НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2018 Том 18 № 2 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2018 Vol. 18 No 2 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.043

МЕТОДОЛОГИЯ ПРОЕКТИРОВАНИЯ, РАЗРАБОТКИ И СОПРОВОЖДЕНИЯ ДОМЕННЫХ СЕМАНТИЧЕСКИХ ПОРТАЛОВ НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ

М.А. Навроцкий[a], Н.А. Жукова[a], Д.И. Муромцев[a], Н.Г. Мустафин[b,c]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b] СПИИРАН, Санкт-Петербург, 199178, Российская Федерация
[c] Санкт-Петербургский государственный электротехнический университет (ЛЭТИ), Санкт-Петербург, 197376, Российская Федерация
Адрес для переписки: d.muromtsev@gmail.com

Информация о статье
Поступила в редакцию 10.01.18, принята к печати 15.02.18
doi: 10.17586/2226-1494-2018-18-2-286-298
Язык статьи – русский

Ссылка для цитирования: Навроцкий М.А., Жукова Н.А., Муромцев Д.И., Мустафин Н.Г. Методология проектирования, разработки и сопровождения доменных семантических порталов научно-технической информации // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 2. С. 286–298. doi: 10.17586/2226-1494-2018-18-2-286-298

Аннотация

Предмет исследования. Рассмотрена проблема построения и поддержки образовательных процессов за счет создания специализированных семантических порталов на основе технологии открытых связанных данных. Такие порталы содержат траектории освоения информации, а также связанные с этими траекториями данные и знания. В статье исследована проблема поддержки такого рода траекторий в технической области. Именно в этой области потребность в предлагаемых порталах наиболее высока. **Метод.** В качестве метода структурирования и представления данных и знаний семантического портала выбран онтологический инжиниринг. В основу построения индивидуальной образовательной траектории положена 4-уровневая иерархическая модель, включающая уровни инкапсуляции и композиции контента, определенные в терминах компетенций, а также уровни порождения и композиции конкретных программ обучения. Моделирование компетенций выполнено в терминах теории графов, а в качестве доступа к данным порталов использован язык запросов SPARQL. **Основные результаты.** С использованием предложенных методов разработано два семантических портала научно-технической информации: для поддержки образовательного процесса в Университете ИТМО и для Центра речевых технологий. Рассмотрены результаты опытной эксплуатации разработанных порталов. **Практическая значимость.** Внедрение предложенного подхода и использование представленной методики позволяет разрабатывать доменные семантические порталы научно-технической информации и использовать их для получения знаний пользователем по индивидуальным траекториям. Использованный стек технологий и методов позволяет легко адаптировать и повторно использовать уже разработанные элементы семантических порталов для различных предметных областей.

Ключевые слова
онтология, открытые связанные данные, интеграция открытых данных, поиск в открытых данных, LOD, SPARQL

DESIGN, DEVELOPMENT AND MAINTENANCE METHODOLOGY OF DOMAIN SEMANTIC PORTALS OF SCIENTIFIC AND TECHNICAL INFORMATION

M.A. Navrotskiy[a], N.A. Zhukova[a], D.I. Mouromtsev[a], N.G. Mustafin[b,c]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
[b] SPIIRAS, Saint Petersburg, 199178, Russian Federation
[c] Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197376, Russian Federation
Corresponding author: d.muromtsev@gmail.com

Article info
Received 10.01.18, accepted 15.02.18
doi: 10.17586/2226-1494-2018-18-2-286-298
Article in Russian

For citation: Navrotskiy M.A., Zhukova N.A., Mouromtsev D.I., Mustafin N.G. Design, development and maintenance methodology of domain semantic portals of scientific and technical information. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 2, pp. 286–298 (in Russian). doi: 10.17586/2226-1494-2018-18-2-286-298

Abstract

Subject of Research. The paper deals with the problem of development and support of educational processes through the creation of specialized semantic portals based on Linked Open Data. These portals contain the individual educational paths, and data and knowledge associated with these paths. The paper studies the problem of maintenance and support of individual educational paths in the technical domain. This is exactly the domain with the highest need for the proposed portals. **Method.** Ontological engineering is chosen as a method of storage and presentation of data and knowledge for semantic portal. The basis for an educational individual path building is four-level hierarchical model, including the levels of encapsulation and composition of content, defined in terms of competences, as well as the levels of generation and composition of specific training programs. Competence modeling is performed in terms of graph theory, and SPARQL query language is used for data access. **Main Results.** Two semantic portals of scientific and technical information were developed with the use of the proposed methods: the first portal supports the educational process at ITMO University; the second portal supports Speech Technologies Center (STC). The results of experimental operation of the developed portals are considered. **Practical Relevance.** Implementing of the proposed approach and the use of the presented technique gives the possibility to develop domain semantic portals of scientific and technical information and apply them for acquiring of knowledge on individual educational paths by the user. The applied stack of technologies and methods makes it easy to adapt and reuse already developed elements of semantic portals for different domains.

Keywords

ontology, Linked Open Data, open data integration, open data search, LOD, SPARQL

Введение

Высокий уровень развития сферы информационных технологий (ИТ) и ее динамичность определяют непрерывное появление и совершенствование многих эффективных методологий разработки программных систем. Методологии могут относиться к одному или нескольким этапам жизненного цикла программного обеспечения (ПО). Каждая из методологий может иметь несколько модификаций. Достаточно высоко востребованы гибкие методологии [1], применяемые при разработке ПО. Большой интерес вызывают методологии непрерывной интеграции процессов разработки, тестирования, развертывания программных систем и др. В основном эти методологии ориентированы на создание типовых программных систем продуктового уровня. В целом существующие методологии позволяют достаточно быстро создавать и модифицировать информационные системы [1] при появлении новых требований, изменении условий решения задач. Изменения часто вызваны необходимостью поддержки вновь разработанных прикладных моделей, методов и средств. Большинство из них являются инновационными, требуют специальной подготовки для их эффективного использования на практике.

В настоящее время процессы повышения квалификации в предметных областях поддержаны очень ограниченно. Как правило, обучение предусматривает привлечение многих преподавателей, формирование специализированных курсов. Стоимость постановки новых курсов оказывается очень высокой, а срок их жизни коротким, что приводит к неоправданным расходам. Альтернативным подходом является поиск и освоение информации самими специалистами. Однако для этого требуется применять многие поисковые системы и другие средства доступа к информации. Кроме того, объем информации часто оказывается очень большим, может предоставляться много данных, напрямую не связанных с изучаемыми вопросами, отсутствовать часть необходимой информации. Отдельную проблему составляет уровень сложности и способы представления информации. Регулярно наблюдаются ситуации, когда информация не отвечает ожиданиям специалиста.

Проблема построения и поддержки образовательных процессов может быть решена за счет создания и внедрения вспомогательного ПО. К такому ПО относятся системы поддержки профессионального развития прикладных специалистов. Их основной задачей является формирование процессов для освоения новых компетенций, приобретения новых умений и навыков специалистами предметных областей.

Подобное ПО может быть реализовано в виде веб-порталов. Особенность этих порталов состоит в их наполнении. Они содержат траектории освоения информации (ТОИ) [2], а также связанные с этими траекториями данные и знания. В качестве основного источника информации и научных знаний можно рассматривать пространство открытых данных (Linked Open Data, LOD). Для работы с научными знаниями применимы технологии научного веба [3]. При поиске научных знаний в LOD должны учитываться особенности траекторий, а также возможности специалистов, осваивающих эти траектории, требования предприятий и организаций, в которых специалисты работают, и другое. В статье рассматривается проблема поддержки ТОИ в технической области. Именно в этой области потребность в предлагаемых порталах наиболее высока. В других, более консервативных областях подобные порталы также могут оказаться востребованными, но для их создания, как правило, требуется разработка формализованных описаний этих областей.

Таким образом, можно считать актуальной задачу создания и использования семантических порталов научно-технической информации (СП НТИ). Это новый класс информационных систем, для которых необходимо создание новых методологий проектирования, создания и сопровождения.

Анализ потребностей и возможностей создания семантических порталов научно-технической информации

Область применения порталов – это средние и небольшие домены. Такие домены, как правило, обладают двумя свойствами. Во-первых, у организаций и предприятий недостаточно средств для создания и проведения традиционных обучающих курсов. Во-вторых, для доменов могут быть построены формализованные описания, как правило, в виде онтологий.

Рассмотрим примеры предметных областей и решаемых в них задач:

– разработка и использование обучающих программ для различных заинтересованных сторон при продвижении продуктов на рынке;
– формирование программ подготовки спортсменов с учетом индивидуальных особенностей и имеющихся средств;
– повышение квалификации сотрудников при переходе в смежную предметную область;
– различные корпоративные системы обучения, в том числе для обучения нового персонала.

Кроме того, задача обучения является актуальной для образовательных учреждений, в первую очередь – высших учебных заведений.

Для создания СП НТИ необходимо наличие технологий создания порталов, технологий управления знаниями. Рассмотрим основные особенности перечисленных технологий и дадим оценку их текущему состоянию.

Современные семантические технологии включают онтологические модели представления знаний и языки запросов для работы с этими моделями, например, SPARQL [3]. Онтологии используются для формализованного описания данных и знаний семантических порталов. Один портал может относиться к одной или нескольким смежным предметным областям. Предоставляемая информация, как правило, содержит множественные ссылки, обеспечивающие связанность данных. Связи могут устанавливаться не только внутри одного портала, но и между разными порталами.

К технологиям управления знаниями относятся представление, публикация и распространение знаний. Для накопления и работы со знаниями используются семантические хранилища данных и знаний (например, CKAN [1]). Средства работы с хранилищами обеспечивают доступ к данным через HTTP-протокол или точки доступа SPARQL.

Результаты проведенного анализа источников данных и технологий работы с ними [1] показали следующее.

1. Современный Веб содержит огромное количество данных и знаний о различных предметных областях. Эту информацию можно использовать при поддержке профессионального развития прикладных специалистов.
2. Данные о предметных областях публикуются системами распространения научных публикаций в виде отдельных наборов данных, как отдельные страницы и т.д. Использование этих данных совместно с открытыми связанными данными позволит не только получать разнообразные наборы данных, но и, связывая их между собой, порождать новые знания.
3. Не каждый пользователь может работать с имеющимися источниками по многим причинам, среди которых – сложности доступа к источнику, отсутствие информации об источнике, отсутствие у источника интерфейса для поиска и др.
4. У пользователя отсутствует механизм персонализации поиска по этим источникам.
5. Не существует какой-либо единой точки доступа к источникам.

Таким образом, состав данных и знаний является достаточным для создания СП НТИ, однако для их использования в обучающих целях требуется развитие существующих технологий работы с ними.

Постановка задачи

Определим место порталов среди других классов информационных систем (рис. 1). При позиционировании порталов интерес представляют информационно-управляющие системы, семантические порталы и корпоративные порталы.

Информационно-управляющая система – это цифровая система контроля или управления. Такие системы используются для решения задач, связанных с необходимостью принятия решений (часто в реальном времени) [2].

Семантический портал – это специфический ресурс, который можно рассматривать как дополнение к инструментам управления знаниями, средствам семантической обработки данных и информации, основанных на технологии семантического веба [3, 4]. Портал предусматривает формализацию предметных областей и бизнес-процессов с помощью метаописаний и онтологий.

Корпоративный портал – это системно организованная совокупность средств передачи данных, информационных ресурсов, протоколов взаимодействия, аппаратно-программного и организационно-методического обеспечения, ориентированная на удовлетворение потребностей пользователей в информационных услугах и ресурсах образовательного характера [5].

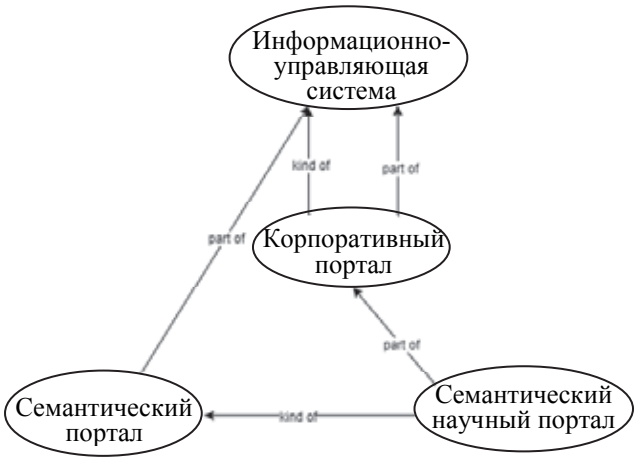

Рис. 1. Место портала относительно существующих классов систем

Используя преимущества существующих классов систем, можно построить семантические научные порталы, являющиеся расширением корпоративных порталов и ориентированные на использование семантических технологий. При этом предусматривается наполнение семантических порталов научной и образовательной информацией.

Для создания и использования СП НТИ необходимо сформулировать и решить частные задачи. К этим задачам относятся:

- поиск информационных элементов, необходимых для решения некоторой прикладной задачи в заданных условиях;
- связывание информационных элементов между собой, а также с прикладными процессами;
- построение информационной модели освоения новых знаний и ее уточнение в соответствии с образовательными траекториями обучения и имеющимися компетенциями, навыками, умениями обучаемого специалиста.

Создаваемый портал должен удовлетворять следующим функциональным требованиям:

- обеспечивать учет интересов пользователей при поддержке индивидуальных образовательных цепочек;
- обеспечивать поддержку образовательного процесса в разных предметных областях в соответствии с образовательной траекторией.

К нефункциональным требованиям можно отнести:

- использование онтологических моделей для описания данных;
- использование существующих технологий семантического веба;
- портируемость порталов в различные предметные области;
- использование LOD-источников как источников научных данных и знаний.

С учетом современных технологических возможностей можно рекомендовать:

- в качестве языка описания онтологических моделей использовать язык OWL 2;
- в качестве доступа к данным порталов использовать язык запросов SPARQL.

Для создания и использования семантического портала в предметной области необходимо, чтобы для этой предметной области существовала онтологическая модель предметной области и поддерживались порталы открытых данных.

Обобщенные модели семантических порталов научно-технической информации и их основные особенности

Определим основные понятия и группы понятий, связанные с поддержкой СП НТИ. К ним относятся:

- иерархия понятий веб, семантического и научного веба;
- информационный контент порталов;
- контекст, в котором рассматривается контент;
- траектория приобретения знаний.

На рис. 2 представлена структура понятий и связи между ними.

Дадим определения основных понятий [6].

Источник – источник данных или знаний (веб-ресурс с открытым доступом – Википедия, семантический портал, библиотека).

Публикации, ученые, проекты и т.д. – информационный контент, предоставляемый информационными источниками.

Поисковый запрос – запрос пользователей, интерпретируемый с учетом контекста для получения новых знаний.

Поисковая выдача – информационный контент, отвечающий требованиям пользователя.

Рис. 2. Структура понятий семантических порталов научно-технической информации и связи между ними

Научно-образовательный процесс можно рассматривать как некоторую цепочку (траекторию), в каждом узле которой предметный специалист, проходящий обучение (обучающийся), получает определенные навыки и знания [6]. Уровни освоения определяются компетентностными моделями. Под компетентностными моделями понимается множество компетенций, которыми владеет обучающийся. Получение компетенций обеспечивается за счет прохождения индивидуальной образовательной траектории [7]. Индивидуальная траектория – это персональный путь получения знаний одним специалистом.

Информационный контент – это опубликованные научные и образовательные данные и знания (научные публикации, образовательные курсы и т.д.). Контекст – условия получения информационного контента, требования к процессу освоения знаний. В настоящее время при реализации процессов обучения контент и контекст учитываются ограниченно.

К основным взаимосвязям можно отнести связь образовательных и доменных знаний и связь между образовательными процессами и процессами работы с формализованными знаниями.

Несмотря на то, что образовательные процессы развиты значительно лучше, чем процессы освоения знаний в доменах, эффективность процесса работы со знаниями даже образовательных процессов остается достаточно низкой. Для повышения эффективности необходимо увеличить число используемых источников знаний. Это проблему предлагается решить за счет создания семантического портала. При этом будет сформирована единая точка доступа к ресурсам научного веба, определены стандартизованные методы, поддержана система поиска.

Концептуальная модель портала представлена на рис. 3.

Данная модель определяет процесс получения знаний пользователем. Модель включает:

- модель требований – описывает требования всех заинтересованных сторон, включая самого обучающегося, образовательное учреждение, стандарты;
- модель обучаемого, содержащая метаданные, модель знаний обучаемого, модель интересов обучаемого;
- модель образовательного курса, описывающая образовательный курс, для освоения которого строится индивидуальная образовательная траектория.

Совокупность знаний пользователя представляет собой граф знаний [8]. Граф знаний пользователя можно представить как множество его компетенций. При этом узлами графа будут компетенции, а ребрами – связи между компетенциями.

В основу построения индивидуальной образовательной траектории положена четырехуровневая иерархическая модель (рис. 4).

На Уровне 0 (М0) находятся элементарные модули, которые инкапсулируют контент для обучения. На базе этих модулей строятся сервисы нижнего уровня или элементарные сервисы. В качестве результата функционирования модулей Уровня 0 выступают элементарные компетенции, необходимые обучаемому для освоения последующего материала.

На Уровне 1 (М1) находятся модули, построенные на основе композиции модулей Уровня 0. Модули этого уровня можно рассматривать как общие процессы, определенные в терминах компетенций.

На Уровне 2 (М2) находятся модули, построенные на основе композиции модулей Уровня 1. Эти модули можно рассматривать как каркасы. Каркасы – это некоторый обобщенный курс, который не предназначен для обучения. Он предназначен только для порождения конкретных программ обучения.

На Уровне 3 (М3) находятся модули, построенные на основе композиции модулей Уровня 2. Это конкретные образовательные программы, на базе которых могут разрабатываться новые программы.

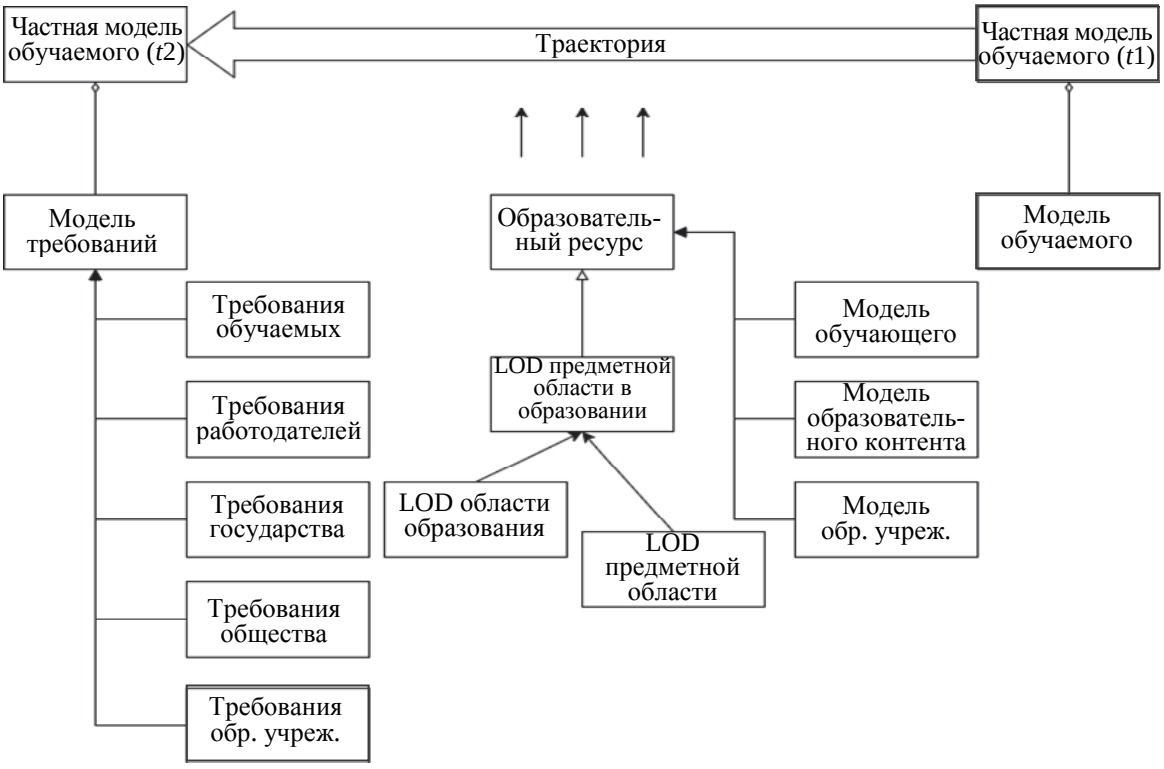

Рис 3. Концептуальная модель портала

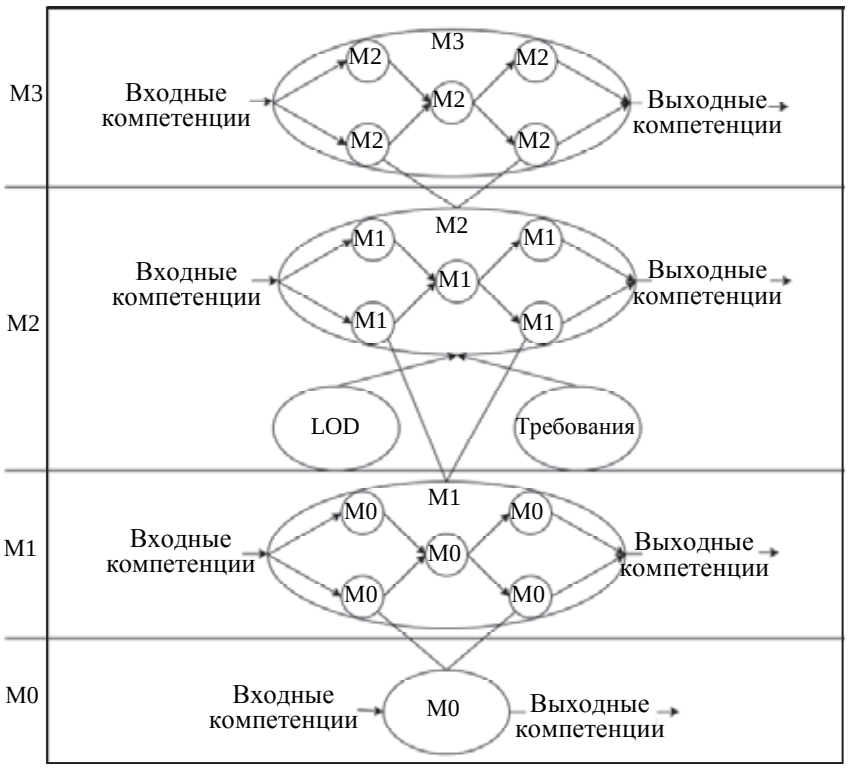

Рис. 4. Иерархическая модель образовательной траектории

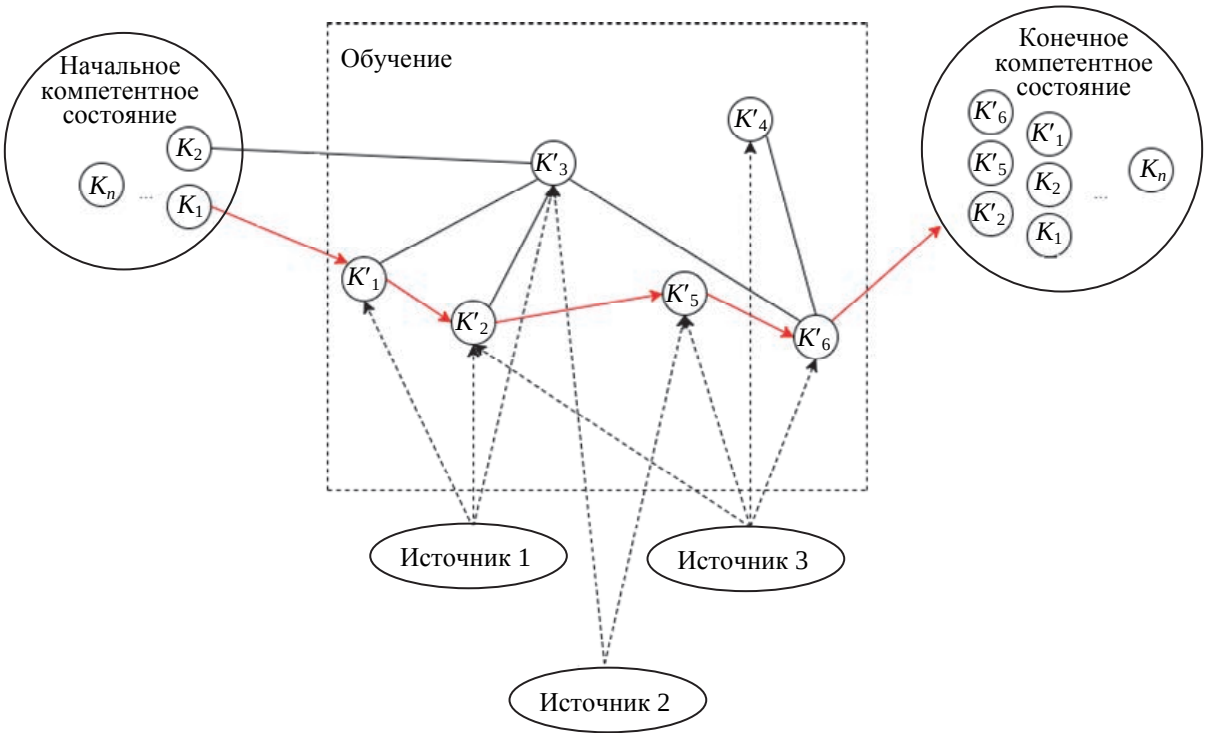

Рис. 5. Процесс преобразования компетентностных моделей

Рассмотрим процесс получения знаний пользователем через преобразование компетентностных моделей (рис. 5).

Имеются два состояния: начальное компетентностное состояние G_1 и конечное компетентностное состояние G_2. Начальное компетентностное состояние описывает знания и навыки пользователя через множество его компетенций и связей между ними: $G_1 = \{S, R\}$, где $S = \{s_i\}$ – множество компетенций пользователя, $R = \{r_q\}$ – множество связей (отношений) между компетенциями. В соответствии с [9], это можно представить в виде графа: $G_1 = (V_1, E_1)$, где $E_1 = \{K_1, K_2, K_3\}$ – вершины графа (компетенции), $V_1 = \{K_1K_2, K_1K_3\}$ – ребра графа (отношения между компетенциями).

Конечное компетентностное состояние определяется на основе требований, которые предъявляются к процессу потребления знаний. Это состояние содержит такой набор компетенций и отношений между ними, который обучающийся должен иметь в конце процесса обучения. Конечное состояние представляется как $G_2 = \{S', R'\}$, где $S' = \{s'_i\}$ – множество компетенций, $R' = \{r'_q\}$ – множество отношений между компетенциями.

Тогда процесс получения знаний пользователем является преобразованием графа G_1 в граф G_2 (рис. 6), где F – функция перехода. Более детально переход показан на рис. 7.

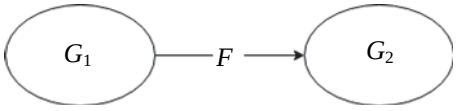

Рис. 6. Переход между состояниями

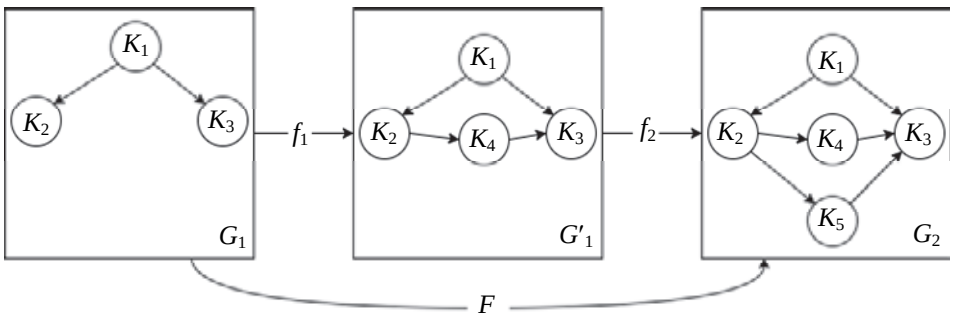

Рис. 7. Детализация процесса перехода между состояниями

Для определения перехода (трансформации графа) необходимо определить различия между графами [8] G_2 и G_1: $G(V,E) = G_2(V_2,E_2)/G_1(V_1,E_1)$.

Преобразование графа представляет собой последовательное применение правил $F = (f_1, f_2)$. При этом получаем $G_1 \underset{f_1}{\Rightarrow} G_1' \underset{f_2}{\Rightarrow} G_2$. На каждом из шагов преобразования выполняется объединение графов. При этом строится множество альтернатив, состоящих из миноров исходного графа. Далее графы оцениваются по сложности преобразований (количество новых связей), и выбираются варианты с минимальной сложностью. Главным условием в данном случае будет наличие у объединяемых графов одинаковых вершин. Тогда преобразование графа можно представить в виде: $G_1' = G_1 \cup G' = G_1'(V_1 \cup V', E_1 \cup E')$. Здесь $G' = (V', E')$, $V' = \{K_2, K_3, K_4\}$, $E' = \{K_2K_4, K_4K_3\}$.

Методы проектирования, разработки и сопровождения доменных порталов научно-технической информации

Предлагаемые методы проектирования, разработки и сопровождения основаны на следующих идеях.

1. Идея использования открытых данных (Linked Open Data, LOD). Идея заключается в том, что в качестве источников данных для СП НТИ следует использовать источники открытых данных. Использование LOD позволяет избежать дублирования данных на серверах портала – данные публикуются один раз только на сервере источника. Также такой подход позволяет решить проблему актуальности данных на портале – данные обновляются на источнике.
2. Идея использования семантических технологий. Идея заключается в использовании онтологических моделей для работы с данными и знаниями предметных областей. Это позволяет разрабатывать модели данных и знаний различной структуры, связывать данные и знания из разных источников. Кроме того, семантические технологии имеют хорошую инструментальную поддержку. За счет этого обеспечивается достаточно высокая надежность работы порталов и низкая стоимость их создания и сопровождения.
3. Идея объединения онтологий в рамках общей системы онтологий. В результате объединения выстраивается связанная система из различных онтологических моделей. При этом могут порождаться новые знания предметной области через связывание существующих знаний. Кроме того, использование множества онтологий, которые описывают различные данные и их источники, позволяет перестраивать информационные потоки, поступающие в портал.
4. Адекватность модели контента и контекста. Адаптация контента и контекста осуществляется для каждого конечного пользователя. Это может быть учет интересов и знаний пользователя, решаемых им задач предметной области и др.
5. Идея использования источников данных с высоким уровнем доверия и качества информации. В качестве таких источников могут выступать порталы открытых данных университетов, научных организаций, порталы организаций с высоким качеством публикуемой информации, которые могут быть отнесены к источникам научного веба.
6. Идея абстрагирования от способов доступа к источникам данных научного веба. Некоторые порталы предоставляют доступ к данным через стандартные интерфейсы. СП НТИ могут использовать эти интерфейсы. Часть порталов имеет собственные интерфейсы. Это приводит к необходимости разработки специализированных модулей для доступа к данным каждого из них. Данные могут иметь собственную уникальную структуру, форматы данных не всегда соответствуют стандартам, могут быть неформализованными. Технологии научного веба позволяют обеспечить единый формат доступа к данным через использование языка запросов [10]. Современные инструментальные средства, реализующие этот интерфейс, позволяют получать данные в любом требуемом формате.

При проектировании СП НТИ предлагается использовать следующий метод.

1. Определение предметной области портала: формируются ключевые слова, которые описывают предметную область портала, что позволяет уточнять источники данных и получаемые наборы данных. Предметная область портала описывается при помощи множества ключевых слов. Ключевые слова заносятся в онтологию портала.
2. Определение требований к порталу: учитываются различные требования – от требований образовательных стандартов и организации до требований к размещению портала и его развертыванию. При этом определяются архитектура портала, требования к визуализации данных, используемые технологии, требования к размещению портала. Для определения требований используются, как правило, официальные документы, например, техническое задание. При описании учитываются принятые в области стандарты.
3. Формирование требований к источникам данных: определяется язык представления наборов данных, лицензии и способ доступа к данным, предметная область источников данных, частота обновления данных, тип доступа к данным и т.д.

Таким образом, в результате проектирования формируется онтологическое описание портала.

Метод разработки СП НТИ включает в себя следующие шаги.

1. Определение источников данных. На основании сформированных ранее требований к источникам данных формируется список источников данных. Для этого предусматривается служебный сервис, входящий в состав СП НТИ.
2. Разработка сервисов, расширяющих возможности портала. Несмотря на то, что СП НТИ содержит базовый набор сервисов, при разработке новых порталов существует возможность добавлять новые сервисы. При создании сервисов должны учитываться требования к используемым технологиям, в частности, протоколам обмена данными.
3. Конфигурация сервисов платформы для разработки портала. На данном шаге формируется описание запросов портала к источникам. Описание включает в себя имя поля в источнике, тип поля, фильтр поля. Описание запросов заносится в онтологию портала.
4. Развертывание и запуск портала. Архитектура системы представляет собой множество изолированных сервисов. При развертывании и запуске портала необходимо это учитывать. Предлагается использовать средства управления контейнерами для запуска всех сервисов портала.

Метод сопровождения СП НТИ предусматривает:

1. уточнение источников данных, добавление или удаление источников данных. На данном этапе при необходимости вносятся изменения в требования к источникам данных. В этом случае необходимо внести новые требования в онтологию и запустить служебный сервис формирования списка источников данных. Портал перестроится на работу с новыми источниками;
2. уточнение способов поиска данных, адаптации данных и слияния данных. В процессе использования портала пользователь приобретает новые знания и навыки, что должно учитываться при поиске. Если требуется уточнить способы адаптации данных для конечного пользователя, то пользователь может уточнить свои интересы и области знаний через интерфейс портала. Если требуется уточнить способы слияния данных, то вносятся изменения в онтологию портала, в описание запросов к источникам данных.

Предложенный метод разработки СП НТИ позволяет создавать порталы, обладающие следующими характеристиками.

1. Существующие порталы рассматриваются как возможные источники данных (так как каждый из них имеет интерфейс для доступа к данным), поиск выполняется по нескольким источникам.
2. Существуют сервисы, которые публикуют техническую информацию в открытом виде, но при этом не являются порталами, в которых данные описываются в семантических форматах. Однако эти системы имеют интерфейс для доступа к данным. Их можно также использовать как источники данных при поиске ненаучной информации.
3. Если пользователь предоставляет свои данные (область интересов, множество навыков и знаний), то поисковую выдачу можно персонализировать и выдавать пользователю только те данные, которые ему интересны, и те, которые ему полезны.
4. Система формирует начальное поле знаний для пользователя, а также позволяет последовательно увеличивать глубину погружения пользователя в предметную область.
5. Сохраняя статистику по каждому поиску и набору интересов пользователей, можно получить информацию о том, какие области и направления наиболее интересуют пользователей, уточнять источники данных для поиска.
6. Для описания научно-технической информации можно использовать онтологические открытые словари.

Метрики для доменных порталов научно-технической информации

Для оценки разрабатываемого портала применимы стандартные метрики качества ПО. Большое значение имеют выбранные источники данных.

Основные метрики источников данных приведены ниже.

1. Использование открытых источников (LOD-источники). Можно использовать HTML-, REST-источники, но наибольшую полезность имеют именно LOD-источники. Метрика определяет способ представления данных и знаний в источнике.
2. Лицензии, по которым можно использовать данные с LOD-источников. Метрика определяет возможность использовать данные различными организациями.
3. Поддержка источниками требуемых языков. Большинство источников предоставляют данные на английском языке. Метрика определяет требования организации к языку представления данных в источнике.
4. Обновляемость данных в источнике. Метрика определяет частоту обновления данных, которая может быть различной для различных предметных областей.
5. Стабильность источника данных. Метрика определяет стабильность ресурсов источника, а именно стабильность доступа к данным, частоту отказов в работе сервисов.
6. Протокол доступа к данным источника. Метрика определяет формат доступа к данным – формат запроса, формат ответа.

7. Авторитетность публикуемых источников данных. Метрика определяет уровень качества данных и доверия к ним.

Используя значения описанных метрик можно определить подклассы порталов.

Разработанные семантические научные порталы

С использованием предложенных методов разработаны два семантических портала научно-технической информации: для поддержки образовательного процесса в Университете ИТМО и для Центра Речевых Технологий (ЦРТ) (г. Санкт-Петербург).

Портал для поддержки образовательного процесса в Университете ИТМО представляет собой систему, которая ищет в LOD-источниках образовательные данные и выдает их пользователю с учетом его интересов. В качестве источников были выбраны Open University Data Portal; DBpedia; Wikidata; University of Southampton Data Portal. Входными данными являются навыки и знания обучающихся. Выходные данные представляют собой веб-страницу, которая предоставляет по запросу пользователя следующие данные: определение термина предметной области на двух языках, список научных публикаций, список исследовательских проектов, список ученых и дополнительные ссылки на близкие темы.

В рамках сотрудничества с ЦРТ был разработан портал, формирующий граф знаний по ключевому слову. В качестве источников были выбраны DBpedia, Wikidata, Wiktionary. Входными данными является ключевое слово, по которому требуется получить информацию. Выходными данными является текстовый файл, который содержит следующие данные: описание на английском языке; список синонимов; к какому классу относится; фасет; основная категория темы.

Пример работы системы

В качестве примера представлен прототип СП НТИ, который был разработан в рамках сотрудничества с ЦРТ.

В примере рассматривается процесс получения данных (построение графа знаний) по ключевому слову. Внешняя система отправляет запрос к СП НТИ с ключевым словом и получает граф знаний по этому слову в виде формализованного текстового файла. Рассмотрим пример работы портала с ключевым словом «Интеллектуальный анализ данных» («Data mining»). Специалист подразделения описывает, какие данные и из каких источников требуется получить системе, которая работает с порталом. Для описания используется онтологическая модель. Также специалист, используя служебный сервис портала, определяет список источников данных – DBpedia, Wikidata, Wiktionary.

В таблице дано описание получаемых полей и их типов.

В рамках данного примера не рассматривается процесс адаптации данных, поскольку выбранные источники данных полностью определили предметную область, и персонализации поиска не требовалось. Аналогичным образом был упрощен процесс слияния данных: каждый источник выдает такие наборы данных, которые не пересекаются между собой.

Как было описано выше, выходным является файл, который в качестве данных содержит список синонимов, класс входного слова, фасет, основная категория темы. Пример выходного файла приведен на рис. 8.

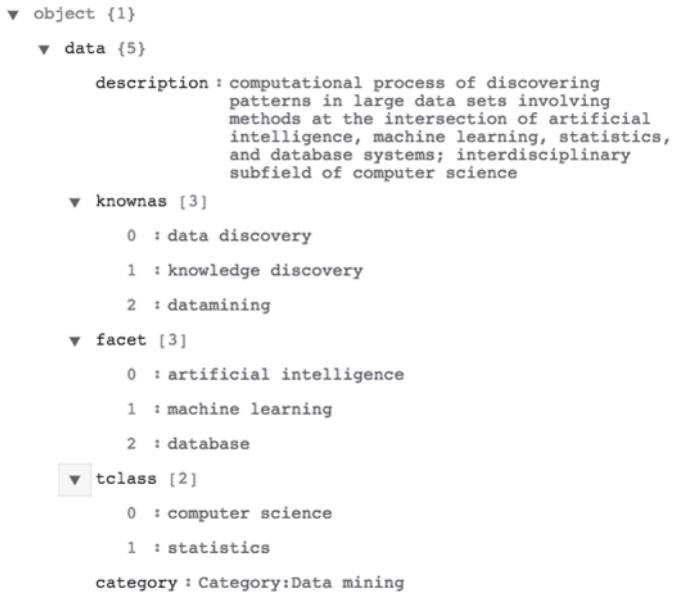

Рис. 8. Пример выходного файла

Поле	Имя поля	Тип в источнике	Источник	Фильтр	Значение фильтра
Описание на английском языке	engdesc	dbo:article	DBpedia	По языку	английский
Класс	Tclass	rdf:type	Wikidata	нет	нет
Список синонимов	knownas	rdfs:label	Wktionary	нет	нет
Фасет	Facet	wd:facetof	Wiktionary	нет	нет
Основная категория темы	tmcat	rdfs:label	Wiktionary	По дополнительному полю	keyword

Таблица. Описание полей (фрагмент)

Количественные оценки

Рассмотрим результаты опытной эксплуатации для СП, который был разработан для Университета ИТМО. Были выделены следующие метрики, по которым проводилась оценка портала:

– число использованных источников;
– сложность запроса;
– время выполнения запроса;
– объем полученных данных;
– связанность данных в ответе.

Рассмотрим каждую из них.

Число использованных источников. Количество использованных источников выбиралось из базы при помощи SWRL-правил, и их количество равно 4. На данный момент не поддерживается получение открытых образовательных курсов по той причине, что не был найден удобный открытый источник. С одной стороны, увеличение источников позволит покрыть предметную область более полно, но с другой стороны, большое количество источников увеличит время обработки запроса, что может быть критично.

Сложность запроса. При обработке запроса пользователя система делает последовательные запросы к разным источникам. Такой подход позволяет избежать сложных вложенных запросов и реализовать механизм генерации запросов, но сказывается на скорости обработки запросов. Общее число запросов – 9. Можно вычислить общее количество получаемых свойств – 44, при этом количество необязательных полей – 18, количество фильтров – 17 (в основном – это языковые фильтры и фильтры с использованием регулярных выражений).

Время выполнения запроса. Время выполнения запроса пользователя можно представить в виде $t = \sum_{i=1}^{n} lod_i + rdr$, где lod_i – время выполнения запроса к источникам данных; rdr – время рендеринга ответа клиенту. Время рендеринга ответа примерно постоянно. Время выполнения всех запросов к LOD-источникам велико и зависит от количества источников. Проведя серию измерений этого параметра, было получено среднее время, за которое пользователь получает наборы данных – 37 с. Для уменьшения времени обработки данных предложено несколько путей решения:

1. уточнение запроса к источнику данных;
2. параллельное выполнение запросов к источникам (например, через использование специальных программных средств [11]);
3. кэширование наборов данных [12].

Были выбраны подходы 1 и 3, так как реализация подхода 2 требует значительных изменений в архитектуре СП и алгоритмов взаимодействия между сервисами.

Для уточнения запросов был реализован механизм перевода запроса на английский язык, что уменьшало время обработки запроса, хотя добавлялся новый запрос. Для ускорения запроса перевода был добавлен фильтр по длине поля, по которому осуществлялся поиск: запрос осуществлялся к DBpedia, и поиск проводился с использованием регулярных выражений по полю rdfs:label. Дополнительно был добавлен фильтр, который сравнивал длину этого поля и ключевого слова. Кроме того, был добавлен механизм кэширования ответов с использованием программного средства Redis, который осуществляет быстрый доступ к данным и имеет простой механизм подключения. Частота обновления была выбрана 1 раз в месяц. Используя описанные способы, удалось сократить время обработки запроса до 3,5 с.

Объем полученных данных. Объем полученных данных определяется ключевыми параметрами, которые будут показаны пользователю. Кроме ключевых параметров, есть дополнительные параметры, которые используются для фильтрации или выявления зависимостей между параметрами. При формировании запроса выделяется 34 параметра, которые будут отображены пользователю.

Связанность данных в ответе. Связанность данных определяется количеством связей между параметрами в ответе. В данном случае имеется 51 связь.

Был проведен анализ зависимостей выделенных метрик от количества источников и сложности запроса. На рис. 9 представлена зависимость времени выполнения запроса от сложности запроса, на рис. 10 –

зависимость объема полученных данных от сложности запроса, на рис. 11 – зависимость связанности данных от сложности запроса, на рис. 12 – зависимость указанных метрик от количества источников.

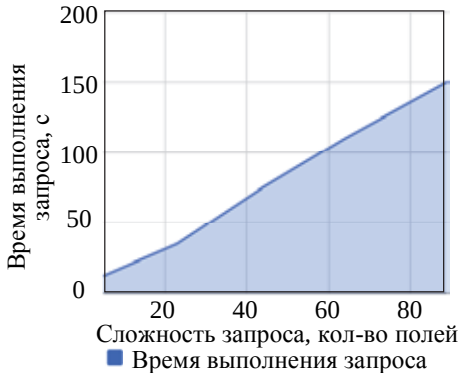

Рис. 9. Зависимость времени выполнения запроса от сложности запроса

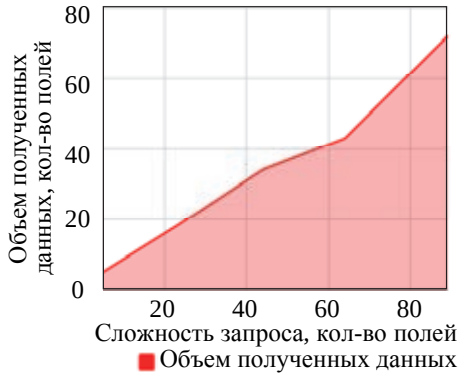

Рис. 10. Зависимость объема полученных данных от сложности запроса

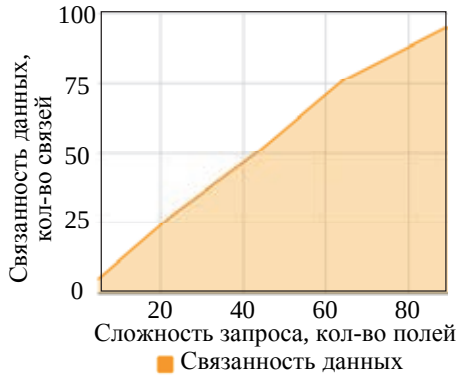

Рис. 11. Зависимость связанности данных от сложности запроса

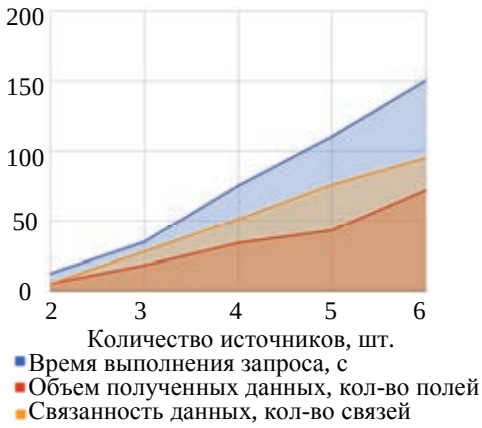

Рис. 12. Зависимость описанных метрик от количества источников

Заключение

Представленная методика была опробована для разработки портала, позволяющего получать данные о ключевых словах. Был проведен анализ количественных оценок портала. При этом в качестве дальнейшего направления развития предполагается:

1. разработка типовых сервисов, которые можно было бы использовать при построении семантических порталов научно-технической информации;
2. разработка решений, которые могли бы позволить реализовать непрерывную адаптацию таких порталов для конечного пользователя с учетом постоянного изменения его интересов, уровня знаний и требований.

Литература

1. Cockburn A. Agile Software Development. Boston: Addison-Wesley, 2002. 304 p.
2. Ключев А.О., Кустарев П.В., Ковязина Д.Р., Петров Е.В. Программное обеспечение встроенных вычислительных систем. СПб.: СПбГУ ИТМО, 2009. 212 с.
3. Maedche A., Staab S. Ontology learning for the semantic web // IEEE Intelligent Systems. 2001. V. 16. N 2. P. 72–79. doi: 10.1109/5254.920602
4. Sure Y., Erdman M., Angele J. et al. OntoEdit: Collaborative ontology development for the semantic web // Lecture Notes in Computer Science. 2002. V. 2342. P. 221–235.
5. Морковкин Д.Е. Организационное проектирование системы управления знаниями // Образовательные ресурсы и технологии. 2013. № 2 (3). С. 74–80.
6. Фурин А.Г. Сущность индивидуальной образовательной траектории: институциональный аспект // Общество: политика, экономика, право. 2016. № 9. С. 41–43.
7. Сытина Н.С. Формирование индивидуальной образовательной траектории студента как условие профессионального развития будущего педагога // Педагогический журнал Башкортостана. 2012. № 3. С. 67–71.
8. Ehrig H., Ermel C., Golas U., Hermann F. Graph and Model Transformation. Springer, 2015. 472 p.
9. Rozewski P., Kusztina E., Tadeusiewicz R., Zaikin O. Intelligent Open Learning Systems. Springer, 2011. 257 p. doi: 10.1007/978-3-642-22667-0
10. Buil-Aranda C., Hogan A., Umbrich J., Vanderbussche P.Y. SPARQL web-querying infrastructure: Ready for action? // Lecture Notes in Computer Science. 2013. V. 8219. P. 277–293. doi: 10.1007/978-3-642-41338-4_18
11. Videla A., Williams J.J.W. RabbitMQ in Action: Distributed Messaging for Everyone. Manning Publ., 2012. 312 p.
12. Carlson J.L. Redis in Action. Manning Publ., 2013. 320 p.

References

1. Cockburn A. Agile Software Development. Boston, Addison-Wesley, 2002, 304 p.
2. Klyuchev A.O., Kustarev P.V., Kovyazina D.R., Petrov E.V. *Embedded Computing Software*. St. Petersburg, SPbSU ITMO Publ., 2009, 212 p. (In Russian)
3. Maedche A., Staab S. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 2001, vol. 16, no. 2, pp. 72–79. doi: 10.1109/5254.920602
4. Sure Y., Erdman M., Angele J. et al. OntoEdit: Collaborative ontology development for the semantic web. *Lecture Notes in Computer Science*, 2002, vol. 2342, pp. 221–235.
5. Morkovkin D.E. Organizational design of knowledge management system. *Educational Resources and Technologies*, 2013, no. 2, pp. 74–80. (In Russian)
6. Furin A.G. The essence of individual learning trajectory: institutional perspective. *Society: Politics, Economics, Law*, 2016, no. 9, pp. 41–43. (In Russian)
7. Sytina N.S. Formation of the individual educational trajectory of the student as the condition of professional development of the future teacher. *Pedagogicheskii Zhurnal Bashkortostana*, 2012, no. 3, pp. 67–71. (In Russian)
8. Ehrig H., Ermel C., Golas U., Hermann F. *Graph and Model Transformation*. Springer, 2015, 472 p.
9. Rozewski P., Kusztina E., Tadeusiewicz R., Zaikin O. *Intelligent Open Learning Systems*. Springer, 2011, 257 p. doi: 10.1007/978-3-642-22667-0
10. Buil-Aranda C., Hogan A., Umbrich J., Vanderbussche P.Y. SPARQL web-querying infrastructure: Ready for action? *Lecture Notes in Computer Science*, 2013, vol. 8219, pp. 277–293. doi: 10.1007/978-3-642-41338-4_18
11. Videla A., Williams J.J.W. *RabbitMQ in Action: Distributed Messaging for Everyone*. Manning Publ., 2012, 312 p.
12. Carlson J.L. *Redis in Action*. Manning Publ., 2013, 320 p.

Авторы

Навроцкий Михаил Александрович – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 57190969016, ORCID ID: 0000-0003-2323-8196, m.navrotskiy@gmail.com

Жукова Наталия Александровна – кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 56406142300, ORCID ID: 0000-0001-5877-4461, nazhukova@mail.ru

Муромцев Дмитрий Ильич – кандидат технических наук, доцент, заведующий кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 55575780100, ORCID ID: 0000-0002-0644-9242, d.muromtsev@gmail.com

Мустафин Николай Габдрахманович – кандидат технических наук, доцент, старший научный сотрудник, СПИИРАН, Санкт-Петербург, 199178, Российская Федерация; профессор, Санкт-Петербургский государственный электротехнический университет (ЛЭТИ), Санкт-Петербург, 197376, Российская Федерация, Scopus ID: 56406142300, ORCID ID: 0000-0001-5986-7221, nikolay.mustafin@gmail.com

Authors

Mikhail A. Navrotskiy – postgraduate, ITMO University, Saint Petersburg 197101, Russian Federation, Scopus ID: 57190969016, ORCID ID: 0000-0003-2323-8196, m.navrotskiy@gmail.com

Natalia A. Zhukova – PhD, Associate Professor, ITMO University, Saint Petersburg 197101, Russian Federation, Scopus ID: 56406142300, ORCID ID: 0000-0001-5877-4461, nazhukova@mail.ru

Dmitry I. Mouromtsev – PhD, Associate Professor, Head of Chair, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 55575780100, ORCID ID: 0000-0002-0644-9242, d.muromtsev@gmail.com

Nikolay G. Mustafin – PhD, Associate Professor, Senior scientific researcher, SPIIRAS, Saint Petersburg, 199178, Russian Federation; Professor, Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197376, Russian Federation, Scopus ID: 56406142300, ORCID ID: 0000-0001-5986-7221, nikolay.mustafin@gmail.com