DOI:10.33942/sit.nes004 УДК 004.9

ИШТЕП ЧЫГУУ ТЕХНОЛОГИЯСЫ АЛУУ МЕТАДАННЫХ ДОКУМЕНТТЕРДИН ИЧИНЕН ПРОЦЕССИНДЕ МААЛЫМАТТЫК СИСТЕМАЛАР ИНТЕГРАЦИЯЛОО

Садырмекова Ж.Б.[1], Самбетбаева М.А.[2]

[1]*Евразия улуттук университети Л. Н.Гумилева НурСултан, Казакстан, janna_1988@mail.ru,*
[2]*Маалыматтык-эсептөө технологиялар институту, Алматы, Казакстан, madina_jgtu@mail.ru*

Издөө: *Маалыматтык колдоо системасын илимий изилдөөлөрдү оперируют менен публикациями, электрондук документтер жана коллекциями, онтологическими жазуулар. Мындай ресурстар издөө жана идентификациялык көйгөйлөрдүн айынан жетүүгө кыйын болушу мүмкүн. Семантикалык маалыматтык ресурстардын ортосундагы байланыштар алардын маанилүүлүгүн жогорулатат жана маалыматтык издөө жана идентификациялоо үчүн кошумча мүмкүнчүлүктөрдү берет.*

Изилдөөнүн жаңылыгы түзүлөт аракет жасаган интеграциялоо жана өркүндөтүү, бир катар актуалдуу жана келечектүү технологияларды түзүү үчүн жаңы муундагы маалымат системаларын колдоо илимий-билим берүү иши. Берененин базиси (апробация) апробацияланган ыкмаларды жана ыкмаларды түзөт, алар аркылуу жалпы милдетти коюу гана эмес, айрым жаңы элементтерди бөлүп көрсөтүү да мүмкүн болот.

Системасын түзүүнүн максаты болуп саналат түзүү фактографической базасын изилдөө артыкчылыктуу тартипте, ал сочетает кызыкчылыктарын персоналиев жана стимулдар аларды натыйжалуу ишине бардык деңгээлде уюмдун илимий жана билим берүү иши.

Беренеде онтологиялык ыкманын жардамы менен илимий-билим берүү ишинде колдонулуучу маалыматтык системанын моделин түзүү принциптери көрсөтүлөт. Фейербахтын философиясы болгон ролдору, ар дайым аларга окшош. Сунушталат мамиле автоматташтыруу жөнүндө маалыматтарды чогултуу жана илимий ишин ката жаатындагы билимдерди, ал бириктирет ыкмалары метапоиска жана маалымат алуу,базирующиеся карата онтологиях. Ар кандай булактардан алынган жана изилдөөнүн натыйжасы болуп саналган илимий маалыматтарды интеграциялоонун маанилүү милдеттеринин бири катары маалыматты сактоонун, издөөнүн жана иштеп чыгуунун бирдиктүү технологиялык платформасын түзүүнүн маанилүүлүгү негизделет.

Негизги сөздөр: *маалыматтык система, онтология, семантическая интеграциялоо, алуу метаданных, репозиторий, OWL, Protégé, RDF, Dspace.*

РАЗРАБОТКА ТЕХНОЛОГИИ ИЗВЛЕЧЕНИЯ МЕТАДАННЫХ ИЗ ДОКУМЕНТОВ В ПРОЦЕССЕ ИНТЕГРАЦИЙ ИНФОРМАЦИОННЫХ СИСТЕМ

Садирмекова Ж.Б. [1], Самбетбаева М.А.[2]

[1]*Евразийский национальный университет им. Л.Н.Гумилева г.Нур-Султан, Казахстан*
[2]*Институт информационных и вычислительных технологии, г. Алматы, Казахстан*

Аннотация. Информационные системы поддержки научных исследований оперируют с *публикациями, электронными документами и коллекциями, онтологическими описаниями. Подобные ресурсы могут оказаться труднодоступными ввиду поисковых и идентификационных проблем. Семантические связи между информационными ресурсами*

повышают их значимость и способствуют предоставлению дополнительных возможности для информационного поиска и идентификации.

Новизна исследования заключается в попытке интегрирования и усовершенствования ряда актуальных и перспективных технологий для создания нового поколения информационных систем поддержки научно-образовательной деятельности. Базис статьи составляют апробированные подходы и методы, посредством которых становится возможным не только постановка общей задачи, но и выделение отдельных новых элементов.

Целью создания системы является формирование фактографической базы исследования в приоритетном порядке, который сочетает интересы персоналиев и стимулы к их эффективной работе на всех уровнях организации научной и образовательной деятельности.

В статье демонстрируются принципы создания модели информационной системы, используемой в научно-образовательной деятельности с помощью онтологического подхода. Обсуждается состояние онтологий как инструмента семантической интеграции. Предлагается подход к автоматизации сбора информации о научной деятельности в заданной области знаний, который объединяет методы метапоиска и извлечения информации,базирующиеся на онтологиях. Обосновывается важность создания единой технологической платформы хранения, поиска и обработки информации как одной из важнейших задач интеграции научных данных, получаемых из различных источников и являющихся результатом исследований.

Ключевые слова: *информационная система, онтология, семантическая интеграция, извлечения метаданных, репозиторий, OWL, Protégé, RDF, Dspace.*

DEVELOPMENT OF TECHNOLOGY FOR EXTRACTING METADATA FROM DOCUMENTS DURING THE INTEGRATION OF INFORMATION SYSTEMS

Sadirmekova Zh.B.[1], Sambetbayeva M.A.[2]

[1]L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan, janna_1988@mail.ru,
[2] Institute of Information and Computing Technologies, Almaty, Kazakhstan, madina_jgtu@mail.ru

Abstract. *Information systems that support research work with publications, electronic documents and collections, and ontological descriptions. Such resources may be difficult to access due to search and identification problems. Semantic links between information resources increase their significance and contribute to providing additional opportunities for information search and identification.*

The novelty of the research is an attempt to integrate and improve a number of current and promising technologies to create a new generation of information systems to support scientific and educational activities. The article is based on proven approaches and methods that make it possible not only to set a General task, but also to identify individual new elements.

The purpose of the system is to create a factual base of research in a priority order that combines the interests of individuals and incentives for their effective work at all levels of the organization of scientific and educational activities.

The article demonstrates the principles of creating a model of an information system used in scientific and educational activities using an ontological approach. The state of ontologies as a tool of semantic integration is discussed. An approach to automating the collection of information about scientific activities in a given field of knowledge is proposed, which combines methods of meta-search and information extraction based on ontologies. The article substantiates the importance of creating a single technological platform for storing, searching and processing information as one of the most important tasks of integrating scientific data obtained from various sources and resulting from research.

Keywords: *information system, ontology, semantic integration, metadata extraction, repository, OWL, Protégé, RDF, Dspace.*

Введение

В настоящее время разрабатываются и используются два типа информационных систем (ИС) – документальные и фактографические. Документальные ИС представляют собой хранилища документов, снабженных метаданными, посредством которых осуществляется классификация и поиск документов. Фактографические ИС накапливают и хранят данные в виде множества экземпляров одного или нескольких типов структурных элементов (информационных объектов); каждый из таких экземпляров или некоторая их совокупность отражают сведения по какому-либо факту, отдельно взятому событию в отрыве от других сведений и фактов [1].

Наиболее востребованным средством информационного обеспечения научно-образовательной деятельности становятся ИС, которые включают в себя возможности обоих вышеназванных типов информационных систем. Такие ИС способствуют удовлетворению информационных потребностей квалифицированного пользователя сообразно схеме «документ – рассуждение – факт». Заметим, что эта схема соответствует RDF-схеме связанных данных [2].

Информационная система для научно-образовательной деятельности (ИСНОД) хранит информацию о сотрудниках и их публикациях, о конференциях и проектах, участниками которых являлись научные сотрудники, а также сведения об организациях, связанных с конкретными научными проектами, различных видах научных изданий и проч.[3]

Модель информационной системы

Одной из ключевых задач статьи является построение модели, точно представляющей программную систему. В данной работе для построение модели используется онтологический метод *(ONT)* [4,5]. Известно несколько подходов к определению подобного понятия, но общепринятой трактовки не выявлено до сих пор, так как в зависимости от каждой конкретной задачи удобно интерпретировать этот термин по-разному: от нестрогого определения вплоть до описаний онтологий в понятиях и конструкциях логики и математики. Мы, в свою очередь, будем рассматривать онтологию как формальную спецификацию разделяемой концептуализации, которая имеет место в некотором контексте предметной области. При этом под концептуализацией мы понимаем не только сбор понятий, но и всю касающуюся их информацию: свойства, отношения, ограничения, аксиомы и утверждения, необходимые для описания и решения задач в избранной предметной области.

Информационные объекты описывают основные *классы сущностей* научного информационного пространства, такие как *Организация, Персона, Научная деятельность, Публикация, Научные мероприятия, Учебный курс, Раздел науки,*

*Компетенция, Географическое место, Сборник материалов конференции*и т. д., а также связи между ними.

Извлечение информации

Для заполнения контента ИСНОД собирается информация из таких источников,как сайты организаций, ассоциаций, проектов и конференций, порталы знаний, социальные научные сети и др. Как было сказано выше, из этих источников извлекается информация *о Проектах, Организациях, Персонах, Конференциях*, т.е. обо всех базовых классах онтологии научной деятельности, кроме информации о *Публикациях.* Информация о *Публикациях* извлекается из репозиторие (Dspace), которые были созданы авторами [6,7].

Модуль извлечения информации осуществляет анализ интернет-ресурсов, скачанных по ссылкам. Документы в сети Интернет могут быть представлены в различных форматах (HTML, DOC, PDF, TXTи другие). Основным форматом для представления информации в Интернет является HTML. Для извлечения метаданных публикации из репозиториев в пакетном режиме осуществляется экпорт данных в формате XML. Предлагаемые методы извлечения информации *о Проектах, Организациях, Персонах, Конференциях* ориентированы на работу с HTML-страницами а информация о *Публикациях* ориентированы на работу с XML-документами.

Для облегчения анализа HTML-страница и XML-документа ресурса представляется ввиде DOM-дерева в соответствии со стандартом DOM (Document Object Model), регламентирующим способ представления содержимого документа (в частности, HTML-страницы и XML документы) посредством набора объектов [8,9]. На основе соответствующего шаблона выполняется анализ DOM-дерева каждой страницы и извлечение описанной этим шаблоном информации.

Шаблон представляет собой XML-документ , в котором для объектов, отношений и атрибутов онтологии указаны маркеры, сигнализирующие о расположении данного объекта, отношения или атрибута.В шаблонах для каждого типа извлекаемой информации указываются обработчики, реализующие алгоритмы обхода и анализа соответствующих фрагментов интернет-старниц.

Важно отметить, что информация о сущностях, представляющих интерес для пользователей ИСНОД, может быть задана различными способами. Например, информация о проекте, может быть представлена сайтом проекта, разделом сайта организации или персоны или публикацией, описывающей проект. Для каждого из этих способов представления на основе класса онтологии *Проект* строится отдельный шаблон.

Как было сказано выше, извлекаемая из интернет-ресурсов информация представляется в виде семантической сети информационных объектов, т.е.

ориентированного мультиграфа. Интеграцию полученного графа в ИСНОД выполняет модуль занесения информации [10].

На сегодняшний день реализованы все основные компоненты данной подсистемы и разработаны методы извлечения информации о *Проектах, Персонах, Организациях* и *Мероприятиях*, включая сопутствующие шаблоны и обработчики, реализующие информацию о публикациях.

Заключение

Информационная система для научно-образовательной деятельности позволяет исследователям значительно сократить время, требуемое для обеспечения доступа к интересующей их информации. При этом эффективность использования каждой конкретной ИС напрямую зависит от полноты корректности представленной в ней информации. Добиться такой полноты можно за счет автоматизации процесса сбора информации. Для этих целей разрабатывается подсистема сбора информации из сети Интернет.

Работа поддержана грантом финансирования научных и (или) научно-технических исследований на 2018-2020 гг. МОН РК.

ЛИТЕРАТУРА

[1] ***Fedotov A.M., Tusupov J.A., Sambetbayeva M.A., Fedotova O.A., Sagnayeva S.K., Bapanov A.A., Tazhibaeva S.Z.*** *Classification model and morphological analysis in multilingual scientific and educational information systems // Journal of Theoretical and Applied Information Technology. - 2016. - Vol.86, issue 1, - P.96-111.*

[2] ***Fedotov A.M., Tusupov J.A., Sambetbayeva M.A., Sagnayeva S.K., Bapanov A.A., Nurgulzhanova A.N., Yerimbetova A.S.*** *Using the thesaurus to develop it inquiry systems // Journal of Theoretical and Applied Information Technology. - 2016. - Vol.86, issue 1, - P.44-61.*

[3] ***Загорулько Ю.А., Загорулько Г.Б., Боровикова О.И.*** *Технология создания тематических интеллектуальных научных интернет-ресурсов, базирующаяся на онтологии // Программная инженерия. 2016. Т. 7. № 2. С. 51-60.*

[4] ***Zagorulko Y., Borovikova O., Zagorulko G.*** *Pattern-Based Methodology for Building the Ontologies of Scientific Subject Domains. In: New Trends in Intelligent Software Methodologies, Tools and Techniques. Proceedings of the 17th International Conference SoMeT_18. H. Fujita and E. Herrera-Viedma (Eds.). Series: Frontiers in Artificial Intelligence and Applications, Vol. 303. Amsterdam: IOS Press, 2018. P. 529–542.*

[5] ***Садирмекова Ж.Б., Самбетбаева М.А., Еримбетова А.С.****. // Построения распределенных интегрируемых информационных систем поддержки научной деятельности// Вестник Казахской академии транспорта и коммуникаций имени М.Тынышпаева//Алматы. №4.2019.С.192-202.*

[6] ***Садирмекова Ж.Б., Самбетбаева М.А., Дайырбаева Э.Н.****. // Проблемы построения семантической интероперабельной цифровой библиотечной системы// Вестник Казахской академии транспорта и коммуникаций имени М.Тынышпаева//Алматы. №4.2019.С.202-212.*

*[7] Zh****.B. Sadirmekova, O.L.Zhizhimov, D.A. Tussupov, M.A. Sambetbayeva****//Requirements for information system to support scientific and educational activities// CEUR Workshop Proceedings (DICR-2019), //Novosibirsk, Russia,. 2019. 44-47pp.*

[8] ***Садирмекова Ж.Б., Жижимов О.Л., Тусупов Д.А. , Самбетбаева М.А****.// Требования, предъявляемые к информационным системам поддержки научно-образовательной деятельности // Сборник XVII Международный научно-практической конференции «Распределенные информационно-вычислительные ресурсы: цифровые двойники и большие данные» (DICR'2019)// Новосибирск, 2019. С. 171-177.*

[9] ***Садирмекова Ж.Б.*** *// Институциональные репозитории открытого доступа // Сборник Международной научно-практической интернет-конференции «Тенденции и перспективы развития Науки и образования в условиях глобализации»// Переяслав-Хмельницкий, 2019. С. 482-487.*

[10] ***Zh.B.Sadirmekova , J.A.Tussupov, M.A.Sambetbayeva*** *// Development of distributed integrated information support systems for scientific activity// Второй Международный научный конференция «Ситуация, язык, речь. Модели и приложения» («Situation, Language, Speech. Models & Applications» — SLS 2019//Рим-Италия. 2019. С. 24-25.*