Технология семантического структурирования контента научных электронных библиотек*

© С.И. Паринов

Центральный экономико-математический институт РАН
sparinov@gmail.com

© М.Р. Когаловский

Институт проблем рынка РАН
kogalov@cemi.rssi.ru

Аннотация

Семантическое структурирование контента научных электронных библиотек и поддержка в явном виде воплощающих его связей между информационными объектами открывает новые возможности для научного творчества и существенно повышает информативность библиотек. В сочетании с категоризацией поддерживаемых семантических связей это порождает многослойную семантическую сетевую структуру, на основе которой становятся возможными качественно новые наукометрические измерения и исследования структурных свойств корпуса научных знаний, представленного в электронных библиотеках. В докладе обсуждается общий подход к решению этой проблемы, предлагается технология его реализации в среде информационного пространства Соционет.

1 Введение

Ситуация, которая сложилась в настоящее время в области развития технологий электронных библиотек (ЭБ), хорошо иллюстрирует, на наш взгляд, действие закона перехода количества в качество. Контент большого количества уже созданных разными организациями научных ЭБ постепенно интегрируется на уровне метаданных. Один из примеров такого объединения метаданных из ЭБ предоставляет система Соционет. Подобная интеграция приводит к появлению научных информационных пространств (ИП), основанных на федеративных принципах. Новым качеством, возникающим в результате этих изменений, является предоставление пользователям единых интерфейсов доступа к интегрированным разнородным научным данным на основе стандартизации метаданных, а также унификация способов доступа к научной информации в ИП, способов ее извлечения из ИП для обработки и использования [14].

С другой стороны, уже много лет ведутся исследования в области анализа семантики связей между научными материалами. Системным обобщением этих результатов стало появление комплекса онтологий SPAR, обеспечивающего достаточно детальную категоризацию отношений, которые могут возникать между научными материалами в электронном виде, и воплощающих их связей. Важным результатом также является появление семантического раздела в модели научных данных CERIF [4]. Обзор основных результатов этих исследований и разработок приведен в разделе 2.

Соединение этих двух достижений: 1) создание средств и сервисов научных ИП, представляющих интегральный контент ЭБ; и 2) разработка классификаторов отношений и семантических словарей, позволяющих выражать существование определенных связей и отношений между объектами научного ИП; порождает важное новое качество. Становится возможным разработка технологий семантического структурирования контента ЭБ. Рассмотрению этих новых возможностей посвящен раздел 3.

В этом разделе рассмотрены основные виды научной деятельности, в процессе которых ученые создают отношения между научными материалами, и которые, как следствие, могут быть зафиксированы созданием семантических связей между объектами ИП. Описаны разработанные авторами статьи категории и словари свойств семантических связей. Рассмотрены технические, организационные, а также этические особенности создания семантических связей между объектами ИП.

Отмечается, что представление связей различных семантических категорий образует над множеством объектов научного ИП многослойную структуру. В частности, могут поддерживаться слои, отображающие структуру продуцирования научных ре-

Труды 13ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

зультатов и другие содержательные отношения между научными публикациями, связи оценки публикаций научными сотрудниками, связи между компонентами научных произведений, связи научно-организационного характера (научное учреждение – сотрудники-авторы публикаций, авторы – публикации) и др.

Раздел 4 посвящен техническим деталям реализации технологий семантического структурирования в среде системы Соционет, которая является уникальным полигоном для отработки подобных нововведений.

В заключении кратко перечислены основные преимущества, которые получает научное сообщество от реализации предлагаемых решений.

2 Семантические связи и электронные научные публикации

Анализ электронных научных публикаций в составе крупных электронных библиотек в части выявления и классификации отношений, которые могут существовать между разделами научной статьи или между исходной статьей и цитируемыми материалами ведется уже достаточно давно.

Например, на основе программного обеспечения компании Ксерокс, ведутся работы по распознаванию и классификации используемых в научных статьях языковых конструкций (для английского языка и отдельных научных дисциплин). Эти исследования позволили эмпирическим путем выявить некоторые устойчивые виды семантических отношений, создаваемые авторами как между разделами внутри научной статьи, так и с цитируемыми в статье материалами [1, 2]. Эмпирическая классификация поводов цитирования (семантики связей цитирования) в научных статьях проведена также в [9]. В этой работе выделен ряд их типичных значений: "слабость цитируемого подхода", "автор использует цитируемую работу как основу или начальную точку" и др. Другой подход к развитию семантических связей реализуется в исследованиях модульности научных документов [3].

Известна также рекомендация консорциума W3, получившая название SKOS (Simple Knowledge Organization System) [8], в которой предлагается модель связывания научных данных, адаптированная для компьютерной обработки. В частности, SKOS включает контролируемые структурные словари семантических значений для связывания научных данных.

В различных научных дисциплинах (в первую очередь биология и медицина) были предприняты попытки разработать более подробную категоризацию отношений между научными текстами. Наиболее известными результатами этих попыток являются онтология SWAN (Semantic Web Applications in Neuromedicine) [10], а также CiTO (Citation Typing Ontology) [6], DoCo (Document Components Ontology) [7] и др. В дальнейшем все эти частные разработки были систематизированы, дополнены и объединены в единый комплекс под названием SPAR (Semantic Publishing and Referencing Ontologies) [5], включающий взаимосвязанную совокупность онтологий различного назначения.

Независимо от этого, в рамках разработки концептуальной модели научных данных CERIF (Common European Research Information Format) [4], ведутся работы по развитию стандартизованной формальной семантики для отображения отношений между объектами научных информационных систем (CRIS).

Уже имеющиеся результаты по выявлению и классификации отношений, которые могут существовать между научными произведениями, в том числе отражающими результаты исследований, создают хорошую основу для разработки технологий семантического структурирования контента научных электронных библиотек.

3 Семантические связи между объектами научных электронных библиотек и информационных пространств

3.1 Научные информационные пространства

Становится все более распространенным явлением интеграция контента ЭБ отдельных организаций в виде объединения метаданных ресурсов, хранящихся в ЭБ, и созданием на этой основе единого каталога. Один из популярных подходов к решению этой задачи основан на технологии Инициативы открытых архивов (www.openarchives.org). Для подобных интегрированных информационных систем в международной литературе принято использовать термин *Information and Data Space* (DIS). В российских публикациях близкий термин *информационное пространство* (ИП) используется с начала 2000-х годов (см. например [13]).

Интеграция метаданных отдельных ЭБ и появление научных ИП является закономерным явлением, т.к. это обеспечивает множество полезных возможностей как разработчикам, так и пользователям - членам научного сообщества. Главным положительным моментом формирования ИП на основе контента отдельных ЭБ является предоставление пользователям единых интерфейсов доступа к интегрированным разнородным научным информационным ресурсам на основе стандартизации метаданных, а также унификация способов доступа к научной информации в ИП и способов ее извлечения для обработки и использования [14].

По определению, научные ИП могут включать все разнообразие научных информационных ресурсов, которые создаются научным сообществом в электронном виде. Например, ИП Соционет (socionet.ru) включает 16 типов информационных объектов, которые можно разбить на две большие группы: 1) объекты, содержание которых представляет результаты исследований и научные выходы (типы paper, article, book, chapter, citation, artifact); и 2) все

другие объекты, не являющиеся научными выходами в прямом смысле, к числу которых относятся персональные профили ученых (тип person), профили научных организаций (тип institution), научные новости (тип news), научные комментарии (тип comment) и т.д.

Каждый информационный объект ИП имеет уникальный идентификатор. На основе этих идентификаторов сервисы ИП позволяют не только просматривать информационные объекты и описывающие их метаданные, но и обеспечивают различные способы использования и обработки объектов, в том числе для формирования разнообразных связей между объектами ИП.

3.2 Научная деятельность и создание семантических связей между объектами

Научное творчество и создание научных произведений (НП) сопровождается рядом типичных процессов, которые, в частности, приводят к установлению учеными определенных смысловых связей между новым НП и уже существующим корпусом научных знаний. В традиционной научной практике это делается, например, с помощь научного цитирования. В таких случаях связь цитирования определяется внутри научного текста в соответствии с общепринятыми правилами.

Как только НП предстают в виде объектов ИП, у разработчиков контента ИП появляются новые возможности как по форме создания связей, так и по их семантическому содержанию. В отличие от традиционной формы связей (например, цитат, определенных внутри научного текста) сервисы ИП позволяют создавать связи, которые являются внешними по отношению научным текстам. Данные о связях в этом случае могут быть включены в метаданные соответствующего объекта ИП или могут существовать как самостоятельные объекты ИП. Кроме этого, электронная форма фиксации связи между двумя объектами ИП допускает включение в ее параметры различного семантического содержания. Все это существенно обогащает информативность связей, упрощает их распознавание и обработку.

Новый подход к формированию связей между объектами ИП как внешних сущностей не исключает их традиционного использования внутри документа в соответствии с правилами оформления научных цитат. Специально созданные сервисы ИП позволяют авторам электронных документов включать в текст правильно оформленные научные цитаты, которые выполняют роль гиперссылок (или указателей) на связи, описанные как самостоятельные информационные объекты вне данного документа.

В электронной среде научного ИП понятие связи между объектами ИП может быть унифицировано. При этом техническая реализация связей может быть осуществлена разными способами: внутри текста документа (как традиционная цитата) или как внешний объект, ассоциированный со связываемыми документами (или их определенными частями) с помощью параметров соответствующей связи.

Необходимо теперь уточнить, что приводит к появлению связей между объектами научного ИП. На наш взгляд, основными процессами научного творчества, при которых между объектами ИП формируются связи различных категорий, являются следующие:

1. *Процесс логического развития и преобразования уже существующего научного знания (научный вывод).* При этом автор создает связи между собственным НП, содержащим его научный результат, и уже существующими НП, которые были им творчески преобразованы и/или получили логическое развитие вследствие его усилий. Использованные автором результаты исследований, следовательно, являются научным основанием, предоставляют данные или метод получения нового научного знания. В подобных случаях корректность полученного результата напрямую зависит от корректности использованных НП, от правильности их понимания и применения автором.

В науке бывают ситуации, когда по прошествии некоторого времени результаты определенных исследований опровергаются. В таких случаях полезным является наличие в научном ИП сведений о связях НП, содержащего некорректный результат, с другими, в которых он определен как основание для их получения, или признается, подтверждается другими авторами. Это позволяет автоматически формировать уведомления авторам таких связанных НП. Подобные сигналы помогают научному сообществу оперативно пересмотреть и сделать ревизию электронного корпуса научных результатов.

Таким образом, данный процесс научного творчества порождает связи, которые полезны для содержательного пересмотра научных знаний, в случае признания недостоверными (или сомнительными) определенных результатов исследований.

2. *Процесс присоединения НП, представляющего новые научные результаты, к существующему корпусу научных знаний*. При этом происходит установление связей НП с другими существующими НП, которые хотя и не являются основанием для получения результатов данного НП, но имеют с ним некоторые иные отношения.

В данном случае не только автор может устанавливать связи между своим НП и уже существующим в ИП корпусом научных знаний (например, с уже существующими родственными и близкими НП), но и члены научного сообщества в своих НП могут фиксировать свое отношение к НП данного автора. Подобные связи могут выражать признание, поддержку и т.п., а также негативное отношение к определенным результатам НП (сомнение, несогласие, обвинение в плагиате и др.).

С учетом того, что связи научного вывода уже выделены в самостоятельную группу (см. п.1 выше), мы предлагаем среди связей, устанавливаемых в процессе присоединения нового НП к корпусу научных знаний, зафиксировать следующие основные группы (подробнее об этих группах связей см. в следующих разделах статьи):

а) связи использования, не требующие пересмотра НП, в отличие от ситуаций, описанных в п. 1 выше (например, использование как источника информации или как авторитетного мнения и т.п.);

б) иерархические и ассоциативные связи, определяющие, что данное НП содержит частный случай результата, изложенного в другом НП, или, наоборот, представляет собой концептуальное обобщение результатов других НП; также возможны другие содержательные ассоциации между результатами, представленными в рассматриваемых НП;

в) связи, характеризующие профессиональные мнения или оценки (согласие, признание, подтверждение, обсуждение, несогласие, получает поддержку от, плагиат, насмешка, пародирование и др.)

г) связи между компонентами НП (с другой его версией, с разделом другого НП, с приложением к этому НП, с иллюстрацией к НП и другие связи между частями НП).

3. *Процесс научной оценки и высказывания мнений учеными о существующих научных выходах и результатах с помощью сервисов ИП.* Этот процесс является особым случаем предыдущего. Его отличие заключается в том, что помимо связей между НП, здесь формируются также связи между личностями ученых, представленных в ИП их персональными профилями, и оцениваемыми НП. При этом возможен весь спектр оценок: от позитивной (признание, поддержка и др.) до негативной (сомнение, несогласие, обвинение в плагиате и др.).

4. *Процесс систематизации, классификации и упорядочивания содержания корпуса научных знаний.* По сути, это процесс аналитической переработки множества уже существующих результатов науки. Основным продуктом этого процесса являются новые связи, обнаруживаемые учеными между уже известными результатами, которые представлены в анализируемых НП. Типичными видами данного процесса являются: написание научных обзоров, классификация и создание тематических указателей научных публикаций в конкретных областях науки т.п.

Результаты этого процесса могут отображаться в ИП путем формирования связей не только между информационными объектами, представляющими НП, но и объектами, которые представляют участников научной деятельности (между авторами и их НП, между научными организациями и их сотрудниками – авторами представленных в ИП НП и т.д.).

Итак, результаты описанных выше четырех основных процессов научного творчества отображаются в ИП следующим образом:

- С одной стороны, существует множество НП (корпус научных знаний), которые наряду с персональными профилями ученых, профилями организаций и другими научными материалами (новости, комментарии и т.п.) представлены как объекты научного ИП.
- С другой стороны, существует множество научных организаций и исследователей – действующих лиц ИП, научная деятельность которых проявляется в форме: а) создания новых НП (пополнение корпуса научных знаний), что приводит к появлению новых объектов научного ИП; и б) создания новых и коррекция существующих семантических связей между объектами научного ИП.
- Индивидуальное научное творчество ученых в создании семантических связей между объектами ИП приводит к расширению и изменению многослойной семантической структуры ИП, различные слои которой формируются из связей различных категорий.
- Формируемые в ИП семантические связи категоризируются в соответствии с их функциональным характером. Каждой категории связей соответствует некоторый слой в многослойной структуре семантических связей между объектами ИП, а также один или несколько словарей свойств, которыми могут обладать связи данной категории.

На наш взгляд, рассмотренные выше четыре вида процессов научной деятельности позволяют выделить следующие категории связей (подробнее они рассматриваются в п. 3.3):

1) связи научного вывода; такие связи обеспечивают идентификацию научных результатов, требующих пересмотра при опровержении «родительского» результата;
2) связи с НП, использованными при получении нового результата исследований, не являющиеся связями научного вывода;
3) связи, характеризующие профессиональные мнения или оценки ученых о конкретных НП;
4) иерархические и ассоциативные связи между НП;
5) связи между компонентами НП;
6) научно-организационные связи.

Следует отметить, что предложенная категоризация связей не является исчерпывающей. Научная практика может формировать новые типы отношений между объектами ИП, что соответственно, приведет к появлению новых слоев в многослойной структуре связей информационных объектов ИП.

3.3 Категории связей и словари их свойств

В данном разделе рассматриваются разработанные авторами словари, определяющие свойства ориентированных бинарных семантических связей между объектами научного ИП. Каждый словарь соответствует только одной из рассмотренных выше категорий семантических связей. Словари создавались для применения в системе Соционет. По мнению авторов, предложенная классификация свойств связей и созданные словари могут найти применения в других научных электронных библиотеках и ИП.

При разработке словарей использованы упоминавшиеся выше онтологии SPAR (в частности онтологии CiTO и DoCo), спецификация SKOS консор-

циума W3C, онтология проекта SWAN, а также один из разделов CERIF, посвященный семантике связей. При этом словари 1, 2, 4 и 5 описывают свойства связей между НП, словарь 3 – связи между НП, а также между персональными профилями ученых и НП, а словари из раздела 6 – свойства связей научно-организационного характера, учитывающие специфику отечественных научно-исследовательских организаций.

Словари организованы следующим образом. Они разбиты по категориям связей, перечисленным в предыдущем разделе, на 6 групп. В последней 6-й группе приведено несколько словарей для различных подкатегорий, в остальных – по одному словарю в группе. Для каждого свойства связей в словарях приведено его название (на русском языке), его оригинальное английское название в использованном источнике, если он имеется, а также в скобках указание на этот источник. Термин «целевой» в названиях определяет объект, на который направлена соответствующая связь.

1) Словарь свойств связей научного вывода

- *Заимствует основополагающие идеи в целевом* - obtain background from (CiTO)
- *Развивает целевой* - updates (CiTO)
- *Подтверждается целевым* - cites as evidence (CiTO)
- *Подтверждает целевой* - confirms (CiTO)
- *Уточняет целевой* - qualifies (CiTO)
- *Исправляет целевой* - corrects (CiTO)

2) Словарь свойств связей использования

- *Содержит утверждения/факты целевого* - contains assertion from (CiTO)
- *Использует данные из целевого* - uses data from (CiTO)
- *Использует метод из целевого* - uses method from (CiTO)
- *Опровергает целевой* - refutes (CiTO)
- *Совершает плагиат целевого* - plagiarizes (CiTO)

3) Словарь свойств связей мнений и оценок

- *Позитивно оценивает целевой* - agrees with (CiTO), supports (CiTO), respondsPositivelyTo (SWAN), credits (CiTO), consistentWith (SWAN)
- *Негативно оценивает целевой* - critiques (CiTO), disagrees with (CiTO), responds NegativelyTo (SWAN), inconsistentWith (SWAN), disputes (CiTO), parodies (CiTO), ridicules (CiTO)
- *Нейтрально оценивает целевой* – responds NeutrallyTo (SWAN)

4) Словарь свойств иерархических и ассоциативных связей между НП

- *Расширяет целевой* - extends (CiTO), broader (SKOS)
- *Сужает целевой* - narrower (SKOS)
- *Родственный целевому* - related (SKOS), relevantTo (SWAN)
- *Альтернативен целевому* - alternativeTo (SWAN)

5) Словарь свойств связей между компонентами НП

- *Дублирующая копия целевого*
- *Новая редакция целевого*
- *Ранняя редакция целевого*
- *Аудио/видео версия текстового целевого*
- *Текстовая версия аудио/видео целевого*
- *Презентация текстового целевого*
- *Часть целевого* - isPartOf (DoCo), paragraph (DoCo), part (DoCo), section (DoCo)
- *Включение целевого как части* - hasPart (DoCo)
- *Абстракт целевого* - abstract (DoCo)
- *Оглавление целевого* - table of contents (DoCo)
- *Предисловие или введение целевого* - foreword (DoCo), preface (DoCo)
- *Приложение к целевому* - appendix (DoCo)
- *Библиография целевого* - bibliography (DoCo)
- *Глоссарий целевого* - glossary (DoCo)

6) Словари свойств научно-организационных связей

6.1) Свойства связей «персона – персона»

- *Научный руководитель* - Mentor (CERIF)
- *Административный руководитель* - Manager (CERIF)

6.2) Свойства связей «персона – организация»

- *Директор* - Director (CERIF)
- *Заместитель директора* - Deputy Director (CERIF)
- *Руководитель подразделения* - Head of Department(CERIF), Group Leader (CERIF)
- *Сотрудник* - Employee (CERIF)
- *Главный научный сотрудник*
- *Ведущий научный сотрудник*
- *Старший научный сотрудник* - Senior Researcher (CERIF)
- *Научный сотрудник* - Researcher (CERIF)
- *Младший научный сотрудник* - Junior Researcher (CERIF)
- *Докторант*
- *Аспирант*
- *Стажер*
- *Профессор* - Professor (CERIF)
- *Доцент* - Assistant Professor (CERIF)

6.3) Свойства связей «персона – публикация»

- *Автор* - Author (CERIF)
- *Редактор* - Editor (CERIF)

- *Рецензент* - Reviewer (CERIF)
- *Переводчик* - Translator (CERIF)

6.4) Свойства связей «организация – публикация»

- *Обладатель прав* – Intellectual Property Rights Claim (CERIF)
- *Издатель* - Publisher (CERIF)
- *Организация-автор* - Author (CERIF)

Предлагаемая категоризация и описанные свойства семантических связей объектов научного ИП отражают субъективную точку зрения авторов. К сожалению, пока не существует общепринятых научным сообществом стандартов, в достаточно полной мере охватывающих рассматриваемую сферу. Однако, как указывалось выше, наш подход основывается на обобщении известных попыток концептуального и онтологического моделирования в области научной и издательской деятельности. Кроме того, мы исходим из характера той информации, которую было бы желательно получать, анализируя корпус научных знаний, представленных в ИП.

3.4 Семантические связи как объекты ИП

В электронных библиотеках традиционно с помощью прямых гиперссылок поддерживаются связи между каталогами информационных объектов и описываемыми в них информационными объектами. Аналогично поддерживаются связи цитирования, связи с профилями авторов и организаций, и некоторые другие. Для этого в ЭБ имеются метаданные, описывающие информационные объекты, их авторов (профили авторов), коллекции информационных ресурсов, организации – места работы авторов (профили организаций) и др. В таком случае ссылки между информационными объектами представляются как атрибуты метаданных, описывающих информационные объекты. С использованием таких ссылок возможно анализировать структуру связей, осуществлять наукометрические измерения, визуализировать структуру связей.

Однако при таком традиционном способе создания связей в ЭБ, как правило, явным образом не отражается семантика связей. Например, для связи цитирования одного информационного объекта с другим отсутствует информация, характеризующая цель цитирования, оценку цитируемой работы и другие характеристики. Рассмотрим в общем виде модель связей между объектами научного ИП, устраняющую это ограничение.

Связи могут представляться в ИП двумя способами. При использовании первого способа данные, описывающие связи, содержатся в метаданных одного из связываемых объектов, например, в метаданных исходного объекта связи. Однако поскольку в ИП, построенных 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 объектами ИП. При втором способе создаваемые связи представляются как самостоятельные объекты ИП. 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 ИП, поскольку при этом не затрагиваются метаданные связываемых объектов.

Описание связи в обоих представлениях должно включать уникальный идентификатор целевого объекта связи, а также может включать атрибут, характеризующий семантику связи, различного рода комментарии и пр. Если связь создается как самостоятельный объект ИП, то ее описание дополнительно к уже перечисленному должно включать уникальный идентификатор данного объекта в ИП; уникальный идентификатор пользователя, создающего данную связь; уникальный идентификатор исходного объекта связи в ИП (мы рассматриваем ориентированные бинарные связи), а также даты создания или изменения связи. Для описания семантики связи указывается свойство связи, выбираемое из поддерживаемых контролируемых словарей, возможный состав которых рассматривался в п. 3.3. Полномочия на установление связей между объектами ИП предоставляются только зарегистрированным в системе пользователям, что обеспечивает автоматическую фиксацию идентификатора пользователя, создающего связи.

Процедура установления связи между двумя информационными объектами ИП может быть реализована по-разному. Далее описана ее реализация в системе Соционет. При этом использован второй способ представления связей.

Множество параметров, влияющих на установление связи, включает: а) тип исходного объекта связи; б) тип целевого объекта связи; в) множество категорий связей, учрежденных в системе для заданной пары типов объектов ИП; г) множество словарей свойств связей, предусмотренных в системе для связей заданной категории; д) множество свойств связей в словаре, выбранном для установления связи между объектами заданного типа.

Рассматриваемая процедура состоит из следующих шагов:

1) Пользователь выбирает пару связываемых объектов ИП.

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 свойств связей для включения в систему. Это предложение вступит в силу только после одобрения администратором ИП.

3) Если подходящая категория связей выбрана, то открывается соответствующий словарь свойств связей.

4) Если в словаре имеется подходящее свойство связи, характеризующее требуемое семантическое отношение между заданной парой объектов, то пользователь его выбирает. Если же такое свойство отсутствует, пользователь может предложить подходящее свойство связей для пополнения данного словаря. Предложение вступит в силу только после одобрения его администратором ИП или соответствующего словаря.

5) По желанию пользовать может привести в описании связи комментарий, объясняющий принятые им решения при ее создании.

6) Сформированный информационный объект 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 ИП. При этом также осуществляется проверка непротиворечивости семантики новой связи с уже существующими связями между данными объектами, установленными тем же пользователем.

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

Поддержка развитой структуры семантических связей в достаточно представительном научном ИП позволяет в результате их анализа осуществлять наукометрические измерения, использовать технологии «живых» публикаций [15, 16], 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 научного ИП может устанавливать связи между любыми его информационными объектами. При определении семантики связей, их создатель выражает свое субъективное мнение, которое в некоторых случаях может вызывать несогласие или протест как авторов объектов, которые участвуют в данных связях, так и других членов научного сообщества. Например, могут вызывать протесты случаи, когда устанавливаются семантические связи, несущие негативную оценку некоторого научного выхода (опровержение, высмеивание, обвинение в плагиате и т.п.).

Как известно, научная истина устанавливается в процессе борьбы 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, на наш взгляд, 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).

Пользователи научного ИП создают связи в своем личном (закрытом от свободного доступа) пространстве. Такое пространство и сервисы для его использования предусматривается для авторов или администраторов информационных ресурсов в системе Соционет и называется их Личной зоной. Создаваемые в Личной зоне объекты, представляющие связи, предлагаются далее для включения в общедоступное ИП. Связи становятся частью общедоступного научного ИП только после их одобрения модератором.

4 Средства создания и обработки семантических связей в Соционет

Для того чтобы создаваемые рассмотренным способом информационные объекты, представляющие семантические связи в научном ИП, стали его полноценной частью, необходимо иметь в системе сервисы: а) обеспечивающие создание связей в соответствии с рассмотренной выше процедурой; б) осуществляющие автоматический мониторинг создаваемых связей и всех их изменений; в) визуализирующие связи в системе навигации ИП; г) выполняющие сбор и накопление наукометрической статистики, позволяющей анализировать распределения связей по их семантическим свойствам для заданных видов научных информационных объектов (статей, монографий, презентаций докладов, авторефератов диссертаций и др.), по заданным авторам и исследовательским организациям; д) анализирующие топологию структуры связей в ИП с учетом их семантических категорий и конкретных свойств.

Пилотный вариант перечисленного комплекса сервисов реализуется в настоящее время в системе Соционет. Далее кратко рассматриваются некоторые вопросы, связанные с их реализацией.

4.1 Средства создания новых словарей свойств связей и отдельных новых свойств в словарях

Содержание словарей, описывающих свойства семантических связей между информационными объектами научного ИП, в силу своей новизны представляет собой предмет для научных дискуссий. Для того чтобы представители научного сообщества, пользующиеся системой Соционет, имели возможности для формирования приемлемых словарей свойств связей, в системе разработаны механизмы, которые, с одной стороны, позволяют создавать и включать в научное ИП альтернативные словари, а с другой – осуществлять пополнение и развитие уже существующих словарей.

В этих целях в Личной зоне зарегистрированного пользователя Соционет существует возможность создавать коллекции объектов типа *metrics*. Структура описания объектов этого типа сконструирована специально для разработки словарей свойств связей.

Описание конкретного объекта коллекции типа *metrics, представляющего некоторое свойство связей,* включает, в частности, следующие атрибуты: а) уникальный идентификатор этого объекта в словаре, которому он принадлежит; б) название объекта словаря (свойства связей); в) пояснения и комментарии для данного объекта; г) указание автора и организации, представляющих данный объект словаря; д) ссылки на источники, откуда заимствовано соответствующее значение словаря, например, в нашем случае – онтология CiTO.

Атрибуты "б" и "в" могут иметь значения одновременно на различных языках (например, на русском и английском). Атрибут "д" может содержать библиографические данные источника, откуда позаимствован данный объект, ссылку на его определения в онтологиях, энциклопедиях и т.п. После создания своего словаря пользователь может предложить его для включения в ИП Соционет для общего использования. Администратор Соционет рассматривает поступившее предложение и, если у него нет возражений, то данный словарь добавляется к списку доступных словарей.

Если у пользователей Соционет возникает намерение внести изменения в уже существующий словарь, то это может быть сделано двумя путями: а) послать автору словаря электронное письмо с предложением (данные об авторе словаря указаны в описании коллекции, представляющей этот словарь); или б) создать в своей Личной зоне дополнительный элемент (описать новое свойство связей) для существующего словаря и послать его автору соответствующего словаря через сервисы Личной зоны.

4.2 Средства создания связей

Для создания связей между объектами ИП пользователи Соционет имеют следующие основные возможности: а) при создании информационного объекта его автор может указать связи данного объекта с другими объектами ИП в метаданных, описывающих создаваемый объект; б) автор может отредактировать метаданные своего ранее созданного объекта, в том числе, добавляя, изменяя или удаляя определенные для этого объекта связи; в) пользователь может создать связь между объектами, не являясь их автором; в этом случае он не может изменять метаданные, описывающие эти объекты; создаваемые им связи являются самостоятельными объектами ИП.

В тех случаях, когда пользователь является автором связываемого объекта (случаи "а" и "б" выше) и при этом данный объект является научной публикацией, он имеет возможность, в частности, наряду с другими видами связей создавать и связи цитирования. Для этого нужно не только описать создаваемую связь в метаданных данного объекта, но и обозначить ее в тексте публикации в соответствии с правилами оформления ссылок и цитат в научной литературе.

В Личной зоне Соционет подобная операция возможна для объектов типов *citation* и *artifact*, которые являются в системе разновидностями НП.

Как уже отмечалось, для случаев "а" и "б" данные о связях могут входить в состав метаданных объекта, для которого связь является исходящей. В противном случае данные о связях могут храниться вне метаданных связываемых объектов. При этом они существуют как самостоятельные объекты ИП и соединяются с метаданными объектов, к которым они относятся по указанным в описании объекта-связи уникальным идентификаторам этих объектов.

Для связей, являющихся самостоятельными объектами, в системе Соционет предусмотрен специальный тип информационных объектов *linkage*. Все созданные объекты-связи должны принадлежать какой-либо коллекции объектов этого типа. Связи первоначально создаются пользователем в его Личной зоне в Соционет и хранятся в его личных коллекциях. Пользователь может предложить созданные связи для размещения в публичное ИП в виде полных коллекций или отдельными объектами через специализированные разделы Открытого Архива Соционет.

Операции по созданию внешних связей между объектами ИП системы Соционет соответствуют процедуре, описанной выше в разделе 3.4.

Фактически, предлагаемый подход предусматривает создание в системе Соционет открытого репозитория семантических связей, который дополняет уже много лет функционирующий открытый репозиторий научных статей, материалов, персональных профилей, профилей организаций и т.п.

4.3 Сервисы визуализации, мониторинга и анализа связей

В системе Соционет предусмотрена возможность визуализации связей между информационными объектами ИП. Для каждого объекта в ИП предоставляется навигационное меню, которое позволяет пользователю переключить просмотр информационных ресурсов в режиме навигации по графу связей. При этом для заданного объекта на экране отображаются узлы ближайших связанных с ним объектов, а также данные о свойствах связей, статистика по связанным объектам и т.п. Навигационный граф показывает все существующие связи объекта, как исходящие, так и входящие, а также позволяет их фильтровать по заданным свойствам, например, включать в него только связи заданной категории. Относительно начального узла графа возможен переход по пути любой длины в структуре связей.

В Соционет работает сервис автоматического ежедневного мониторинга изменений связей. Фиксируются факты появления новых связей, а также изменения значений атрибутов существующих. В рамках этой процедуры ежедневно обновляется ин-

декс связей, который используется для построения навигационного графа связей, для построения списков связанных объектов заданной вложенности, а также для поиска и анализа связей по заданным атрибутам.

На основе этого сервиса также разрабатывается система оповещений пользователей по электронной почте о значимых изменениях в связях. Система оповещения конструируется так, чтобы рассылать уведомления: а) автору объекта, если установлена новая связь с его объектом, б) автору объекта – научного произведения при изменении объектов, с ним связанных (например, при изменении статьи, которую автор цитирует в своем произведении) и т.д. [17, 19]. При появлении связи-оценки автору оцениваемого объекта в уведомлении предполагается сообщать характер этой оценки и идентифицировать объект-связь, ее определяющий.

Система рассылает автоматические оповещения сразу после включения созданных связей в общедоступное ИП, т.е. после предварительного одобрения их модератором Соционет.

Авторы объектов ИП - научных произведений, с которыми установлены связи, а также все желающие, имеют возможность выражать свое мнение (согласие, возражение или др.) по поводу семантики и значений других атрибутов созданных связей, представленных в их описаниях.

Подобные профессиональные мнения и оценки ученых о содержании связей также выражаются с помощью средств создания связей, относящихся к категории "мнения и оценки", и также должны пройти через процедуру модерирования. Они становятся доступными в ИП при просмотре всех объектов, которым они посвящены или с которыми связаны (при просмотре исходной связи, которой даны оценки, при просмотре профиля автора такой связи, при просмотре профиля автора, который дал оценку связи и т.п.).

Вся история установления и изменения связей фиксируется в базе данных системы. На этой основе реализован сбор и накопление наукометрической статистики. Система также создает и хранит статистический портрет ученого, иллюстрирующий, когда и какие связи им создавались (исходящие связи), а также аналогичные данные о связях, установленных с НП данного ученого (входящие связи). При этом обеспечивается возможность получить представление о характере распределения семантических свойств как входящих, так и исходящих связей.

Подробнее вопросы формирование из этих данных наукометрической статистики обсуждаются в [11, 12].

Нужно, наконец, отметить, что поддержка в системе Соционет развитой многослойной структуры семантических связей между информационными объектами, являющимися научными публикациями, позволяет получать разнообразную аналитическую информацию о структуре различных областей исследований, вкладе в их развития конкретных ученых, о процессе их эволюции и т.д. Исследования в этой области планируется развивать на основе системы Соционет.

В связи с этим следует здесь упомянуть функциональный модуль SciValSpotlight проекта SciVal компании Elsevier [18]. Этот модуль служит для анализа 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, в отличие от нашего подхода, при этом используются традиционные «немые» связи - связи, не несущие семантики. Наш подход обеспечивает более дифференцированный анализ, результаты которого, учитывают семантику связей.

5 Заключение

Предложенный в данной работе подход к семантическому структурированию контента научных электронных библиотек и информационных пространств обеспечивает существенное обогащение как информационных, так и функциональных возможностей этих важных средств информационной поддержки научных исследований. 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] Fredrik Åström, Ágnes Sándor. Models of Scholarly Communication and Citation Analysis, In: Proc. of ISSI 2009: The 12th Intern. Conf. of the International Society for Scientometrics and Informetrics: Volume 1. http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=1459018&fileOId=1883080

[2] Ágnes Sándor, Aaron Kaplan, Gilbert Rondeau. 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

[3] Anita de Waard; Joost Kircz. Modeling Scientific Research Articles – Shifting Perspectives and Persistent Issues, Proc. of ELPUB 2008 Conf. on Elec-

tronic Publishing - Toronto, Canada - June 2008. http://elpub.scix.net/data/works/att/234_ elpub2008 .content.pdf
[4] CERIF 2008 – 1.2 Semantics, euroCRIS. http://www.eurocris.org/Uploads/Web%20pages/ CERIF2008/Release_1.2/CERIF2008_1.2_ Semantics.pdf
[5] David Shotton. Introduction the Semantic Publishing and Referencing (SPAR) Ontologies. October 14, 2010. http://opencitations.wordpress.com/ 2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/
[6] David Shotton. CiTO, the Citation Typing Ontology. J. of Biomedical Semantics 2010, 1(Suppl 1): S6. http://www.jbiomedsem.com/content/1/S1/S6
[7] David Shotton, Silvio Peroni. DoCO, the Document Components Ontology. 17/02/2011. http://purl.org/ spar/doco/
[8] SKOS - Simple Knowledge Organization System. http://www.w3.org/TR/skos-reference/
[9] Simone Teufel, et al. Automatic classification of citation function. Proc. of the 2006 Conf. on Empirical Methods in Natural Language Processing. http://portal.acm.org/citation.cfm?id=1610091
[10] SWAN (Semantic Web Applications in Neuromedicine) - Scientific Discourse Relationships Onto logy Specification. http://swan.mindinformatics. org/spec/1.2/discourserelationships.html
[11] Когаловский М.Р., Паринов С.И. Метрики онлайновых информационных пространств //Экономика и математические методы. – 2008. – Вып. 2.
[12] Когаловский М.Р., Паринов С.И. Использование связей цитирования для наукометрических измерений в системе Соционет. Соционет: Электронный депонент, 2009. http://socionet.ru/ publication.xml?h=repec:rus:rssalc:web-32
[13] Паринов С.И. СОЦИОНЕТ.РУ как модель информационного пространства 2-го поколения //Информационное общество. - 2001, вып. 1, с. 43-45. http://emag.iis.ru/arc/infosoc/emag.nsf/ BPA/709c3727bab54cf4c3256c01002d2e6e
[14] Паринов С.И. Информационные хабы. Социнет Электронный депонент. http://socionet.ru/ publication.xml?h=repec:rus:mqijxk:9
[15] Паринов С.И., Когаловский М.Р. Технология поддержки электронных научных публикаций как «живых» документов. Труды XI Всероссийской научной конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Петрозаводск, 17-21 сентября 2009 г. – Петрозаводск: КарНЦ РАН, 2009
[16] Паринов С.И., Когаловский М.Р. «Живые» документы в электронных библиотеках //Прикладная информатика. – 2009. - № 6, 2009. Авторская версия: http://socionet.ru/publication. xml?h=repec:rus:isyigw:article-215
[17] Паринов С.И. Концепция виртуальной научной среды "Открытая Наука" //Труды международной суперкомпьютерной конф. "Научный сервис в сети Интернет: суперкомпьютерные центры и задачи", Новороссийск, 20-25 сентября 2010 г. – М.: Изд-во МГУ, 2010, стр. 473-481. Электронная авторская версия: http://socionet.ru/ publication.xml?h=repec: rus:mqijxk:24
[18] SciVal. http://www.elsevier.com/wps/find/ electronicproductdescription.cws_home/720941/ description#description
[19] Sergey Parinov The electronic library: using technology to measure and support Open Science. In Proc. of the World Library and Information Congress: 76th IFLA General Conference and Assembly. 10-15 August 2010, Gothenburg, Sweden. pp. 1-13. Электронная авторская версия: http://socionet.ru/publication.xml?h=repec:rus:mqijxk:25

A technology for semantic structurization of scientific digital library content

Sergey Parinov, Mikhail Kogalovsky

Semantic structuring of digital libraries' contents, which resulted in explicit implementation of semantic linkages among information objects, opens new opportunities for scientific creativity and increase a quality of the libraries' contents. In combination with categorizing of semantic linkages it is establishing multilayer semantic networks over information objects that allow new qualitative scientometrics measurements and investigation on structuring properties of the corpus of science represented by digital libraries contents. The paper discusses proposed approach and specifications of semantic structuring technology and describes a pilot implementation of the technology within Socionet scientific data and information space and services.

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