Семантическое структурирование контента научных электронных библиотек на основе онтологий*

М.Р. Когаловский
Институт проблем рынка РАН
kogalov@cemi.rssi.ru

С.И. Паринов
Центральный экономико-математический институт РАН
sparinov@gmail.com

Предлагается подход к организации информационных ресурсов научных электронных библиотек, предусматривающий их семантическое структурирование. Обогащенный семантическими связями контент электронной библиотеки позволяет проводить наукометрические измерения на основе классификации связей ее информационными объектами, обеспечивает новые возможности для анализа представленного в ней корпуса научных знаний и новые формы интерактивной научной деятельности. В работе обсуждаются основные идеи предлагаемого подхода, свойства семантической структуры связей и использование онтологий для определения их семантики. Рассматриваются также функции механизмов формирования, поддержки и анализа семантической структуры контента библиотеки, реализуемые в системе Соционет.

1. Введение

Коллекции информационных ресурсов традиционных текстовых научных библиотек, как правило, представляют собой совокупности отдельных явным образом не документов. Единственный вид связей, который обычно используется – это связи цитирования. Эти связи представляются в цитирующем документе в неструктурированном виде как список использованной литературы или реже структурированным образом, как в отчетах по грантам РФФИ, где список публикаций представляется в размеченном виде. В любом из этих случаев связи-ссылки являются «немыми». Они не несут какой-либо информации, кроме указания целевого документа и самого факта существования ссылки, не отражают семантики отношения между связываемыми документами. Однако связи цитирования обладают различной семантикой, и она может быть представлена в явном виде в научных электронных библиотеках. Будем далее называть связи с явно выраженной семантикой представляемых ими отношений между связываемыми документами семантическими связями. Заметим, что между содержащимися в научной электронной библиотеке документами могут поддерживаться также семантические связи, отличные от связей цитирования, т.е. явным образом не представленные в исходных документах таких связей. Они могут создаваться на основе мнения эксперта о существовании некоторого отношения между содержанием двух документов в ситуации, когда этот факт не отмечен явным образом в этих документах.

Поддержка семантических связей явным образом приводит к образованию многослойной семантической структуры контента электронной библиотеки, каждый слой которой соответствует некоторому свойству семантических связей. Такая структура может формироваться на основе онтологии связей и может служить источником информации для проведения качественно новых наукометрических измерений, для исследования структурных свойств корпуса знаний в различных областях науки.

Система управления онлайновой научной электронной библиотекой может не только обрабатывать запросы относительно семантической структуры контента, но и

* Работа поддерживается грантами РФФИ 09-07-00378 и РГНФ 11-02-12026-в

располагать механизмами, позволяющими пользователям самостоятельно устанавливать семантические связи. Системные механизмы могут также осуществлять мониторинг состояния структуры связей и автоматически оповещать авторов публикаций о том, что некоторая их работа стала участником вновь учрежденной связи или что ликвидирована существующая связь, в которой эта работа являлась участником.

Такого рода операционная среда и получаемая ее средствами информация открывают новые возможности для развития научных исследований, обеспечивают новые технологии для научной и научно-организационной деятельности. В статье обсуждается подход авторов к созданию такой среды, реализуемый в научном информационном пространстве Соционет.

Остальная часть статьи построена следующим образом. В разделе 2 обсуждается постановка рассматриваемой в статье проблемы и задачи, требующие решения. В разделе 3 анализируются процессы, связанные с научной деятельностью, в результате которых возникают потребности формирования семантических связей между публикациями, а в среде электронной библиотеки – между представленными в ней информационными объектами. Обсуждаются также категории семантических связей, которые при этом порождаются. В разделе 4 рассматриваются свойства структуры семантических связей, ее многослойный характер. В разделе 5 обсуждаются семантика связей, вопросы их классификации, известные онтологии, созданные для этих целей, и онтология, используемая авторами. В разделе 6 приводится функциональная характеристика операционных возможностей механизмов формирования, поддержки и анализа многослойной структуры семантических связей, реализуемых в системе Соционет. В заключении подводятся итоги обсуждения. Завершается статья предложением, реализация которого позволила бы широко использовать обсуждаемые в статье идеи, и списком литературы.

2. Постановка проблемы

Контент традиционных научных электронных библиотек, как правило, состоит из коллекций отдельных не взаимосвязанных документов – электронных версий публикаций, изданных типографским способом, научных отчетов, рабочих записок, рецензий, авторефератов диссертаций, полных текстов диссертационных работ, таблиц научных данных, карт звездного неба и др. Научные электронные библиотеки могут содержать также сведения об исследователях – авторах представленных в них публикаций, об организациях, в которых они работают.

В последние годы на основе библиографических ссылок в публикациях, выпускаемых в авторитетных периодических изданиях, начали создаваться индексы цитирования, которые формируют библиометрическую статистику. Связи цитирования в текстовых публикациях обычно представляются неструктурированным образом в виде списка используемой литературы. Они не являются при этом носителями какой-либо информации, кроме указания целевой публикации ссылки и существования самого факта ссылки. Тем не менее, с фактом цитирования связана некоторая не отраженная таким образом семантика, выражающая отношение автора цитирующего документа к цитируемому источнику или какое-либо иное семантическое отношение между участвующими в связи документами. Как правило, связи цитирования аннотируются в тексте публикации, и в таких случаях семантика связи все-таки описана, но в неструктурированном виде. Это создает значительные сложности для ее анализа. Поэтому практически отсутствуют разработки программного инструментария для этих целей.

Наряду со связями цитирования между документами научных электронных библиотек существуют, как уже отмечалось, разнообразные другие семантические связи. Например, связь может указывать, что ее целевой документ содержит научные

результаты, базирующиеся на результатах, описанных в исходном документе связи, или что в исходном документе связи опровергается результат, изложенный в ее целевом документе. Связь может также указывать, что исходный ее документ является новой редакцией целевого документа или что он представляет собой его составную часть, например, аннотацию. Существует большое разнообразие семантических связей, которые можно при необходимости поддерживать между документами библиотеки.

Определяемые явным и структурированным образом семантические связи между документами в электронной библиотеке могут быть представлены и могут динамически поддерживаться как самостоятельные информационные объекты, содержащие идентификаторы участвующих в них документов и значения других атрибутов. Объекты-связи могут быть типизированы, и их свойства определяются значениями атрибутов, свойственных каждому типу.

В результате такого определения семантических связей между документами библиотеки в составе ее контента порождается многослойная семантическая структура объектов-связей. При этом каждому типу связей соответствует некоторый слой этой структуры, который может наряду с полной структурой связей служить для наукометрических измерений и анализа. В частности, могут поддерживаться слои, отображающие структуру продуцирования научных результатов и другие содержательные отношения между научными публикациями, например, связи оценки публикаций научными сотрудниками, связи между компонентами научных публикаций, связи научно-организационного характера (научное учреждение – сотрудники-авторы публикаций, авторы – публикации) и др.

Анализ структуры таких связей в научной электронной библиотеке позволяет решать ряд задач, связанных с поддержкой научно-организационной деятельности, позволяет авторам публикаций более продуктивно использовать имеющиеся в электронной библиотеке научные информационные ресурсы, дает возможность извлекать из контента библиотеки ценную информацию, не содержащуюся в отдельных документах. Например, можно получать полезные наукометрические сведения, а также сведения, основанные на анализе топологии структуры связей, которые достаточно сложно получить иным путем. Исследование топологии связей научных публикаций позволяет, в частности, анализировать процесс формирования научных направлений и школ, влияние публикаций тех или иных исследователей на формирование научных направлений или теорий. Поддержка семантической структуры связей обеспечивает также дополнительные пути доступа пользователей к информационным объектам библиотеки. Другое направление, где необходима поддержка семантических связей между документами электронной библиотеки, – это технология «живых» публикаций, подробно рассмотренная в работе [6, 7].

Для эффективного использования тех новых возможностей, которые обеспечиваются благодаря поддержке в онлайновой электронной библиотеке многослойной структуры семантических связей представленных в ней документов, необходимо, чтобы система управления электронной библиотекой удовлетворяла определенным требованиям. В частности, она должна быть способна не только обрабатывать запросы относительно семантической структуры контента, но и располагала механизмами, позволяющими пользователям самостоятельно устанавливать, модифицировать или удалять семантические связи в рамках их полномочий, а также обеспечивать мониторинг состояния структуры семантических связей. Такие механизмы позволяют автоматически оповещать авторов документов библиотеки о том, что некоторая их работа стала участником вновь учрежденной связи, что ликвидирована существующая связь, в которой она являлась участником, или что изменились значения ее атрибутов.

Семантическое структурирование контента научных электронных библиотек представляет значительно больший интерес, если оно поддерживается на представительном репозитории научных документов. Одним из популярных подходов к созданию крупных репозиториев научных публикаций, позволяющих интегрировать коллекции ряда научных и образовательных учреждений, является подход, основанный на технологии открытых архивов. Поддержка и исследование семантической структуры в создаваемом на ее основе крупном интегрированном контенте дает возможность изучать структуру результатов научных исследований не только отдельных научных коллективов или школ, но и целых направлений науки и областей знаний.

Возможности формирования в научных электронных библиотеках явно представленных структурированным образом многоаспектных семантических связей между содержащимися в них научными документами в сочетании с методами мониторинга изменений структуры этих связей и основанными на такой структуре новыми функциональными возможностями является, на наш взгляд, весьма перспективным новым направлением развития научных электронных библиотек.

Для эффективного использования обсуждаемых возможностей необходимо решить следующие задачи:

- разработать способы и конкретные форматы представления семантических связей между документами электронной библиотеки в виде самостоятельных информационных объектов;
- классифицировать семантические связи, которые целесообразно поддерживать в научных электронных библиотеках, и построить онтологию связей;
- оценить, какие операционные возможности должна обеспечивать система управления научной электронной библиотекой для того, чтобы извлекать в достаточно полной мере ту информацию, которая содержится в структуре семантических связей представленных в ней документов.

Данная статья посвящена обсуждению предлагаемого авторами подхода к решению этих задач, а также опыта его реализации на платформе крупной отечественной онлайновой научно-образовательной электронной библиотеки Соционет [3, 4], основанной на технологии открытых архивов и содержащей информационные ресурсы социально-экономической тематики. Эта библиотека функционирует уже более десяти лет и приобрела в последние годы статус де-факто институциональной электронной библиотеки Отделения общественных наук РАН. В ней содержатся также публикации ряда образовательных учреждений и других организаций.

Соционет является полигоном для проведения исследований в области перспективных технологий электронных библиотек. Постоянно проводятся работы по расширению разнообразия представляемых в этой системе информационных ресурсов и развитию функциональности механизмов управления библиотекой. Основные идеи данной работы сформировались на основе наших ранних публикаций [1, 2, 5, 11].

Дальнейшее обсуждение в этой статье будет в значительной мере отражать особенности подхода, принятого авторами для развития функциональных возможностей и системных механизмов Соционет, обеспечивающих поддержку и использование семантических связей.

3. Источники информации для структурирования контента библиотеки

В онлайновой электронной библиотеке, располагающей необходимой функциональностью, пользователи могут не только инициировать запросы на получение интересующих их документов или информации, генерируемой на основе структуры семантических связей представленных в библиотеке документов, но и самостоятельно в

модерируемом администратором системы режиме создавать новые объекты-связи, модифицировать или удалять ранее созданные ими такие информационные объекты.

Возникает вопрос, чем руководствуется пользователь, предпринимая действия, направленные на пополнение или модификацию структуры семантических связей в контенте библиотеки. Естественно, устанавливая новые или модифицируя существующие ранее установленные им семантические связи, пользователь исходит из той информации, которую он черпает, являясь непосредственным участником ряда процессов, свойственных научной и научно-организационной деятельности. Рассмотрим основные виды этих процессов.

Процессы систематизации, классификации и упорядочения корпуса научных знаний. Исследование возникающей научной проблемы начинается с поиска и аналитической переработки уже существующих релевантных научных результатов. Исследователь выявляет основополагающие публикации, работы, которые базируются на их результатах и развивают их, и на этой основе синтезирует структуру взаимосвязей изученных им публикаций. Важными результатами процессов рассматриваемого вида являются аналитические обзоры состояния исследований, классификаторы известных публикаций и тематические указатели литературы для конкретных областей науки или направлений исследований.

Процессы научной оценки опубликованных результатов. Значительную часть своего времени сложившиеся ученые затрачивают на подготовку рецензий, отзывов и других документов, содержащих оценку различных научных произведений. Такие оценочные материалы часто публикуются в периодике или представляются в электронном виде. В онлайновых электронных библиотеках, содержащих такого рода «оценочные» произведения, выраженные в них оценки могут быть явно представлены с помощью семантических связей между оценивающими и оцениваемыми произведениями. Эти связи могут выражать позитивное или негативное отношение, обвинение в плагиате и пр. Если в электронной библиотеке содержатся описания профилей ученых (информационных объектов, представляющих их характеристики), то оценка научного произведения может быть выражена и без необходимости создания специального оценивающего произведения путем установления связи, характеризующей нужную оценку, между профилем автора оценки и оцениваемым произведением.

Процессы продуцирования нового научного знания и создания материализующих его публикаций. В своих публикациях, представляющих научному сообществу новые научные результаты, автор по необходимости должен охарактеризовать соотношение содержащихся в его работе результатов с опубликованными ранее им или другими авторами: на какие известные публикации опирался автор в своей работе, из каких работ он использовал представленные в них результаты или приведенные в них данные и т.п. Эти отношения могут отображаться автором или другими пользователями онлайновой библиотеки в виде соответствующих семантических связей между публикациями.

Процессы создания научных произведений. Эти процессы имеют отношение не только к деятельности научных сотрудников, оформляющих результаты своих исследований, но и к издательской деятельности. Они могут осуществляться индивидуально или коллективом исследователей. При коллективной работе, как правило, разными авторами, совместно подготавливающими научное произведение, создаются отдельные его фрагменты, из которых, в конечном счете, формируется законченная публикация. При этом некоторые фрагменты, как и публикация в целом, могут иметь различные версии. Отдельные фрагменты созданного произведения могут продолжать существовать как самостоятельные информационные объекты, например, аннотации статей или монографий. Связи между фрагментами готовящейся публикации и/или их версиями и публикацией в целом рождаются в этих процессах.

Научно-организационные процессы. Функционирование научных учреждений и научных коллективов обеспечивается научно-организационной деятельностью, в процессе которой формируются и могут отображаться в электронных библиотеках различного рода связи между научными организациями и их сотрудниками, сотрудниками и опубликованными ими работами, организацией в целом и ее подразделениями, тематикой исследований и подразделениями организации, их выполняющими и т.д. Связи такого рода, поддерживаемые в научных электронных библиотеках, могут использоваться для продуцирования различной полезной информации.

На основе информации, которую получает пользователь электронной библиотеки в результате участия в процессах перечисленных видов, он может придти к выводу о целесообразности установления тех или иных семантических связей между документами электронной библиотеки. Кроме того, он может явным образом представить в библиотеке семантически проинтерпретированные связи цитирования, содержащиеся в текстах документов библиотеки. Все эти семантические связи подразделяются на ряд категорий. Ограничимся здесь рассмотрением только некоторых из них.

Связи научного вывода. Такие связи отображают идейные зависимости и соотношения представленных в связываемых документах научных результатов, полученных исследователями в процессе развития научного знания. Например, связь этой категории может указывать, что результаты, обсуждаемые в ее исходном документе, базируются на результатах, рассматриваемых в целевом документе связи. Связи этой категории требуют особой обработки механизмами мониторинга состояния структуры семантических связей контента библиотеки. Так, возможна ситуация, когда устанавливается некоторая связь, такая что результат, обсуждаемый в исходном документе этой новой связи опровергает результаты, представленные в ее целевом документе. В такой ситуации механизмы мониторинга должны оповещать авторов всех документов, аналогичных рассмотренным в этом примере, поскольку вопрос о достоверности изложенных в них результатов становится открытым.

Оценочные связи. Связи такого рода позволяют представить в библиотеке разнообразные оценки (в зависимости от конкретного используемого типа связей) целевых документов связей. Оценка может обсуждаться в каком-либо содержащемся в библиотеке документе, например, в рецензии на изданную монографию, или исходить от имени некоторого зарегистрированного в библиотеке пользователя, мнение которого отражает не специально подготовленный им документ, включенный в библиотеку, а лишь тип устанавливаемой им связи. Соответственно, исходными документами связей этой категории могут быть документы библиотеки или профили пользователей. С помощью связей данной категории можно указать, в частности, что некоторый пользователь не согласен с мнением автора целевого документа, в нем выраженным, или что рецензия на целевой документ, представленная в исходном документе, в целом положительная.

Структурные и версионные связи. Связи данной категории представляют отношения между фрагментами создаваемых или уже созданных документов либо между их версиями. К числу таких связей относятся, например, связи, у которых: 1) исходный документ содержит иллюстрацию к тексту, содержащемуся в целевом документе; 2) исходный документ - это аннотация статьи, содержащейся в целевом документе; 3) исходный документ является переводом на иностранный язык или новой редакцией статьи, представленной целевым документом связи.

Научно-организационные связи. Эта категория включает связи, информация о которых может быть полезной в научно-организационной деятельности организации, являющейся владельцем данной электронной библиотеки. Как уже отмечалось, в электронной библиотеке могут содержаться профили организаций (и/или их подразделений), профили авторов документов, документы, представляющие научные

публикации или научные данные и т.п. Соответственно, на этой основе могут поддерживаться разнообразные семантические связи: между персонами (связи между профилями авторов, например, руководитель-аспирант), между персонами и организацией (связи между профилями авторов и организаций, характеризующие статус автора в организации), между персонами и документами (между профилями авторов и документами) и др.

Более подробно предлагаемая нами категоризация семантических связей информационных объектов в научных электронных библиотеках рассматривается в работе [8].

4. Свойства семантической структуры связей

Обсуждаемая в данной работе структура семантических связей, формируемая и поддерживаемая над контентом электронной библиотеки, порождается бинарными ориентированными семантическими связями между информационными объектами библиотеки, являющимися документами в ее коллекциях информационных ресурсов, а также профилями организаций и авторов документов – сотрудников организаций.

Ранее отмечалось, что семантические связи, определяемые в библиотеке явным образом в виде структурированных данных, представляются и могут динамически поддерживаться как самостоятельные информационные объекты. Информационные объекты-связи категоризируются, как было описано выше, и в рамках каждой категории типизируются в соответствии с их семантикой. Таким образом, каждый экземпляр устанавливаемых в библиотеке связей относится к какой-либо категории, а в рамках категории к какому либо типу связей этой категории. Свойства экземпляров объектов-связей задаются значениями атрибутов, определенных для соответствующих типов связей. Между двумя информационными объектами библиотеки может быть определено несколько связей одной или нескольких категорий.

Каждому экземпляру объекта-связи при его создании присваивается некоторое значение уникального идентификатора, а значения его атрибутов наряду с другими возможными свойствами указывают категорию и тип представляемой им связи, идентификатор пользователя, который создает этот объект (устанавливает эту связь), идентификаторы исходного и целевого информационных объектов библиотеки, участвующих в данной связи, дату ее установления.

Семантическая структура связей, поддерживаемая в электронной библиотеке, динамична. Могут устанавливаться новые, а также обновляться или ликвидироваться существующие связи – мнения авторов связей могут изменяться с течением времени. Динамичность структуры связей обусловлена и пополнением контента библиотеки новыми информационными объектами – потенциальными участниками связей.

Как уже указывалось, семантические связи информационных объектов в библиотеке (документов и профилей пользователей и организаций) подразделяются на категории, о которых говорилось выше, и каждой категории соответствует некоторый набор типов связей. Эти наборы представляются в виде словарей типов связей. При необходимости словари могут дополняться в процессе функционирования системы новыми типами связей.

В некоторых категориях связей могут существовать типы связей с противоречивой семантикой. Например, к категории оценочных связей могут относиться связи между документами, одни из которых выражают одобрение или согласие исходного документа с целевым, а другие - опровержение результатов, рассмотренных в целевом документе. Естественно, что между двумя публикациями не могут быть одновременно определены связи этих двух типов, установленные одним и тем же пользователем. Возникновение таких ситуаций должны предотвращать системные механизмы библиотеки. В то же время,

вполне возможны семантически противоречивые связи между двумя информационными объектами, установленные разными пользователями.

Системные механизмы должны обеспечивать выполнение и некоторых других ограничений на создание, обновление и ликвидацию экземпляров связей, участниками которых являются профили авторов или организаций. Для выполнения таких операций пользователь должен обладать необходимыми полномочиями.

Каждая семантическая категория связей между информационными объектами библиотеки и каждый относящийся к ней тип связей образует некоторый слой в структуре связей. Таким образом, в электронной библиотеке, механизмы которой обладают рассматриваемой функциональностью, поддерживается многослойная структура семантических связей принадлежащих ей информационных объектов, которая при достижении достаточной ее представительности становится весьма значимым полигоном для анализа и поддержки научной и научно-организационной деятельности.

Особого внимания заслуживает вопрос о семантике связей, и он обсуждается в следующем разделе.

5. Семантика связей

Выше мы рассмотрели предлагаемый нами подход к решению первой из задач, указанных в разд. 1. Перейдем теперь к обсуждению подхода к решению второй задачи – к решению вопроса о том, какие типы семантических связей между информационными объектами целесообразно поддерживать в научной электронной библиотеке.

Известны попытки систематической классификации семантических связей между единицами информационных ресурсов и/или их компонентами, предпринятые для использования их результатов в электронных библиотеках, издательских системах, представления знаний в среде Семантического Веб. Рассмотрим наиболее известные разработки в этой области.

Специалистами в области биомедицины из Оксфордского и Болонского университетов разработан модульный онтологический комплекс SPAR (*the Semantic Publishing and Referencing Ontologies*) [14, 17]. Он состоит из восьми независимых повторно используемых детализированных онтологий, которые позволяют описывать семантику библиографических объектов, а также их отношений со связями цитирования, с библиографическими записями, с компонентами документов, а также с различными аспектами процесса научных публикаций. Фактически, они представляют собой таксономии и описаны в языках OWL2 DL и RDF консорциума W3C. Первые четыре из них (FaBiO, CiTO, BiRO and C4O) полезны для описания библиографических объектов, библиографических записей и источников в списках литературы в публикациях, связей цитирования, контекстов цитирования и их связей с релевантными разделами цитируемых публикаций, а также для организации библиографических записей и ссылок в библиографиях, упорядоченных списках источников и в библиотечных каталогах. Четыре остальных онтологии (DoCO, PRO, PSO and PWO) служат для создания структурированных управляемых словарей компонентов документов, ролей публикаций, состояний публикаций и потоков работ в издательских процессах.

Специалистами в области нейромедицины из Главного госпиталя в Массачусетсе и Медицинской школы в Гарварде разработана онтология SWAN (*Semantic Web Applications in Neuromedicine*) [12]. Как и SPAR, эта онтология состоит из набора онтологий-модулей. Онтологии, входящие в состав SWAN, также описаны на языке описания онтологий OWL DL. Как указывается в спецификации SWAN, цель этой онтологии – обеспечение в рамках Семантического Веб комфортной среды, называемой авторами *социально-технической экосистемой*, которая позволяет создавать и сохранять семантический контекст научных

коммуникаций, обеспечивает доступ к нему, его интеграцию, а также обмен неструктурированной или слабоструктурированной цифровой научной информацией.

К рассматриваемому направлению примыкает также рекомендация SKOS (*Simple Knowledge Organization System*) [18] консорциума W3C. Спецификация SKOS предназначена для поддержки использования систем организации знаний, таких как тезаурусы, схемы классификации, таксономии и рубрикаторы (*Subject Heading Systems*) в среде Семантического Веб. Для этой цели определяется концептуальная схема (в спецификации она называется *общей моделью данных*) для совместного использования и связывания систем организации знаний средствами Веб. Унификация концептуальной схемы, определяемой спецификацией SKOS, создает возможности для относительно нетрудоемкой интеграции существующих систем организации знаний в Семантический Веб.

Следует отметить здесь важную тенденцию конструирования сложных онтологий, предназначенных для достаточно широкой сферы применения: они строятся по модульному принципу. Такой подход облегчает их повторное использование. В различных конкретных ситуациях может не существовать потребности в использовании полной онтологии. Тогда используется только нужный ее модуль. При этом модульность облегчает также интеграцию с другими онтологиями. Примером такой интеграции может служить комплекс SPAR, в котором использованы элементы SWAN. В свою очередь, в SWAN используется SKOS.

Следует, наконец, упомянуть также имеющий отношение к обсуждаемому в этом разделе вопросу проект CERIF [9], который в 1980-1990-е годы реализовывался при поддержке Европейской комиссии, а в 2000 г. был передан ею под опеку международной научной организации euroCRIS. Главная цель этого проекта фактически заключается в создании стандарта так называемой *полной модели данных (Full Data Model*), которая рассматривается как единая основа создания информационных систем (*Current Research Information Systems*, CRIS) для поддержки научно-организационной деятельности в разных странах и научных организациях. Благодаря стандартизации модели данных обеспечивается интероперабельность таких систем. В последнее время в составе CERIF была предложена спецификация стандартизованной семантики полной модели данных [10], иначе говоря, онтология, определяющая систему терминов для обозначения сущностей этой модели и связей между ними.

Рассмотренные результаты в области классификации возможных семантических связей между научными документами или другими объектами научной деятельности могут использоваться в качестве основы для семантического структурирования контента научных электронных библиотек. В разработке прототипа механизмов семантической структуризации контента в системе Соционет мы использовали именно фрагменты рассмотренных онтологий – CiTO, DoCo, SWAN, SKOS и CERIF. Наиболее существенную часть нашей гибридной онтологии составляют фрагменты из CiTO и DoCo.

Онтология CiTO (*the Citation Typing Ontology*) [13, 15] обеспечивает возможности для характеристики природы связей цитирования, как фактологических (например, «цитирует как источник данных» или «цитирует как основоплагающую»), так и риторических (например, «уточняет» или «опровергает»). При этом учитываются как непосредственные и явные связи цитирования, так и косвенные и неявные. Онтология DoCO (*the Document Components Ontology*) [16] классифицирует составные части документов. Она предоставляет структурированный управляемый словарь их компонентов. Это, например, «Введение», «Обсуждение», «Благодарности», «Список использованных источников», «Приложение» и т.д.

Важно здесь отметить, что результаты указанных исследований могут быть использованы для категоризации некоторых видов связей на множестве научных

документов, включающих не только научные тексты. Это обстоятельство имеет в нашем случае существенное значение, поскольку, как отмечалось ранее, нас интересуют не только связи между текстовыми документами электронной библиотеки или документами, содержащими научные данные, но и связи, участниками которых являются также профили организаций и их сотрудников – авторов и пользователей библиотеки.

Более подробно онтология связей, используемая нами в системе Соционет, обсуждается в работе [8].

6. Функции механизмов поддержки и использования связей в Соционет

Для формирования в электронной библиотеке и продуктивного использования описанной многослойной структуры семантических связей информационных объектов ее контента необходимо, чтобы система управления библиотекой включала механизмы, предоставляющие необходимые операционные возможности. Кратко рассмотрим состав и функции таких механизмов, которые реализуются в системе Соционет, и проиллюстрируем их примерами.

Механизмы формирования и поддержки словарей связей. В реализуемом в системе Соционет прототипе не предполагается поддерживать формальные спецификации онтологии связей. Поддерживаемая онтология имеет модульную структуру. Каждый модуль соответствует некоторой категории связей, и для него поддерживается управляемый словарь относящихся к нему типов связей. Рассматриваемые механизмы позволяют системному администратору формировать и модифицировать эти словари. Пользовательский интерфейс механизмов создания связей предоставляет доступ к словарям и справочной информации, необходимой для их корректного использования.

Механизмы управления связями. Эти механизмы позволяют авторизованному пользователю создавать в модерируемом режиме связи между информационными объектами библиотеки. Как указывалось выше, связи создаются как информационные объекты специального типа. При создании новой связи используются управляемые словари типов связей. Новая связь создается только при условии, если она не противоречит уже существующему набору связей между информационными объектами – ее участниками. Механизмы управления связями позволяют также ликвидировать существующие связи и обновлять значения их атрибутов. В частности, может быть изменен и тип существующей связи. В рассматриваемой группе механизмов важное место занимает механизм мониторинга состояния структуры связей. При появлении новой связи, удалении связи или некоторых изменениях атрибутов связей этот механизм генерирует сообщения авторам документов - участников таких связей, стимулируя тем самым их реакцию на эти события.

Механизмы обработки запросов. Эти механизмы выполняют довольно большой набор функций, позволяющих получать разнообразную информацию о структуре связей в библиотеке. Прежде всего, это статистическая информация. Можно, в частности, запросить количество связей заданных типов или некоторой категории, исходящих из данного информационного объекта библиотеки либо входящих в него. Например, можно узнать, какое имеется количество положительных или негативных оценок данной работы.

Другая группа запросов позволяет получить перечень информационных объектов библиотеки, связанных с заданным объектом как исходным или целевым в связях заданных типов или категорий. Запросы этого вида позволяют, например, выяснить, на результаты каких публикаций опирается некоторая конкретная работа или, наоборот, в каких публикациях получены результаты, основанные на данной работе. При этом можно учитывать как непосредственные, так и транзитивные связи. В качестве критерия отбора интересующих пользователя связей или его компонента может также использоваться идентификация автора связей.

Важную группу запросов составляют операции над полным графом связей. Здесь можно решать множество различных задач, связанных как с анализом топологии графа и вычленением подграфов с заданными свойствами, так и с визуализацией подграфов. Например, можно вычленить и визуализировать из многослойной структуры связей слой, соответствующий связи некоторого типа, такой как связь, указывающая на использование одного документа библиотеки как основополагающего для других документов. Можно также запросить подграф, образованный связями, относящимися к категории развития научных результатов, и указать, что ему должна принадлежать некоторая имеющаяся в библиотеке общепризнанная основополагающая публикация в некоторой области исследований. Полученный подграф будет характеризовать логику развития данной области науки, конечно, если в библиотеке будет достаточно основательно представлены публикации, относящиеся к этой области. Еще одним примером операций над полным графом связей библиотеки является операция вычленения из него подграфа связей, установленных данным пользователем, возможно, с указанием в запросе также категории или конкретного типа связей.

Отметим, наконец, что визуализация графа связей или некоторого его подграфа может быть использована для навигации в структуре связей и просмотра свойств отдельных связей и участвующих в них документов.

Заключение

Формирование и поддержка семантических связей, хотя эти операции распределяются между пользователями и не являются слишком трудоемкими для каждого из них, тем не менее, требуют определенных затрат. Однако эти затраты окупаются теми новыми возможностями, которые обеспечиваются в электронных научных библиотеках, использующих рассмотренные технологии.

Семантическое структурирование контента научных электронных библиотек с явным структурированным представлением семантических связей между содержащимися в них информационными объектами открывает новые возможности для наукометрического анализа и исследований истории развития научного знания в конкретных направлениях науки. В отличие от традиционных индексов цитирования, наукометрические измерения в такой среде приобретают смысловую дифференциацию.

Функционирующая в онлайновом режиме электронная библиотека, обладающая рассматриваемыми возможностями организации информационных ресурсов и операционными средствами, вместе с тем, становится полигоном для использования новых форм коммуникаций в научном сообществе. Она позволяет обеспечить оперативный обмен мнениями, своеобразные дискуссии относительно представленных в электронной библиотеке научных документов в виртуальном информационном пространстве, осуществляемые без обременительных организационных и административных усилий. Создание комфортной онлайновой среды для поддержки такой деятельности может стимулировать активное участие в ней заинтересованных исследователей.

Прототип обсуждаемой в этой работе информационной среды в настоящее время частично разработан и развивается в рамках системы Соционет. Важно отметить, что при этом используется унифицированная технология как для формирования, поддержки и использования структуры семантических связей контента электронной библиотеки, так и для поддержки «живых» публикаций в электронной библиотеке [6, 7].

Замечание. ***О семантике связей цитирования***

Наукометрические измерения, основанные на связях цитирования между научными публикациями, активно развиваются в последние годы. Создан ряд авторитетных

национальных и международных индексов цитирования. Однако осуществляемые в них в настоящее время измерения основаны на «немых» связях – библиографических ссылках в списках литературы публикуемых работ, которые никак не отражают семантики представляемых ими связей между публикациями.

На наш взгляд, важным шагом в развитии функциональных возможностей индексов цитирования, а также технологии, обсуждаемой в данной работе, стало бы создание научным сообществом стандарта классификатора библиографических ссылок, а также принятие таких норм в издательском деле, которые предписывают авторам представляемых для публикации работ индексировать на его основе ссылки, содержащиеся в послестатейном списке использованных источников и непосредственно в тексте работы (в подстрочных и концевых сносках) аналогично тому, как это делается с индексированием публикаций по УДК и ББК. При этом наряду с «универсальным» для всех областей знаний классификатором ссылок могут быть созданы специализированные классификаторы для некоторых областей, учитывающие их специфику.

Литература

1. Когаловский М.Р., Паринов С.И. Метрики онлайновых информационных пространств //Экономика и математические методы. – 2008. – Вып. 2.
2. Когаловский М.Р., Паринов С.И. Использование связей цитирования для наукометрических измерений в системе Соционет. Электронный депонент Соционет, 2009. http://socionet.ru/ publication.xml?h=repec:rus:rssalc:web-32
3. Паринов С.И. СОЦИОНЕТ.РУ как модель информационного пространства 2-го поколения //Информационное общество. - 2001, вып. 1, с. 43-45. http://emag.iis.ru/arc/infosoc/emag.nsf/BPA/709c3727bab54cf4c3256c01002d2e6e
4. Паринов С.И. Информационные хабы. Электронный депонент Соционет? 2007. http://socionet.ru/publication.xml?h=repec:rus:mqijxk:9
5. Паринов С.И. Концепция виртуальной научной среды "Открытая Наука" //Труды международной суперкомпьютерной конференции "Научный сервис в сети Интернет: суперкомпьютерные центры и задачи", Новороссийск, 20-25 сентября 2010 г. – М.: Изд-во МГУ, 2010, стр. 473-481. Электронная авторская версия: http://socionet.ru/ publication.xml?h=repec: rus:mqijxk:24
6. Паринов С.И., Когаловский М.Р. «Живые» документы в электронных библиотеках //Прикладная информатика. – 2009. - № 6. Авторская версия: http://socionet.ru/publication.xml?h=repec:rus:isyigw:article-215
7. Паринов С.И., Когаловский М.Р. Технология поддержки электронных научных публикаций как «живых» документов. Труды XI Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2009», Петрозаводск, 17-21 сентября 2009 г. – Петрозаводск: КарНЦ РАН, 2009.
8. Паринов С.И., Когаловский М.Р. Технология семантического структурирования контента научных электронных библиотек. Представлена на XIII Всероссийскую научную конференцию «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2011», Воронеж, октябрь 2011.
9. CERIF 2008 - Final Release (1.2). http://www.eurocris.org/Index.php?page=CERIF2008&t=19
10. CERIF 2008 – 1.2 Semantics. euroCRIS, November 22, 2010. http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_Semantics.pdf
11. Parinov S. The electronic library: using technology to measure and support Open Science. In Proc. of the World Library and Information Congress: 76th IFLA General

Conference and Assembly. 10-15 August 2010, Gothenburg, Sweden. pp. 1-13. Авторская версия: http://socionet.ru/publication.xml?h=repec:rus:mqijxk:25

12. Semantic Web Applications in Neuromedicine (SWAN) Ontology. W3C Interest Group Note, 20 October 2009. http://www.w3.org/TR/2009/NOTE-hcls-swan-20091020/

13. Shotton D. CiTO, the Citation Typing Ontology. J. of Biomedical Semantics 2010, 1(Suppl 1): S6. http://www.jbiomedsem.com/content/1/S1/S6

14. Shotton D. 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/

15. Shotton D., Peroni S. CiTO, The Citation Typing Ontology, v2.0. – 2011. http://purl.org/spar/cito/

16. Shotton D., Peroni S. DoCO, the Document Components Ontology. – 2011. http://speroni.web.cs.unibo.it/cgi-bin/lode/req.py?req=http:/purl.org/spar/doco

17. Shotton D. and Peroni S. Semantic annotation of publication entities using the SPAR (Semantic Publishing and Referencing) Ontologies /Beyond the PDF Workshop, La Jolla, 19 January 2011.
http://imageweb.zoo.ox.ac.uk/pub/2010/Publications/Shotton&Peroni_semantic_annotation_of_publication_entities.pdf

18. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 August 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/