М. Р. Когаловский, С. И. Паринов, Классификацияи использование семантических связей между информационными объектами в научных электронных библиотеках, *Информ. и её примен.*, 2012, том 6, выпуск 3, 32–42

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 3. С. 32–42

КЛАССИФИКАЦИЯ И ИСПОЛЬЗОВАНИЕ СЕМАНТИЧЕСКИХ СВЯЗЕЙ МЕЖДУ ИНФОРМАЦИОННЫМИ ОБЪЕКТАМИ В НАУЧНЫХ ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ*

М. Р. Когаловский[1], С. И. Паринов[2]

Аннотация: Обсуждается подход, обеспечивающий повышение информационной ценности контента научной электронной библиотеки благодаря поддержке классифицированных семантических связей между содержащимися в ней информационными объектами. Рассматривается реализация предлагаемого подхода на основе отечественной системы Соционет, объединяющей большое число научных электронных библиотек и являющейся де-факто институциональным исследовательским информационным пространством Отделения общественных наук Российской академии наук.

Ключевые слова: электронная библиотека; информационный объект; коллекция информационных ресурсов; семантическая связь; классификатор связей; онтология; наукометрия

1 Введение

Информационные объекты, содержащиеся в научных электронных библиотеках (статьи, книги, персональные профили авторов, профили организаций и др.), имеют многие другие связи (отношения) друг с другом, помимо обычно поддерживаемых средствами управления информационными ресурсами и отображаемых программными интерфейсами электронных библиотек. Большое число видов связей из-за неразвитости технологий пока остается за рамками электронных библиотек, часто только в сознании исследователей. Такие связи являются ненаблюдаемыми и не фиксируются в цифровой форме.

К обычно отображаемым связям между информационными объектами электронных библиотек относятся, например, связи между научными публикациями (статья, книга и т. п.) и персональными профилями их авторов, а также с профилями организаций, в которых были получены соответствующие результаты исследований. Кроме этого, статьи, снабженные кодами тематических классификаторов, имеют связи с тематическими рубриками классификаторов научных дисциплин. Все чаще в электронных библиотеках встречаются связи между статьями и комментариями их читателей. В некоторых крупных электронных библиотеках поддерживаются связи цитирования, профили авторов и связи публикаций с ними. Механизмы для этой цели имеются, в частности, в электронной библиотеке ACM (Association for Computing Machinery) [1], научной библиотеке eLibrary РФФИ [2], в Академии Google (Google Scholar) [3]. Однако в этих и в других аналогичных случаях связи цитирования не несут никакой информации, кроме самого факта цитирования, не характеризуют семантики отношения между цитирующим и цитируемым текстовым документом. Будем называть такие связи «*немыми*».

Связи между информационными объектами в электронных библиотеках, в том числе и связи цитирования, обладают семантикой, и она может быть явно описана способом, доступным для пользователей и системных механизмов, и продуктивно использована. На ее основе может, в частности, формироваться более дифференцированная наукометрическая статистика, учитывающая позитивное, негативное или иное отношение автора цитирующего документа к цитируемому. Связи с явно описанной семантикой будем называть далее *семантическими связями*.

Как уже отмечалось, между содержащимися в научной электронной библиотеке информационными объектами могут поддерживаться не только семантические связи, такие как «немые» связи цитирования и другие, упоминаемые ранее. Например, семантическая связь может создаваться для выражения мнения автора одного из информационных объектов или экспертов о существовании некоторого отношения между контентом двух информационных объектов в ситуации, даже когда этот факт не отмечен явным образом в контенте рассматриваемых информационных объектов.

* Работа поддерживается РГНФ, проект 11-02-12026-в.

[1] Институт проблем рынка Российской академии наук, kogalov@cemi.rssi.ru

[2] Центральный экономико-математический институт Российской академии наук, sparinov@gmail.com

Явным образом описанные и поддерживаемые в библиотеке семантические связи могут быть *классифицированы* на основе характера отношений между информационными объектами — участниками связей. Введение классификации приводит к образованию многослойной семантической структуры контента электронной библиотеки, каждый слой которой соответствует некоторому классу семантических связей. Такая структура может служить источником информации для проведения качественно новых наукометрических измерений, для исследования структурных свойств корпуса знаний в различных областях науки, представительным образом отраженного в контенте электронной библиотеки.

Онлайновый режим функционирования научной электронной библиотеки позволяет реализовать ее систему управления таким образом, чтобы не только поддерживалась семантическая структура контента и обрабатывались пользовательские запросы, касающиеся ее характеристик, но и предоставлялась пользователям возможность самостоятельно в децентрализованном режиме описывать и создавать семантические связи. Могут быть также предусмотрены мониторинг состояния структуры связей и автоматическое оповещение авторов информационных объектов о том, что некоторый их объект стал участником вновь учрежденной связи или что ликвидирована существующая связь, в которой этот объект являлся участником. Благодаря этому автор информационного объекта, получивший указанное оповещение, стимулируется тем самым реагировать на эту ситуацию, если событие, о котором он информируется, противоречит его представлениям.

Таким образом, в онлайновой электронной библиотеке, в которой поддерживаются классифицированные семантические связи, может быть обеспечен комплекс новых возможностей:

- поддержка многослойной структуры семантических связей;
- создание новых связей и аннулирование существующих связей не только администраторами информационных ресурсов, но и пользователями системы в децентрализованном режиме;
- формирование дифференцированной по классам семантических связей статистики связей, в частности, касающейся связей цитирования;
- оповещение авторов представленных в электронной библиотеке информационных объектов об их включении в новые связи или об аннулировании некоторых связей, в которых они были участниками.

Обладающая такими возможностями информационная среда обеспечивает качественно новые технологии для научной и научно-организационной деятельности, открывает новые возможности для коммуникаций в научном сообществе. В предлагаемой статье обсуждается подход авторов к созданию такой среды, реализуемый в научном информационном пространстве Соционет [4].

Остальная часть статьи организована следующим образом. В разд. 2 уточняется постановка рассматриваемой в статье проблемы и предлагаются пути ее решения. В разд. 3 обсуждается вопрос о классификации связей и дается краткий обзор известных исследований в рассматриваемой области. В разд. 4 обсуждаются принципы представления семантических связей в электронной библиотеке как самостоятельных информационных объектов. Свойства семантических связей рассматриваются в разд. 5. В разд. 6 обсуждается реализация предлагаемого в статье подхода в среде системы Соционет. В заключении подводятся итоги обсуждения.

2 Уточнение постановки проблемы

Коллекции информационных ресурсов традиционных научных электронных библиотек состоят из множества объектов определенных типов: 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. Коллекции могут содержать также сведения об авторах представленных в них публикаций, об организациях, в которых они работают, и информационные объекты других типов.

В последние годы на основе библиографических ссылок, содержащихся в публикациях, которые выпускаются в авторитетных периодических изданиях, начали создаваться индексы цитирования, обеспечивающие формирование библиометрической статистики. 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. Однако с фактом цитирования связана еще и некоторая не отображаемая при этом семантика, выражающая отношение автора цитирующего документа к цитируемому источнику или какое-либо иное семантическое отношение между цити-

рующей и цитируемой публикацией. 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, и в существующих системах такой анализ обычно не производится.

Наряду со связями цитирования между информационными объектами научных электронных библиотек существуют разнообразные другие семантические связи. Например, связь может указывать, что ее целевой информационный объект содержит научные результаты, базирующиеся на результатах, описанных в исходном объекте связи, или что в исходном объекте связи опровергается результат, изложенный в ее целевом объекте. Связь может также указывать, что ее исходный информационный объект является новой редакцией целевого объекта или представляет собой его составную часть, например аннотацию.

Существует большое разнообразие семантических связей, которые можно при необходимости поддерживать между информационными объектами в библиотеке. Эти связи выявляются в результате участия представителей научного сообщества в процессах, реализующих их научную и научно-организационную деятельность. К числу основных видов таких процессов можно отнести процессы систематизации, классификации и упорядочения корпуса научных знаний (например, при подготовке аналитических обзоров), процессы научной оценки опубликованных результатов (рецензирование работ), процессы продуцирования нового научного знания, процессы создания научных произведений, научно-организационные процессы. Именно на основе информации, рождающейся в результате участия в процессах перечисленных видов, пользователь электронной библиотеки может прийти к выводу о целесообразности создания тех или иных семантических связей между представленными в ней некоторыми информационными объектами.

Определяемые явным и структурированным образом семантические связи могут быть представлены и могут динамически поддерживаться как самостоятельные информационные объекты электронной библиотеки. Такие объекты содержат идентификаторы участвующих в них информационных объектов и значения других атрибутов. Объекты-связи могут быть классифицированы, и их свойства определяются значениями атрибутов, специфических для каждого класса.

В результате определения явным образом описанных классифицированных семантических связей, как уже отмечалось, порождается многослойная семантическая структура контента библиотеки. При этом каждому классу связей соответствует некоторый слой этой структуры, который наряду с полной структурой связей может 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 частями 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 также 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 информационных объектах. 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 структуры семантических связей обеспечивает также дополнительные (навигационные) пути доступа пользователей к информационным объектам библиотеки. Другое направление, где необходима поддержка семантических связей между информационными объектами электронной библиотеки, — это технология «живых» публикаций, подробно рассмотренная в [5, 6].

Для эффективного использования новых возможностей, которые обеспечиваются благодаря поддержке в онлайновой электронной библиотеке многослойной структуры семантических связей представленных в ней информационных объектов, 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 располагать механизмами, позволяющими пользователям самостоятельно устанавливать, модифицировать или удалять семантические связи в рамках их полномочий, а также обеспечивать мониторинг изменений состояния структуры семантических связей. Механизмы мониторинга позволяют автоматически оповещать авторов информацион-

ных объектов о том, что некоторый их информационный объект стал участником вновь учрежденной связи или что ликвидирована существующая связь, в которой он являлся участником, либо об изменениях значений ее атрибутов.

Семантическое структурирование контента научных электронных библиотек представляет значительно больший интерес, если оно поддерживается на представительном репозитории научных информационных объектов. 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 [7]. Поддержка и исследование семантической структуры в создаваемом на ее основе крупном интегрированном контенте, формируемом на федеративных принципах рядом исследовательских организаций, дает возможность изучать структуру результатов научных исследований не только отдельных научных коллективов или школ, но и целых направлений науки и областей знаний.

Обеспечение возможностей поддержки в научных электронных библиотеках явно представленных классифицированных семантических связей между содержащимися в них информационными объектами в сочетании с методами мониторинга изменений структуры этих связей и основанными на такой структуре новыми функциональными возможностями является, по мнению авторов, весьма перспективным новым направлением развития научных электронных библиотек. Для эффективного практического использования обсуждаемых возможностей необходимо решить следующие задачи:

- разработать способы и конкретные форматы представления семантических связей между информационными объектами электронной библиотеки в виде самостоятельных информационных объектов специального типа;
- создать классификатор семантических связей, которые целесообразно поддерживать в научных электронных библиотеках;
- определить операционные возможности, которые должна обеспечивать система управления научной электронной библиотекой для того, чтобы извлекать в достаточно полной мере информацию, содержащуюся в структуре семантических связей представленных в ней информационных объектов.

В данной работе обсуждается предлагаемый авторами подход к решению этих задач и его реализация в среде крупного отечественного онлайнового научно-образовательного пространства, поддерживаемого системой Соционет [4], основанного на технологии открытых архивов и содержащего большой объем информационных ресурсов по социально-экономической тематике. Соционет функционирует уже более десяти лет и приобрела в последние годы де-факто институциональный статус в Отделении общественных наук. Информационное пространство Соционет содержит также публикации ряда образовательных учреждений и других организаций. Соционет стала полигоном для проведения исследований в области перспективных технологий электронных библиотек. Постоянно проводятся работы по расширению разнообразия типов представляемых в этой системе информационных ресурсов и развитию функциональности механизмов управления библиотекой. Основные идеи данной работы сформировались на основе более ранних публикаций [8–11] и были детально представлены на конференции RCDL-2011 [12].

3 Классификация связей и известные работы в данной области

Проблемы структуризации крупных коллекций информационных ресурсов электронных библиотек и классификации семантических связей в последние годы привлекают большое внимание исследователей. Известны попытки систематической классификации семантических связей между единицами информационных ресурсов и/или их компонентами, предпринятые для использования в электронных библиотеках, издательских системах, для представления знаний в среде Семантического Веба. Рассмотрим наиболее известные разработки в этой области.

Прежде всего следует упомянуть работы по распознаванию и классификации используемых в научных статьях языковых конструкций (для английского языка и отдельных научных дисциплин), проводимые средствами программного обеспечения компании Xerox. Они позволили эмпирическим путем выявить некоторые устойчивые виды семантических отношений, существующих как между разделами внутри научной статьи, так и между статьей и цитируемыми в ней материалами [13, 14]. Эмпирическая классификация поводов цитирования в научных статьях предлагается также в [15]. В этой работе выделен ряд их типичных вариантов: «слабость цитируемого подхода», «автор использует цитируемую работу как основу или исходную точку» и др. Другой подход к развитию

классификации семантических связей реализуется в исследованиях модульности научных документов [16].

К рассматриваемому направлению примыкает также рекомендация SKOS (Simple Knowledge Organization System) [17] консорциума W3C. Эта спецификация предназначена для поддержки использования систем организации знаний, таких как тезаурусы, схемы классификации, таксономии и рубрикаторы (Subject Heading Systems) в среде Семантического Веба. Для этой цели определяется концептуальная схема (в спецификации она называется *общей моделью данных*) для совместного использования и связывания систем организации знаний средствами Веба. Унификация концептуальной схемы, определяемой спецификацией SKOS, создает возможности для относительно нетрудоемкой интеграции существующих систем организации знаний в Семантический Веб.

Специалистами в области биомедицины из Оксфордского и Болонского университетов разработан модульный онтологический комплекс SPAR (the Semantic Publishing and Referencing Ontologies) [18, 19]. Он состоит из восьми независимых повторно используемых детализированных онтологий. Фактически каждая из них представляет собой таксономию, описанную на языках OWL2 DL и RDF консорциума W3C. Первые четыре из них (FaBiO — FRBR-aligned Bibiographic Ontology, где FRBR — Functional Requirements for Bibliographic Records); CiTO — Citation Typing Ontology; BiRO — Bibliographic Reference Ontology; C4O — Citation Counting and Context Characterization Ontology) 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 ссылок в библиографиях, в списках источников и в библиотечных каталогах. Остальные онтологии (DoCO — Document Components Ontology; PRO — Publishing Roles Ontology; PSO — Publishing Status Ontology; PWO — Publishing Workflow Ontology) служат для создания структурированных управляемых словарей классов компонентов документов, ролей публикаций, состояний публикаций и потоков работ в издательских процессах.

В Главном госпитале Массачусетса и в Медицинской школе в Гарварде разработана онтология SWAN (Semantic Web Applications in Neuromedicine) [20]. 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 Веба 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.

Нужно отметить здесь важную тенденцию конструирования сложных онтологий, предназначенных для достаточно широкой сферы применения: они строятся по модульному принципу. Такой подход облегчает их повторное использование. Обычно не требуется использовать полную онтологию и берется 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 (Common European Research Information Format) [21], 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 проекте CERIF уделяется большое внимание семантическим аспектам созданной модели. Для этой цели разработаны онтологии CERIF [22, 23].

Рассмотренные результаты в области классификации возможных семантических связей между научными публикациями и/или другими продуктами научной деятельности могут использоваться в качестве основы для семантического структурирования контента научных электронных библиотек. В разработке классификатора семантических связей в обсуждаемом в этой работе проекте авторы использовали фрагменты рассмотренных онтологий — CiTO, DoCo, SWAN, SKOS и CERIF. Наиболее существенную часть предлагаемого классификатора определяют фрагменты онтологий CiTO и DoCo.

Онтология CiTO [24, 25] обеспечивает возможности для характеристики природы связей цитиро-

вания, как фактологических (например, «цитирует как источник данных» или «цитирует как основополагающую»), так и риторических (например, «уточняет» или «опровергает»). При этом учитываются как непосредственные и явные связи цитирования, так и косвенные и неявные. Онтология DoCO [26] классифицирует составные части документов. Она предоставляет структурированный управляемый словарь классов их компонентов, например «Введение», «Обсуждение», «Благодарности», «Список использованных источников», «Приложение» и т. д.

Результаты рассмотренных исследований могут быть использованы для классификации некоторых видов связей на множестве не только текстовых научных информационных объектов. Это обстоятельство имеет в рассматриваемом случае существенное значение, поскольку, как отмечалось ранее, интерес представляют также связи, участниками которых являются профили организаций и их сотрудников — авторов и пользователей библиотеки, а также информационные объекты других типов, не являющиеся текстовыми документами.

Классификатор связей в системе Соционет предусматривает разбиение множества семантических связей информационных объектов (текстовых информационных объектов, профилей пользователей и организаций и информационных объектов других типов) на категории (оценочные связи, научно-организационные связи и др.). Каждой категории соответствует некоторый набор классов связей. Эти наборы представляются в виде словарей классов связей. При необходимости в процессе функционирования системы может пополняться состав категорий и словари могут дополняться новыми классами связей. Более подробно принципы организации и содержание классификатора семантических связей, используемого в системе Соционет, обсуждается в [12].

Помимо рассмотренных выше работ, появляются также новые публикации, посвященные затронутой проблеме. Однако авторам не известны проекты, в которых реализован описанный выше комплекс возможностей использования классифицированных семантических связей между информационными объектами научных электронных библиотек.

4 Семантические связи как информационные объекты библиотеки

В электронных библиотеках традиционно с помощью гиперссылок поддерживаются связи между каталогами и описываемыми в них информационными объектами. В системе Соционет таким же образом поддерживаются связи цитирования, связи с профилями авторов и организаций и некоторые другие. Для этого в Соционет имеются метаданные, описывающие информационные объекты, их авторов (профили авторов), коллекции информационных ресурсов, организации — места работы авторов (профили организаций) и др. В таком случае связи между информационными объектами представляются как атрибуты метаданных, описывающих эти информационные объекты. С использованием связей такого вида можно анализировать структуру связей, осуществлять наукометрические измерения, визуализировать структуру связей.

Однако при таком традиционном способе представления связей явным образом не отображается семантика связей. Например, для связи цитирования одного информационного объекта с другим отсутствует информация, характеризующая цель цитирования, оценку цитируемой работы и другие характеристики. Предлагаемая далее модель связей между информационными объектами научной электронной библиотеки устраняет это ограничение.

В общем случае связи могут представляться двумя способами. При использовании первого способа, представленного выше, данные, описывающие связи, содержатся в метаданных одного из связываемых объектов, например в метаданных исходного объекта связи. Однако поскольку в электронной библиотеке, построенной на федеративных принципах, изменять метаданные может только их автор или уполномоченный автором администратор информационных ресурсов, то при этом способе только они и могут создавать связи этого объекта с другими информационными объектами.

При втором способе создаваемые связи представляются как самостоятельные информационные объекты. Такой способ является более универсальным и предпочтительным, так как он охватывает все многообразие возможных ситуаций, обеспечивает более богатые возможности анализа структуры связей, которые значительно проще реализуются, и он позволяет создавать связи любому пользователю, поскольку при этом не затрагиваются недоступные ему метаданные связываемых объектов.

Описание связи в обоих представлениях должно включать уникальный идентификатор целевого объекта связи, а также может включать атрибуты, характеризующие семантику связи, различного рода комментарии и пр. Если связь создается как самостоятельный информационный объект,

то ее описание в дополнение к уже перечисленному должно включать: уникальный идентификатор данного объекта-связи в библиотеке; уникальный идентификатор пользователя, создающего данную связь; уникальный идентификатор исходного объекта связи (рассматриваются ориентированные бинарные связи), а также дата создания или изменения связи. Для описания семантики связи используется имя класса связи, выбираемое из поддерживаемых контролируемых словарей, а также значения свойств конкретного экземпляра связи, определяемые пользователем. Полномочия на создание связей между информационными объектами предоставляются только зарегистрированным в библиотеке пользователям, что обеспечивает автоматическую фиксацию идентификатора пользователя, создающего связи, при его входе в систему.

Рассмотрим процедуру создания связи между двумя информационными объектами в системе Соционет, в которой реализованы оба способа представления связей. При первом способе создание связи осуществляется автором исходного информационного объекта связи или его представителем. Рассмотрим процедуру второго способа.

Множество параметров, влияющих на создание связи, включает: тип исходного объекта связи; тип целевого объекта связи; множество категорий связей, учрежденных в системе для заданной пары типов связываемых объектов; множество словарей классов связей, предусмотренных в системе для связей заданной категории; множество классов связей в словаре, выбранном для создания связи между объектами заданных типов.

Рассматриваемая процедура состоит из следующих шагов:

1. Пользователь выбирает пару связываемых информационных объектов.
2. Из множества категорий связей, предусмотренных в системе для выбранной пары типов объектов, выбирается конкретная категория. Если подходящей категории не существует, пользователь имеет возможность предложить новую категорию и предоставить соответствующий ей словарь классов связей для включения в систему. Это предложение вступит в силу только после одобрения администратором системы.
3. Если подходящая категория связей выбрана, то открывается соответствующий словарь классов связей.
4. Если в словаре имеется подходящий класс связей, характеризующий требуемое семантическое отношение между заданной парой объектов, то пользователь его выбирает. Если же такой класс отсутствует в данном словаре, пользователь может предложить подходящий класс связей для пополнения данного словаря. Предложение вступит в силу только после одобрения его администратором системы или соответствующего словаря.
5. По желанию пользовать может привести в описании связи комментарий, объясняющий мотивы ее создания.
6. Сформированный информационный объект-связь сохраняется. При этом система запрашивает у пользователя, в какую его коллекцию следует поместить созданный объект, а также уникальный идентификатор этого объекта в соответствующей коллекции.

Рассмотренная процедура обеспечивает создание информационного объекта, представляющего требуемую связь среди других объектов библиотеки. При этом также осуществляется проверка непротиворечивости семантики новой связи с уже существующими связями между данными объектами, созданными тем же пользователем.

Хотя формирование семантических связей между информационными объектами требует определенных трудозатрат, в результате информативность контента научной электронной библиотеки существенно повышается. Создаются также дополнительные возможности для анализа семантической структуры контента.

Поддержка развитой структуры семантических связей в научной электронной библиотеке с достаточно представительным контентом позволяет в результате их анализа осуществлять наукометрические измерения, использовать технологии «живых» публикаций [5, 6], а также получать качественно новую информацию о развитии научных знаний в конкретных областях исследований и о вкладе отдельных ученых.

В описанной выше процедуре предполагается, что любой зарегистрированный пользователь научной электронной библиотеки может создавать связи между любыми ее информационными объектами. При определении семантики связей их создатель выражает свое субъективное мнение, которое в некоторых случаях может вызывать несогласие или протест как авторов объектов, участвующих в данных связях, так и других членов научного сообщества. Например, могут вызывать протесты случаи, когда устанавливаются семантические связи, несущие негативную оценку некоторого научного произведения (опровержение, высмеивание, обвинение в плагиате и т. п.).

Как известно, научная истина устанавливается в борьбе мнений. Поэтому если научное сообщество начинает использовать подобные технические средства, то с учетом потенциального конфликта интересов научная среда должна предоставлять ученым равные права и одинаковый доступ к использованию этих средств, а также надежную фиксацию профессиональной и социально-этической ответственности ученого за характер использования им данных средств.

Для выполнения данных принципов, по мнению авторов, крайне важно обеспечить модерирование всех создаваемых связей с точки зрения соблюдения авторами научной этики, а также наличия в создаваемых связях признаков добавленной научной «стоимости» или научного вклада (исключение связей с чисто эмоциональным или ненаучным содержанием).

В системе Соционет пользователи создают связи в своем личном (закрытом от свободного доступа) пространстве. Такое пространство с сервисами для его использования предусматривается для авторов или администраторов информационных ресурсов в системе и называется их Личной зоной. Создаваемые в Личной зоне объекты-связи предлагаются далее для включения в общедоступные информационные ресурсы. Они становятся общедоступными только после одобрения модератором.

5 Свойства семантических связей

Обсуждаемая в данной работе структура семантических связей, формируемая и поддерживаемая над контентом электронной библиотеки, порождается бинарными ориентированными семантическими связями между информационными объектами библиотеки, составляющими ее коллекции информационных ресурсов.

Как уже отмечалось, 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 классифицируются в соответствии с их семантикой. Таким образом, каждый экземпляр создаваемых в библиотеке связей относится к какой-либо категории, а в рамках категории — к какому-либо классу связей этой категории. Свойства экземпляров объектов-связей задаются значениями атрибутов, определенных для соответствующих классов 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 класс представляемой им связи, идентификатор пользователя, который создает этот объект-связь, идентификаторы исходного и целевого информационных объектов библиотеки, участвующих в данной связи, дату ее создания.

Структура семантических связей, поддерживаемых в электронной библиотеке, динамична. Могут создаваться 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.

В некоторых категориях связей существуют классы связей с противоречивой семантикой. Например, к категории оценочных связей относятся связи между информационными объектами, которые выражают одобрение или согласие исходного объекта с целевым, а также связи, выражающие опровержение результатов, представленных в целевом информационном объекте. Естественно, что между двумя информационными объектами не могут одновременно существовать связи этих двух классов, созданные 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, созданные разными пользователями. Системные механизмы должны контролировать выполнение и некоторых других ограничений на создание, обновление и ликвидацию экземпляров связей. К ним относятся, в частности, ограничения доступа — для выполнения таких операций пользователь должен обладать необходимыми полномочиями.

Каждая семантическая категория связей между информационными объектами библиотеки и каждый относящийся к ней класс связей, как уже отмечалось, образуют некоторые слои 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 ана-

лиза и поддержки научной и научно-организационной деятельности.

6 Реализация предлагаемого подхода в системе Соционет

Для формирования в электронной библиотеке и продуктивного использования 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 предусмотрены в системе Соционет.

Механизмы формирования и поддержки словарей связей. Классификатор семантических связей в системе Соционет, как уже отмечалось, имеет модульную структуру и представлен в виде совокупности управляемых словарей.

Основу разработанных словарей составляют упоминавшийся выше комплекс онтологий SPAR (в частности, онтологии CiTO и DoCo), спецификация SKOS консорциума W3C, онтология проекта SWAN, а также один из разделов CERIF, определяющий семантику связей.

Каждый из словарей соответствует некоторой предусмотренной в классификаторе категории связей и содержит имена относящихся к ней классов 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 классов связей. Новая связь создается только при условии, если ее создание не нарушает заданных ограничений (см. разд. 5).

В системе Соционет поддерживается множество типов информационных объектов — статьи, монографии, диссертации или авторефераты диссертаций, профили пользователей, профили организаций, рубрикаторы, научные артефакты, цитаты, информационные объекты-связи и т. п. Для каждой пары типов информационных объектов допустимы только определенные классы связей. При попытке создания конкретного экземпляра связи проверяется его допустимость.

Механизмы управления связями позволяют также ликвидировать существующие связи и обновлять значения их атрибутов. Можно, например, изменить текст комментария. 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 информационных объектов — участников таких связей, стимулируя тем самым их реакцию на эти события.

Фактически предлагаемый подход предусматривает создание в системе Соционет наряду с уже много лет функционирующим открытым крупным репозиторием метаданных научных статей, монографий, персональных профилей, профилей организаций и других информационных объектов также и открытого репозитория семантических связей, который является ценным информационным источником структурного анализа представленного в системе корпуса научных знаний.

Механизмы обработки запросов. Эти механизмы выполняют довольно большой набор функций, позволяющих получать разнообразную информацию о структуре связей в библиотеке. Прежде всего, это статистическая информация. В системе Соционет имеются развитые сервисы для наукометрических измерений. Они обсуждаются подробно в [8, 27]. Измерения на основе структуры семантических связей существенно обогащают аналитические возможности системы. Можно, например, запросить количество связей заданного класса или некоторой категории, исходящих из данного информационного объекта библиотеки либо входящих в него. Например, можно узнать, сколько имеется положительных или негативных оценок данной работы.

Другая группа запросов позволяет получить перечень информационных объектов, связанных с заданным объектом как исходным или целевым в связях заданных классов 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 связей или одного из термов критерия может также использоваться идентификатор автора связей. Таким образом может быть получена разнообразная аналитическая информация о структуре различных областей исследований, вкладе в их

развитие конкретных ученых, о процессе эволюции интересующих областей знаний и т. д. Исследования в этой области планируется развивать на основе системы Соционет.

Следует здесь упомянуть проект SciVal компании Elsevier [28]. Функциональный модуль SciVal Spotlight созданного компанией программного продукта позволяет осуществлять анализ научной деятельности исследовательского учреждения или страны в целом, на основе которого может оцениваться эффективность исследований и могут приниматься стратегические решения. Принятый в этом интересном проекте подход основан на анализе структуры связей цитирования публикаций субъектов научной деятельности, поддерживаемых в индексе цитирования Scopus. Однако при этом используются традиционные «немые» связи — связи, не несущие семантики. В этом смысле предлагаемый авторами подход выгодно отличается, так как обеспечивает более дифференцированный анализ, результаты которого учитывают семантику связей.

Механизмы визуализации и анализа графа связей. Важную группу запросов составляют запросы операций над полным графом связей. Здесь можно решать различные задачи, связанные как с анализом топологии графа и вычленением подграфов с заданными свойствами, так и с визуализацией подграфов полного графа. Например, можно вычленить и визуализировать из многослойной структуры связей слой, соответствующий связи некоторого класса, такой как связь, указывающая на использование одного информационного объекта как основополагающего для других объектов. 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, если, конечно, в библиотеке будут достаточно основательно представлены публикации, относящиеся к этой области. Еще одним примером операций над полным графом связей библиотеки является операция вычленения из него подграфа связей, созданных данным пользователем, возможно, с указанием в запросе также категории или конкретного класса связей.

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

7 Заключение

В работе предложен подход, обеспечивающий повышение информационной ценности контента научной электронной библиотеки путем поддержки в ней классифицированных семантических связей между содержащимися в ее коллекциях информационными объектами, а также создания механизмов управления связями и обработки информации, носителями которой они являются.

Реализующая этот подход технология позволяет более эффективно использовать существующий корпус электронных знаний благодаря визуализации семантических связей между научными произведениями, навигации в такой многослойной семантической структуре, созданию основы для получения качественно новых наукометрических измерений, а также для структурного исследования электронного корпуса научных знаний.

Предлагаемая технология обеспечивает также естественный механизм мотивации научных коммуникаций в исследовательском сообществе в процессе создания и обсуждения новых научных результатов. Она хорошо согласуется также с технологией «живых» публикаций, для поддержки которой применимы реализующие ее механизмы.

Литература

1. ACM Digital Library. http://dl.acm.org.
2. Научная электронная библиотека eLibrary.ru. http://elibrary.ru.
3. Академия Google. http://scholar.google.com.
4. *Паринов С. И.* СОЦИОНЕТ.РУ как модель информационного пространства 2-го поколения // Информационное общество, 2001. Вып. 1. С. 43–45.
5. *Паринов С. И., Когаловский М. Р.* Технология поддержки электронных научных публикаций как «живых» документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL-2009): Труды XI Всеросс. науч. конф. (Петрозаводск, 17–21 сентября 2009). — Петрозаводск: КарНЦ РАН, 2009. С. 53–58.
6. *Паринов С. И., Когаловский М. Р.* «Живые» документы в электронных библиотеках // Прикладная информатика, 2009. № 6. С. 123–131.
7. Open Archives Initiative. http://www.openarchives.org.
8. *Когаловский М. Р., Паринов С. И.* Метрики онлайновых информационных пространств // Экономика и математические методы, 2008. Вып. 2. С. 108–120.
9. *Когаловский М. Р., Паринов С. И.* Использование связей цитирования для наукометрических измерений в системе Соционет // Депонировано в Соционет, 2009. http://socionet.ru/publication.xml?h=repec:rus:rssalc:web-32.

10. *Parinov S.* The electronic library: using technology to measure and support Open Science // World Library and Information Congress: 76th IFLA General Conference and Assembly Proceedings. (10–15 August 2010, Gothenburg, Sweden). P. 1–13.
11. *Паринов С. И.* Концепция виртуальной научной среды «Открытая Наука» // Научный сервис в сети Интернет: суперкомпьютерные центры и задачи: Труды междунар. суперкомпьютерной конф. (Новороссийск, 20–25 сентября 2010). — М.: МГУ, 2010. С. 473–481.
12. *Паринов С. И., Когаловский М. Р.* Технология семантического структурирования контента научных электронных библиотек // Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL'2011): Труды XIII Всеросс. науч. конф. (Воронеж, 19–22 октября 2011). — Воронеж: ВГУ, 2011. С. 197–206.
13. *Åström F., Sándor Á.* Models of scholarly communication and citation analysis // ISSI 2009: 12th Conference (International) of the International Society for Scientometrics and Informetrics Proceedings. Vol. 1. http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=1459018&fileOId=1883080.
14. *Sándor Á., Kaplan A., Rondeau G.* Discourse and citation analysis with concept-matching, Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.7518&rep=rep1&type=pdf.
15. *Teufel S., Siddharthan A., Tidhar D.* Automatic classification of citation function // 2006 Conference on Empirical Methods in Natural Language Processing Proceedings. http://portal.acm.org/citation.cfm?id=1610091.
16. *De Waard A., Kircz J.* Modeling scientific research articles — shifting perspectives and persistent issues // ELPUB 2008 Conference on Electronic Publishing Proceedings. — Toronto, Canada, 2008. http://elpub.scix.net/data/works/att/234_elpub2008.content.pdf.
17. SKOS — Simple Knowledge Organization System. http://www.w3.org/TR/skos-reference.
18. *Shotton D.* Introducing 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.
19. *Shotton D., Peroni S.* Semantic annotation of publication entities using the SPAR (Semantic Publishing and Referencing) ontologies // Beyond the PDF Workshop. La Jolla. January 19, 2011. http://imageweb.zoo.ox.ac.uk/pub/2010/Publications/Shotton&Peroni_semantic_annotation_of_publication_entities.pdf.
20. Semantic Web Applications in Neuromedicine (SWAN) ontology. W3C Interest Group Note. October 20, 2009. http://www.w3.org/TR/2009/NOTE-hcls-swan-20091020.
21. CERIF 2008 — Final Release (1.2). http://www.eurocris.org/Index.php?page=CERIF2008&t=19.
22. CERIF-2008-1.3 Ontology. http://spi-fm.uca.es/neologism/cerif#.
23. CERIF-2008-1.3 Semantic Vocabulary. http://spi-fm.uca.es/neologism/semcerif#.
24. *Shotton D., Peroni S.* CiTO, the Citation Typing Ontology, v. 2.0. http://purl.org/spar/cito.
25. *Shotton D.* CiTO, the Citation Typing Ontology // J. Biomedical Semantics, 2010. Vol. 1. Suppl. 1. P. 6. http://www.jbiomedsem.com/content/1/S1/S6.
26. *Shotton D., Peroni S.* DoCO, the Document Components Ontology. http://speroni.web.cs.unibo.it/cgi-bin/lode/req.py?req=http:/purl.org/spar/doco.
27. *Когаловский М. Р., Паринов С. И.* Использование связей цитирования для наукометрических измерений в системе Соционет. Депонировано в Соционет, 2009. http://socionet.ru/publication.xml?h=repec:rus:rssalc:web-32.
28. SciVal. http://www.elsevier.com/wps/find/electronicproductdescription.cws_home/720941/description#description.