Когаловский М.Р.
к.т.н., доцент
Институт проблем рынка РАН
(Москва)

Паринов С.И.
д.т.н.
Центральный экономико-математический институт РАН
(Москва)

Технология семантически обогащаемых научно-образовательных электронных библиотек

Публикуется в сборнике трудов конференции "Экономическая эффективность информационных бизнес-систем", Экономический факультет МГУ, 16 апреля 2015 г.

Аннотация

Обсуждается разработанная авторами уникальная инновационная технология для научно-образовательных электронных библиотек, основанная на возможностях открытой публикации и открытого доступа к их информационным объектам, а также децентрализованной декларации пользователями в онлайновом режиме семантических связей между ними с использованием многоаспектной таксономии связей. Новая технология обеспечивает динамическое обогащение семантики контента библиотеки, реализацию в сообществе ее пользователей новых форм научных коммуникаций и научной деятельности в духе открытой науки, поддерживает «живые» публикации, наукометрию, более содержательную по сравнению с традиционной, а также навигационный доступ к информационным объектам по структуре семантических связей, сформировавшейся над контентом библиотеки. Рассматриваемая технология реализована как составная часть крупной исследовательской информационной системы Соционет.

Ключевые слова: электронная библиотека, семантическая связь, таксономия связей, система Соционет, инновационная технология

Введение

Обсуждаемая в предлагаемом докладе технология развивается и реализуется авторами под влиянием двух важных тенденций в разработках научно-образовательных электронных библиотек и информационных систем другого назначения. Это – все более широкое признание принципов *открытой науки,* а также разработки специальных онтологий для отображения с помощью семантических связей различного рода отношений между информационными объектами контента электронных библиотек.

Принципы открытой науки направлены на повышение эффективности научной деятельности. Хотя они пока еще четко не сформулированы, широко признаны ключевые из них - открытый доступ к результатам исследований и открытое их использование, открытые коммуникации авторов и читателей их работ, открытое рецензирование публикаций и возможность публичного обсуждения мнений экспертов, открытая статистика результативности научной деятельности отдельных исследователей и научных коллективов в целом. Среди ряда выполняемых в этой области проектов привлекает внимание деятельность рабочей группы Open Science Initiative (The National Science Communication Institute, Сиэтл, США). Ее «дорожная карта» опубликована в документе [12].

Онтологии, определяющие возможные семантические связи между представленными в электронных библиотеках научными публикациями и/или другими информационными объектами, позволяют, в частности, явным образом представлять мотивы цитирования одних публикаций в других, связи между версиями и/или частями публикаций, харак-

теризовать соотношение объемов охвата обсуждаемых проблем в связываемых публикациях, декларировать различные другие отношения между ними. В последние годы для указанных целей разработан ряд модульных комплексов онтологий и отдельных онтологий, формальных, специфицированных в языке описания онтологий OWL, и неформальных, представленных в виде таксономий семантических связей.

Технология, разработанная авторами, использует фрагменты модульных комплексов онтологий *SPAR* (*The Semantic Publishing and Referencing Ontologies*) [17] и *SWAN* (*Semantic Web Applications in Neuromedicine*) [16], рекомендации *SKOS* (*Simple Knowledge Organization System*) [18], проект CRediT [10] открытого стандарта классификатора существенных для коллективно выполняемого исследования ролей его участников. Используются также элементы модели научных данных *CERIF (Full Data Model*) [8], развиваемой европейской организацией euroCRIS (http://www.eurocris.org/), и онтологии [9], определяющей систему терминов, обозначающих сущности этой модели и отношений между ними.

Разработка обсуждаемой технологии стимулировалась стремлением авторов создать конструктивный инструментарий, обеспечивающий реализацию на практике принципов открытой науки в научно-образовательных электронных библиотеках [14]. Базовые принципы разработки включают:

- открытую возможность публикации результатов исследований в электронной библиотеке и открытый доступ к содержащимся в ней публикациям, к описаниям семантических связей и к другим информационным объектам;
- открытую возможность декларации семантических связей допустимых классов между информационными объектами библиотеки;
- возможность декларации семантических связей децентрализовано в онлайновом режиме на протяжении всего времени функционирования библиотеки в отличие от традиционного подхода, используемого в семантических электронных библиотеках, при котором семантика контента библиотеки определяется ее разработчиками априори, до начала функционирования системы;
- многоаспектность семантики декларируемых связей, позволяющая представлять в явном виде спектр различного рода отношений, прежде всего научного характера, между информационными объектами библиотеки;
- наличие развитых механизмов поддержки активности пользователей библиотеки, в том числе, сервиса оповещения авторов публикаций об использовании их информационных объектов как участников созданных связей или авторов созданных связей о какие-либо изменениях связанных ими объектов со стороны их авторов.

Полигоном для реализации разработанной технологии послужила крупная научно-образовательная информационная система Соционет [7], функционирующая уже почти полтора десятилетия. Система основана на *технологии открытых архивов* (Open Archive Initiative, OAI) [11]. Она импортирует метаданные из репозиториев более 1700 отечественных и зарубежных открытых архивов исследовательских, образовательных и библиотечных учреждений и, в свою очередь, предоставляет ресурсы собственного репозитория метаданных для импорта другим системам и т.д. В настоящее время информационное пространство Соционет включает около 2.5 млн. информационных объектов, содержит описания более 8 миллионов семантических связей. Реализацию рассматриваемой технологии рассмотрим далее на примере этой системы.

Организация информационных ресурсов

Рассмотрим, прежде всего, кратко организацию информационных ресурсов системы Соционет. Как уже отмечалось, система основана на технологию открытых архивов. Это означает, что публикации и другие информационные объекты, доступ к которым предоставляется пользователям, непосредственно в системе не хранятся. Ими владеют различные организации, которые поддерживают их на собственных ресурсах в Вебе и предоставляют к ним открытый доступ. Механизмами системы поддерживаются описания

этих информационных ресурсов, создаваемые их владельцами либо их представителями. Описания представляются стандартизованным образом с использованием одного из допустимых форматов (Дублинское ядро, MARC и др., а также языка XML) и содержат гиперссылки на соответствующие им информационные объекты в Вебе. Описания являются *представителями* соответствующих информационных объектов в системе. Именно с ними имеют прежде всего дело пользователи. Если при ознакомлении с описанием какого-либо объекта он вызвал интерес у пользователя, то можно получить к нему доступ, перейдя по гиперссылке, содержащейся в его описании. Семантические связи между информационными объектами представляются в библиотеке как *связи между их описаниями* и сами представляются, как правило, в виде самостоятельных информационных объектов, которые, в свою очередь, также могут участвовать в других связях.

Предполагается, что информационные объекты системы *типизируются*. Описание каждого типа информационных объектов содержит специфический для него набор атрибутов. Среди допустимых типов объектов *book* (книга), *article* (статья), *paper (*отчет, рабочая записка, тезисы доклада и др.), *news* (новость), *comment* (комментарий), *thesis* (диссертация), *person* (персона – автор информационного объекта или пользователь библиотеки), *institution* (организация) и др. В отличие от информационных объектов других типов, персоны, организации и связи представляются только их описаниями. У них нет иного представления в Вебе. Описания персон и организаций называются их *профилями.* Описания информационных объектов одного типа могут группироваться в *коллекции*, которые представляют для пользователей виртуальные коллекции описываемых информационных объектов. Коллекции также имеют свои описания. Описания всех коллекций и составляющих их информационных объектов составляют *репозиторий метаданных* открытого архива и именно с ним работают фронтальные интерфейсы пользователей системы.

Система электронной библиотеки, основанная на технологии OAI, помимо интерфейсов конечных пользователей для доступа к его репозиторию метаданных, поиска и просмотра в нем описаний требуемых информационных объектов и доступа к ним путем навигации по гиперссылкам, содержащимся в описывающих их метаданных, обладает также механизмом *сборщика метаданных* (Metadata Harvester) из других открытых архивов и программным интерфейсом (API) для сборщиков метаданных других открытых архивов, которые могут импортировать метаданные из репозитория метаданных данного архива с помощью запросов, определенных стандартным протоколом OAI-PMH [19].

Использование технологии открытых архивов обеспечивает для электронной библиотеки внутреннюю и внешнюю интеграцию информационных ресурсов. *Внутренняя интеграция* обеспечивается поддержкой централизованного репозитория метаданных, описывающих информационные ресурсы различных владельцев, распределенные в среде Веба. *Внешняя интеграция* обеспечивается благодаря интероперабельности ресурсов различных библиотек, основанных на технологии OAI, за счет стандартизации описаний информационных объектов и поддержки каждой из них стандартного протокола OAI-PMH.

Семантика связей информационных объектов библиотеки

Ключевое значение для широты спектра новых возможностей, обеспечиваемых для пользователей системы предлагаемой технологией, имеет степень разнообразия классов семантических связей, определяемых поддерживаемой в системе *онтологией связей*. В нашем случае множество допустимых классов семантических связей определяется многоаспектной таксономией, базирующейся на указанных выше онтологиях и включающей собственные дополнения, разработанные авторами. Механизмы, реализующие создание, поддержку и использование таксономии, позволяют расширять множество классов допустимых связей. Допускаются только *бинарные ориентированные связи*.

Классы, входящие в состав разработанной таксономии, можно условно разделить на три категории: научные связи, структурные связи, связи принадлежности. *Научные связи* определяют характер использования или развития результатов, обсуждаемых в одной

публикации, в другой, характеризуют профессиональную оценку данной публикации некоторым пользователем либо в другой публикации, научную близость публикаций и др. отношения между ними. *Структурные связи* определяют информационные объекты как составные части некоторой публикации, как ее различные версии или варианты представления (например, текст доклада и его презентация), связывают абстракты публикаций и аннотации их фрагментов, актуализирующие либо оценивающие их содержание. Наконец, *связи принадлежности* определяют авторство публикаций и других информационных объектов, аффилиацию авторов, вклады соавторов в создание коллективных публикаций.

Используемая таксономия семантических связей имеет *двухуровневую организацию*. Классы верхнего уровня включают семантически близкие подклассы связей, например, связей развития и дополнения результатов исследований, оценочных связей, связей между компонентами и/или версиями либо представлениями публикаций, связей мнений о существующих связях, связей, характеризующих вклады соавторов коллективных публикаций в их подготовку и др. Каждому классу верхнего уровня таксономии соответствует *контролируемый словарь*, а классам второго уровня – значения из этого словаря, представляющие имена этих классов, являющихся подклассами соответствующего класса верхнего уровня.

Соционет допускает расширение таксономии связей путем создания и поддержки дополнительных контролируемых словарей. Подробнее реализованная в системе базовая таксономия семантических связей описана в работе [1].

Инновационные возможности предлагаемых технологий

Соционет обеспечивает для пользователей реализованную развитым образом функциональность систем, основанных на технологии открытых архивов, прежде всего, открытую публикацию научно-образовательных информационных ресурсов и открытый доступ к ним, формирование статистики доступа и др. Вместе с тем, реализация в среде системы обсуждаемой технологии дополнительно обеспечивает ряд новых инновационных возможностей. Авторам неизвестны другие электронные библиотеки с таким функциональным потенциалом. Рассмотрим кратко эти новые возможности.

Создание, поддержка, расширение при необходимости и обеспечение использования таксономии семантических связей. Как уже указывалось, таксономия представляется в системе в виде набора контролируемых словарей связей. Расширение поддерживаемой в системе таксономии может осуществляться созданием дополнительных словарей.

Декларация семантических связей и формирование семантической структуры контента системы. Авторизованные пользователи могут декларировать в онлайновом режиме семантические связи между информационными объектами системы, в частности, между двумя научными публикациями или между собственными профилями и публикациями. Допустимые классы связей определяются поддерживаемой в системе таксономией связей. Выбранный пользователем класс связей несет *информацию о семантике отношения* между связываемыми информационными объектами, представляемого создаваемой связью. Создаваемые семантические связи представляются как обычные информационные объекты, и пользователь имеет возможности открытой их публикации в системе. В результате совместной деятельности пользователей по декларации семантических связей в своего рода режиме социальной сети [3] в системе порождается динамически изменяемая *многослойная сеть семантических связей* между информационными объектами, каждый слой которой соответствует некоторому классу связей [5].

Коммуникация представителей сообщества пользователей библиотеки. Возможности открытой декларации и публикации семантических связей в системе обеспечивают *научные коммуникации* пользователей в виртуальной среде библиотеки. Действительно, создание конкретной семантической связи служит *сообщением* автору участвующей в ней публикации, которое является носителем информации, определяемой классом и описанием этой связи. Поскольку автор связываемой публикации является зарегистрированным в

системе пользователем, на адрес его электронной почты, указанный в его профиле, может быть направлено сообщение, которое уведомит о появлении интересующей его связи. Поскольку создавать связи и другие информационные объекты могут только авторизованные пользователи, автор сообщения может быть идентифицирован по персональному профилю создателя связи. Публичная доступность созданных семантических связей и передаваемых с их помощью сообщений обуславливает *ответственное отношение* к ним их авторов. Эти сообщения могут носить оценочный характер – своего рода рецензии целевой публикации связи, могут информировать авторов публикаций о близких к ним работам, в которых используются и каким образом представленные в них результаты либо получены аналогичные результаты, могут отражать рекомендации авторам публикации по развитию их работ. Подробнее коммуникационные возможности рассматриваемой технологии обсуждаются в [2].

Поддержка «живых» документов. Привлекательной возможностью для авторов электронных публикаций является возможность актуализировать их содержание в соответствии с изменением и углублением его представлений об исследуемой проблеме. Такого рода электронные документы, содержание которых корректируется на протяжении времени, стали называться *«живыми» документами* [6]. Для поддержки «живых» документов в контексте обсуждаемых технологий недостаточно указывать в их заголовке дату новой версии или ее номер. Проблема в том, что такой документ может быть участником семантической связи, которая, например, характеризует его оценку. Эта оценка может стать неадекватной после внесения автором изменений в документ. Решение проблемы заключается в том, что при внесении изменений в информационный объект – участник связи автор связи получает уведомление об этом событии и может при необходимости внести соответствующее изменение в описание этой связи или удалить ее. Актуализация «живого» документа может осуществляться также с помощью связей, поддерживающих аннотирование фрагментов абстрактов публикаций. Другое средство поддержки «живого» документа – отсылка с помощью связей соответствующего класса от старой к новой версии документа.

Формирование статистики дифференцированно по классам связей. Такая статистика автоматически генерируется для отдельной публикации, для всех публикаций некоторого автора или организации в целом. Благодаря этому в системе доступен статистический портрет деятельности в этой среде ее пользователей, в частности, авторов представленных в системе публикаций.

Семантическое обогащение ссылок цитирования и новые наукометрические измерения. Библиографические ссылки на цитируемые источники в научных публикациях не несут какой-либо информации о мотивах их цитирования. Однако на таких «немых» связях базируются наукометрические измерения, которые, по сути, учитывают лишь только факт цитирования. В результате высокий индекс цитирования может иметь публикация, содержащая грубые ошибки и в связи с этим вызвавшая активную реакцию научного сообщества. Предлагаемые технологии предусматривают возможность семантического обогащения «немой» ссылки цитирования путем отнесения такой связи к соответствующему классу таксономии связей, характеризующему конкретный мотив цитирования. На этой таких семантизированных связей цитирования и других семантических связей в системе генерируются *наукометрические показатели*, более дифференцированные и более содержательные по сравнению с традиционными [4, 15].

Контекстная визуализация семантических связей. Для созданных в системе семантических связей обеспечивается контекстная визуализация. При просмотре описания конкретного информационного объекта пользователь может видеть все входящие и исходящие связи этого объекта. Предусмотрена *фильтрация* по классам связей.

Навигация пользователя в контенте библиотеки по его семантической структуре. Контекстная визуализация сети семантических связей позволяет пользователю наряду с поиском требуемых ему информационных объектов с помощью поисковых сервисов си-

стемы осуществлять доступ к ним путем семантической навигации по структуре связей, осуществляя пошаговый переход от описания одного информационного объекта к описанию связанного с ним объекта по выбранной исходящей семантической связи. Как уже отмечалось, для рассмотрения могут быть отфильтрованы только связи нужного класса.

Реализация технологии в Соционет

Технология, обеспечивающая для пользователей научной электронной библиотеки описанную выше функциональность, реализована, как уже указывалось в среде системы Соционет, которая, вместе с тем, обладает развитыми средствами поддержки технологии открытых архивов, средствами генерации статистики доступов к ее информационным объектам, а также многоаспектным пользовательским интерфейсом. Система обменивается ресурсами репозитория метаданных с крупнейшей международной системой RePEc.

Для реализации новой технологии в состав механизмов Соционет включены дополнительные функциональные модули. Главными из них являются следующие:

- *Средства создания и использования контролируемых словарей семантических связей*, представляющих поддерживаемую в системе таксономию связей, и матрицы допустимости классов связей. Эта матрица определяет классы связей, допустимые для заданной пары типов связываемых информационных объектов.
- *Средства создания, изменения и удаления семантических связей*. Предусмотрен ряд пользовательских интерфейсов для создания конкретных семантических связей. Наряду с универсальным интерфейсом на все случаи, предусмотрены различающиеся своими функциями интерфейсы для авторов публикаций, представленных в системе, и для их читателей [13].
- *Сервис уведомления*, активизирующийся при создании какой-либо конкретной семантической связи или изменении ее свойств (например, ее класса или комментария в ее описании). Сервис направляет уведомление по электронной почте автору публикации или ранее созданной связи, которая становится участницей новой создаваемой связи. Уведомление позволяет адресату получить необходимую информацию о семантике создаваемой или изменившейся связи, ее авторе и т.п. Сервис уведомления стимулирует ответную реакцию авторов публикаций - участников связей, а также пользователей системы – создателей связей, ставших участниками других связей. Тем самым этот системный механизм является *движителем коммуникационного процесса*. Сервис уведомления имеет ряд параметров настройки, управляющих характером его функционирования.
- *Средства генерации статистики связей и наукометрических показателей*. Система автоматически генерирует для каждого информационного объекта статистику входящих и исходящих связей. Для связей, отражающих научные отношения между информационными объектами системы, такие статистические данные представляют собой новые наукометрические показатели.

Заключение

Основные механизмы описанной технологии реализованы и доступны пользователям системы Соционет. Авторы продолжают развивать разработанную технологию.

Литература

1. Когаловский М.Р., Паринов С.И. Таксономия семантических связей информационных объектов системы Соционет. http://socionet.ru/publication.xml?h=RePEc:rus:rssalc:web-53

2. Когаловский М.Р., Паринов С.И. Научные коммуникации в среде семантически обогащаемых электронных библиотек //Программная инженерия. – 2015. - № 4. – С. 31-38.

3. Когаловский М.Р., Паринов С.И. Технологии социальной сети для создания семантических связей информационных объектов в научной электронной библиотеке //Программирование. МАИК/Наука «Интерпериодика». - 2014.- Т. 40. - № 6. – С. 22-33.

4. Когаловский М.Р., Паринов С.И. Новый источник данных для наукометрических исследований. XV Всероссийская научн. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2013. Ярославль, Россия, 14-17 октября 2013 г.». г. Ярославль: Ярославский университет, 2013. – С. 107-117.

5. Когаловский М.Р., Паринов С.И. Семантическое структурирование контента научных электронных библиотек на основе онтологий. В кн.: "Современные технологии интеграции информационных ресурсов: сборник научных трудов». Санкт-Петербург: Президентская библиотека им. Б.Н. Ельцина, 2011. – С. 26-45.

6. Паринов С.И., Когаловский М.Р. «Живые» документы в электронных библиотеках //Прикладная информатика. – 2009. - № 6 (24). – С. 123-131.

7. Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайновых сервисов //Российский научный электронный журнал «Электронные библиотеки». - 2003. - Том 6. - Вып. 1. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part1/PLP

8. CERIF 1.3 Full Data Model (FDM): Introduction and Specification. euroCRIS, 2012. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf

9. CERIF 1.3 Semantics: Research Vocabulary. CERIF Task Group, euroCRIS, 2012. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3 Semantics.pdf

10. Liz Allen, Amy Brand, Jo Scott, Micah Altman and Marjorie Hlava. Credit where credit is due. Nature/ International weekly journal of science. Vol. 508, Issue 7496, April 2014. http://www.nature.com/polopoly_fs/1.15033!/menu/main/topColumns/topLeftColumn/pdf/ 508312a.pdf

11. Open Archives Initiative. http://www.openarchives.org/

12. Open Science Initiative Working Group. Mapping the Future of Scholary Publishing. First Edition? January 2015. http://nationalscience.org/wp-content/uploads/2015/02/OSI-report-Feb-2015.pdf

13. Parinov S. Semantic enrichment of research outputs metadata: new CRIS facilities for authors. Submitted to MTSR 2014, 8th Metadata and Semantics Research Conference, 27-29 November 2014, Karlsruhe, Germany.

14. Parinov S. Towards a Semantic Segment of a Research e-Infrastructure: necessary information objects, tools and services. Metadata and Semantics Research, Communications in Computer and Information Science. J. M. Dodero, M. Palomo-Duarte, P. Karampiperis, Eds. Springer. Vol. 343, 2012, pp. 133-145. http://socionet.ru/pub.xml?h=RePEc:rus:mqijxk:30

15. Parinov S., Kogalovsky M. Semantic Linkages in Research Information Systems as a New Data Source for Scientometric Studies. Scientometric. Vol. 98, Issue 2 (2014), pp. 927-943.

16. Semantic Web Applications in Neuromedicine (SWAN) Ontology. W3C Interest Group Note, 20 October 2009. http://www.w3.org/TR/hcls-swan/

17. Shotton D. Open Citations and Related Work. Introduction the Semantic Publishing and Referencing (SPAR) Ontologies. October 14, 2010. http://opencitations.wordpress.com/2010/ 10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/

18. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 August 2009. http://www.w3.org/TR/skos-reference/

19. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. Document Version 2015-01-08.
http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Kogalovsky M.R.
Candidate of Science (Tech.), Associate Prof.
Market Economy Institute of RAS
(Moscow, Russia)

Parinov S.I.
Doctor of Science (Tech.)
Central Economics and Mathematics Institute of RAS
(Moscow, Russia)

Technology for Semantic Enrichable Scientific and Educational Digital Libraries

Abstract

We present an innovative technology for scientific and educational digital libraries, which provides to their users a unique set of facilities. It is based on the open publication and access to information objects of the digital library. In addition library users can declare online semantic linkages of some classes between content information objects by using multifold taxonomy of semantic relationships. New technology provides dynamic enrichment of the library content semantics, as well as new forms of scientific communications and cooperation in user community of the library that is in line with the Open Science approach. It supports also the «live» publications, more intensional scientometrics, and navigation through semantic linkage structure created over the digital library content. The presented technology is implemented as a part of the large research information system Socionet.

Keywords: digital library, semantic linkage, taxonomy of semantic relationships, Socionet research information system, innovative technology