Новый источник данных для наукометрических исследований

© М.Р. Когаловский

Институт проблем рынка РАН

kogalov@cemi.rssi.ru

© С.И. Паринов

Центральный экономико-математический институт РАН

Москва

sparinov@gmail.com

Аннотация

Обсуждается подход к созданию новых источников данных для наукометрических исследований, основанный на технологии семантического структурирования контента научных электронных библиотек, который был предложен авторами в ранее опубликованных работах. Рассматриваются основные принципы предлагаемого подхода, а также результаты его реализации в среде системы Соционет. В качестве нового наукометрического источника данных рассматриваются коллекции семантических связей между различными информационными объектами из контента системы. Семантические связи классифицируются с помощью заданной таксономии, представляемой в виде набора контролируемых словарей. Такие источники данных позволяют проводить более многоаспектные по сравнению с традиционными наукометрические исследования представленного в системе корпуса научных знаний. По мнению авторов, создание глобальных автономных репозиториев семантических связей, интегрирующих коллекции связей из различных научных электронных библиотек в анализируемых областях знаний, является перспективным направлением развития наукометрии. Работа поддержана РФФИ, проект 12-07-00518-а и РГНФ, проект 11-02-12026-в.

1 Введение

В настоящее время важную роль в оценке деятельности исследовательских организаций и ученых стали играть наукометрические измерения, результаты которых существенным образом влияют на формирование их научного авторитета, а также и на распределение финансовых ресурсов, предназначенных для поддержки исследовательских программ и отдельных проектов.

Сложившаяся практика оценки научной результативности ученых и научных периодических изданий базируется, в частности, на использовании *индексов цитирования*. Нужно отметить, что термин *индекс цитирования* имеет два значения. Это, одной стороны, библиографическая информационная система, предназначенная для наукометрических измерений, в которой регистрируются связи цитирования, определяемые пристатейными списками источников, между научными публикациями из охватываемого ими корпуса периодических изданий. С другой стороны – это имеющий несколько разновидностей наукометрический показатель, представляющий собой статистическую оценку количества цитирований научных публикации какого-либо автора (традиционное простое количество цитирований, индекс Хирша и др.).

Более значимыми считаются статьи, опубликованные в журналах с более высоким импакт-фактором, а также индексируемые признанными международными системами - индексами цитирования SCOPUS, Web of Science, Web of Knowledge, Springer и др. Активно ведется работа по формированию отечественной системы РИНЦ, которая пока, к сожалению, не располагает достаточно представительным корпусом публикаций для наукометрических измерений, но охватываемое ею научное публикационное пространство интенсивно развивается.

Несомненно, количественные характеристики внимания представителей научного сообщества к той или иной публикации, а также интегральные характеристики внимания к публикациям конкретного автора или периодического издания, представляемые индексами цитирования, являются важными показателями качества научной деятельности. Однако одной из слабых сторон традиционной практики наукометрических измерений является их базирование на связях цитирования с неопределенной явным образом семантикой. Мы называем такие связи *«немыми»* [6], т.к. они сами по себе не несут какой-либо информации, характеризующей, например, мнение автора цитирующей работы о цитируемом источнике или цель цитирования. В связи с

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

этим возможны такие парадоксальные ситуации, когда высоким количеством цитирований обладает статья, содержащая грубые ошибки и/или принципиальные заблуждения, касающиеся обсуждаемой проблемы, и в связи с этим вызывающая активный отклик научного сообщества.

Избежать указанных ситуаций и вместе с тем существенно обогатить информационную основу наукометрических исследований позволяет использование в системах - индексах цитирования семантических связей между научными публикациями. Мы называем *семантическими* связи с явным образом декларированной семантикой. Такие связи могут создаваться, поддерживаться и использоваться для наукометрических исследований и в научных электронных библиотеках, а также в других научных информационных системах, обладающих необходимыми для этого механизмами.

Актуальность такого подхода отмечалась, в частности, еще в работе [5]. Такие более информативные источники для наукометрических исследований позволяют генерировать не только традиционные наукометрические показатели, но и, что более важно, получать новые многоаспектные более утонченные количественные и качественные характеристики отдельных публикаций или групп публикаций, авторов публикаций, а также научных организаций в целом. Исследуя структуру семантических связей, можно выявлять формирующиеся направления в науке, исследовать историю их развития, получать другие полезные результаты.

В последние годы для явного описания семантики связей разработан ряд специальных онтологий, учитывающих не только различные классы связей цитирования, но и связей другой природы, таких как «автор-публикация», «организация-автор», «публикация-фрагмент публикации» (аннотация, оглавление, предисловие, библиография и т.п.), связи между ее версиями, вариантами представления и др. Одним из ранних проектов в этой области является комплекс онтологий SPAR (The Semantic Publishing and Referencing Ontologies) [28, 32]. Следует упомянуть также онтологию SWAN (Semantic Web Applications in Neuromedicine) [26], созданную специалистами в области нейромедицины, рекомендацию SKOS (Simple Knowledge Organization System) [33] консорциума W3C. Работы в этой области ведутся в европейской ассоциации euroCRIS в рамках рабочей группы CERIF (The Common European Research Information Format) [12], а также другими научными коллективами.

На основе указанных онтологий реализуется ряд исследовательских проектов, в которых предусматривается явная спецификация семантики связей. Известны, например, проекты Nanopub.org [16] и SiteULike.org [29]. Авторами данной работы с использованием разработанных онтологий конструктивизирован и развит подход, представленный в уже упоминавшейся статье [5]. При этом учитываются не только связи цитирования и обеспечиваются новые функциональные возможности для наукометрических исследований по сравнению с известными проектами, прежде всего, за счет иного способа представления семантических связей [6-8, 24, 25]. Описания семантических связей представляются в данном проекте как самостоятельные информационные объекты, а не встраиваются в метаданные связанных публикаций. Особенности принятых авторами решений, некоторые новые результаты в развитии и реализации ранее опубликованного подхода, а также функции необходимых для этого механизмов системы Соционет [10, 23] обсуждаются в следующих разделах статьи.

В разделе 2 обсуждаются возможные способы представления семантических связей (их описаний) в контенте научной электронной библиотеки. Аргументируются достоинства их представления как самостоятельных информационных объектов. Раздел 3 посвящен рассмотрению методов создания и описания семантических связей. В разделе 4 обсуждаются вопросы использования онтологий связей и основанных на них контролируемых словарей для описания семантики связей, кратко характеризуются опубликованные проекты онтологий связей, результаты которых используются в данной работе. В разделе 5 рассматривается организация совокупности определенных в электронной библиотеке семантических связей как нового источника данных для наукометрических исследований. Наконец, в разделе 6 описаны результаты реализации предлагаемого подхода в среде системы Соционет. В заключении обсуждаются элементы новизны обсуждаемого подхода и кратко представлены предполагаемые направления развития данного проекта.

2 Представление семантических связей в электронных библиотеках

В научной электронной библиотеке для каждой представленной в ней публикации и объектов других типов, например, профилей авторов или организаций, существует некоторый информационный объект (описатель), содержащий совокупность определяющих его свойства метаданных. Если необходимо учредить и явным образом отразить существование связи между данным и некоторым другим объектом, например, связи цитирования, то такая связь также явным образом должна быть описана. Явная спецификация семантических связей между информационными объектами, составляющими контент научной электронной библиотеки, обогащает представленный в ней корпус научных знаний. Описатель связи (ее метаданные) является ее представителем в информационной системе, как и описатели связываемых объектов.

В научных электронных библиотеках представляют интерес прежде всего *бинарные ориентированные семантические связи*. Далее будут рассматриваться именно такие связи. Информационный объект, из которого исходит связь, будем называть *исходным объектом связи*, а на который связь направлена – ее *целевым объектом*.

В известных авторам проектах [16, 29] для описания факта существования такой связи вводятся дополнительные метаданные в описателе исходного информационного объекта описываемой связи. Для семантической связи значения таких атрибутов характеризуют, в частности, не только идентификатор ее целевого объекта, но и ее семантику, а также другие свойства этой связи. При таком *«встроенном»* способе представления связей между существующими в библиотеке информационными объектами создавать связи в системе может только лицо, обладающее полномочиями обновления описателей этих объектов. Обычно такими полномочиями наделяются администраторы информационных ресурсов, но не пользователи электронной библиотеки.

Другой, *«автономный»*, способ представления связей между информационными объектами, который предложен и развивается в работах авторов данной статьи [6-8], предусматривает поддержку описателей связей в системе как *самостоятельных информационных объектов*. Такой автономный описатель связи содержит метаданные, описывающие уникальные идентификаторы связываемых информационных объектов, семантику связи, идентификатор учредившего связь лица, дату ее создания и некоторые другие необходимые характеристики. При использовании такого способа представления связей учреждать их в системе могут не только администраторы информационных ресурсов, но и обычные зарегистрированные пользователи системы. Таким образом, открывается возможность для нового вида научной деятельности, позволяющая ученым в онлайновом режиме высказывать свои мнения о представленных в электронной библиотеке публикациях, оценивать описываемые в них результаты, характеризовать семантику научных отношений между различными публикациями. Можно, например, создавая связь, указать, что исходная ее публикация использует методы или данные, описанные в целевой публикации этой связи, или что в исходной публикации обнаружен плагиат результатов, представленных в целевой публикации связи и т.п.

Представление семантических связей в научной электронной библиотеке как самостоятельных информационных объектов обеспечивает ряд преимуществ по сравнению со «встроенным» способом их представления. Действительно, на основе контента электронной библиотеки может быть создан автономный по отношению к ее контенту *репозиторий семантических связей*, который может служить новым источником данных для наукометрии. Такие репозитории, созданные в различных библиотеках, могут интегрироваться, и благодаря этому формировать глобальные репозитории семантических связей для той или иной области знаний. На этом пути возможно создание достаточно представительных источников данных для наукометрии в различных областях науки. Если электронные библиотеки построены на основе технологии открытых архивов OAI (Open Archives Initiative) [21], интеграция созданных в их среде репозиториев семантических связей легко обеспечивается точно так же, как и интеграция контентов самих электронных библиотек.

Кроме того, как уже отмечалось, автономное представление связей позволяет поддерживать виртуальную социальную среду *для совместной деятельности* пользователей-ученых в качестве экспертов, добровольно высказывающих мнения и оценки, касающиеся представленных в электронной библиотеке публикаций и других научных информационных ресурсов. Эта деятельность дополняет традиционную практику анонимного рецензирования печатных изданий, а открытость высказанных оценок для научного сообщества позволяет ответным образом реагировать на них, способствует более высокой ответственности и объективности их авторов. Возможности для такой деятельности обеспечиваются, например, сервисом F1000Research [17] проекта FACULTYof1000. Однако подход, обсуждаемый в этой статье, предусматривает поддержку также структурированных данных в форме семантических связей, которые позволяют учитывать высказанные экспертами оценки в наукометрических исследованиях.

В обсуждаемом здесь подходе используется именно «автономный» способ представления семантических связей.

3 Методы описания и создания семантических связей

Будем далее предполагать, таким образом, что семантические связи между информационными объектами контента электронной библиотеки представляются как «автономные» информационные объекты. При этом возможны три метода описания и создания таких объектов.

Первый из них основан на компетенции ученых-экспертов и предусматривает «ручное» описание и создание ими связей при поддержке имеющегося в библиотеке механизма со специальным пользовательским интерфейсом. Создавая связь, эксперт указывает в ее описателе идентификаторы исходного и целевого объектов связи. Используя контролируемые словари (таксономии), основанные на поддерживаемой онтологии связей, он специфицирует класс, к которому относится данная связь. В описателе связи также указываются идентифицирующие данные эксперта, дата создания связи, при необходимости и комментарий [24, 25].

Второй метод применим в случае, когда исходный информационный объект связи является текстовой публикацией. При этом требуется предварительная его обработка. Она заключается в том, что эксперт просматривает не пристатейный список литературы, а текст публикации, и выявляет встречающиеся в нем библиографические ссылки на использованные источники. Анализируя контекст каждой ссылки и используя контролируемые словари семантических связей, он осуществляет ее *онтологическое аннотирование* [3]. Эту работу, конечно же, может выполнить и сам автор данного текста.

Такой размеченный текст далее обрабатывается специальным механизмом системы, и для каждой аннотированной ссылки генерируется описатель соответствующей связи, помещаемый в некоторую коллекцию связей, которая представлена в библиотеке. При использовании данного метода, как и рассмотренного выше, созданные связи обладают авторством, которое приписывается эксперту или автору публикации, выполнившему указанную обработку ее текста. Проблема онтологического аннотирования библиографических ссылок в научных публикациях подробно рассматривается в работе [3].

Наконец, третий метод может рассматриваться как автоматизированный вариант второго. В последние годы активно развиваются исследования, посвященные анализу эмоциональной окраски, тональности текста [2, 14, 15, 20, 22]. Это направление исследований называется в зарубежной литературе *Sentiment analysis* (или *Opinion mining*). Задача такого анализа заключается в определении мнения автора анализируемого текста относительно предмета обсуждения. Если применить методы *Sentiment analysis* к окрестности внутритекстовой ссылки, т.е. к ее контексту, как это делается в работе [34], то можно выявить мнение автора о цитируемой работе. Используя такой метод, можно тем самым автоматизировать ту работу по онтологическому аннотированию ссылок, которую во втором подходе выполняет эксперт. Далее, аналогично предыдущему, на основе таких аннотаций можно автоматически генерировать описатели связей. Их автором является, естественно, автор текста, содержащего аннотированные ссылки. Конечно же, возможности такой «диагностики» семантики ссылок ограничиваются лишь классами оценочных ссылок.

В настоящее время в системе Соционет реализован только первый - экспертный *«ручной»* метод создания семантических связей как самостоятельных информационных объектов.

4 Описание семантики связей

Основой для реализации пользовательских инструментов, позволяющих создавать семантические связи между информационными объектами и оперировать ими, стали разработанные в последние годы онтологии семантических связей. В частности, это онтологии связей между объектами научной сферы деятельности (публикациями различных типов, темами исследований, учеными, исследовательскими организациями и др.). В этих онтологиях определяются иерархии классов связей, соответствующие различного рода отношениям, в которых могут состоять информационные объекты – участники связей. Рассмотрим их кратко.

Среди основательно проработанных проектов следует назвать, прежде всего, модульный комплекс онтологий SPAR (*the Semantic Publishing and Referencing Ontologies*) [28, 32], созданный сотрудниками Оксфордского и Болонского университетов. SPAR включает восемь независимых онтологий, позволяющих описывать семантику библиографических объектов, а также их отношений. Эти онтологии описаны средствами языков OWL2 DL и RDF консорциума W3C. Первые четыре из них (FaBiO, CiTO, BiRO and C4O) позволяют описывать библиографические объекты, библиографические записи и источники в списках литературы публикаций, а также связи цитирования, контексты цитирования и их связи с релевантными разделами цитируемых публикаций. Четыре остальных онтологии (DoCO, PRO, PSO and PWO) могут использоваться для описания семантики компонентов документов, ролей и состояний публикаций, потоков работ в издательских процессах.

Другой заслуживающий внимания проект в рассматриваемой области – модульный комплекс онтологий SWAN (*Semantic Web Applications in Neuromedicine*) [26], разработанный в Главном госпитале Массачусетса и Медицинской школе Гарварда. Авторы характеризуют его назначение как обеспечение в Семантическом Вебе комфортной среды - *социально-технической экосистемы*, которая позволяет создавать и сохранять семантический контекст научных коммуникаций, обеспечивает доступ к нему, его интеграцию, а также обмен неструктурированной и слабоструктурированной цифровой научной информацией. Онтологии комплекса описаны в его спецификации средствами языка описания онтологий уровня OWL DL.

Примерно в то же время консорциумом W3C была принята рекомендация SKOS (*Simple Knowledge Organization System*) [33], предназначенная для поддержки систем организации знаний - тезаурусов, схем классификации, таксономий и рубрикаторов (*Subject Heading Systems*) - в среде Семантического Веба. SKOS определяет концептуальную схему, называемую *общей моделью данных,* служащую для совместного использования и связывания систем организации знаний средствами Веба. Благодаря унифицированной концептуальной схеме SKOS упрощается интеграция существующих систем организации знаний в Семантический Веб.

Следует отметить, что принцип модульности организации таких сложных комплексных онтологий, как SPAR и SWAN, облегчает их повторное использование. В некоторых приложениях нет необходимости использовать полную онтологию. Тогда могут использоваться отдельные ее модули. Облегчается также интеграция онтологий. Так, в комплексе SPAR используются элементы SWAN, а в SWAN используется SKOS.

Существенный вклад в рассматриваемую область вносит также европейская научная общественная организация euroCRIS, инициировавшая и развивающая проект CERIF. Одним из главных результатов этого проекта является создание унифицированной концептуальной схемы, называемой в материалах проекта *полной моделью данных* (*Full Data Model*) [11]. Эта модель рассматривается как единая основа создания информационных систем (*Current Research Information Systems*, CRIS) для поддержки научно-организационной деятельности в разных странах и

различных научных организациях. Благодаря стандартизации концептуальной схемы обеспечивается интероперабельность таких систем. В последнее время в рамках проекта CERIF была предложена спецификация стандартизованной семантики полной модели данных [12], онтология, определяющая систему терминов для обозначения сущностей этой модели и отношений между ними.

Рассмотренные онтологии могут быть использованы для создания онтологии семантических связей информационных объектов конкретной научной электронной библиотеки, адекватной характеру представленных в ней информационных ресурсов и ее функциональных механизмов, в частности, ее наукометрического аппарата. Структура семантических связей, определенная на контенте электронной библиотеки, является *многослойной* [6-8]. Каждый ее слой соответствует некоторому классу связей, определенному в используемой онтологии. Семантическая структура контента библиотеки может использоваться как источник данных для наукометрических исследований, а также служить основой семантической навигации в ее контенте. Для практического использования на основе используемой онтологии для конкретной электронной библиотеки может быть сформирована *таксономия семантических связей*, представленная в виде набора *контролируемых словарей* имен классов семантических связей.

В электронной библиотеке может поддерживаться несколько таксономий связей, основанных на разных онтологиях, точно так же как для рубрикации информационных объектов может использоваться несколько рубрикаторов научно-технической информации. Так, в системе Соционет поддерживаются рубрикатор ГРНТИ [1] и классификатор JEL (Journal of Economic Literature Classification System) [18]. На основе их рубрик по запросам генерируется наукометрическая статистика [9].

При спецификации пользовательских запросов в электронной библиотеке по умолчанию может использоваться основная встроенная в систему таксономия с ее набором управляемых словарей. В противном случае используемая таксономия должна специфицироваться в запросе.

5 Описания семантических связей - источник данных для наукометрии

Организованные совокупности семантических связей между информационными объектами контента научной электронной библиотеки, созданные описанными выше методами и средствами, служат более информативным источником данных для наукометрических исследований по сравнению с традиционной наукометрией, которая базируется на множестве «немых» ссылок цитирования.

Описания семантических связей могут использоваться наукометрическими сервисами библиотеки, продуцирующими разнообразные показатели как на основе предоставляемой описанием связей информации о том, какие информационные объекты связаны, так и информации о семантике существующих связей.

Семантические связи, организованные в виде коллекций информационных объектов специального типа, могут составлять подмножество контента электронной библиотеки. Они могут быть также организованы в виде самостоятельного информационного ресурса, сосуществующего с ее контентом.

Представляется заманчивой интеграция таких информационных ресурсов, созданных и поддерживаемых в различных научных электронных библиотеках, и формирование открытых (свободно доступных) глобальных репозиториев семантических связей для наукометрических исследований [24, 25] в различных областях знаний. Такие источники данных более полно представляют корпус научных знаний рассматриваемой области науки. На их основе можно формировать адекватную наукометрическую статистику и другие характеристики состояния данной области знаний.

Интеграция ресурсов семантических связей различных научных электронных библиотек может быть осуществлена известными методами виртуальной или материализованной интеграции данных из множества информационных источников [4, 19, 36]. Интеграция относительно легко реализуется при условии базирования библиотек-источников на технологии Инициативы открытых архивов OAI и протоколе OAI-PMH [21, 35].

В обоих случаях может возникнуть проблема неоднородности онтологий семантических связей, используемых в библиотеках–источниках. Для ее решения возможны два подхода. При первом подходе таксономия интегрированного репозитория строится как объединение таксономий, применяемых в электронных библиотеках–источниках. В такой ситуации интегрированный источник представляет собой фактически федерацию семантически неоднородных наборов связей. Статистические запросы при этом будут учитывать только связи, семантика которых определена указанной в запросе таксономией. При втором подходе осуществляется семантическая интеграция всех охватываемых наборов связей, представленных в библиотеках-источниках. Для интегрированного репозитория создается некоторая общая таксономия таким образом, чтобы было возможно определить отображения таксономий библиотек-источников в эту общую таксономию. Анализ и обработка контента интегрированного глобального репозитория связей осуществляется при этом на основе общей таксономии связей.

Помимо возможности обработки глобального репозитория семантических связей средствами созданных для него наукометрических сервисов, целесообразно обеспечить интерфейс прикладного программирования (API) для доступа к его контенту. Таким образом, обеспечивается возможность разработки разнообразных приложений, оперирующих контентом глобального репозитория семантических связей.

6 Реализация предлагаемого подхода в среде системы Соционет

В рассматриваемом проекте в качестве среды реализации предлагаемого подхода используется система Соционет. В настоящее время реализованы механизмы системы, обеспечивающие описание, создание, хранение, модификацию, удаление и просмотр семантических связей, формирование коллекций связей. Кроме того, реализованы основные средства создания и поддержки контролируемых словарей семантических связей. Реализован также ряд статистических сервисов, обрабатывающих семантику связей и генерирующих новые наукометрические показатели. Рассмотрим несколько подробнее основные особенности уже реализованных средств и сервисов.

Создание и организация связей в системе. Совокупности связей поддерживаются в системе в форме коллекций информационных объектов специального типа *linkage*. Наряду с таким «автономным» вариантом представления семантических связей поддерживается и встроенный вариант, который, однако, реализован лишь для обеспечения полноты возможностей при дальнейшем развитии системы. Пока реализованы только средства для «ручного» создания связей экспертом-пользователем системы (см. разд. 3). К коллекциям связей применимы все имеющиеся в системе функциональные возможности управления коллекциями любого типа данных.

Как уже указывалось, в Соционет поддерживаются бинарные ориентированные семантические связи. Информационными объектами–участниками связей могут быть объекты различных видов (электронные монографии, статьи в периодике, диссертации и авторефераты диссертаций, классификаторы, авторы публикаций, исследовательские или образовательные учреждения и др.). Среди объектов-участников связей в Соционет могут быть и библиографические описания публикаций, возможно, дополненные аннотациями. Эти объекты в системе относятся к типу *artifact.* Коллекциями таких объектов могут представляться тематические библиографии.

Информационные объекты-участники связей могут быть внутренними для системы (содержатся в ее контенте) или внешними. Внешние объекты не содержатся в контенте системы. Они доступны в Вебе по их адресу (URL). Допустимость внешних информационных объектов, а также публикаций, представленных их библиографическими описаниями (объекты типа *artifact*), в качестве участников связей позволяет охватить семантическими связями все доступное ученым цифровое научное информационное пространство.

Метаданные коллекций связей являются составной частью репозитория метаданных системы Соционет, основанной на технологии *открытых архивов OAI.* Поэтому этот фрагмент репозитория может быть представлен в виде самостоятельного репозитория метаданных, обеспечивая организацию совокупности коллекций связей в системе как самостоятельного открытого архива. Этот архив при необходимости может интегрироваться с аналогичными архивами связей других электронных библиотек на основе протокола OAI-PMH. Таким образом могут формироваться глобальные репозитории семантических связей, более представительные полигоны для наукометрических исследований. Если интегрируемые архивы семантических связей не основаны на единой стандартной для них таксономии связей, то, как уже отмечалось, возникнет необходимость решения проблемы преодоления неоднородности таксономий.

Связи в системе Соционет могут создавать зарегистрированные в ней пользователи. Такие пользователи имеют в системе свои профили и, тем самым, они идентифицируемы как авторы создаваемых ими семантических связей. Имеющиеся в Соционет средства мониторинга состояния связей в необходимых случаях (например, при появлении в системе семантически противоречивых созданных разными авторами связей между информационными объектами некоторой пары) могут оповещать их сообщениями по электронной почте, адрес которой должен указываться в профиле пользователя, формируемом в процессе его регистрации в системе.

При создании связи в Соционет формируется ее описатель (метаданные связи), включающий следующие метаданные: уникальный идентификатор связи, тип и идентификатор ее исходного объекта, тип и идентификатор целевого объекта (или URI для внешнего целевого объекта), описание ее семантики - класс таксономии, к которому она относится, дату ее создания, уникальный идентификатор, идентификатор автора связи и при необходимости его комментарий [24, 25].

В зависимости от типов информационных объектов-участников создаваемой связи она может принадлежать только к какому-либо классу таксономии, соответствующему этой паре типов ее объектов-участников. Однако для заданной пары объектов может быть создано несколько связей. Разные авторы связей могут создать несколько связей одного класса. Один и тот же автор не может создать несколько связей одного класса для заданной пары объектов, но имеет возможность создать несколько связей разных классов.

В некоторых ситуациях при создании связи включаются системные механизмы, которые автоматически генерируют другие связи. Это имеет место, например, в случае, когда целевой объект создаваемой связи, являющийся текстом статьи, одновременно состоит в связях с другими объектами, которые характеризуются связями как копии этой статьи. Тогда, если создаваемая экспертом связь является оценочной для такого объекта, то автоматически будут генерироваться аналогичные оценочные связи того же исходного объекта с объектами, представляющими другие копии.

Автоматическая генерация явно описанных связей может осуществляться также как побочный эффект при создании связи, порождающей транзитив-

ные отношения между информационными объектами в системе (см. пример ниже).

Спецификация семантики связей. В инструментарии создания и использования семантических связей в системе Соционет для спецификации семантики создаваемых связей использована *гибридная онтология*, являющаяся расширением объединения некоторых фрагментов онтологий CiTO [27, 30], DoCo [31], SWAN [26], SKOS [33] и CERIF [11, 12]. На ее основе создана двухуровневая таксономия классов семантических связей, представленная в виде набора контролируемых словарей, каждый из которых соответствует одному из классов верхнего уровня иерархии классов таксономии, а значения в словарях соответствуют подклассам этих классов.

Действующая версия системы Соционет поддерживает набор контролируемых словарей семантических связей, включающий словари: связей научного вывода; связей использования; связей между компонентами научного произведения; а также между его версиями или копиями; связей научных оценок (оценочных связей); иерархических и ассоциативных связей между публикациями; связей объектов вида «персона-персона», «персона-организация», «персона-публикация». Более подробно используемая в Соционет таксономия семантических связей и представляющие ее контролируемые словари описаны в работе [8].

Важно здесь отметить, что созданная для Соционет таксономия семантических связей позволяет классифицировать связи не только между объектами-научными текстами. Это обстоятельство имеет существенное значение, поскольку, участниками связей в системе могут быть как научные публикации и их компоненты, так и наборы научных данных, организации и их сотрудники – авторы информационных объектов и пользователи системы, а также информационные объекты других типов, представленные в контенте системы.

В системе Соционет реализован механизм расширения таксономии семантических связей пользователями путем дополнения классов к существующим словарям и создания новых контролируемых словарей в режиме модерирования администратором системы.

Сервисы системы для операций с семантическими связями. Эти сервисы выполняют довольно большой набор функций, позволяющих получать разнообразную информацию о структуре связей в библиотеке. Прежде всего, это статистическая информация. Ряд таких сервисов уже реализован в системе. Реализованы возможности семантической навигации по слою структуры связей, соответствующему заданному классу связей. Предстоит также реализация ряда других сервисов, осуществляющих исследование топологии графа связей и позволяющих на этой основе получать ряд полезных результатов, характеризующих состояние и генезис представленного в системе корпуса научных знаний.

Поскольку, как уже отмечалось, в системе могут одновременно поддерживаться несколько таксономий семантических связей, при обращении к таким сервисам необходимо указывать на основе какой из них или какого конкретного контролируемого словаря должна проводиться обработка пользовательского запроса. Например, если пользователя интересует статистика оценочных связей, то в запросе должен быть указан словарь или класс какого-либо словаря научных оценок из той или иной поддерживаемой в системе альтернативной таксономии.

Характер генерируемой по запросам статистической информации может быть весьма разнообразным. Например, можно запросить для конкретного информационного объекта количество входящих или исходящих из него связей (т.е. связей, в которых данный объект участвует как целевой или, соответственно, как исходный), относящихся к некоторому классу верхнего уровня, т.е. к некоторому контролируемому словарю таксономии в целом, или к его подклассу, т.е. к одному из значений в заданном словаре. Содержательная интерпретация полученной статистики, естественно, зависит от заданного словаря. Здесь возможно большое количество вариантов: сколько имеется позитивных или негативных мнений о данной статье, в каком количестве работ используются предложенные в ней методы или содержащиеся в ней научные данные, сколько обнаружено случаев плагиата данной статьи и т.д.

При формировании статистики для некоторых классов таксономии могут учитываться транзитивные связи этих классов. Например, существует связь между статьей A - исходным объектом связи и статьей B – целевым объектом этой связи, указывающая, что в B предлагается более широкое обсуждение проблемы, которой посвящена A. Кроме того, существует связь между статьей B как исходным объектом и статьей C – целевым объектом этой связи, которая также указывает, что в C предлагается более широкий взгляд на предмет обсуждения по сравнению с B. Тогда фактически существует явно не описанная транзитивная связь A с C с той же семантикой. Это обстоятельство должно учитываться при формировании статистики статей, в которых более широко обсуждается проблема, рассматриваемая в статье A.

Статистические запросы могут быть обобщены на все множество информационных объектов заданного типа или на все связи класса верхнего уровня таксономии (на указанный словарь в целом). Например, может запрашиваться статистика мнений обо всей совокупности монографий из какой-либо коллекции или общее количество работ, в которых высказано позитивное мнение о данной работе.

Поскольку в описателях связей указывается дата их создания, возможны запросы статистики, относящейся к заданным промежуткам времени или, что более сложно, динамической статистики (временных рядов некоторых статистических показателей).

Другая группа запросов позволяет получить перечень конкретных информационных объектов биб-

лиотеки, связанных с заданным объектом как исходным или целевым в связях заданных классов. Содержательные интерпретации получаемого при этом результата также различаются в зависимости от используемого словаря или конкретного его класса связей. 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 не только непосредственные, но и транзитивные связи. В критерии отбора интересующих пользователя связей может также использоваться идентификатор автора связей.

Используя цепочки связей вида «автор-публикация» + «публикация-публикация» можно получить количество публикаций, в которых выражено негативное или позитивное отношение к публикациям данного автора, либо список таких публикаций. С использованием более длинных цепочек вида «организация-сотрудник (автор)» + «автор-публикация» + «публикация-публикация» можно получить аналогичные сведения для интересующей организации, агрегированные по всем ее сотрудникам – авторам представленных в системе работ.

Наконец, важную группу запросов составляют запросы операций 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. Так, можно вычленить и визуализировать из многослойной структуры семантических связей слой, соответствующий связям некоторого класса, например, указывающего на использование одной публикации из контента системы как основополагающей для других публикаций. Можно также запросить подграф, образованный связями, относящимися к классу 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 контенте системы 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 конкретного класса связей.

7 Заключение и направления дальнейшей работы

Мы рассмотрели подход, позволяющий использовать коллекции или репозитории семантических связей информационных объектов контента научных электронных библиотек, расширенного внешними информационными объектами, как источник данных для новых нетрадиционных многоаспектных наукометрических исследований. Прототип необходимых для этого программных средств реализован в системе Соционет, и продолжается работа по развитию его функциональных возможностей в рассмотренных направлениях.

Некоторые идеи рассмотренного подхода опубликованы другими авторами практически одновременно с нашими ранними работами в этой области. Однако рассмотренный здесь подход, по мнению авторов, обладает существенной степенью новизны по отношению к известным работам.

Одно из главных отличий заключается в представлении семантических связей как самостоятельных информационных объектов специального типа данных *linkage*, которые хранятся автономно от объектов-участников связей, а не встраиваются в их описатели. Отчуждение описателей связей от описателей связываемых объектов позволяет строить коллекции связей, формировать на их основе репозитории связей, которые могут интегрироваться с другими репозиториями. Благодаря этому становится возможным формирование представительных глобальных репозиториев семантических связей для различных областей знаний. Может быть обеспечено также *повторное использование* коллекций семантических связей как информационного ресурса для наукометрических исследований, более содержательных по сравнению с традиционной наукометрией, основанной на «немых» связях цитирования.

Важно также, что благодаря этому связи может создавать лицо, не являющееся создателем информационных объектов и/или метаданных (описателей) информационных объектов-участников связей. Это могут делать любые пользователи, зарегистрированные в системе, тем самым развивая семантическую структуру ее контента. Если описание связи встроено в метаданные информационного объекта (обычно исходного объекта связи), то без вторжения в них создавать связи невозможно. А эта операция доступна только владельцу метаданных рассматриваемого объекта.

Другое достоинство предлагаемого подхода заключается в том, что в нашем случае используется более многоаспектная онтология и основанная на ней таксономия семантических связей. Связываемыми информационными объектами могут быть не только научные публикации, но и научные информационные объекты иных типов, а также объекты, представляющие авторов публикаций, пользователей системы, организации, в рамках которых выполнялись публикуемые работы и т.п. Поэтому естественно, что онтология связей Соционет определяет более богатое множество отношений, воплощаемых семантическими связями, поддерживаемыми в системе. Благодаря этому, анализируя структуру таких связей, можно получать различные новые количественные и качественные результаты, которые не позволяют получать существующие индексы цитирования. Важно отметить, что при этом предусматривается возможность использования одновременно нескольких альтернативных таксономий связей.

Рассматриваемый подход обеспечивает создание и динамическое развитие пользователями системы семантической структуры ее контента, которая представляет собой своего рода его *«семантический ореол»* (Semantic Halo [13]). Благодаря ему пользователи получают информационно насыщенное представление о структуре мнений ученых по поводу существующих в системе информационных объектов. Вместе с тем, обеспечивается *семантическая навигация* по контенту системы, которая может осуществляться по слоям семантической структуры, соответствующим классам используемой таксономии, и создает комфортные условия для доступа пользователей к информационным ресурсам системы.

Отметим, наконец, еще одну важную особенность обсуждаемого в данной работе подхода. Он открывает возможности для новых форм научной деятельности, воплощаемой в виртуальной среде онлайновой системы. Электронная библиотека, в которой реализован обсуждаемый подход, представляет собой фактически социальную сеть, в среде которой совместно действуют представители научного сообщества. Результатами их деятельности являются представленные в явном виде мнения о научных публикациях, а также развивающаяся семантическая структура контента системы, позволяющая использовать новые методы наукометрических исследований. Представление семантических связей в системе как самостоятельных информационных объектов позволяет декларировать мнения о них точно так же, как и относительно других информационных объектов. Такая поддержка мнений о мнениях образует своеобразный дискуссионный форум в среде системы.

Авторы намерены продолжить работу по развитию инструментария, реализующего рассмотренный подход, в нескольких направлениях. В частности, предполагается создать механизм поддержки альтернативных таксономий семантических связей. Планируются также эксперименты по интеграции репозиториев семантических связей. Наконец, будут созданы дополнительные сервисы для наукометрических исследований, учитывающие новые информационные возможности семантически структурированного контента системы.

Литература

[1]. ГРНТИ – рубрикатор научно-технической информации. Редакция 2007 года. – URL: http://www.grnti.ru/ [Дата обращения 15 июля 2013 г.]

[2]. Клековкина М.В., Котельников Е.В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики. Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г. Переславль-Залесский, Институт программных систем РАН, 2012. – С. 118-123.

[3]. Когаловский М.Р. Онтологическое аннотирование библиографических ссылок в научных публикациях и его использование в наукометрии //Информационные ресурсы России (в печати).

[4]. Когаловский М.Р. Методы интеграции данных в информационных системах. Депонент Соционет, 2010.http://socionet.ru/pub.xml?h=RePEc:rus:rssalc:web-39 [Дата обращения 15 июля 2013 г.]

[5]. Когаловский М.Р., Паринов С.И. Использование связей цитирования для наукометрических измерений в системе Соционет. Институт проблем рынка РАН, Центральный экономико-математический институт РАН, 2009. Электронный депонент Соционет. http://socionet.ru/publication.xml?h=repec:rus: rssalc:web-32 [Дата обращения 15 июля 2013 г.]

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

[7]. Когаловский М.Р., Паринов С.И. Классификация и использование семантических связей между информационными объектами в научных электронных библиотеках //Информатика и ее применения. 2012. Т. 6. Вып. 3. С. 32-42.

[8]. Паринов С.И., Когаловский М.Р. Технология семантического структурирования контента научных электронных библиотек. Труды XIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2011. Воронеж, 19-22 октября 2011 г.». – г. Воронеж: Воронежский государственный университет, 2011. – С. 94-103.

[9]. Когаловский М.Р., Паринов С.И. Наукометрические измерения в электронных библиотеках на основе рубрикаторов научной информации //Электронные библиотеки (электронный журнал). 2012. Т. 15, № 6. http://www.elbib.ru/ index.phtml?page=elbib/rus/journal/2012/part6/KP [Дата обращения 15 июля 2013 г.]

[10]. Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайновых сервисов //Российский научный электронный журнал «Электронные библиотеки». – 2003. – Том 6. – Вып. 1. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part1/PLP [Дата обращения 15 июля 2013 г.]

[11]. CERIF 1.3 Full Data Model (FDM): Introduction and Specification. euroCRIS, 2012. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf [Дата обращения 15 июля 2013 г.]

[12]. CERIF 1.3 Semantics: Research Vocabulary. CERIF Task Group, euroCRIS, 2012. http://www. eurocris.org/Uploads/Web%20pages/CERIF-1.3/ Specifications/CERIF1.3_Semantics.pdf [Дата обращения 15 июля 2013 г.]

[13]. Dix A., Levialdi S. & Malizia A. Semantic halo for collaboration tagging systems. In the Social Navigation and Community-Based Adaptation Technologies Workshop-June 20th, 2006.

[14]. Feldman R. Techniques and Applications for Sentiment Analysis. Communications of the ACM, April 2013, vol. 56, no. 4, pp. 82-89.

[15]. Galassini C., Malizia A., and Bellucci A. An approach for developing intelligent systems for sentiment analysis over social networks. Intelligent Systems and Control /742: Computational Bioscience, J.F. Whidborne, P. Willis, G. Montana, Eds. Cambridge, United Kingdom, July 11 – 13, 2011.

[16]. Groth P., Gibson A., Velterop J. The Anatomy of a Nano-publication. Information Services and Use 30(1/2) (2010). http://iospress.metapress.com/ content/ftkh21q50t521wm2/ [Дата обращения 15 июля 2013 г.]

[17]. F1000Research. http://f1000research.com/ [Дата обращения 15 июля 2013 г.]

[18]. Journal of Economic Literature (JEL) Classification System. – URL: http://www.aeaweb.org/jel/jel _class_system.php[Дата обращения 15 июля 2013 г.]

[19]. Kalinichenko L.A., Briukhov D.O., Skvortsov N.A., Zakharov V.N. Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections. Вторая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Протвино, 2000, с. 78-90.

[20]. Karlsson S. About LogEc, 2011. http://logec.repec. org/about.htm [Дата обращения 15 июля 2013 г.]

[21]. Open Archives Initiative. http://www. Openarchi ves.org/ [Дата обращения 15 июля 2013 г.]

[22]. Pang B., Lee L. Opinion Mining and Sentiment Analysis //Foundations and Trends in Information Retrieval. 2008. Volume 2, Issue 1-2. January 2008, pp. 1-135. http://dl.acm.org/citation.cfm?id= 1454712 [Дата обращения 15 июля 2013 г.]

[23]. Parinov S. The electronic library: using technology to measure and support Open Science. In: Proceedings of the World Library and Information Congress: 76th IFLA General Conference and Assembly, Gothenburg, Sweden, August 10-15, 2010. http://www.ifla.org/files/hq/papers/ifla76/155-parinov-en.pdf [Дата обращения 15 июля 2013 г.]

[24]. Parinov S. Open Repository of Semantic Linkages. In: Proceedings of 11th International Conference on Current Research Information Systems e-Infrastructure for Research and Innovations (CRIS 2012), Prague, 2012. http://socionet.ru/publication. xml?h=repec:rus:mqijxk:29 [Дата обращения 15 июля 2013 г.]

[25]. Parinov S. Towards a Semantic Segment of a Research e-Infrastructure: necessary information objects, tools and services. Metadata and Semantics Research, Communications in Computer and Information Science. J. M. Dodero, M. Palomo-Duarte, P. Karampiperis, Eds. Springer, vol. 343, 2012, pp. 133-145. http://socionet.ru/pub.xml?h= RePEc:rus: mqijxk:30 [Дата обращения 15 июля 2013 г.]

[26]. Semantic Web Applications in Neuromedicine (SWAN) Ontology. W3C Interest Group Note, 20 October 2009. http://www.w3.org/TR/2009/NOTE -hcls-swan-20091020/ [Дата обращения 15 июля 2013 г.]

[27]. Shotton D. CiTO, the Citation Typing Ontology. J. of Biomedical Semantics 2010, 1(Suppl 1): S6. http://www.jbiomedsem.com/content/1/S1/S6 [Дата обращения 15 июля 2013 г.]

[28]. 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-refere ncing-spar-ontologies/ [Дата обращения 15 июля 2013 г.]

[29]. Shotton D. Use of CiTO in CiteULike. 2010. http://opencitations.wordpress.com/2010/10/21/use -of-cito-in-citeulike/[Дата обращения 15 июля 2013 г.]

[30]. Shotton D., Peroni S. CiTO, The Citation Typing Ontology, v2.0. – 2011. http://purl.org/spar/cito/ [Дата обращения 15 июля 2013 г.]

[31]. 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 [Дата обращения 15 июля 2013 г.]

[32]. Shotton D., 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/Publicatio ns/Shotton& Peroni_semantic_annotation_of_ pub lication_entities.pdf [Дата обращения 15 июля 2013 г.]

[33]. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 August 2009. http://www.w3.org/TR/skos-reference/ [Дата обращения 15 июля 2013 г.]

[34]. Small H. Interpreting maps of science using citation context sentiments: a preliminary investigation. Scientometrics, Springer Netherlands, Volume 87, Issue 2, 2011, pp. 373-388

[35]. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. Document Version 2008-12-07T20:42:00Z. http://www.openarchives.org/OAI/2.0/openarchive sprotocol.htm [Дата обращения 5 июля 2013 г.]

[36]. Wache H., Vogele T., Visser U., Stuckenschmidt H., Schuster G., Neumann H. and Hubner S. Ontology-Based Integration of Information — A Survey of Existing Approaches. In Proceedings of

IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, 2001, pp. 108-117. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.7857[Дата обращения 15 июля 2013 г.]

New data source for scientometric studies

Mikhail R. Kogalovsky, Sergey I. Parinov

The paper is dedicated to discussion of an approach to creation of new data sources for the scientometric researches, based on technology of semantic structuring the content of scientific digital libraries. The technology was described by authors in earlier published their papers. The basic principles of offered approach and the results of its implementation in the environment of Socionet system are considered. Mentioned new scientometric data source is collection of semantic linkages between various information objects of the system content. The linkages are classified by given taxonomy that is represented by a set of semantic controlled vocabularies. Such data sources allow more multifold analysis in comparison with traditional scientometric researches of the scientific knowledge corpus supported by the system. According to authors opinion the creation of the global autonomous repositories of semantic linkages which integrate the linkage collections from various scientific digital libraries in analyzed knowledge areas is perspective direction of scientometrics evolution. This research is supporting by RFBR, project 12-07-00518-a, and RFH, project 11-02-12026-в.