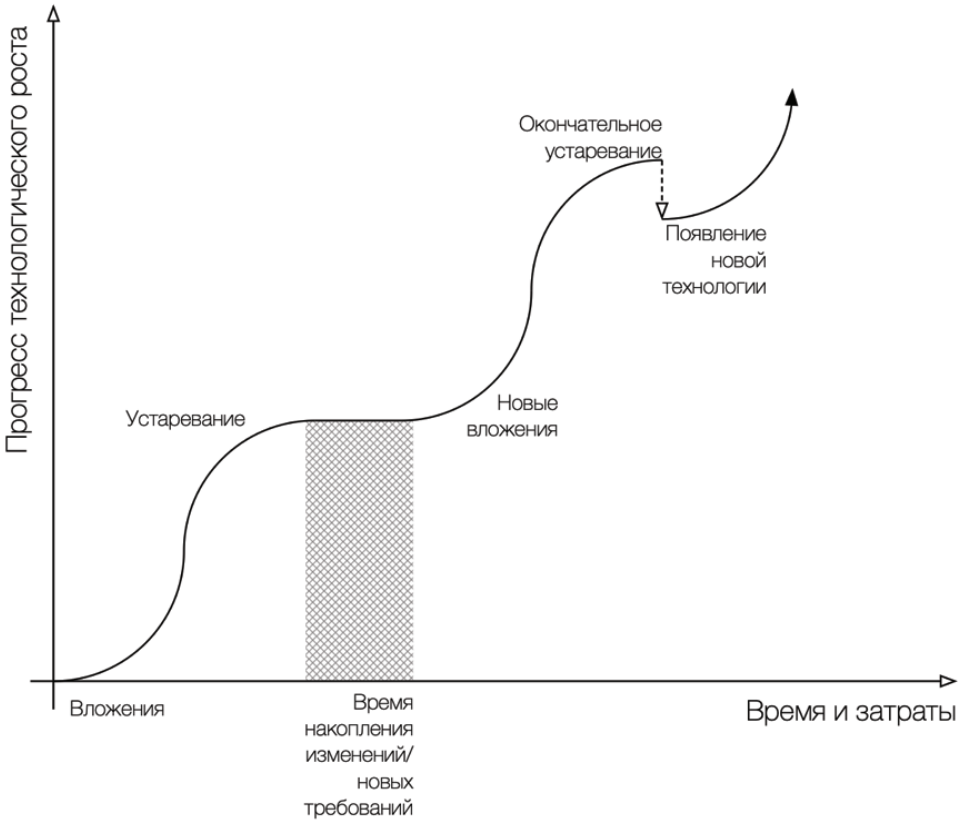

Рис. 3. Прогресс технологического устаревания технологий

нием является утверждение, при котором данные, представляющие информацию в документоведческой терминологии, сопоставляются с термином «реквизит документа», а для систем электронного документооборота — с метаданными [16]. Это весьма распространённая, но ошибочная практика.

По существу, реквизитам НПА на уровне международных спецификаций терминологически соответствует понятие «*контекст*». Метаданным же соответствуют сведения о структурах контекста и контента, например, способ кодирования идентификатора или даты в реквизитах.

Контекст может содержать как свой атрибутный состав, так и информационные структуры, описывающие, например, связи (электронных) документов между собой.

Метаданные как термин в самом общем виде могут интерпретироваться лишь как информация о самих информационных структурах (например, об особенностях их структур). Использование же термина «метаданные» для описания контекста не вполне корректно, хотя и довольно широко распространено.

Используемые в работе понятия применимы к любому направлению цифровой трансформации документооборота, но это представляется особенно актуальным, когда речь идёт о нормативных правовых актах.

Наиболее важным документом за рамками технологических международных спецификаций и документов, за исключением ГОСТ и ГОСТ Р, в части особенных понятий в области машиночитаемости права за последние несколько лет можно назвать «*Концепцию развития технологий машиночитаемого права*»[6], в которой предлагается цифровой стандарт, подсистема цифровой платформы, консолидирующая различные представления документа, содержащего *требования* (стандарта), в состав которого входят объекты, предназначенные как для человеко-ориентированного представления (классическое текстовое представление и карта требований), так и для программной обработки в составе систем (подсистем), решающих различные прикладные задачи (*рис*. 4) (специализированные прикладные карты/слои данных)[7].

Так, в Концепции приведена *классификация машиночитаемости* в соответствии со следующей иерархией:

1. Формирование *образов* (первичная оцифровка) документов. Определяется авторами как получение графических образов документов (например, путем сканирования или фотографи-

[6] Концепция развития технологий машиночитаемого права, утв. Правительственной комиссией по цифровому развитию, использованию информационных технологий для улучшения качества жизни и условий ведения предпринимательской деятельности, протокол от 15 сентября 2021 г. № 31. URL: https://www.economy. gov.ru/material/file/792d50ea6a6f3a9c75f95494c253ab99/31_15092021.pdf (дата обращения: 25.11.2022).

[7] Умные (SMART) стандарты: возможности и перспективы // Техэксперт. URL: https://cntd.ru/industriya-4-0.ru/actual (дата обращения: 25.11.2022).

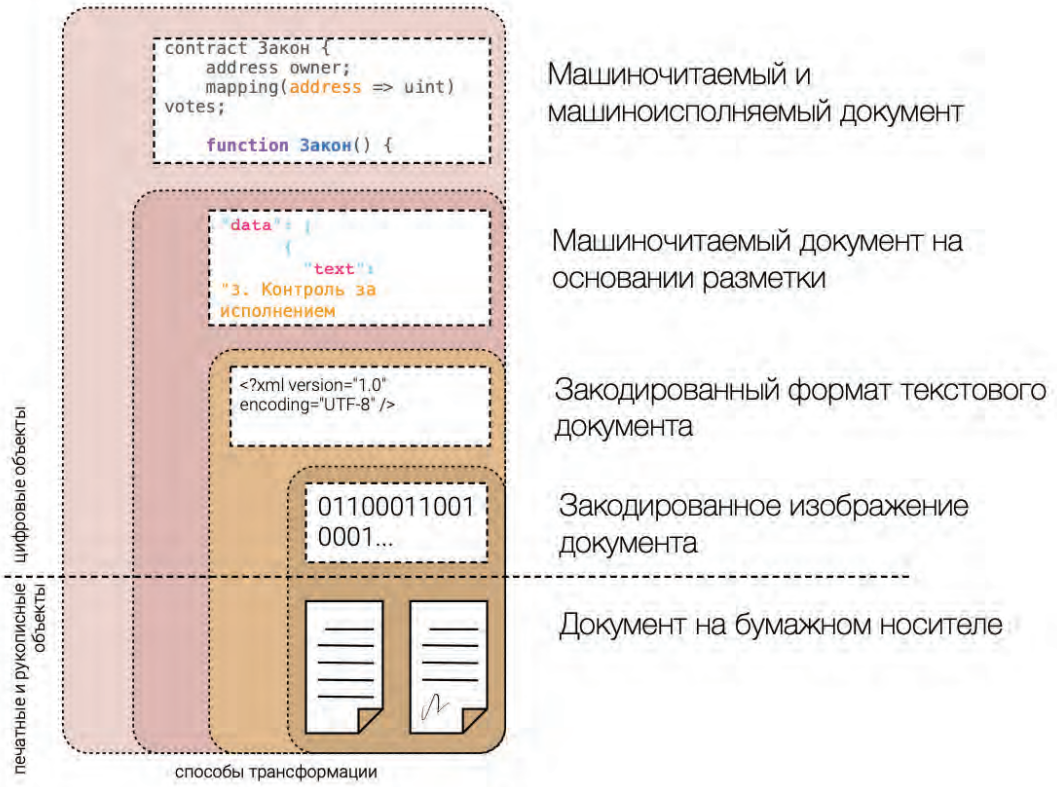

Рис. 4. Типы машиночитаемых документов

рования), их хранение и использование в виде изображений страниц.

2. Формирование *текстового цифрового документа* — создание полнотекстовых документов путём распознавания текста из образов документов либо путём создания или ввода текста в традиционных форматах текстовых редакторов, таких как *ODT*[8], *LaTeX, OOXML MS Word*[9], *Apple Pages, Bred, Emacs, Nano* и др., либо не использующих разметку, либо применяющих как открытые, так и проприетарные традиционные форматы визуальной разметки документа, задающей его структуру, шрифты, графические либо иные элементы, но не предназначенной для маркировки смысловых элементов для машинной обработки. При этом следует заметить, что конвертация между такими форматами, включая создание pdf- и xml-документов, сама по себе не может быть отнесена к технологиям машиночитаемого права.
3. Формирование *машиночитаемого документа*. Определяется как обособление смысловых элементов документа с помощью разметки, явным образом относящей содержание определенных в документе полей к отдельным типам данных.
4. Формирование *машиночитаемого/машиноисполняемого программного кода*. Представляет собой построение на базе онтологий и норм программного кода (в том числе самоисполняемых контрактов [14]), который автоматизирует совершение действий, необходимых для реализации прав и (или) предотвращает случайные или намеренные ошибки в процессе такой реализации.

Данный подход к классификации вполне обоснован, однако однозначно требует уточнений в части формулировок и определений с инженерной точки зрения. В частности, необходимо учесть следующие особенности.

1. Большое количество редакций одних и тех же правовых актов (часто до нескольких сотен редакций на один НПА), по сути, представляющие собой исторические версии одного и того же акта, которые тем не менее требуется использовать одновременно, например, для определения, какая норма права действовала на определённую дату.
2. Процесс принятия НПА и процесс его опубликования, предполагающий работу сразу с несколькими несовпадающими наборами *требований* к структурам правовых актов.

[8] ГОСТ Р ИСО/МЭК 26300-2010. Информационная технология. Формат Open Document для офисных приложений (OpenDocument) v1.0 // Электронный фонд правовой и нормативно-технической информации. URL: https://docs.cntd.ru/document/1200083207 (дата обращения: 25.11.2022).

[9] ISO/IEC 29500-1:2016 Information technology — Document description and processing languages — Office Open XML File Formats — Part 1: Fundamentals and Markup Language Reference // International Organization for Standardization. URL: https://www.iso.org/standard/71691.html (дата обращения: 25.11.2022).

3. Наличие «изменяющих» НПА и логика получения актуальных редакций. Подавляющее большинство НПА сейчас опубликовывается в формате «изменяющих» документов. Это актуально для всех существующих НПА, в которые вносятся изменения. На федеральном уровне таких НПА большинство. Один изменяющий НПА может вносить изменения сразу в несколько других НПА, причём определять постепенное вступление отдельных положений в силу. Как следствие, сам изменяющий документ в электронной форме не представляет интереса для анализа, скажем, средствами машинного обучения, поскольку содержит не сами нормы права, а семантические конструкции, изменяющие нормы права в других НПА.
4. Наличие феномена формулировки нормы права в нескольких предложениях либо нескольких норм права в рамках одного предложения [1, 12]. Грамматическое предложение и нормативно-правовое предписание — не одно и то же. Безусловно, решение вопроса определённости содержания диспозиции правовых норм (право, обязанность или запрет) изначально основано на применении правил русского языка. Однако формулировка нормы права может быть сформулирована во всей полноте в рамках нескольких предложений.
5. Необходимость сохранения сведений о прочих НПА, даже если подобные НПА не содержатся или по каким-то причинам не могут содержаться в текущем технологическом решении. Например, ссылка на документ ограниченного распространения (или структурный элемент) обязана быть ссылкой с корректной атрибуцией даже при условии, что подобного документа в текущей базе данных быть не может.
6. Формализованные *требования* к идентификации структурных элементов, в т. ч. к возможным принципам нумерации структурных элементов, включая неявные, исключающие типовые подходы к нумерации, принятые в большинстве обычных текстовых редакторов.

Рассмотрим для наглядности *простой пример*: внесение в существующий нормативный правовой акт изменений другим нормативным правовым актом (*рис*. 5).

Даже на этом простом примере видно, что получающиеся цифровые объекты актуальной (новой) редакции НПА никогда не существуют в виде отдельного подписываемого документа (за исключением существования в технологическом виде в самом нормотворческом процессе). Однако они обязаны существовать, поскольку машино- и человекочитаемость производных электронных «квазидокументов» актуальных (новых) редакций нормативных правовых актов так же важна, как и первичного нормативного правового акта.

Всё вышеперечисленное приводит к тому, что классификация, представленная в *«Концепции развития технологий машиночитаемого права»*, пригодна не только для анализа принимаемых и подписываемых электронных документов НПА, но и для многочисленных редакций этих НПА, которые в большинстве случаев являются аутентичными «квазидокументами», чаще всего — производными от применения изменяющих документов.

Это важное отличие, заключающееся в том, что первичные аутентичные электронные документы также имеют место быть в рамках цифровой трансформации. Причем наибольший интерес представляют не они сами, а получающиеся в результате применения изменений новые редакции изменяемых НПА. Следовательно, задача классификации форматов машиночитаемых НПА (и их редакций) имеет сугубо практическое назначение, например, для определения *требований* к форматам обмена данными и/или требованиям к построению информационных систем.

Однозначно можно выделить *двойственную* природу электронных репрезентаций правовых актов и редакций правовых актов, представляющую собой сочетание таких информационных объектов, как:

- идентифицирующие сведения о правовом акте;
- контекст НПА (связи с другими НПА, сведения об опубликовании и др.);
- контент отдельных редакций НПА;
- контекст редакций отдельных НПА (начало и окончание сроков действия; информация о том, на основании каких НПА были внесены изменения в структурные элементы и др.).

В силу этого один отдельный НПА едва ли можно представить универсальным образом как один информационный объект, кроме самых тривиальных случаев. Скорее, НПА можно представить как *сложную информационную структуру*, содержащую целый набор различных сведений, в совокупности представляющих собой исчерпывающие сведения о конкретном НПА.

Поэтому *«Концепция развития технологий машиночитаемого права»* требует уточнения в той части, которая касается описания *требования* к форматам *контента* редакций НПА. Почему именно контента? Потому что преимущественно именно эти данные имеет смысл анализировать с точки зрения анализа текстов.

Функциональные ограничения в рамках классификации в Концепции смешаны с инженерными спецификациями и определениями. Например, относится ли инженерная система, оперирующая образами правовых актов в формате *PDF/A* с текстовым слоем, к уровню машиночитаемости 1-го уровня? Или она относится всё же ко 2-му уровню? А система, хранящая сведения об НПА в формате *XML* — она относится к 2-му или к 3-му (или даже 4-му) уровню? Очевидно, что упоминание конкретного формата не гарантирует отнесение инженерной системы и формата к определённому уровню.

Для устранения этих неопределённостей *предлагается* описание классификации из Концепции, уточнённое в инженерном смысле. При этом речь идёт

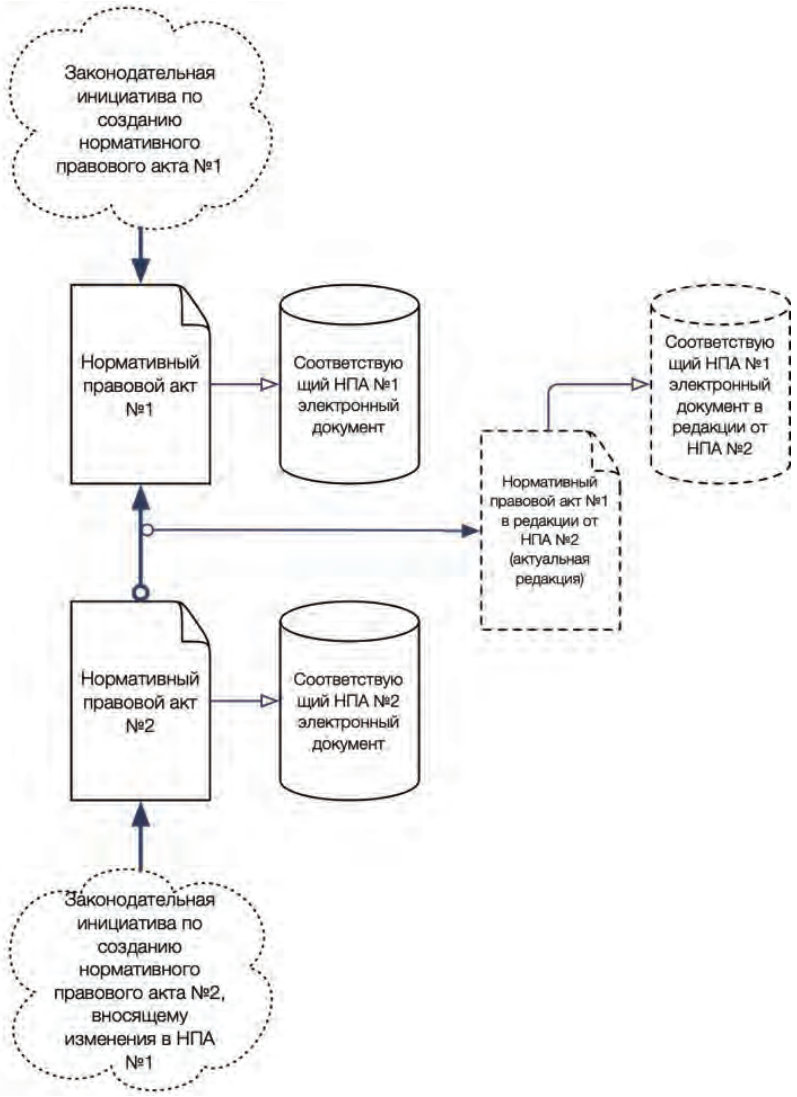

Рис. 5. Внесение изменений в нормативный правовой акт

в бóльшей степени о классификации *уровня машиночитаемости форматов аутентичных электронных документов НПА*, чем о технологиях обработки машиночитаемых НПА или применении машиночитаемых НПА для решения конкретных задач.

В соответствии с принятой Концепцией предлагается 4 основных *уровня* машиночитаемых форматов правовых актов, определяемых по способу определения контента аутентичных электронных документов НПА: 1 — графического представления; 2 — документального представления; 3 — машиночитаемого документа; 4 — искусственного языкового представления.

Уровень графического представления

Включает форматы *контента* аутентичных электронных документов, предназначенные для хранения растровых и векторных графических изображений печатных форм НПА. Чаще всего является форматом для хранения изображений с определяемыми заранее соглашениями по управлению умолчаниями (способе определения цветовых пространств, шрифтов и права на их использование, цветовых моделей). Такие форматы используются при сканировании, отправке факсов, распознавании текста, в полиграфии, широко поддерживаются графическими приложениями. К данным форматам относятся все форматы хранения растрового и векторного изображения, а также контейнеры для них, такие как *PDF/A* (*Portable Document Format/A*).

Машиночитаемый аутентичный электронный документ-редакция НПА, представленный как объект графического представления, может эффективно существовать либо как цифровой объект архивного хранения, либо как изображение в цифровой форме, пригодное для копирования, печати, передачи по каналам связи.

Исторически так сложилось, что наиболее распространенный в Российской Федерации формат для хранения растрового изображения документа НПА — битовое поле в формате *TIFF*. Этот формат, как и многие другие «технологически нейтральные» форматы хранения битовых полей, позволяет определить, как должно

выглядеть изображение НПА, если мы рассматриваем его как «рисунок». «Нейтральное» соглашение (в рамках формата) о способе кодирования в этом случае является определяющим для использования будущими пользователями. Так, «технологическая нейтральность» битового поля обеспечивается кодированием заголовка в специальном формате *TIFF*, что позволяет позже использовать его уже формализованным образом.

Функционально не существует какого-либо различия в конкретном способе кодирования растровой информации для применения в форматах аутентичных электронных документов НПА. Какой бы ни был выбран способ кодирования, машиночитаемость этого уровня во всех случаях преследует только одну цель — перевод в первоначальную цифровую форму существующих на бумаге НПА с целью передачи их по каналам связи и/или тиражирования.

Наиболее распространённый формат архивного хранения при этом может опираться и на более структурно-сложные форматы, например, основанные на *PostScript* или *PDF*. *PDF/A* — стандарт *ISO* 19005-1:2005[10] — основан на описании стандарта *PDF* версии 1.4 от *Adobe Systems Inc*. Несмотря на то, что формат изначально разрабатывался как коммерческий, впоследствии он был стандартизирован как «технологически нейтральный» на уровне международной спецификации. Этот формат активно развивается и имеет подмножества в виде *PDF/A*-1*a* и *PDF/A*-1*b*.

Файлы в формате *PDF/A* самодостаточны. Вся информация, необходимая для того, чтобы каждый раз отображать документ в неизменном виде, внедрена в файл/цифровой объект. Формат *PDF/A* на уровне соглашений исключает использование информации из внешних источников (код, шрифтовые программы или гиперссылки).

Другие ключевые элементы для соответствия *требованиям PDF/A* включают:

- запрещён аудио- и видеоконтент;
- запрещено внедрение кода на *JavaScript* и команд на запуск исполняемых файлов;
- все используемые шрифты должны быть внедрены, а также легально доступны для неограниченного универсального отображения;
- цветовые пространства определяются независимым от устройства способом;
- внутреннее шифрование отдельных объектов не разрешено;
- предписывается использование основанных на стандартах метаданных.

Существует подуровень графического представления в виде форматов допечатной подготовки (как правило, на базе *PostScript*, *PDF/A*) с отдельным «текстовым слоем». Несущественная разница с полностью графическими форматами заключается в том, что в ряде случаев это несколько облегчает извлечение текстовой информации, сокращая издержки на задачи *OCR* (оптическое распознавание символов).

Несмотря на то, что *PDF/A* как формат кодирования имеет возможность существования в виде отдельного «текстового слоя», он не может рассматриваться на документальном уровне, поскольку предполагает ограниченную функциональную выгоду в связи с неуправляемым способом размещения символов, широким распространением в документах таблиц, формул, метаинформации и др., что невозможно корректно извлечь из «текстового слоя».

Здесь также важно указать, что соответствие *требованиям PDF/A* для форматов аутентичных электронных документов включает требования, но не может ограничить составителя подобного документа включить не входящие в спецификацию данные. Подобное отношение к спецификации предполагает, что будущие пользователи могут корректно интерпретировать только те данные, которые относятся к спецификации, но будут вынуждены игнорировать те данные, значение которых не стандартизировано.

Основная задача форматов уровня машиночитаемости графического уровня представления НПА — сохранение оригинального облика документа, максимально близкого к той первоначальной печатной форме, в которой этот документ был подписан и оформлен как печатный документ.

Основной *недостаток* — невозможность сколько-нибудь «продвинутых» машинных операций, кроме самых примитивных (базовых), таких как копирование, отображение, передача по каналам связи, вывод на печать. Данный уровень машиночитаемости идеально подходит для задач архивного хранения аутентичных электронных документов и хранения эталонных печатных форм как структурных элементов редакций правовых актов.

Уровень документального представления

В связи с широким характером применения офисных пакетов производства текстов для подготовки правовых актов применение форматов офисных пакетов может рассматриваться как отдельный уровень/слой машиночитаемых НПА.

Офисные форматы *Office Open XML* (*OOXML*, *DOCX*, *XLSX*, *PPTX*, приняты как международная спецификация) или *OpenDocumentFormat* (*ODF*) — *OASIS Open Document Format for Office Application* (1 мая 2006 г. принят как международный стандарт) обладают рядом существенных преимуществ в виде универсальности, широкого распространения, формализованности и поддержки в виде ПО, способного работать с подобными форматами данных.

Оба формата очень похожи с инженерной точки зрения — оба являются контейнерами *ZIP*, объединя-

[10] ISO 19005-1:2005. Document management — Electronic document file format for long-term preservation — Part 1 : Use of PDF 1.4 (PDF/A-1) // International Organization for Standardization. URL: https://www.iso.org/standard/38920.html (дата обращения: 25.11.2022)

ющими файловую структуру, содержащую *XML* файлы с содержимым, манифесты (метаданные) и вложения.

Форматы обладают следующими недостатками для применения в качестве формата аутентичных электронных документов НПА:

- невозможность проведения формальной верификации и валидации документа НПА;
- избыточность по ширине формата/оформлению;
- невозможность применить *требования* к структурным элементам;
- невозможность сохранения вспомогательных информационных слоев или контекста аутентичного электронного документа НПА;
- невозможность определения своих метаданных.

Машиночитаемый документальный уровень удобен для работы над небольшими и обособленными НПА, например, такими как приказы, распоряжения, прочие правовые акты (см. *рис*. 2), не связанные с другими актами и не требующие последующего автоматического анализа или извлечения фактов из тела документа или другой автоматической обработки.

Основное *достоинство* документального уровня машиночитаемости — возможность использования открытых и распространённых офисных форматов хранения.

Основной *недостаток* — в невозможности формальной проверки корректности правового акта, в отсутствии возможности формализованно хранить контекст, метаданные (сведения о соглашениях и структурах). То есть такой уровень формата можно считать малопригодным к процедуре формальной проверки данных на соответствие модели предметной области. Это делает использование подобных форматов ограниченно пригодным в рамках машиночитаемости, поскольку ширина инженерных решений по поддержанию всего многообразия возможностей делает применение этих форматов «не нейтральным».

Уровень машиночитаемых документов

Перевод цифрового объекта, представляющего аутентичный электронный документ НПА, в форму, пригодную для передачи по каналам связи или сохранения, называется *сериализацией*, а обратный процесс — *десериализацией*.

Способ сериализации (процесс перевода структуры данных в байтовую последовательность) цифровых объектов может предполагать использование различных соглашений или спецификаций. Структура данных может быть переведена в последовательность байтов, в строковые представления (тоже последовательность байтов, но уже в рамках соглашений о кодировании), в более высокоуровневые форматы соглашений о кодировании (байтовых последовательностей) *JSON, XML, CBOR, ANS.1*, *MessagePack* и ещё несколько десятков других «технологически нейтральных» спецификаций и большое количество частных реализаций.

С одной стороны, способ кодирования информации имеет важное значение, поскольку обязан учитывать порядок следования байтов при определении типов данных, последовательность следования бит в байтах, способ кодирования строк, чисел, дат, имён и др. Это имеет решающее значение, поскольку программно-аппаратные платформы, на которых будет вестись обработка данных, в момент создания самогó сериализованного цифрового объекта неизвестны. Поэтому с точки зрения «технологической нейтральности» и понятий архивного использования подобные форматы обязаны опираться на зрелые и точно формализованные решения, лучше всего — на международные, широко используемые спецификации с длительным жизненным циклом и существующие в рамках деятельности международного института по стандартизации.

С другой стороны, почти любой зрелый формат сериализации позволяет осуществлять прозрачное преобразование в другие форматы с использованием большого количества инструментальных средств разработки и языков программирования. В силу этого принципиальной разницы в выборе конкретной технологии сериализации с этой точки зрения нет (среди проработанных решений), если эта технология инженерно обоснована с точки зрения «технологической нейтральности». Подавляющее большинство зрелых технологических стандартов сериализации «интероперабельны» между собой. И хотя все они и имеют друг перед другом некоторые преимущества (некоторые более компактны, другие легче читаются и др.), практически все они могут позволить осуществлять трансформацию машиночитаемых данных друг в друга без потерь, чаще всего в автоматическом режиме.

Не следует путать понятие сериализации/десериализации и качество «схемы данных». Сериализованные машиночитаемые данные могут существенно отличаться друг от друга по качеству и пригодности к машинному анализу в зависимости от качества «схемы данных». Чем глубже НПА проработан как информационный объект, тем больше с ним можно выполнить машинных операций.

Все подобные «технологически нейтральные» машиночитаемые форматы НПА определяются строгими *требованиями* к структуре данных, которые в обязательном порядке ***должны*** иметь возможность строгого формального контроля, *во-первых*, на соответствие самому формату кодирования, а, *во-вторых*, на формальное соответствие закодированных данных установленной ранее ***схеме данных*** на основании таких технологий, как *XML Schema Definition, JSON Schema* и др. в зависимости от способа реализации.

Схема возможных данных для применения в предметной области является определяющим показателем. Чем точнее она соответствует решаемым задачам, тем более «машиночитаемым» является такое представление данных. Так, для НПА на минимальном уровне машиночитаемости следует выделять как минимум все виды структурных элементов НПА (часть, пункт,

подпункт, статья, параграф и др.), учитывать способы кодирования структурных элементов, способы кодирования ссылок на НПА, определять способы хранения и представления штампов, подписей, реквизитов и всего остального атрибутного состава НПА, сведения об изменяющих документах, способы хранения формул, изображений, карт, эталонных форм (шаблонов) документов, определять, на каких методических рекомендациях основана структура документа, является ли он подлинным (аутентичным) или «квазидокументом» (полученным путём сборки из изменяющего документа НПА), и так далее.

Это связано с тем, что с этого уровня представления данных возможно оперирование типами данных языков программирования, и аутентичный электронный документ представлен в программе не как большой строковый литерал, а как информационное «дерево» объектов с удобным доступом к отдельным структурным элементам контента и контекста. Использование «технологически нейтрального» верифицируемого формата позволяет проверить, соответствует ли структура документа *требованиям*, строить модели расширенного контекста, активно применять модели валидации и цифрового преобразования.

Формальный контроль при этом может осуществляться на основании описания структуры данных без каких-либо сведений о предметной области, автономно и не требуя в обязательном порядке ПО, которое умеет работать с этим типом данных. Так, формальное определение *требований* к структуре НПА может быть проверено теперь программными способами без каких-либо сложных эвристик, на основании только описанной заранее схемы.

Машиночитаемые форматы позволяют гарантированно определять структурные элементы НПА, вносимые в них изменения, производить декомпозицию НПА на соответствующие логике изменяющих документов элементы. Машиночитаемые документы в форматах уровня ***должны*** позволять осуществлять анализ НПА без зависимости от какого-то специфического ПО и ***должны*** определять подобный формат полно. То есть подобные форматы ***не должны*** позволять существовать избыточно широким, неточным или двусмысленным определениям. Любой элемент структуры данных НПА ***должен*** иметь однозначную формулировку, подкрепленную соответствующей схемой проверки. Достижение подобного уровня машиночитаемости позволяет без применения сложных технологий решать задачи производства изменяющих документов и производства актуальных редакций из изменяющих документов в автоматическом режиме.

Принцип хранения данных при использовании офисных форматов данных предполагает сериализацию данных в *XML* форматы. Вопреки расхожему мнению, хранение сведений в формате *XML* не является автоматическим решением проблем упорядочивания данных. Различие хранений данных в специализированной разметке и в документоориентированных форматах — в том, что схемой данных в документоориентированных форматах фактически является заранее заданная структура в *OpenXML* или *OpenDocument Format,* определяющая способ размещения данных на странице для печати, а не нормативная предметная область. То есть сама спецификация *OpenXML* или *ODF* является негативным фактором, влияющим на качество машиночитаемости. Частные машиночитаемые форматы данных отличаются тем, что здесь схемой данных является сама модель предметной области НПА, что позволяет программам воспринимать сериализованные данные максимально точно относительно предметной области, а не способа отображения информации на листе бумаги.

Так, машиночитаемые форматы документов подобного уровня могут расширять *требования* к элементам данных, содержащим сведения о грамматической структуре предложений, выделением частей предложения, определением форматов для редакций НПА, полученных путём применения изменяющих документов.

Такой уровень формализации машиночитаемых документов, который позволяет взаимодействовать с банком НПА как с единым информационным полем, отличается тем, что ***должен*** позволять связывать разные редакции машиночитаемых НПА в единую информационную модель. Машиночитаемые документы такого уровня могут вносить формальные *требования* для поддержки произвольных «информационных слоёв», выделяющими структурно интересующие сущности, — например, названия организаций, должности, регионы и др.

Дальнейшее развитие цифровой схемы данных отличается тем, что их машиночитаемые форматы позволяют определять структурно формулировки норм права. Форматы на этом уровне ***должны*** позволять однозначно определять (находить, выделять из текста документа) норму права, даже если она сформулирована в нескольких разделённых грамматических конструкциях, находящихся в разных структурных элементах НПА.

В этой связи основной цифровой объект такого формата данных — норма права. Подобный объект ***должен*** быть определён на основании «нейтральных» форматов, а способ кодирования данных должен позволять извлекать нормы права без применения специализированных программных средств, средствами самого формата хранения.

При этом способы кодирования и представления должны позволять определять произвольно сложные (возможные) структуры соответствующих цифровых объектов, соответствующих нормам права.

Уровень искусственных языковых представлений

Машиночитаемые НПА на этом уровне являются исполняемыми программами, написанными на формально строгих специализированных искусственных языках программирования со всеми свойственными

языкам программирования особенностями: полнотой, синтаксическими *требованиями*, возможностью статического анализа и др.

Машиночитаемые на этом уровне представления НПА нужны для автоматизированного или автоматического применения норм права, решения задач машинопроектируемости НПА и других перспективных задач.

Заключение

В настоящий момент в Российской Федерации не существует сколько-нибудь обоснованного технологического стандарта или проекта в высокой степени готовности на уровне непосредственного программирования машиночитаемых/машиноисполняемых правовых актов. Вероятнее всего, подобные технологии могут быть основаны на базе формально-строгих языков *DSL* (*domain specific languages*), оперирующих онтологиями, широко выходящими за рамки одного лишь права. Специалистами широко обсуждается возможность применения технологий *распределённого реестра* [15], который мог бы быть полезен для построения доверенной среды программируемых машиночитаемых правовых актов. Тем не менее само применение технологий распределённого реестра не является универсальным средством для решения задач машиночитаемости и машиноисполняемости, не снимает всех вопросов формализации перевода правовых норм в цифровую логику.

Рецензент: ***Запольский Сергей Васильевич,*** *доктор юридических наук, профессор, заслуженный юрист Российской Федерации, главный научный сотрудник Института государства и права Российской академии наук, г. Москва, Российская Федерация.*
E-mail: zpmoscow@mail.ru

Литература

1. Ершов В.В. Правовое и индивидуальное регулирование общественных отношений : монография. М. : РГУП, 2018. 628 с.
2. Ершов В.В. Регулирование правоотношений : монография. М. : РГУП, 2020. 564 с.
3. Ловцов Д.А. Системология правового регулирования информационных отношений в инфосфере : монография. М. : РГУП, 2016. 316 с. ISBN 978-5-93916-505-1.
4. Ловцов Д.А. Системология научных исследований. М. : НЦПИ при Минюсте РФ, 2018. 76 с. ISBN 978-5-90167-53-3.
5. Ловцов Д.А. Информационная теория эргасистем. Тезаурус : монография. М. : Наука, 2005, 248 с. ISBN 5-02-033779-X.
6. Ловцов Д.А. О концепции комплексного подхода // Философские исследования. 2000. № 4. С. 158—174.
7. Ловцов Д.А. Концепция комплексного «ИКС»-подхода к исследованию сложных правозначимых явлений как систем // Философия права. 2009. № 5. С. 40—45.
8. Ловцов Д.А., Ермаков И. В. Классификация и модели нетрадиционных информационных каналов в эргасистеме // НТИ РАН. Сер. 2. Информ. процессы и системы. 2005. № 2. С. 1—7.
9. Ловцов Д.А., Богданова М.В., Лобан А.В., Паршинцева Л.С. Статистика (компьютеризированный курс) / Под ред. Д.А. Ловцова. М. : РГУП, 2020. 400 с. ISBN 978-5-93916-834-2.
10. Поддубный Н. В. Самоорганизующиеся системы: онтологический и методологический аспекты : автореф. дис. … д-ра филос. наук: 09.00.01. Ростов н/Д., 2000.
11. Теория государства и права / Под ред. В.М. Корельского, В.Д. Перевалова. М. : Норма, 2002. 616 с.
12. Трапезникова Е.В. Лингвистическая конструкция диспозиции нормы права // Юридическая наука и практика: Вестник Нижегородской академии МВД России. 2021. № 1 (53). С. 251—257.
13. Федосеев С.В. Применение современных технологий больших данных в правовой сфере // Правовая информатика. 2018. № 4. С. 50—58. DOI: 10.21681/1994-1404-2018-4-50-58 .
14. Федосеев С.В. Информационные и программные аспекты разработки и применения смарт-контрактов // Правовая информатика. 2021. № 3. С. 27—35. DOI: 10.21681/1994-1404-2021-3-27-35 .
15. Черных А.М. Технологии распределённого реестра и защита документооборота судебной системы // Правовая информатика. 2020. № 4. С. 20—28. DOI: 10.21681/1994-1404-2020-4-20-28 .
16. Юмашева Ю.Ю. От машинно-ориентированных к электронным документам: эволюция объектов и терминов // Документ. Архив. История. Современность. 2017. № 17. С. 417—445.

CLASSIFICATION OF DATA FORMATS FOR ELECTRONIC DOCUMENTS FOR SERIALISATION OF LEGAL REGULATIONS IN MACHINE-READABLE FORM

Stanislav Lakhtin, *Head of Project Group of the Committee for the Development of a Secure Environment in the Field of Digital Technologies, Moscow, Russian Federation.*
E-mail: e8686@yandex.ru

Vladimir Tsimbal, *Dr.Sc. (Technology), Professor, Honoured Scientist of the Russian Federation, Leading Researcher at the Institute of Engineering Physics, Serpukhov, Russian Federation.*
E-mail: tsimbal@mail.ru

Andrei Amelenkov, *Ph.D. (Economics), Associate Professor, Head of Division of the Moscow State Linguistic University, Moscow, Russian Federation.*
E-mail: amelioncov@linguanet.ru

Keywords: *legal informatisation, digital transformation of the legal field, machine readability of law, legal regulation, levels, requirements, serialisation.*

Abstract

Purpose of the paper: creating the research and methodological basis for the concept of developing machine-readable law technologies.

Methods used: system analysis, classification, clusterisation, formal grammars.

Study findings: a justification is given for a classification of machine-readability levels. Technological requirements for formats for digital objects serialisation, presentation of legal standards, documents and regulations are determined. A justification is given for following conclusions: the machine-readability of law depends of the level of information technologies development and cannot go beyond their boundaries, and the technologies of direct programming of machine-readable/machine-executable legal regulations can probably be based on formally strict DSL (domain specific) languages using ontologies that go far beyond the boundaries of law alone.

References

1. Ershov V.V. Pravovoe i individual'noe regulirovanie obshchestvennykh otnoshenii : monografiia. M. : RGUP, 2018. 628 pp.
2. Ershov V.V. Regulirovanie pravootnoshenii : monografiia. M. : RGUP, 2020. 564 pp.
3. Lovtsov D.A. Sistemologiia pravovogo regulirovaniia informatsionnykh otnoshenii v infosfere : monografiia. M. : RGUP, 2016. 316 pp. ISBN 978-5-93916-505-1.
4. Lovtsov D.A. Sistemologiia nauchnykh issledovanii. M. : NTsPI pri Miniuste RF, 2018. 76 pp. ISBN 978-5-90167-53-3.
5. Lovtsov D.A. Informatsionnaia teoriia ergasistem. Tezaurus : monografiia. M. : Nauka, 2005, 248 pp. ISBN 5-02-033779-X.
6. Lovtsov D.A. O kontseptsii kompleksnogo podkhoda. Filosofskie issledovaniia, 2000, No. 4, pp. 158–174.
7. Lovtsov D.A. Kontseptsiia kompleksnogo "IKS"-podkhoda k issledovaniiu slozhnykh pravoznachimykh iavlenii kak sistem. Filosofiia prava, 2009, No. 5, pp. 40–45.
8. Lovtsov D.A., Ermakov I.V. Klassifikatsiia i modeli netraditsionnykh informatsionnykh kanalov v ergasisteme. NTI RAN. Ser. 2. Inform. protsessy i sistemy, 2005, No. 2, pp. 1–7.
9. Lovtsov D.A., Bogdanova M.V., Loban A.V., Parshintseva L.S. Statistika (komp'iuterizirovannyi kurs). Pod red. D.A. Lovtsova. M. : RGUP, 2020. 400 pp. ISBN 978-5-93916-834-2.
10. Poddubnyi N.V. Samoorganizuiushchiesia sistemy: ontologicheskii i metodologicheskii aspekty : avtoref. dis. … d-ra filos. nauk: 09.00.01. Rostov n/D., 2000.
11. Teoriia gosudarstva i prava. Pod red. V.M. Korel'skogo, V.D. Perevalova. M. : Norma, 2002. 616 pp.
12. Trapeznikova E.V. Lingvisticheskaia konstruktsiia dispozitsii normy prava. Iuridicheskaia nauka i praktika: Vestnik Nizhegorodskoi akademii MVD Rossii, 2021, No. 1 (53), pp. 251–257.
13. Fedoseev S.V. Primenenie sovremennykh tekhnologii bol'shikh dannykh v pravovoi sfere. Pravovaia informatika, 2018, No. 4, pp. 50–58. DOI: 10.21681/1994-1404-2018-4-50-58 .
14. Fedoseev S.V. Informatsionnye i programmnye aspekty razrabotki i primeneniia smart-kontraktov. Pravovaia informatika, 2021, No. 3, pp. 27–35. DOI: 10.21681/1994-1404-2021-3-27-35 .

15. Chernykh A.M. Tekhnologii raspredelennogo reestra i zashchita dokumentooborota sudebnoi sistemy. Pravovaia informatika, 2020, No. 4, pp. 20–28. DOI: 10.21681/1994-1404-2020-4-20-28 .
16. Iumasheva Iu.Iu. Ot mashinno-orientirovannykh k elektronnym dokumentam: evoliutsiia ob"ektov i terminov. Dokument. Arkhiv. Istoriia. Sovremennost', 2017, No. 17, pp. 417–445.