

Формирование новых экземпляров в онтологии научной и учебно-методической информации

Е. Ю. Шклярова, С. Ю. Землянская
Донецкий национальный технический университет, г. Донецк
E-mail: zsaae07@gmail.com

Аннотация

Данная научная статья представляет исследование, посвященное разработке и применению онтологии научной и учебно-методической информации. В статье рассматриваются важные аспекты развития онтологии, представлена методика формирования новых экземпляров в существующей модели. Этот процесс позволяет расширить онтологию, улучшить процессы поиска и анализа информации. Представлен программный модуль, разработанный с использованием библиотеки RDFLib, который позволяет автоматически создавать новые экземпляры на основе данных, извлеченных из научных публикаций.

Введение

В данной научной статье рассматривается онтология научной и учебно-методической информации [1]. Онтология представляет собой формальную спецификацию понятий, классов и связей между ними, которая помогает в описании и организации знаний в семантическом виде.

Одним из важных аспектов в развитии онтологии является формирование новых экземпляров в существующей модели. Это позволяет уточнить и расширить семантическую модель предметной области, обогатить онтологию новыми концептами и связями, а также улучшить процессы поиска и анализа информации. Формирование новых экземпляров может осуществляться путем автоматического извлечения данных из научных статей, учебников и других источников. Однако существуют ограничения и проблемы, связанные с неоднозначностью и разнообразием источников данных, сложностью извлечения информации из различных типов материалов, а также необходимостью постоянного обновления и поддержки онтологии.

Цель данной статьи заключается в представлении методики формирования новых экземпляров в онтологии научной и учебно-методической информации. Рассмотрены этапы анализа существующей онтологии и добавления данных в онтологию. Представлены инструменты и библиотеки, используемые для работы с онтологическими данными, и проведено сравнение некоторых из них.

Методика, описанная в статье, предоставит практические рекомендации и примеры формирования новых экземпляров в онтологии научной и учебно-методической информации, а также поможет преодолеть

возможные проблемы и сложности, связанные с этим процессом

Методика формирования новых экземпляров

Первый этап методики формирования новых экземпляров в онтологии заключается в анализе уже существующей онтологии. В ходе этого анализа определяются ее текущая структура, содержание концептов, связей и атрибутов (рис. 1). В онтологии присутствуют следующие классы в соответствии с потребностями, такими как определение статей автора, извлечение полезных данных из документов и выявление статей с определенной тематикой:

- класс «Человек» предназначен для идентификации авторов статей и связи их с соответствующими документами;

- класс «Документ» представляет собой основной объект, содержащий информацию о научных статьях. К нему могут быть привязаны различные свойства и атрибуты, такие как заголовок, ключевые слова, аннотация и т.д.;

- класс «Структурная часть документа» имеет абстрактный характер и используется только в качестве базового класса для наследования другими классами, связанными со структурой документа;

- класс «Дисциплина» представляет собой класс для описания дисциплины. Он используется для классификации статей по конкретным областям знаний;

- класс «Научное мероприятие» предназначен для идентификации событий, связанных с научной деятельностью, таких как конференции, семинары и прочие мероприятия

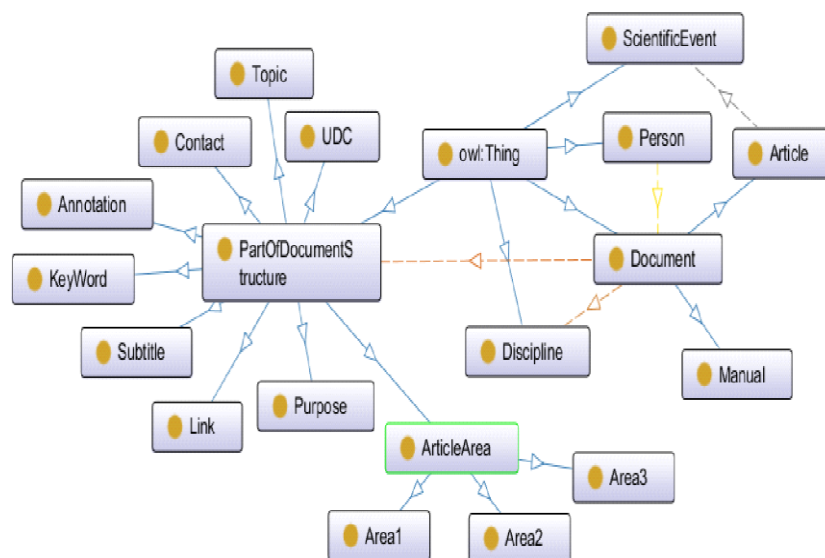


Рисунок 1 –Онтологическая модель кафедры вуза

Следующий этап включает процесс автоматического извлечения данных из научных статей. Это осуществляется с помощью методов обработки естественного языка. В процессе извлечения данных была получена информация, которая может быть использована для формирования новых экземпляров в онтологии, таких как имена авторов, УДК, список литературы, контактная информация, название статьи, аннотация, классификации и связи между понятиями.

После извлечения данных из текста статьи необходимо добавить эти данные в онтологию в виде конкретных экземпляров классов, так называемых индивидов. Каждый экземпляр обладает своими уникальными характеристиками и свойствами, которые могут быть определены в соответствии с его классом. Свойства определяют отношения и атрибуты, которыми обладает экземпляр.

Экземпляры создаются, модифицируются и удаляются в процессе работы с онтологией, что позволяет динамически расширять и изменять модель данных в соответствии с требованиями и новыми знаниями о предметной области. Для этого применяются алгоритмы сравнения и сопоставления, которые позволяют определить соответствия между извлеченными данными и существующими концептами в онтологии [2]. Это помогает установить, какие данные могут быть использованы для создания новых экземпляров или дополнения существующих концептов.

Алгоритм сравнения и сопоставления для добавления новых экземпляров в онтологию включает следующие шаги:

- создание графа для работы с онтологией и загрузка существующей онтологии;

- определение пространства имен для онтологии;
- создание нового индивида определенного класса;
- назначение уникального идентификатора и других свойств новому индивиду;
- проверка существования индивидов классов с заданными значениями;
- добавление отношений между индивидами при помощи свойств объекта;
- сохранение изменений в онтологии.

Алгоритм позволяет эффективно добавлять новые данные в онтологию и устанавливать связи между существующими и новыми индивидами, расширяя и обогащая семантическую модель предметной области.

Для работы с онтологиями можно использовать библиотеки Python, которые обеспечивают доступ к онтологическим данным и позволяют выполнять операции чтения, записи и модификации. Среди них можно выделить: RDFLib, Owlready2 и PyOWL. Проведем их сравнение

RDFLib [3] является одной из наиболее популярных библиотек для работы с RDF (Resource Description Framework) в Python. Она предоставляет функциональность для создания, чтения и записи RDF-графов, которые являются основой для представления онтологических данных. С помощью RDFLib можно устанавливать соединение с онтологией, извлекать данные, добавлять новые экземпляры и связи, а также выполнять запросы SPARQL для извлечения информации из онтологии.

Owlready2 [4] является еще одной популярной библиотекой для работы с онтологиями в Python. Она предоставляет простой и удобный интерфейс для доступа к

онтологическим данным, основанным на стандарте OWL (Web Ontology Language). Owlready2 позволяет создавать экземпляры классов, определять и изменять свойства объектов, а также выполнять запросы SPARQL-DL для извлечения информации из онтологии [5]. Библиотека также обеспечивает интеграцию с базами данных, такими как SQLite, для хранения и поиска данных.

PyOWL [6] – это Python-обертка для библиотеки OWL API, которая предоставляет функции для работы с онтологиями OWL в Python. OWL API является мощным инструментом для работы с онтологиями OWL и предоставляет поддержку различных версий OWL, таких как OWL 2 и OWL 1.1 [7].

Результаты сравнения библиотек приведены в табл. 1.

Таблица 1 – Сравнение библиотек для работы с онтологиями

Достоинства	Недостатки
RDFLib	
Широкий спектр функциональности, включая создание, чтение и запись RDF-графов, поддержку различных форматов сериализации RDF (таких как XML, JSON, N3), выполнение запросов SPARQL и тд	RDFLib имеет несколько сложный и громоздкий интерфейс, особенно для новичков в работе с RDF и онтологиями
Поддержка различных RDF-сериализаций и форматов	В некоторых случаях производительность RDFLib может быть ниже, чем у некоторых других библиотек
Хорошая документация и активное сообщество пользователей	
Owlready2	
Легкая в использовании библиотека с простым и понятным API	Могут возникать проблемы с производительностью при работе с большими онтологиями
Поддержка различных версий OWL (например, OWL 2) и возможность работы с различными типами онтологических объектов (классы, свойства, индивиды и т. д.)	Некоторые функциональности могут быть ограничены или недостаточно развиты, поскольку библиотека нацелена на простоту использования
Имеет возможность интеграции с другими библиотеками Python	
PyOWL	
Обеспечивает прямое взаимодействие с OWL API	Имеет более сложный API по сравнению с другими библиотеками, и его использование может потребовать более глубокого понимания OWL и его концепций
Обеспечивает широкий спектр функций, таких как чтение, запись, редактирование и вывод логики	

Все рассмотренные технологии являются достаточно функциональными для взаимодействия с онтологиями. Однако, учитывая несущественные недостатки, простоту использования, надежность и широкую поддержку в сообществе RDFLib, её выбор представляется предпочтительным. Эта библиотека предоставляет удобные инструменты для работы с онтологиями и позволяет эффективно обрабатывать RDF-графы.

С использованием библиотеки RDFLib был разработан программный модуль, который

позволяет формировать новые экземпляры концептов онтологии, таких как научные статьи, учебники, авторы, направления и других, представленных в общей структуре онтологической модели.

Для формирования новых экземпляров модуль использует семантические правила и шаблоны, определенные в онтологии. Это позволяет системе автоматически создавать новые экземпляры на основе доступных данных и связей в онтологии.

Для рассмотренной ранее онтологии мы сможем получать следующие элементы онтологии:

- Экземпляры классов. Модуль может извлекать конкретные экземпляры людей, такие как ФИО авторов статей, конкретных дисциплин, к которым относятся статьи, а также экземпляры частей документа, таких как УДК, аннотация, название, список литературы и т.д.

- Атрибуты экземпляров: свойства или характеристики, которые присущи каждому конкретному экземпляру.

- Информация о взаимодействии и связях между различными экземплярами объектов в онтологии. Например, модуль может извлекать информацию о связи между конкретной статьей и ее авторами (свойство "isAuthor"), между статьей и ее документом (свойство "hasDocument"), а также между частями документа и самим документом (свойство "PartOfDocument").

В процессе формирования новых экземпляров модуль использует проверки и ограничения, определенные в онтологии, чтобы гарантировать соблюдение правил и структуры данных. Это поможет избежать ошибок и обеспечить согласованность информации в онтологии.

Алгоритм сравнения и сопоставления для добавления новых экземпляров в онтологию может включать следующие шаги:

Шаг 1: Импорт и чтение онтологии

- Импортирование онтологии, с которой будут сопоставляться новые данные, с использованием RDFLib.

- Чтение онтологии для получения существующих классов и свойств.

Шаг 2: Сопоставление данных с онтологией

- Итерация по каждому триплету из новых данных и проверка существования сопоставляемых концептов и свойств в онтологии.

- Если сопоставление не найдено, создание новых экземпляров и связей на основе данных из триплета.

- Если сопоставление найдено, обновление существующих экземпляров или связей в соответствии с данными из триплета.

Шаг 3: Запись изменений в онтологию

- Запись изменений и добавленных экземпляров в онтологию с использованием RDFLib.

Шаг 4: Проверка согласованности и корректности данных

- Проверка добавленных экземпляров и связей на соответствие правилам и ограничениям, заданным в онтологии.

- Если данные не соответствуют правилам, применение соответствующих

корректировок или уведомление о возможных проблемах.

Рассмотрим пример формирования новых экземпляров онтологии при добавлении информации о статье «Проектирование автоматизированной системы онлайн-поиска попутчиков» [8], опубликованной в сборнике материалов конференции ИСУКМ-2021.

После извлечения данных модуль добавляет новые экземпляры в онтологию. Перечень добавленных экземпляров, сформированный в результате извлечения данных, представлен на рис. 2, а подробная информация о каждом новом экземпляре – в табл. 2. После добавления экземпляров производится проверка данных и связей для обеспечения их корректности и соответствия заданной онтологической структуре.

На рис. 3 приведена подробная информация об элементах, появившихся в онтологии при добавлении статьи.



Рисунок 2 – Новые добавленные экземпляры онтологии

Для проверки правильности результатов после добавления экземпляров в онтологию, мы можем выполнить DL-запросы (Description Logic queries) [9, 10]. На рис. 4.а показан результат выполнения запроса DL-Query для определения авторов добавленной статьи. Этот запрос помогает убедиться, что связи между статьей и ее авторами были установлены корректно.

Таблица 2 – Информация о новых экземплярах

Наименование	Тип	Свойство данных	Значение свойства
udc1	UDC	udcValue	007.51
author2	Person	firstName	С
		lastName	Землянская
		middleName	Ю
author3		firstName	Е
		lastName	Мащенко
		middleName	Н
annotation1	Annotation	annotationValue	Шклярова Е. Ю., Землянская С. Ю., Мащенко Е. Н. Проектирование автоматизированной системы онлайн-поиска попутчиков. В статье обосновывается актуальность разработки автоматизированной системы онлайн-поиска попутчиков. Приведено детальное описание архитектуры разрабатываемой системы. Особое внимание уделено проектированию информационной и функциональной модели системы. Для демонстрации возможностей системы представлен прототип, базирующийся на предложенной архитектуре, рассмотрены дальнейшие направления развития системы.
key1	KeyWord	keyWordValue	система
key2	KeyWord	keyWordValue	попутчик
key3	KeyWord	keyWordValue	водитель
key4	KeyWord	keyWordValue	пользователь
key5	KeyWord	keyWordValue	поездка
key6	KeyWord	keyWordValue	поезд
link1	Link	linkValue	BlaBlaCar [Электронный ресурс] URL: https://www.blablacar.com.ua
link2	Link	linkValue	Доедем вместе! [Электронный ресурс] URL: http://www.doedemvmeste.ru
link3	Link	linkValue	Довезу! [Электронный ресурс] URL: http://www.dovezu.ru/
link4	Link	linkValue	CASE-средства для разработки информационных систем / Маклаков С.В – М.:Диалог-МИФИ, 1999.-256 с.
link5	Link	linkValue	Использование методологии моделирования IDEF при формировании структурно - параметрической модели реализации технологий обеспечения эффективного развития промышленных предприятий в условиях постиндустриальной экономики [Текст] / Тебекин А.В. // Вестник Российской таможенной академии. 2015. № 4 (33). С. 96 -103.
contact1	Contact	contactValue	zsaac07@gmail.com
area1	Area1	articleAreaValue	Информационные системы

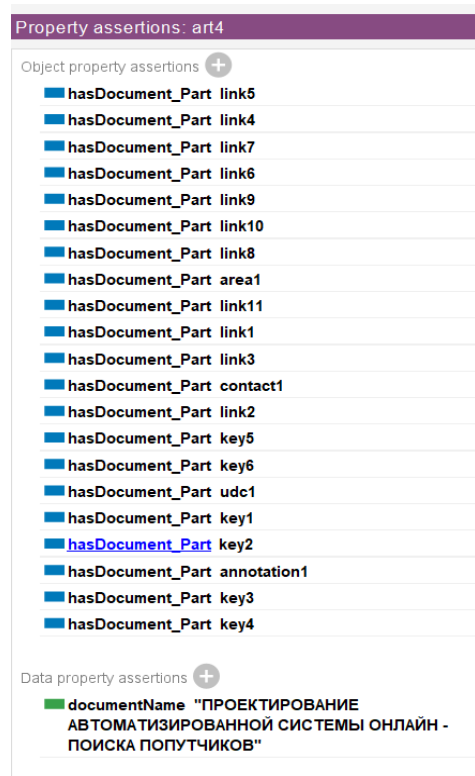
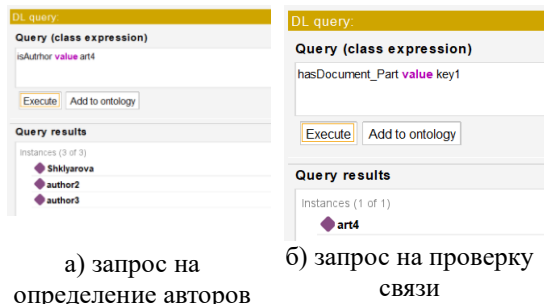


Рисунок 3 – Подробная информация о новых элементах онтологии при добавлении статьи

Далее, на рис. 4.б, представлен запрос, который проверяет правильность установления связей между ключевым словом "key1" и статьями, содержащими это ключевое слово. Результат выполнения запроса позволяет убедиться, что все соответствующие статьи были связаны правильно с ключевым словом.

Наконец, на рис. 5 выполнен запрос, который выводит всех авторов, содержащихся в онтологии. Этот запрос (Person) возвращает список URI всех объектов класса «Человек», позволяя убедиться, что все авторы были успешно добавлены в онтологию.

Результаты выполнения запросов подтверждают правильность работы алгоритмов добавления экземпляров и установления связей в онтологии.



а) запрос на определение авторов

б) запрос на проверку связи

Рисунок 4 – DL-запросы для проверки правильности добавления

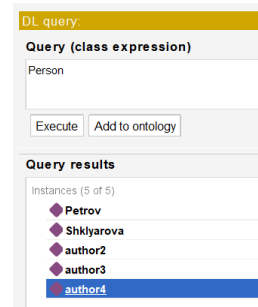


Рисунок 5 – DL-запрос, который выводит всех авторов

Таким образом, алгоритм позволяет расширить онтологию научно-учебных материалов, добавляя новые экземпляры и связи на основе данных из статей, обеспечивая корректность и согласованность с ожидаемой структурой и содержимым онтологии. Модуль для онтологии научно-учебных материалов позволит извлекать экземпляры классов, атрибуты экземпляров и информацию о связях между экземплярами, обогащая онтологию и давая более полное представление о предметной области.

Выводы

В статье была представлена методика формирования новых экземпляров в онтологиях, которая позволяет расширять и обогащать онтологии с учетом новых данных и связей. Процесс автоматического извлечения данных из научных статей и добавления их в онтологию с использованием семантических правил и шаблонов обеспечивает эффективное и надежное обновление онтологической модели. Проверка правильности результатов с помощью DL-запросов позволяет убедиться в корректности добавления экземпляров и установления связей. Эта методика имеет практическое применение для расширения онтологий и обеспечения точной и согласованной информации в информационных системах

Литература

1. Андриевская, Н. К. Онтологический подход в системах обработки данных научных и научно-образовательных организаций / Н. К. Андриевская // Проблемы искусственного интеллекта. – 2020. – № 1(16). – С. 23-36. – EDN WIZMRV.
2. Bird S., Klein E., Loper E. Natural Language Processing with Python // O'Reilly Media, 2009. – URL: https://www.researchgate.net/publication/220691633_Natural_Language_Processing_with_Python (дата посещения: 20.12.2022).

3. Официальная документация RDFLib. – URL: <https://rdflib.readthedocs.io/> (дата посещения: 20.02.2023).
4. Официальная документация Owlready2. – URL: <https://owlready2.readthedocs.io/> (дата посещения: 20.02.2023).
5. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M. *Ontologies and Databases: The DL-Lite Approach* // Cambridge University Press, 2017. – PP. 255-356.
6. Репозиторий PyOWL на GitHub. – URL: <https://github.com/NeuralVedant/PyOWL> (дата посещения: 20.02.2023).
7. Corcho, O., Fernández-López, M., Gómez-Pérez, A. *Methodologies, tools and languages for building ontologies: Where is their meeting point?* // Data & knowledge engineering, 2003. 46(1). – PP. 41-64. – URL: <https://typeset.io/papers/methodologies-tools-and-languages-for-building-ontologies-3o724oh3up> (дата посещения: 06.06.2023).
8. Шклярова, Е. Ю. Проектирование автоматизированной системы онлайн-поиска попутчиков / Е. Ю. Шклярова, С. Ю. Землянская, Е. Н. Машенко // Информатика, управляющие системы, математическое и компьютерное моделирование (ИУСМКМ-2021) : Материалы XII Международной научно-технической конференции в рамках VII Международного Научного форума Донецкой Народной Республики к 100-летию ДонНТУ, Донецк, 26–27 мая 2021 года. – Донецк: Донецкий национальный технический университет, 2021. – С. 115-122. – EDN VVRAMZ.
9. Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E. *Ontology Engineering in a Networked World* // Ontology Engineering in a Networked World, 2012. – PP. 213-233.
10. Платонов А. В., Полещук Е. А. Методы автоматического построения онтологий. – URL: <https://cyberleninka.ru/article/n/metody-avtomaticheskogo-postroeniya-ontologiy/viewer> (дата обращения: 06.06.2023).

Шклярова Е. Ю., Землянская С. Ю. Формирование новых экземпляров в онтологии научной и учебно-методической информации. Данная научная статья представляет исследование, посвященное разработке и применению онтологии научной и учебно-методической информации. В статье рассматриваются важные аспекты развития онтологии, представлена методика формирования новых экземпляров в существующей модели. Этот процесс позволяет расширить онтологию, улучшить процессы поиска и анализа информации. Представлен программный модуль, разработанный с использованием библиотеки RDFLib, который позволяет автоматически создавать новые экземпляры на основе данных, извлеченных из научных публикаций.

Ключевые слова: онтология, модификация онтологии, семантическая модель, новые экземпляры, концепт, сопоставление, целостность методика

Shklyarova E., Zemlyanskaya S. Formation of New Instances in the Ontology of Scientific and Educational Methodological Information. This research article presents a study dedicated to the development and application of an ontology for scientific and educational methodological information. The article examines important aspects of ontology development and introduces a methodology for forming new instances within an existing model. This process allows for the expansion of the ontology and improvement of information search and analysis processes. A software module, developed using the RDFLib library, is presented, which enables the automatic creation of new instances based on data extracted from scientific publications.

Keywords: ontology, ontology modification, semantic model, new instances, concepts, mapping, integrity, methodology.

Статья поступила в редакцию 21.05.2023.
Рекомендована к публикации профессором Скобцовым Ю.А.