

Онтолого-ориентированная информационная система верификации формализованных документов

И. Ю. Балашова, Е. А. Дзюба, Н. А. Попова

Пензенский государственный университет
440026, Россия, г. Пенза, ул. Красная, 40

Аннотация. В работе поднимается проблема повышения эффективности обработки документов, подаваемых в налоговый орган при государственной регистрации предпринимательской деятельности, путем сокращения времени обработки документов за счет автоматизации процесса верификации. Актуальность исследования обусловлена значительным количеством рутинных операций, выполняемых специалистами налогового органа вручную, а также большим объемом входящих документов. В статье рассмотрен процесс обработки документов, подаваемых в налоговый орган при регистрации различных форм для ведения предпринимательской деятельности. На основе проведенного исследования обоснована целесообразность разработки программных средств автоматизированной верификации документов. Выделены особенности данной задачи, требующие использования онтологического подхода к представлению данных. Описана онтология формализованного документа и правил его верификации. Предложен алгоритм верификации документов в рамках построенной онтологической модели. Описана архитектура информационной системы, включающая метаданные, сервер приложения и приложение пользователя. Уровень метаданных представлен совокупностью онтологий, построенных на основе разработанной онтологической модели. Разработана информационная система, обеспечивающая поддержку онтологий, а также автоматизированную верификацию документов. Представлены результаты автоматизированной обработки документов с использованием разработанной системы, подтверждающие снижение временных затрат на верификацию документов.

Ключевые слова: верификация, онтология, семантическая интеграция данных, обработка документов

Поступила 16.01.2024, одобрена после рецензирования 07.02.2024, принята к публикации 09.02.2024

Для цитирования. Балашова И. Ю., Дзюба Е. А., Попова Н. А. Онтолого-ориентированная информационная система верификации формализованных документов // Известия Кабардино-Балкарского научного центра РАН. 2024. Т. 26. № 1. С. 11–20. DOI: 10.35330/1991-6639-2024-26-1-11-20

MSC: 68U35

Original article

Ontology-oriented information system for verification of formalized documents

I.Yu. Balashova, E.A. Dzyuba, N.A. Popova

Penza State University
440026, Russia, Penza, 40 Krasnaya street

Abstract. The paper raises the problem of improving the efficiency of processing documents submitted to the tax authority during the state registration of entrepreneurial activity by reducing the time of document processing through the automation of the verification process. The relevance of the study is due to a

significant number of routine operations performed manually by specialists of the tax authority, as well as a large volume of income documents. The article considers the process of processing documents submitted to the tax authority when registering various forms for business activities. On the basis of the conducted research the expediency of development of software tools for automated verification of documents is substantiated. The peculiarities of this task requiring the use of ontological approach to data representation are highlighted. The ontology of a formalized document and the rules of its verification are described. An algorithm of document verification within the framework of the constructed ontological model is proposed. The architecture of information system including metadata, application server and user application is described. The metadata layer is represented by a set of ontologies built on the basis of the developed ontology model. An information system providing support for ontologies as well as automated document verification is developed. The results of automated document processing using the developed system are presented, confirming the reduction of time costs for verification of documents.

Keywords: verification, ontology, semantic data integration, document processing

Submitted 16.01.2024,

approved after reviewing 07.02.2024,

accepted for publication 09.02.2024

For citation. Balashova I.Yu., Dzyuba E.A., Popova N.A. Ontology-oriented information system for verification of formalized documents. *News of the Kabardino-Balkarian Scientific Center of RAS*. 2024. Vol. 26. No. 1. Pp. 11–20. DOI: 10.35330/1991-6639-2024-26-1-11-20

ВВЕДЕНИЕ

Государственная регистрация является обязательным условием для осуществления предпринимательской деятельности и осуществляется в строго предусмотренном законом порядке органами исполнительной власти, ответственными за регистрацию субъектов предпринимательской деятельности [1]. При регистрации бизнеса в налоговую службу подается установленный законом пакет документов. В налоговой службе документы проходят поэтапную проверку (рис. 1).

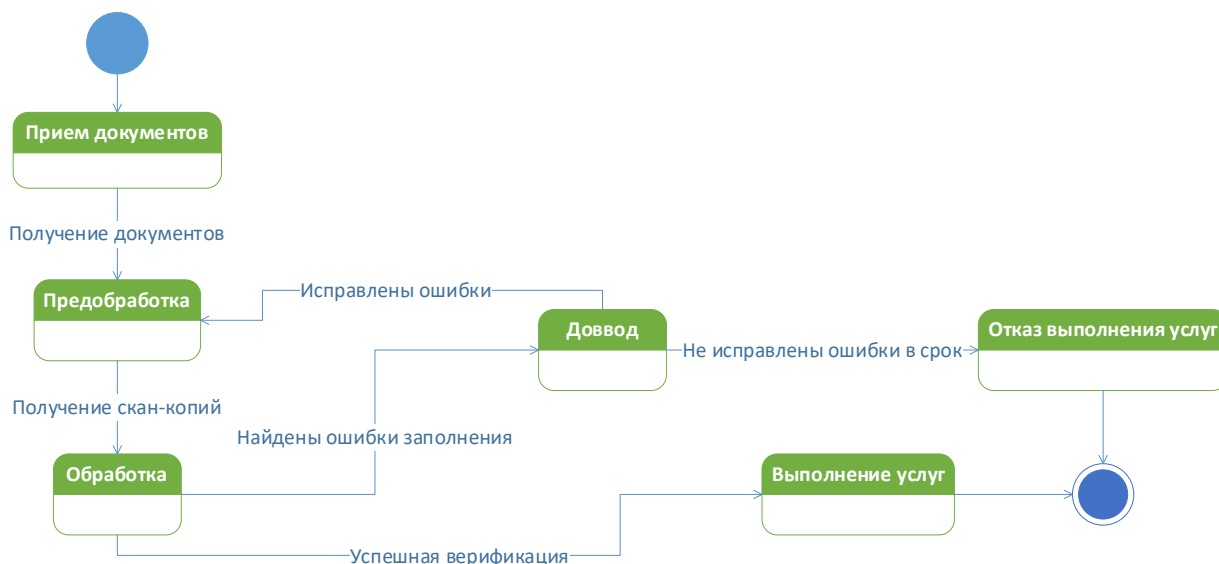


Рис. 1. Состояния обработки документов, подаваемых при регистрации бизнеса

Fig. 1. State of processing documents that are submitted during business registration

Время обработки документов – один из важнейших количественных показателей эффективности документооборота. Верификация – ключевая операция, которая осуществ-

ляется на различных этапах обработки поданного пакета документов. В случае электронной подачи документов частичный контроль за заполнением документов обеспечивается специальным интернет-сервисом Федеральной налоговой службы «Государственная онлайн-регистрация бизнеса». Электронное взаимодействие заявителей и регистрирующего органа повышает эффективность выполнения налоговым органом его функций за счет упрощения и доступности процедуры регистрации для субъектов предпринимательства [2].

При подаче документов в бумажном виде верификация выполняется специалистами службы, при этом данный процесс автоматизирован лишь частично.

Часто задачи верификации на различных этапах существенно различаются, в результате чего верификация одного типа может выполняться неоднократно и занимать при этом значительное время. Автоматизация верификации позволит решить проблему распределения и перекрытия всех типов верификации, сократить время, затрачиваемое специалистом на данном этапе проверки документов, тем самым повысить эффективность процесса государственной регистрации предпринимательской деятельности.

МАТЕРИАЛЫ И МЕТОДЫ

Для повышения эффективности предлагается метод верификации документов, подаваемых при государственной регистрации юридических лиц (ЮЛ) и индивидуальных предпринимателей (ИП), основанный на онтологическом подходе к представлению метаданных. Документы для регистрации ЮЛ и ИП относятся к формализованным документам и имеют четко определенную форму и структуру, утвержденную ФНС России. В электронном виде такие документы представляются в формате XML. Обеспечить поддержку унифицированного представления данных с учетом их семантических свойств позволяет использование онтологии в качестве модели данных [3]. Онтологический подход обеспечивает высокую гибкость моделирования посредством использования стека семантических технологий [4]. Семантические технологии позволяют создавать модели, которые могут быть легко адаптированы к различным ситуациям и требованиям. Также обеспечивается возможность автоматического анализа и обработки данных, что дает возможность автоматически извлекать информацию из базы знаний и применять алгоритмы и инструменты для анализа, классификации и принятия решений.

Управление документами включает в себя также управление метаданными. Использование XML-описания метаданных позволяет включить метаинформацию, несущую машинопонимаемую семантику, в представление данных [5]. XML делает данные более понятными и легко интерпретируемыми для программного обеспечения за счет описания метаинформации в иерархической и структурированной форме. Также XML является широко поддерживаемым форматом, что обеспечивает совместимость и переносимость метаинформации между разными приложениями и платформами. Подобный подход часто используется для построения онтологий при разработке семантически-ориентированных программных систем [6–10]. Выделение данных о документе и его содержимом в онтологию метаданных (рис. 2) позволит организовать автоматическую верификацию формализованного документа.

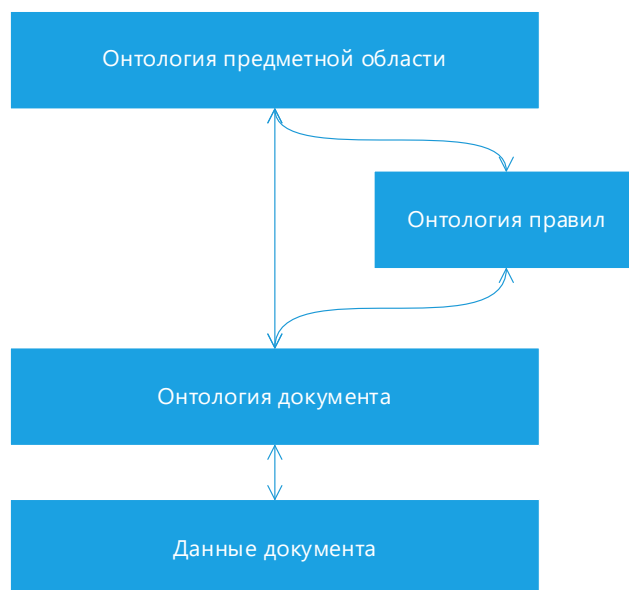


Рис. 2. Многоуровневое представление метаданных документа

Fig. 2. Multilayered representation of document metadata

В метаданных должны присутствовать структурные описания документов, а также правила их верификации, соответствующие различным видам контроля.

ОНТОЛОГИЧЕСКАЯ МОДЕЛЬ ДОКУМЕНТА И АЛГОРИТМ ВЕРИФИКАЦИИ

Построена онтология документа, фрагмент которой представлен на рисунке 3. Совокупность реквизитов, идентифицирующих документ, представлена классом «Форма». Структурные части документа выделены классами «Лист», «Раздел» и «Поле». Условия формирования значения полей заданы классом «Условие». Класс «Данные» описывает совокупность внесенных в документ данных. Критерии заполнения данными структурных единиц документа содержатся в классе «Объем». Класс «Правило» содержит множество правил верификации. Его подклассы соответствуют группам правил, применяемых при различных видах контроля: форматного, логического, справочного и объемного. Класс «Справочник» содержит пути к справочной информации, предназначенной для верификации значений полей документа.

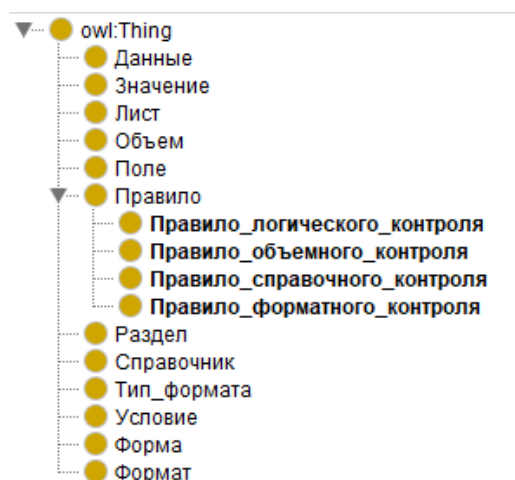


Рис. 3. Фрагмент иерархии классов онтологии «Документ»

Fig. 3. Fragment of class hierarchy of the ontology “Document”

Определен набор $R = \{r_i \mid i = \overline{1, n}\}$ свойств онтологии. Описание важнейших из них представлено в таблице 1.

Таблица 1. Свойства онтологии

Table 1. Ontology properties

r_i	Семантика	Домен свойства	Диапазон значений свойства	Описание
r_1	«включает»	Форма	Лист	Свойство описывает структурную топологию документа
		Лист	Раздел	
		Раздел	Поле	
r_2	«заполняется согласно правилу»	Лист Поле Раздел Форма	Правило	Свойство определяет совокупность правил заполнения соответствующих структурных единиц документа
r_3	«является правилом вида»	Правило	Правило форматного контроля Правило логического контроля Правило справочного контроля Правило объемного контроля	Свойство устанавливает принадлежность правила к соответствующему виду
r_4	«определяет формат»	Правило форматного контроля	Тип формата	Свойство описывает характеристическое свойство правила форматного контроля
r_5	«определяет условие»	Правило логического контроля	Значение	Свойство описывает характеристические свойства правила логического контроля
		Правило логического контроля	Условие	
r_6	«верифицируется по справочнику»	Правило справочного контроля	Справочник	Свойство описывает характеристическое свойство правила справочного контроля
r_7	«должен быть заполнен»	Лист Поле Правило ОК Раздел	Объем	Свойство задает требование заполнения структурных единиц документа

На классах «Правило» и «Форма» онтологии задана функция w – «являться правилом формы», отражающая связь правила и его компонент с заданной формой. В синтаксисе дескрипционной логики ALC (от англ. attributive language with complement) функция w имеет вид:

$$I_{PR} w I_{ФРМ} : ((I_{PRr_3} I_{ФК}) \sqcap (I_{ФКr_4} I_{ФТ})) \sqcup ((I_{PRr_3} I_{ЛК}) \sqcup (I_{ЛКr_5} I_{УСЛ})) \sqcup$$

$$((I_{PRr_3} I_{ЛК}) \sqcup (I_{ЛКr_5} I_{ЗНЧ})) \sqcup ((I_{PRr_3} I_{СК}) \sqcup (I_{СКr_6} I_{СПР})) \sqcup$$

$$((I_{PRr_3} I_{ОК}) \sqcap (I_{ОКr_7} I_{ОБМ})) \sqcap ((I_{ПОЛЕr_2} I_{ПР}) \sqcap (I_{ПОЛЕr_1} I_{РЗДЛ}) \sqcap (I_{РЗДЛr_1} I_{ЛСТ}) \sqcap (I_{ЛСТr_1} I_{ФРМ})),$$

где I_{PR} : «Правило»; $I_{ФРМ}$: «Форма»; $I_{ФК}$: «Правило форматного контроля»; $I_{ФТ}$: «Формат»; $I_{ЛК}$: «Правило логического контроля»; $I_{УСЛ}$: «Условие»; $I_{ЗНЧ}$: «Значение»; $I_{СК}$: «Правило справочного контроля»; $I_{СПР}$: «Справочник»; $I_{ОК}$: «Правило ОК»; $I_{ОБМ}$: «Объем»; $I_{ПОЛЕ}$: «Поле»; $I_{РЗДЛ}$: «Раздел»; $I_{ЛСТ}$: «Лист».

Функция w определяет правило верификации поля документа заданной формы.

Разработан алгоритм верификации на основе представленной онтологической модели, последовательность реализации которой включает следующие шаги:

Шаг 1 – определение формы верифицируемого документа на основе метаданных онтологии.

Шаг 2 – построение набора правил для каждого структурного элемента формы.

Шаг 3 – получение данных о значении структурного элемента загруженного документа по тега, определенным в онтологии документа.

Шаг 4 – оценка соответствия полученного значения правилам верификации.

Шаг 5 – подсчет количества и анализ типа выявленных ошибок.

По результатам выполнения процедуры разбора составляется отчет, в котором сохраняются основные ошибки в документе по всем видам контроля.

РЕЗУЛЬТАТЫ

Предложенный метод реализован в виде системы автоматизированной верификации документов. Архитектура разработки состоит из трех уровней: метаданные, сервер, приложение пользователя. Уровень метаданных включает совокупность онтологий, описывающих документы и правила их верификации. Структура системы приведена на рисунке 4.

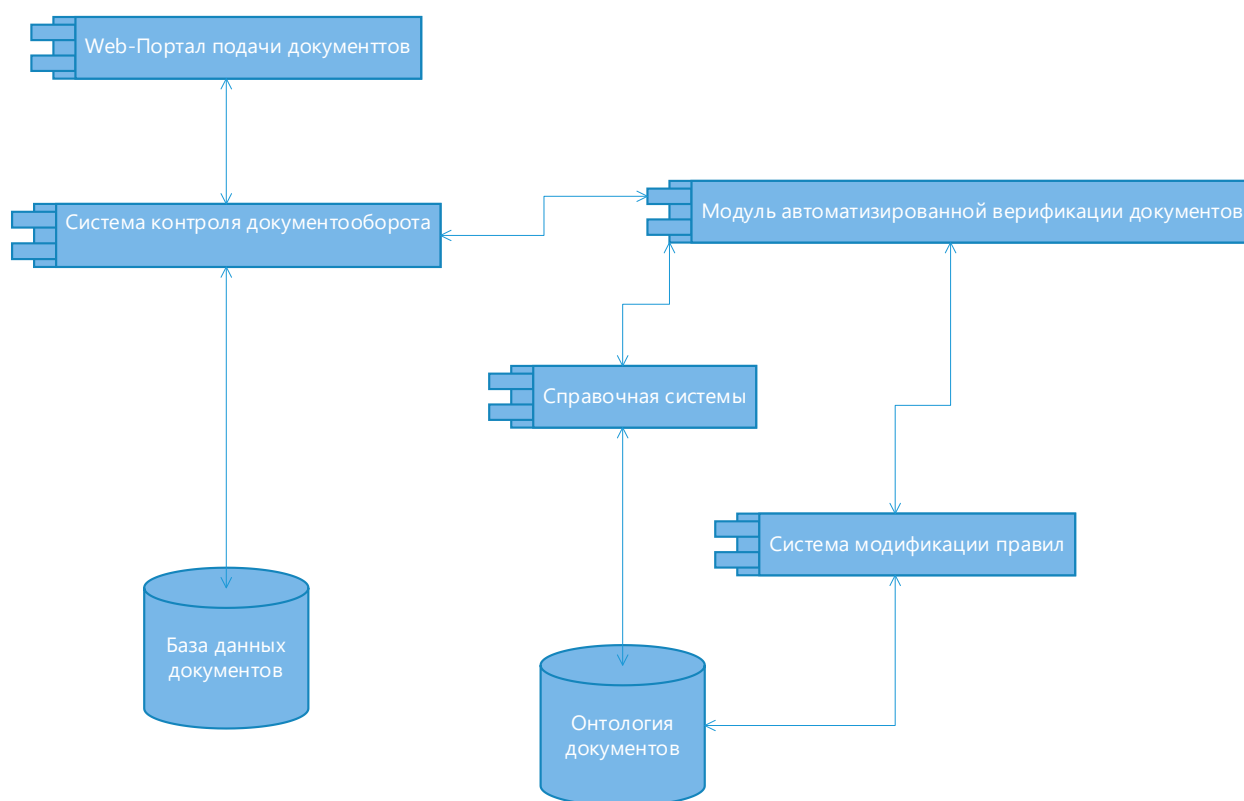


Рис. 4. Структура среды информационной системы

Fig. 4. Structure of the information system environment

Загрузка проверяемого файла происходит в верификаторе, в состав которого входят библиотеки, производящие разбор документа. По результатам работы верификатора осуществляется построение отчета (рис. 5).

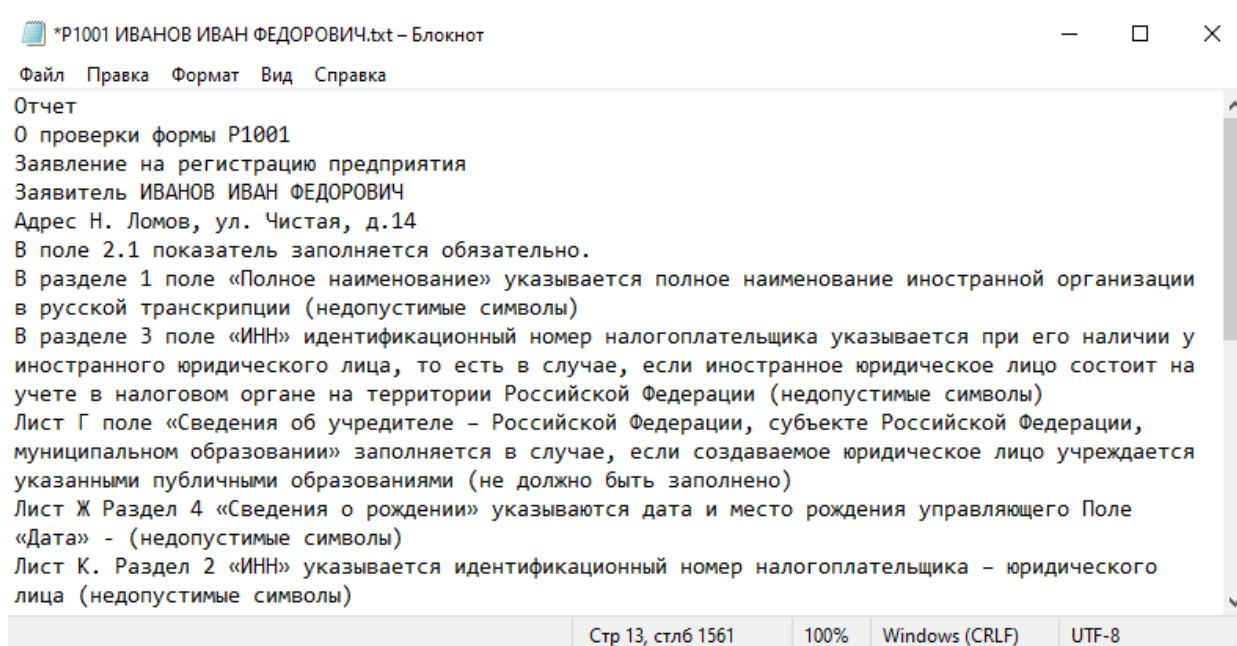


Рис. 5. Фрагмент отчета о результатах верификации документа

Fig. 5. Fragment of the report on the results of document verification

Наличие ошибок является основанием решения об отказе в рассмотрении документов.

ОБСУЖДЕНИЕ

Проведен анализ времени, затрачиваемого на верификацию документов вручную и с помощью разработанной системы. Для этого в отделе налогового органа собрана статистика времени проверки документов за рабочую неделю (табл. 2).

Таблица 2. Среднее время проверки документов вручную

Table 2. Average time for checking an application for registration in manual mode

Показатель	Пн	Вт	Ср	Чт	Пт	В среднем за неделю
Количество проверенных форм (шт.)	7	11	5	6	8	7,4
Среднее время проверки (мин.)	68,5	80	48	67	81,5	69

При подключении программных средств время проверки сократилось почти на 50 % (табл. 3).

Таблица 3. Среднее время проверки документов в информационной системе

Table 3. Average time for verification of an application for registration in the information system

Показатель	Пн	Вт	Ср	Чт	Пт	В среднем за неделю
Количество проверенных форм (шт.)	15	17	14	12	14	14,4
Среднее время проверки (мин.)	32	28,2	34,2	40	34,2	33,7

Как видно, количество проверенных форм при этом возросло почти в два раза.

ЗАКЛЮЧЕНИЕ

Использование разработанной системы в работе налоговой службы позволяет сократить материальные и временные затраты на верификацию документов при регистрации юридических лиц и ИП, что способствует повышению эффективности данной процедуры. Построенная онтологическая модель и процедура верификации могут быть расширены, что обеспечивает легкость внесения модификаций в форму или правила верификации при их изменении уполномоченным органом. Полученные результаты применимы в теоретических и практических исследованиях, связанных с разработкой систем верификации структурированных электронных документов.

СПИСОК ЛИТЕРАТУРЫ

1. Зорина Е. А. Государственная регистрация юридических лиц в РФ: административно-правовой статус субъектов предпринимательской деятельности как стороны регистрационного производства // Человек: преступление и наказание. 2013. № 4(83). С. 125–127.
2. Трофимова Е. В. Электронная легитимация и взаимодействие субъектов предпринимательства с регистрирующими (налоговыми) органами // Вестник Университета имени О. Е. Кутафина (МГЮА). 2020. № 7(71). С. 29–36. DOI:10.17803/2311-5998.2020.71.7.029-036
3. Максимов Н. В., Лебедев А. А. Онтологическая система «знания-деятельность» // Онтология проектирования. 2021. Т. 11. № 2(40). С. 185–211. DOI: 10.18287/2223-9537-2021-11-2-185-211
4. Волчек Д. Г., Романов А. А. Создание и обучение онтологий на основе анализа контекста и метаданных слабоструктурированного контента // Экономика: вчера, сегодня, завтра. 2020. Т. 10. № 1-1. С. 303–312. DOI: 10.34670/AR.2020.91.1.033
5. Воробьев В. И., Монахова Т. В. Методы защиты метаданных в формате XML // Онтология проектирования. 2018. Т. 8. № 2(28). С. 253–264. DOI: 10.18287/2223-9537-2018-8-2-253-264
6. Половикова О. Н. Анализ XML-подхода для описания метаданных и онтологий в Semantic Web // Известия Алтайского государственного университета. 2014. № 1-2(81). С. 119–123. DOI: 10.14258/izvasu(2014)1.2-19
7. Воробьев В. И., Солдаткина А. А. Метод онтологического анализа web-ресурса на основе метаданных // Научный вестник Новосибирского государственного технического университета. 2018. № 3(72). С. 43–58. DOI: 10.17212/1814-1196-2018-3-43-58
8. Щекин А. В. Особенности наследования информации в задачах интеграции систем технической подготовки производства // Онтология проектирования. 2020. Т. 10. № 2(36). С. 201–217. DOI: 10.18287/2223-9537-2020-10-2-201-217
9. Муромцев Д. И., Волчек Д. Г., Романов А. А. Индустриальные графы знаний – интеллектуальное ядро цифровой экономики // Control Engineering Россия. 2019. № 5(83). С. 32–39.
10. Shachnev D. A. Searching for activity results and experts in a given subject area, taking results significance into account // Programmная Ingeneria. 2021. Vol. 12. No. 5. Pp. 260–266. DOI: 10.17587/prin.12.260-266

REFERENCES

1. Zorina E.A. State registration of legal entities in the russian federation: administrative and legal status of business entities as the parts of registration production. *Chelovek: prestuplenie i nakazanie* [Man: crime and punishment]. 2013. No. 4(83). Pp. 125–127. (In Russian)

2. Trofimova E.V. Electronic legitimation and interaction of entrepreneurships with registration (tax) authority. *Courier of the Kutafin Moscow State Law University (MSAL)*. 2020. No. 7(71). Pp. 29–36. DOI: 10.17803/2311-5998.2020.71.7.029-036. (In Russian)
3. Maksimov N.V., Lebedev A.A. Ontological system “knowledge-activity”. *Ontology of designing*. 2021. Vol. 11. No. 2(40). Pp. 185–211. DOI: 10.18287/2223-9537-2021-11-2-185-211. (In Russian)
4. Volchek D.G., Romanov A.A. Creation and training of ontologies based on the analysis of context and metadata of poorly structured content. *Ekonomika: vchera, segodnya, zavtra* [Economics: yesterday, today, tomorrow]. 2020. Vol. 10. No. 1-1. Pp. 303–312. DOI: 10.34670/AR.2020.91.1.033. (In Russian)
5. Vorobiev V.I., Monakhova T.V. Protection of metadata in xml format. *Ontology of designing*. 2018. Vol. 8. No. 2(28). Pp. 253–264. DOI: 10.18287/2223-9537-2018-8-2-253-264. (In Russian)
6. Polovikova O.N. Analysis of XML-based approach to description of semantic web metadata and ontologies. *Izvestiya Altayskogo gosudarstvennogo universiteta* [News of Altai State University]. 2014. No. 1-2(81). Pp. 119–123. DOI: 10.14258/izvasu(2014)1.2-19. (In Russian)
7. Vorobev V.I., Soldatkina A.A. Method of ontological analysis of a web-resource based on metadata. *Science Bulletin of the Novosibirsk State Technical University*. 2018. № 3(72). Pp. 43–58. DOI: 10.17212/1814-1196-2018-3-43-58. (In Russian)
8. Shchekin A.V. The specifics of information inheritance in cad/cam-integration. *Ontology of designing*. 2020. Vol. 10. No. 2(36). Pp. 201–217. DOI: 10.18287/2223-9537-2020-10-2-201-217. (In Russian)
9. Muromtsev D.I., Volchek D.G., Romanov A.A. *Industrial'nye grafy znaniy – intellektual'noe yadro tsifrovoy ekonomiki* [Industrial knowledge graphs – the intellectual core of the digital economy]. *Control Engineering Rossiya* [Control engineering Russia]. 2019. № 5(83). Pp. 32–39. (In Russian)
10. Shachnev D. A. Searching for activity results and experts in a given subject area, taking results significance into account. *Programmnyaya Ingeneria*. 2021. Vol. 12. No. 5. Pp. 260–266. DOI: 10.17587/prin.12.260-266

Информация об авторах

Балашова Ирина Юрьевна, канд. техн. наук, доцент, доцент кафедры «Математическое обеспечение и применение ЭВМ», Пензенский государственный университет;

440026, Россия, г. Пенза, ул. Красная, 40;

irs-80@mail.ru, ORCID: <https://orcid.org/0009-0003-9388-7375>

Дзюба Елена Анатольевна, ст. преподаватель кафедры «Математическое обеспечение и применение ЭВМ», Пензенский государственный университет;

440026, Россия, г. Пенза, ул. Красная, 40;

dzyuba_ea@mail.ru

Попова Наталия Александровна, канд. техн. наук, доцент, доцент кафедры «Математическое обеспечение и применение ЭВМ», Пензенский государственный университет;

440026, Россия, г. Пенза, ул. Красная, 40;

popov.tasha@yandex.ru, ORCID: <https://orcid.org/0000-0001-9713-4897>

Information about the authors

Irina Yu. Balashova, Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Mathematical Support and Computer Use, Penza State University;

440026, Russia, Penza, 40 Krasnaya street;

irs-80@mail.ru, ORCID: <https://orcid.org/0009-0003-9388-7375>

Elena A. Dzyuba, Senior Lecturer of the Department of Mathematical Support and Computer Use, Penza State University;

440026, Russia, Penza, 40 Krasnaya street;

dzyuba_ea@mail.ru

Nataliya A. Popova, Candidate of Technical Sciences, Associate Professor of the Department of Mathematical Support and Computer Use, Penza State University;

440026, Russia, Penza, 40 Krasnaya street;

popov.tasha@yandex.ru, ORCID: <https://orcid.org/0000-0001-9713-4897>