УДК 004.822:004.89:398

ПОДХОД К ПОСТРОЕНИЮ ИНФОРМАЦИОННОЙ ИССЛЕДОВАТЕЛЬСКОЙ СРЕДЫ ДЛЯ АНАЛИЗА НАУЧНЫХ МАТЕРИАЛОВ И СОЗДАНИЯ АННОТИРОВАННЫХ КОРПУСОВ

Серый Алексей Сергеевич
м.н.с., Институт систем информатики им. А.П. Ершова СО РАН,
e-mail: alexey.seryj@iis.nsk.su
630090 г. Новосибирск, проспект Академика Лаврентьева, 6
Гриневич Анна Александровна
к.ф.н., н.с., Институт филологии СО РАН,
e-mail: anngrinevich@gmail.com
630090 г. Новосибирск, ул. Николаева, 8
Лисин Владислав Александрович
Аспирант, Новосибирский Государственный Университет,
e-mail: vladlisin2@gmail.com
630090 г. Новосибирск, ул. Пирогова, 1.

Аннотация. В статье предложен подход к построению исследовательской среды для интеграции информационных ресурсов определенной области знаний и поддержки научных исследований. Особенностью подхода является комбинация в рамках единой информационной системы, основанных на онтологиях средств представления, систематизации и аннотирования интегрированных в систему ресурсов, а также ориентация на совместную работу специалистов над созданием размеченного корпуса. В статье приведен пример применения предложенного подхода для разработки информационной системы.

Ключевые слова: Semantic Web, онтология, информационный ресурс, аннотированный корпус.

Цитирование: Серый А.С., Гриневич А.А., Лисин В.А. Подход к построению информационной исследовательской среды для анализа научных материалов и создания аннотированных корпусов // Информационные и математические технологии в науке и управлении. 2021. № 4 (24). С. 68-76. DOI: 10.38028/ESI.2021.24.4.007.

Введение. В настоящее время имеется запрос на структурирование и формализацию данных в различных научных предметных областях и их представление в общем доступе. В одних областях размер имеющихся данных слишком велик, и аналитических способностей человека недостаточно для работы с ним. В других – большая часть материалов хранится в архивах или распределена по труднодоступным источникам. Кроме того, современная практика научных исследований, как правило, предполагает взаимодействие ученых, работающих в одной или смежных областях науки, обмен данными и результатами исследований. Соответственно, в обоих перечисленных выше случаях требуются систематизация и интеграция информационных ресурсов определенной области знаний в общее научное пространство.

В качестве средств представления знаний во многих случаях применяются онтологические модели предметной области. В последние годы онтологии все чаще применяются и при разработке информационных систем как основа модели данных [1]. В тех областях науки, где информационные ресурсы выступают предметом исследования, как, например, в набирающих популярность исследованиях аргументации [2, 3], помимо систематизации ресурсов требуются также средства их анализа и аннотирования. В рамках различных проектов разработаны инструменты для создания аннотированных корпусов, где размечаются вхождения сущностей и понятий соответствующей предметной области. При

этом предметная область, сущности которой применяются при аннотировании информационных ресурсов, также может быть формализована в виде онтологии. Размеченный корпус, сам по себе, являясь важным научным результатом, может быть использован в качестве обучающих данных в задачах машинного обучения для извлечения сущностей предметной области из неразмеченных материалов.

В статье предлагается подход к разработке информационной исследовательской среды, объединяющей средства формализации совокупности входящих в нее ресурсов и средства создания аннотированных корпусов на основе онтологий научных предметных областей.

Статья организована следующим образом. В разделе 1 приведено описание подхода к разработке информационной исследовательской среды (ИИС). Раздел 2 посвящен описанию данных ИИС и методам их хранения. В разделе 3 показан пример применения предлагаемого подхода при разработке информационной системы. В заключении подводятся итоги и намечаются дальнейшие перспективы исследования.

1. Подход к разработке ИИС. В соответствии с предлагаемым подходом, разработанная на его основе информационная система (ИС) должна решать следующие основные задачи: создание и поддержка репозитория информационных ресурсов, аннотирование – разметка вхождений элементов предметных онтологий в ресурсах ИС, создание и поддержка предметных онтологий на уровне простых операций, обеспечение доступа к ресурсам и данным, навигация.

Для решения поставленных задач в рамках ИС различаются два вида онтологий. Мета-онтология обеспечивает формальное описание коллекции информационных ресурсов и формирует основу модели данных для пользовательского интерфейса, навигации и поиска. Предметная онтология является основой для аннотирования содержащихся в ИС ресурсов. Чаще всего аннотируемыми ресурсами являются тексты, однако размеченный корпус может включать и другие материалы, такие, как аудио и видеозаписи. В зависимости от предметной области и стоящих перед исследователями задач, ИС может содержать одну или несколько предметных онтологий. При этом, мета-онтология всегда единственная.

В соответствии с основными задачами ИС ее архитектура предполагает разделение функциональности на четыре блока или модуля: Предметный блок (**ПБ**), Блок репозитория (**РБ**), Блок аннотирования (**АБ**) и Блок пользовательского интерфейса (**БИ**). На рис. 1 представлена общая архитектура ИС, разработанной в рамках предлагаемого подхода.

Архитектура ИС предполагает разделение пользователей на три группы. Обычные пользователи или *Гости* имеют доступ на чтение ко всем ресурсам системы. Гостям также доступны поисковые механизмы. Задача *Исследователей* состоит в пополнении и поддержке репозитория ресурсов, а также в создании размеченных корпусов. Третья группа пользователей – *Эксперты предметной области*, или просто *Эксперты*, занимаются разработкой и поддержкой предметных онтологий. Группы исследователей и экспертов могут пересекаться. Как показывает практика, пользователь-эксперт в большинстве случаев принимает участие и в создании аннотированных корпусов, т.е. является также исследователем.

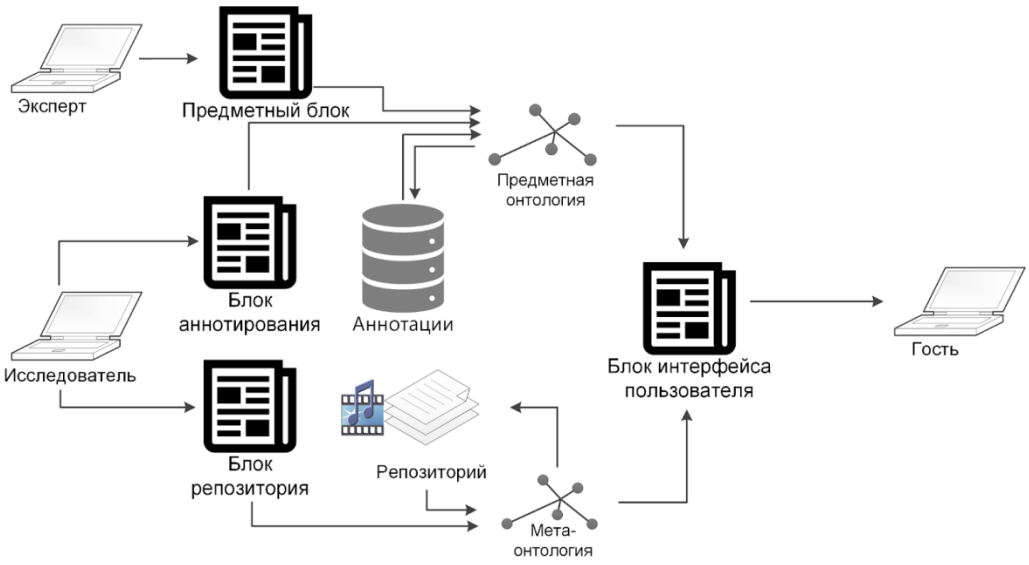

Рис.1. Архитектура информационной системы.

Предметный блок реализует инструменты для работы с предметными онтологиями. Пользователями функциональности **ПБ** являются пользователи-эксперты. Блок аннотирования и Блок репозитория предоставляют пользователям-исследователям возможности создавать, редактировать и аннотировать ресурсы системы, создавая аннотированные корпуса. Согласно рис. 1, **АБ** и **РБ** не могут обращаться к хранилищу данных напрямую, все взаимодействие организовано через соответствующие онтологии. Исследователи аннотируют ресурсы, сопоставляя сущности выбранной предметной онтологии определенным фрагментам. Множество размеченных фрагментов образует аннотацию ресурса. Предполагается, что один ресурс может иметь несколько альтернативных аннотаций.

Блок интерфейса обеспечивает работу графического пользовательского интерфейса, поисковых механизмов и навигации.

2. Разработка хранилища данных. Данные в ИИС представлены репозиторием ресурсов, аннотациями, предметными онтологиями и мета-онтологией. В зависимости от размера системы и количества пользователей существует ряд альтернативных подходов к организации хранилища. В данной работе рассмотрены технологии хранения данных, опробованные авторами при разработке информационных систем.

2.1. Репозиторий ресурсов и аннотации. Репозиторий ресурсов – это хранилище текстовых и мультимедийных материалов. При небольшом размере репозитория файлы могут быть расположены непосредственно в файловой системе.

Более масштабируемым решением является использование документоориентированных хранилищ данных. Одним из наиболее популярных и производительных является MongoDB [4]. Спецификация GridFS позволяет хранить файлы размером более 16 МБ. MongoDB обеспечивает разделение данных, что позволяет поддерживать почти неограниченный репозиторий ресурсов.

2.2. Онтология. Наименее тривиальной задачей при проектировании хранилища данных ИС является организация хранения онтологий. Стандартными средствами в данной области являются хранилища триплетов, такие, как Jena Fuseki [5] и OpenLink Virtuoso [6] с работой через точку доступа и языком запросов SPARQL. Некоторые хранилища триплетов содержат встроенную машину логического вывода. Подобные решения обеспечивают сохранение онтологии в формате RDF-триплетов и доступ к ней посредством стандартных инструментов.

Тем не менее, во многих случаях хранилища триплетов не обладают производительностью, достаточной для обеспечения работы информационной системы в режиме реального времени и в условиях, когда требуется частое обновление данных. В большей степени это верно для SPARQL-точек доступа. Помимо производительности, существуют и другие недостатки, такие, как ограниченная поддержка транзакций, меньшие по сравнению с другими СУБД возможности масштабирования и т.д. По этой и другим причинам многие специалисты в области Semantic Web переходят на альтернативные способы хранения онтологий и связанных данных.

Одной из наиболее привлекательных альтернатив являются графовые базы данных, т.к. онтология хорошо представима в виде графа. Графовые БД обеспечивают высокую производительность и возможности для масштабирования.

При разработке ИС в рамках предлагаемого подхода мы использовали для хранения онтологий графовую базу данных Neo4j [7]. СУБД Neo4j представляет данные в виде помеченного гиперграфа (Labeled Property Graph, LPG), обладает хорошей производительностью, предоставляет удобный интерфейс администратора, простой и мощный язык запросов Cypher. На рис. 2 показаны результаты сравнения производительности хранилища триплетов Jena Fuseki 3.17.0 и СУБД Neo4j 4.2.5 на наиболее частых запросах. Время в миллисекундах.

- Взять дерево классов онтологии (на схеме обозначен как **GT**).
- Взять свойства заданного класса (**GCP**).
- Взять связи заданного класса (**GCR**).
- Взять экземпляры заданного класса (классов) (**GCI**).
- Взять N объектов со сдвигом M (**GNM**).
- Взять свойства заданного экземпляра (**GIP**).
- Поиск экземпляра по заданным значениям свойств (**SI**).

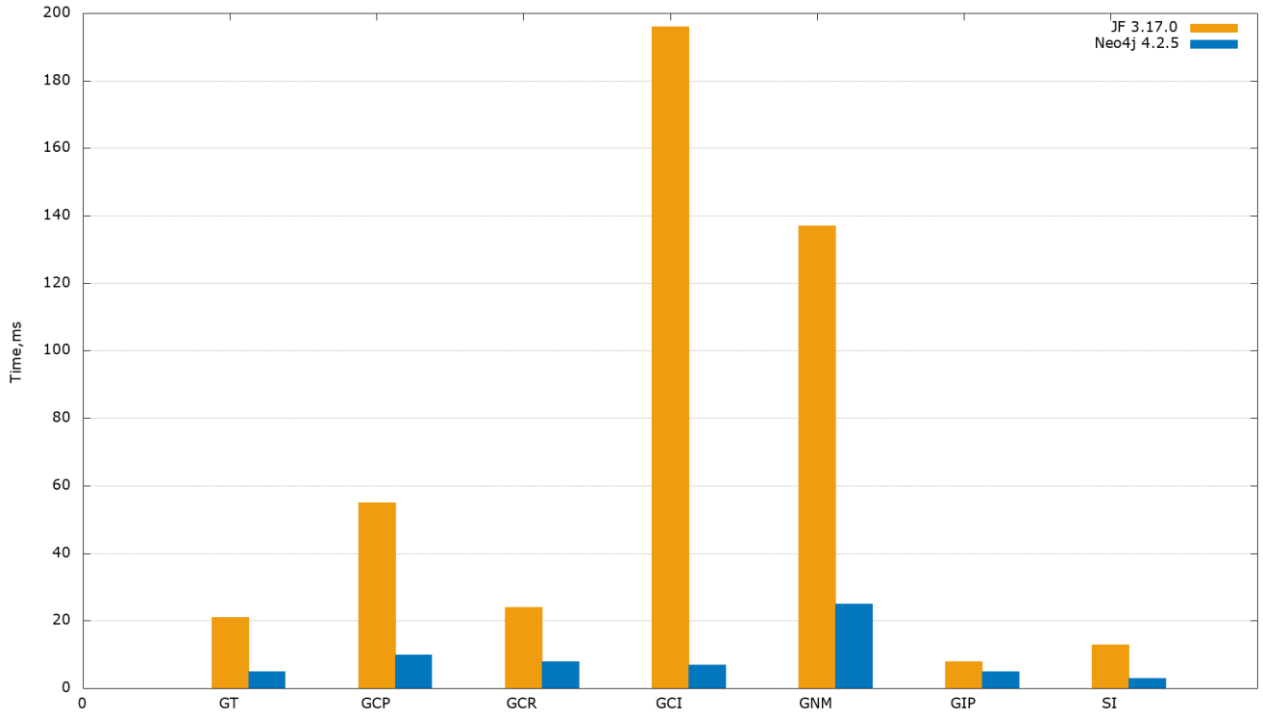

Рис. 2. Сравнительный анализ производительности хранилищ данных.

Сравнение проводилось на компьютере с 32 ГБ RAM и процессором Intel Core i7 на онтологии размером 63025 триплетов.

Из рис. 2 очевидно, что производительность графовой СУБД в большинстве случаев существенно выше, чем у хранилища триплетов. Помимо этого, для СУБД Neo4j имеется большое количество поддерживаемых сообществом драйверов для большинства самых распространенных языков программирования. Свободно распространяемая версия позволяет хранить граф размером до 34 миллиардов вершин. Имеется возможность распределенного хранения больших графов (data sharding). В настоящее время сообщество и разработчики Neo4j активно вовлечены в исследования в области Semantic Web, результатом которых стала разработка официально поддерживаемого плагина Neosemantic [8], предназначенного для работы с онтологиями и связанными данными. В Neo4j предусмотрена настройка SPARQL-точки доступа на основе графовой БД, таким образом, сохраняется возможность предоставления доступа к онтологиям в соответствии с принципами Linked Data.

3. Применение подхода при разработке информационных систем. Предлагаемый подход применяется при разработке информационной системы для поддержки исследований фольклора коренных народов Сибири и Дальнего Востока. В настоящее время в фольклористике существует социальный заказ на оцифровку и доступ в сети Интернет к накопленным материалам, которые хранятся в личных архивах исследователей или разрознены по труднодоступным малотиражным изданиям. В общемировой практике также наблюдается повышенный интерес к применению онтологий в гуманитарной области, в частности, в культурологии и фольклористике. Многие исследователи работают над тем, чтобы оцифровать и представить в общем доступе культурное наследие своего народа [9]. Архивы, музеи и библиотеки предоставляют свои данные в общий доступ, в том числе в виде RDF-триплетов [10].

Систематизация и интеграция фольклорных ресурсов в общее научное пространство – до сих пор нерешенная задача сибирской фольклористики. Создаваемая ИС предназначена для построения исследователями-фольклористами аннотированного корпуса фольклорных текстов с размеченными универсальными культурными константами (универсалиями). Здесь культурные универсалии понимаются в первую очередь как концепты, формирующие картину мира сибирских этносов [11].

Мета-онтология была разработана на основе CIDOC-CRM [12] – формальной модели, созданной для поддержки интеграции, стандартизации и обмена информацией в области культурного наследия. Данная модель является зарегистрированным стандартом ISO [13].

CIDOC-CRM спроектирована таким образом, что основными концептами являются события, связанные со временем действия и местами, в которых они произошли. При разработке мета-онтологии мы выражаем при помощи событий процессы сбора и подготовки фольклорных произведений. Событие представляется своими временем, местом действия, актором и результатом. Акторами являются персоны, задействованные в процессе подготовки материалов: собиратели, переводчики, люди, ответственные за расшифровки аудиозаписей и т.д. В процессе разработки мы добавили некоторые свойства, отсутствующие в оригинальной онтологии CIDOC-CRM для более естественного представления информации. Некоторые из этих свойств были импортированы из онтологий FOAF, DBpedia и GeoNames, остальные – это новые свойства, созданные для данной онтологии. На рис. 3 показан пример описания процесса подготовки фольклорного произведения в виде совокупности событий.

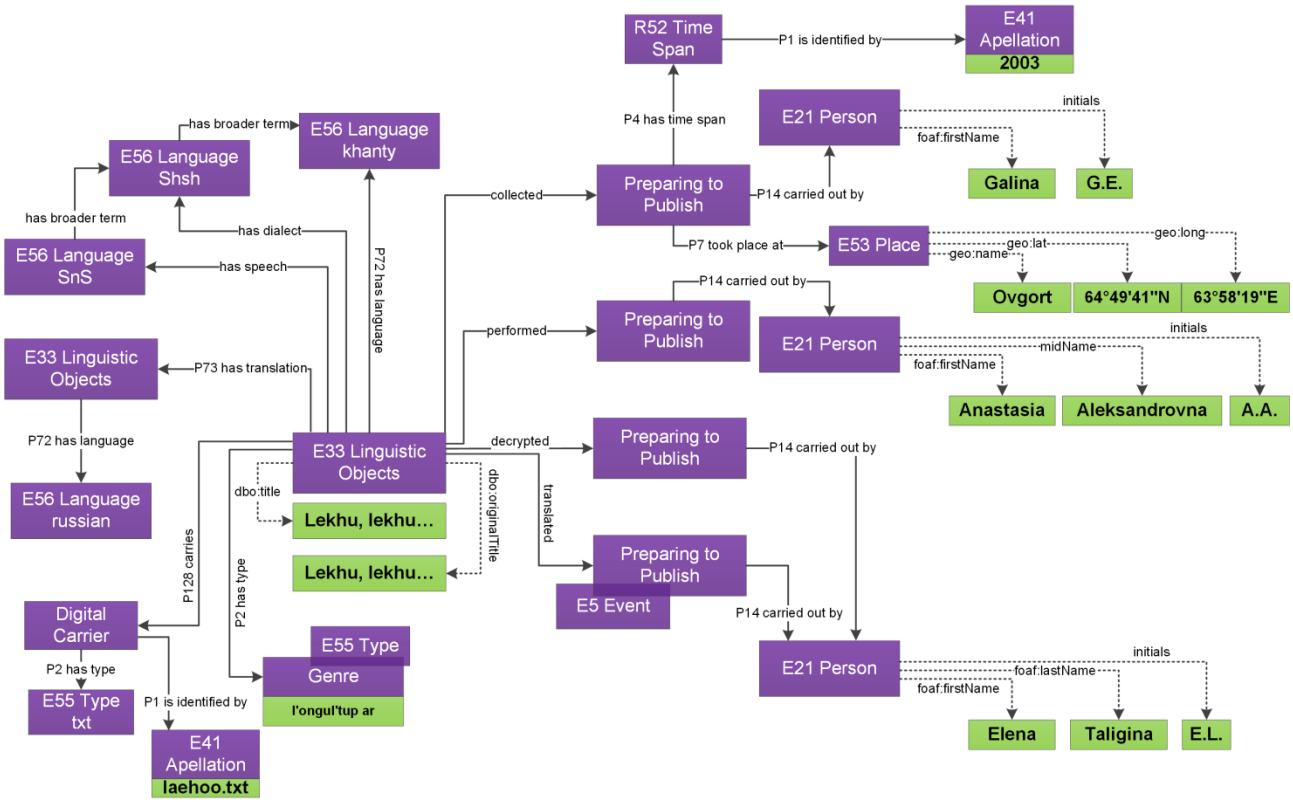

Рис. 3. Представление фольклорного произведения в мета-онтологии на основе стандарта CIDOC.

CIDOC-CRM спроектирована таким образом, что основными концептами являются события, связанные со временем действия и местами, в которых они произошли. При разработке мета-онтологии мы выражаем при помощи событий процессы сбора и подготовки фольклорных произведений. Событие представляется своими временем, местом действия, актором и результатом. Акторами являются персоны, задействованные в процессе подготовки материалов: собиратели, переводчики, люди, ответственные за расшифровки аудиозаписей и т.д. В процессе разработки мы добавили некоторые свойства, отсутствующие в оригинальной онтологии CIDOC-CRM для более естественного представления информации. Некоторые из этих свойств были импортированы из онтологий FOAF, DBpedia и GeoNames, остальные – это новые свойства, созданные для данной онтологии. На рис. 3 показан пример описания процесса подготовки фольклорного произведения в виде совокупности событий.

Для аннотирования текстов была разработана онтология предметной области, содержащая описания лингво-культурологических концептов, распространенных на территории сибирского макрорегиона.

Разработка данной ИС нацелена на формализацию предметной области и интеграцию разрозненных фольклорных материалов на языках сибирских этносов, накопленных за долгое время. Авторы полагают, что такая система может расширить спектр доступных ресурсов и предоставить специалистам новые возможности по анализу культур различных народов, подняв, таким образом, исследования сибирского фольклора на новый уровень.

Заключение. Цель данной работы – описание подхода к разработке информационной исследовательской среды для интеграции научных материалов и создания аннотированных корпусов на основе онтологий научных предметных областей. В работе приведено описание подхода и намечены основные пути его реализации.

В будущем планируется создание на основе предложенного подхода платформы для разработки информационных исследовательских сред. Платформа будет предоставлять базис для разработки ИИС в виде хранилища данных, редактора предметных онтологий, инструментов разметки и базовых функций пользовательского интерфейса.

Благодарности. Работа выполнена при финансовой поддержке РФФИ и Правительства Новосибирской области в рамках Проекта № 20-412-540001.

СПИСОК ЛИТЕРАТУРЫ

1. Загорулько Ю.А., Загорулько Г.Б., Боровикова Ю.А. Технология создания тематических интеллектуальных научных интернет-ресурсов, базирующаяся на онтологии // Программная инженерия. 2016. Т. 7. № 2. С. 51-60. DOI: 10.17587/prin.7.51-60.
2. Online Visualization of Argument. Режим доступа: http://ova.arg-tech.org/ (дата обращения: 01.12.2021).
3. Zagorulko Y., Garanina N., Sery A., Domanov O. Ontology-Based Approach to Organizing the Support for the Analysis of Argumentation in Popular Science Discourse. In: Kuznetsov S., Panov A. Artificial Intelligence. RCAI 2019. Communications in Computer and Information Science. Vol. 1093. Springer. Cham. Pp. 348-362.
4. MongoDB. Режим доступа: https://www.mongodb.com/ (дата обращения: 01.12.2021).
5. Jena Fusek. Режим доступа: https://jena.apache.org/documentation/fuseki2/index.html (дата обращения: 01.12.2021).
6. OpenLink Virtuoso Homepage. Режим доступа: https://virtuoso.openlinksw.com/ (дата обращения: 01.12.2021).
7. Neo4j Homepage. Режим доступа: https://neo4j.com/ (дата обращения: 01.12.2021).
8. Neosemantics Homepage. Режим доступа: https://neo4j.com/labs/neosemantics/ (дата обращения: 01.12.2021).
9. Hyvönen E., Mäkelä E., Kauppinen T. el al: CultureSampo – Finnish culture on the Semantic Web 2.0. Thematic perspectives for the end-user. In Proceedings, Museums and the Web. Pp. 15-18.
10. Marden J., Li-Madeo C., Whysel N., Edelstein J. Linked Open Data for Cultural Heritage: Evolution of Information Technology. In Proceedings of the 31st ACM international conference on Design of communication. Pp. 107-112.
11. Степанов Ю.С. Константы: Словарь русской культуры. Москва.: Издательство «Академический проект». 2001. 990 с.
12. CIDOC Conceptual Reference Model. Режим доступа: http://www.cidoc-crm.org/ (дата обращения: 01.12.2021).
13. CIDOC-CRM ISO page. Режим доступа: https://www.iso.org/standard/57832.html (дата обращения: 01.12.2021).

UDC 004.822:004.89:398

AN APPROACH TO DEVELOPING AN INFORMATION RESEARCH ENVIRONMENT FOR ANALYSIS OF SCIENTIFIC INFORMATION RESOURCES AND CREATING ANNOTATED CORPORA

Alexey S. Sery
Junior Researcher, A.P. Ershov Institute of Informatics Systems of SB RAS
e-mail: alexey.seryj@iis.nsk.su
630090, Novosibirsk, Russia, pr. Acad. Lavrentjev 6
Anna A. Grinevich
Ph.D., Researcher, Institute of Philology of SB RAS
e-mail: anngrinevich@gmail.com
630090, Novosibirsk, Russia, Nikolaeva st. 8
Vladislav A. Lisin
Postgraduate, Novosibirsk State University
e-mail: vladlisin2@gmail.com
630090, Novosibirsk, Russia, Pirogova st., 1.

Abstract. The paper presents an approach to the development of a research environment, facilitating an integration of information resources dedicated to a certain scientific domain and supporting scientific research. The main feature of the approach is combining an ontology-based tools for presenting and annotating scientific information resources within a single information system. The development of the information system is aimed towards the joint work of researchers on the creating annotated corpora of resources. The paper provides an example of the proposed approach being put into practice when developing an information system.

Keywords: Semantic Web, ontology, information system, annotated corpora

Acknowledgements: The paper was prepared based on the results of a study conducted as part of the projects of the Russian Foundation for Basic Research No. 20-412-540001.

REFERENCES

1. Zagorulko Yu.A., Zagorulko G.B., Borovikova O.I. Tekhnologija sozdanija tematicheskikh intellektual'nykh nauchnykh internet-resursov, bazirujushhajasja na ontologii [Technology for building subject-based intelligent scientific internet resources based on ontology] // Programmnaja inzhenerija – Software Engineering. 2016. No. 2. Pp. 51-60. (in Russian). DOI: 10.17587/prin.7.51-60.
2. Online Visualization of Argument. Available at: http://ova.arg-tech.org/ (accessed 01.12.2021).
3. Zagorulko Y., Garanina N., Sery A., Domanov O. Ontology-Based Approach to Organizing the Support for the Analysis of Argumentation in Popular Science Discourse. In: Kuznetsov S., Panov A. Artificial Intelligence. RCAI 2019. Communications in Computer and Information Science. Vol. 1093. Springer. Cham. Pp. 348-362.
4. MongoDB. Available at: https://www.mongodb.com/ (accessed 01.12.2021).
5. Jena Fusek. Available at: https://jena.apache.org/documentation/fuseki2/index.html (accessed 01.12.2021).
6. OpenLink Virtuoso Homepage. Available at: https://virtuoso.openlinksw.com/ (accessed 01.12.2021).

7. Neo4j Homepage. Available at: https://neo4j.com/ (accessed 01.12.2021).
8. Neosemantics Homepage. Available at: https://neo4j.com/labs/neosemantics/ (accessed 01.12.2021).
9. Hyvönen E., Mäkelä E., Kauppinen T. el al: CultureSampo – Finnish culture on the Semantic Web 2.0. Thematic perspectives for the end-user. In Proceedings, Museums and the Web. Pp. 15-18.
10. Marden J., Li-Madeo C., Whysel N., Edelstein J. Linked Open Data for Cultural Heritage: Evolution of Information Technology. In Proceedings of the 31st ACM international conference on Design of communication. Pp. 107-112.
11. Stepanov Yu.S. Konstanty: Slovar' russkoj kul'tury [Constants: The dictionary of the Russian culture] Moskva : Akademicheskij proekt= Academic project. 2001. 990 p. (in Russian).
12. CIDOC Conceptual Reference Model. Available at: http://www.cidoc-crm.org/ (accessed 01.12.2021).
13. CIDOC-CRM ISO page. Available at: https://www.iso.org/standard/57832.html (accessed 01.12.2021)

Статья поступила в редакцию 06.12.2021; одобрена после рецензирования 15.12.2021; принята к публикации 20.12.2021.

The article was submitted 06.12.2021; approved after reviewing 15.12.2021; accepted for publication 20.12.2021.