Семантическое аннотирование информационных ресурсов в научной электронной библиотеке средствами таксономий

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

[1]Институт проблем рынка РАН,
[2]Центральный экономико-математический институт РАН,
Москва

kogalov@gmail.com sparinov@gmail.com

Аннотация. Описана проблема семантического аннотирования фрагментов полных текстов публикаций, а также ссылок цитирования в публикациях научной электронной библиотеки. Предложен таксономический подход к описанию семантики аннотаций. Обсуждены основные понятия, связанные с аннотированием. Представлен ряд таксономий аннотаций, почерпнутых из литературы и опыта собственных разработок авторов. Рассмотрена реализация семантического аннотирования публикаций в научной информационной системе Соционет, которая использует также открытые данные, создаваемые средствами проекта CitEcCyr. На основе данных о содержании цитирований при просмотре публикаций в Соционет автоматически создаются аннотации внутритекстовых ссылок на используемые источники из списков литературы публикаций. Создаваемые аннотации содержат сводную информацию об источниках и статистику их цитирований.

Ключевые слова: информационный ресурс, аннотация, таксономия, цитирование, электронная библиотека, система Соционет, проект CitEcCyr.

Semantic Annotation of Information Resources by Taxonomies in Scientific Digital Library

© M.R. Kogalovsky[1] © S.I. Parinov[2]

[1]Market Economy Institute of RAS,
[2]The Central Economical and Mathematical Institute of RAS,
Moscow

kogalov@gmail.com parinov@gmail.com

Abstract. The paper discusses a semantic annotating problem with focus on full texts of research papers and citation references in publications from scientific digital library. We propose a taxonomy based approach for specifying annotation semantics. We discuss the main concepts of annotation and some annotation taxonomies taken from literature and early created by ourselves. An implementation of semantic annotating approach within the research information system Socionet is presented. This implementation is using also the open citation data created by the CitEcCyr project tools. Based on data about the content of citations while browsing the publications at Socionet automatically annotations are created for in-text references to the sources from the reference lists of publications. Generated annotations contain summary information about the sources and the statistical data about their citations.

Keywords: information resource, annotation, taxonomy, citation, digital library, Socionet system, CitEcCyr project.

1 Введение

Работая с печатным научным текстом, читатель часто делает выписки цитат или других важных для него фрагментов публикации, выделяет их в тексте, делает комментарии на полях. При работе с текстом на компьютере средствами текстовых редакторов все эти возможности также доступны. Так, версии широко распространенного текстового редактора MS Word позволяют идентифицировать фрагменты текста шрифтовым выделением или цветом, связывать с нужными фрагментами комментарии. Выделять фрагменты текста цветом и/или сопровождать их комментариями позволяют также продукты компании Adobe такие как Adobe Reader или Adobe Acrobat и некоторые другие программные

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

средства. К сожалению, средства для таких целей не предусмотрены в стандартных веб-браузерах при просмотре страниц в формате HTML или XML, и для этого нужно использовать другие программные инструменты.

Аннотации, как результаты такой работы с текстом, читатель может создавать для собственных целей и/или для других ученых, в том числе, в процессе совместной работы по подготовке текстового документа или его экспертизе. Деятельность такого рода называется *аннотированием*. В общем случае аннотироваться могут не только тексты, но и информационные ресурсы, представленные в иных средах (графика, аудио, видео).

Аннотирование может осуществляться *в двух формах*. Первая из них заключается в дополнении к свойствам аннотируемого объекта некоторых новых атрибутов, характеризующих его дополнительные ранее не определенные свойства. Это, например, цветовое выделение фрагментов текста, тегирование музыкальных клипов, фотографий в коллекции или статей в Википедии и т. п. Вторая форма аннотирования состоит в создании нового информационного объекта, ассоциируемого с аннотируемым (целевым) объектом (субъектом аннотирования) и несущего некоторую относящуюся к нему информацию, например, комментарий, характеризующий эмоции читателя, связанные с восприятием содержания целевого объекта, или оценку его содержания, различного рода дополнения к нему и т.д. Такие вновь созданные информационные объекты, ассоциируемые с целевыми объектами, называются их *аннотациями*. В англоязычной Википедии [3] аннотацией называются «метаданные (например, комментарий, пояснение, разметка презентации), которые присоединяются к тексту, изображению или другим данным. Часто аннотации ссылаются на конкретную часть исходных данных».

В настоящее время созданы компьютерные технологии, предназначенные для аннотирования информационных объектов Веба, которые представлены в различных видах – тексты, аудио, видео и др. Однако для пользователей научных электронных библиотек и других научных информационных систем особый интерес представляет аннотирование *цифровых текстовых документов.* При этом в качестве целевых информационных объектов могут выступать не только такие документы в целом, но и их отдельные фрагменты.

В ряде развитых электронных библиотек, например, в системе Соционет [11], их информационные ресурсы включают как текстовые документы, так и различного рода связи, отражающие различные отношения между ними. В таких случаях объектами аннотирования могут быть не только фрагменты текстовых документов или документы в целом, но и связи между ними.

Аннотации сами могут представляться в виде связей между аннотатором (автором аннотации, представленным в библиотеке его персональным профилем) и аннотируемым целевым объектом. В таком случае семантика аннотации представляется семантикой этой связи.

Аннотация целевого объекта может иметь различную *семантику*, которая представляется явным или неявным образом. Если семантика представлена явным образом, то такая аннотация называется *семантической.* Соответственно, деятельность, продуктом которой являются такие аннотации, естественно называть *семантическим аннотированием.* Назначение семантической аннотации – специфицировать смысл и некоторые свойства аннотируемого ресурса.

Семантика аннотации может быть выражена *неформально*, неструктурированными метаданными, например, в виде комментария-пояснения на естественном языке, или *формально* с помощью структурированных метаданных, связывая аннотированный ресурс с некоторой семантической структурой конкретной предметной области, например, с микроформатами или с онтологией предметной области коллекции текстовых документов.

При использовании онтологии (в более простом случае – таксономии) для аннотирования используются ее классы и отношения. В случае использования онтологии для формального описания семантики аннотации аннотирование называют *онтологическим*. Могут использоваться и *комбинированные аннотации*, состоящие из формального и неформального компонентов. Например, аннотация может указывать класс таксономии, характеризующий свойство аннотируемого объекта, а также содержать текстовый комментарий на естественном языке, выполняющий аналогичную функцию или характеризующий отношение автора аннотации к целевому объекту.

Использование семантического аннотирования существенным образом обогащает восприятие информационных ресурсов пользователями, помогает интерпретировать контент аннотированных ресурсов пользователям и механизмам систем, оперирующих с ними. Оно также обеспечивает дополнительные возможности для большей полноты и точности поиска информационных ресурсов, для их анализа и обработки в больших коллекциях. На основе коллекций аннотированных научных публикаций семантические аннотации могут также использоваться для генерации различных наукометрических показателей.

Семантическое аннотирование может выполняться *вручную* экспертами, может быть *полуавтоматическим* или полностью *автоматическим,* выполняемым с помощью программных систем-аннотаторов, основанных на извлечении необходимой для этого информации из аннотируемого ресурса. Среди таких систем известны разработки, базирующиеся на наборах данных Open Linked Data (LOD) (см., например, [7]), на DBpedia [6] или Freebase [5].

В настоящей статье обсуждается подход авторов к семантическому аннотированию информационных объектов контента научных электронных библиотек

– полных текстов публикаций и их фрагментов, в частности, ссылок в тексте на используемые источники с их контекстом. Подход реализован и продолжает развиваться при участии авторов в рамках отечественной научной информационной системы Соционет [11]. Семантика аннотаций определяется средствами встроенной в систему таксономии, представленной в виде набора контролируемых словарей. Наряду с публикациями, представленными в системе, в качестве источников субъектов аннотирования используется массив описаний ссылок цитирования, автоматически генерируемый из полных текстов этих публикаций в PDF-формате [4].

Существенно отметить здесь, что специфика таксономии, используемой в нашем случае, ориентирована на описание семантики аннотаций для научных публикаций.

Остальная часть статьи организована следующим образом. В разделе 2 рассмотрен ряд проектов и публикаций, в которых предложены различные варианты таксономий аннотаций, позволяющих описывать те или иные аспекты их семантики. Особое внимание уделяется рекомендациям консорциума W3C по открытому аннотированию, также включающим один из вариантов таксономии аннотаций. В разделе 3 обсуждаются принятый подход к семантическому аннотированию в системе Соционет и его реализация. Раздел 4 посвящен обсуждению инструментария для автоматической генерации на основе полных текстов публикаций, представленных в PDF-формате, аннотаций ссылок на используемые источники. При этом аннотация включает извлеченный из полного текста контекст ссылки. Сгенерированный массив аннотаций ссылок цитирования может далее обрабатываться средствами системы Соционет. Заключение (раздел 5) подводит итоги обсуждения проблемы семантического аннотирования.

2 Таксономии аннотаций

По проблематике аннотирования вообще и семантического аннотирования, в частности, существует обширная литература, посвященная обсуждению различных подходов к аннотированию ресурсов, представленных в различных средах и относящихся к различным областям приложений, созданию стандартов в этой области, разработкам инструментария для автоматизации процесса аннотирования, подходов к семантическому аннотированию на основе различных семантических структур (систем знаний), использованию семантического аннотирования в области информационного поиска и извлечения информации из текстов, для анализа и обработки аннотированных информационных ресурсов.

Здесь мы рассмотрим несколько представленных в литературе, в том числе, разработанных авторами данной статьи подходов к описанию семантики аннотаций на основе их классификации с помощью подходящих таксономий. Иначе говоря, рассмотрим ряд походов к семантическому аннотированию на основе таксономий аннотаций, базирующих на различных их свойствах. Назовем такое аннотирование *таксономическим аннотированием*. Помимо описания семантики аннотаций использование такого подхода позволяет создавать механизмы поиска публикаций и фрагментов публикаций, адекватных потребностям пользователей, в частности, ссылок на используемые источники, а также генерировать на этой основе новые нетрадиционные наукометрические показатели.

Используемая таксономия обычно зависит от предметной области аннотируемых информационных ресурсов, целей аннотатора (эксперта или инструмента аннотирования), характера ресурсов (например, фрагменты текста или ссылки на используемые в нем источники).

Рассмотрим ряд таксономий аннотаций, предлагаемых для использования в научных электронных библиотеках. Прежде всего, обратимся к работе с привлекательным названием *"What are Semantic Annotations?"* [10]. Хотя это название обязывает авторов предложить какое-либо определение понятия *семантическая аннотация*, такого определения в явном виде в статье нет. Однако предложены общий взгляд на аннотирование и некоторая полезная систематизация сферы аннотирования. Предложения авторов статьи базируются на анализе различных подходов к аннотированию ресурсов на примере таких систем, как Semantic Wikis, Semantic Blogs, Tagging. При этом аннотирование рассматривается в общем виде как присоединение определенных данных к некоторой другой порции данных с установлением того или иного отношения между аннотированными и аннотирующими данными. Авторы различают три типа аннотаций – неформальные, формальные и онтологические. *Неформальная аннотация* представляется не на формальном языке и поэтому не является *машино-интерпретируемой* (у авторов – машиночитаемой). Напротив, *формальная аннотация* представляется на формальном языке и благодаря этому *машино-понимаема*. Однако в ней не используются термины онтологии. Наконец, *онтологическая аннотация* (которую авторы, вероятно, и понимают как семантическую) основана на использовании только терминов онтологии, и поэтому она имеет общепонятный смысл в сообществе, разделяющем эту онтологию.

В [10] предложена также общая модель аннотации, в которой предполагается, что аннотация состоит из четырех компонентов: *субъекта аннотации* – аннотируемых данных, ее *объекта* – аннотирующих данных, *предиката*, определяющего тип отношения между объектом и субъектом аннотации, и, наконец, *контекста аннотации,* характеризующего, когда и кем она создана, возможно, период времени или область пространства, где она имеет силу, и т. п. Каждый из этих компонентов может быть формальным или неформальным. Для случая аннотирования ресурсов

Веба понятия формальной и онтологической аннотации определяются более конкретно с использованием URI.

В терминах компонентов общей модели аннотации в цитируемой работе предложены заимствованные авторами из ряда публикаций *критерии* (измерения) для классификации аннотаций. Показано, какие классы аннотаций используются в каждой из анализируемых в начале статьи систем, обладающих средствами аннотирования. Использованы следующие критерии классификации аннотаций:

Ассоциация – способ, которым аннотация ассоциируется с аннотируемым ресурсом – является ли она встроенной в этот ресурс или внешней по отношению к нему и ассоциируется с ним ссылкой из ресурса;

Гранулярность субъекта аннотации – относится ли аннотация к субъекту в целом, к какому-либо его разделу или другой составной его части;

Особенность представления – аннотация относится к самому документу или к понятиям, описанным в нем либо относящимся к нему;

Повторное использование терминологии – использует ли аннотация собственную терминологию или термины из существующих онтологий и тем самым интероперабельна и понятна для других;

Тип объекта – является ли объект аннотации литеральным или текстовым, структурным или онтологическим;

Контекст – контекст аннотации: когда, кем она создана, в какой сфере, какой срок ее действительности и т. п.

Предложенная классификация аннотаций, хотя и не полна, по нашему мнению, полезна для описания не семантики аннотаций, создаваемых в той или иной электронной библиотеке, скорее, функциональных возможностей используемого в конкретной системе подхода к аннотированию и/или конкретных инструментов семантического аннотирования, а также для сопоставления функциональности различных таких подходов/инструментов.

Значимый вклад в создание технологий и инструментария интероперабельного аннотирования, основанного на формальном языке представления аннотаций, вносит деятельность Группы по открытому аннотированию (Open Annotation Group или кратко OAG), функционирующей в последние годы в рамках консорциума W3C. Эта группа разрабатывает спецификации стандарта онтологии (в терминологии группы – *модели данных*), описываемой на языке RDF, и протокола для открытого интероперабельного аннотирования цифровых документов – текстов, графических изображений, аудио, таблиц и других ресурсов, а также их фрагментов.

В настоящее время предложенные группой спецификации приобрели статус рекомендации консорциума [15–17] и рассматриваются как средство для Семантического Веба, хотя некоторые их элементы могут иметь и более широкое применение.

В спецификациях OAG предложена онтология аннотирования, формально определяющая различные виды аннотаций: комментарии, аннотации сущностей (или как теперь принято говорить, вещей), заметок, примеров, опечаток и т. п.

Таблица 1

№/№ п.п.	Мотивация	Пояснение
1.	Оценивание	Аннотация служит для оценки целевого ресурса.
2.	Установка закладки	Аннотация отмечает некоторое указанное ее автором место в тексте целевого ресурса.
3.	Классифицирование	Аннотация используется для классификации целевого ресурса.
4.	Комментирование	Аннотация представляет собой комментарий, относящийся к целевому ресурсу.
5.	Описание	Аннотация служит для описания свойств целевого ресурса.
6.	Редактирование	Аннотация указывает необходимость редактирования целевого ресурса, например, с тем чтобы устранить опечатку.
7.	Выделение маркером	Аннотация указывает намерение ее автора выделить цветом целевой ресурс или его фрагмент для того, чтобы по какой-то причине обратить на него внимание.
8.	Идентификация	Аннотация служит для придания индивидуальности целевому ресурсу путем ассоциирования с ним какого-либо уникального идентификатора, например, URI.
9.	Связывание	Аннотация определяет связь с некоторым ресурсом, имеющим отношение к целевому.
10.	Модерирование	Аннотация служит для указания ценности или качества целевого ресурса, например, для модерирования дискуссий и обсуждений.
11.	Запрашивание	Аннотация содержит вопрос о целевом ресурсе.
12.	Ответ	В аннотации приводится отклик на целевой ресурс.
13.	Создание пометы	Аннотация содержит помету для целевого ресурса.

В контексте данной статьи представляет интерес используемый в онтологии контролируемый словарь мотивов, которыми руководствуется создатель аннотаций. Этот словарь, по существу, может рассматриваться как таксономия мотивов аннотирования, позволяющая явным образом специфицировать их семантику. Классы словаря мотивов аннотирования приведены в таблице 1.

Частным случаем связей между текстовыми документами в электронной библиотеке являются связи цитирования, представляемые в виде ссылок на используемые или упоминаемые в данной публикации источники вместе с контекстами этих ссылок. Такие ссылки, как и другие связи, могут стать субъектами аннотирования наряду с текстовыми документами или их фрагментами. С позиций аннотирования целесообразно различать разные виды ссылок на использованные источники: ссылки с контекстом – цитатой из цитируемого источника, ссылки с иным контекстом и, наконец, ссылки на источники, указанные в списке литературы, но с отсутствующими на них ссылками в тексте.

Для семантического аннотирования ссылок цитирования также могут использоваться таксономии ссылок. В ряде публикаций содержатся предложения подходящих для этого таксономий. Например, в работе [8], посвященной анализу категоризации влияния цитируемых источников на цитирующие публикации, предлагается классификация ссылок цитирования в трех измерениях: *функция* (*Function*), *полярность* (*Polarity*) и *влияние* (*Impact*). Для каждого из этих измерений предложен свой набор классов. Измерению *функция* соответствуют классы, указывающие, что цитируемый источник полезен (*Useful*), отражает противоположную точку зрения (*Contrast*), обладает недостатками (*Weakness)*, вносит поправки *(Correct*), уклоняется (*Hedges*), выражает благодарность (*Acknowledge)*, подтверждение (*Corroboration)*, полемизирует *(Debate*). Для измерения *полярности* предлагаются следующие классы: позитивная (*Positive*), негативная (*Negative*) и нейтральная (*Neutral*). Наконец, для измерения *влияния* предложены такие классы: негативное (*Negative)*, незначительное (*Perfunctory*) и существенное (*Significant)*.

В работе [14] также предложена классификация ссылок цитирования. Используются иные критерии по сравнению с рассмотренными выше. Ссылки классифицируются *по месту в тексте* и ранжируются таким образом, что выше их ранг в разделе с результатами, ниже в обзоре литературы, *по количеству вхождений*, а также *по стилю.* В качестве места в тексте рассматриваются его разделы: абстракт, введение, обзор литературы, методология, результаты/обсуждение, заключение. Возможные варианты стиля: неконкретное упоминание (not specially), конкретное и интерпретирующее упоминание, прямая цитата.

В используемых в настоящее время описаниях ссылок цитирования отсутствуют атрибуты, которые бы позволили отобразить их классификацию по критериям значимости (место в тексте), интенсивности (частотности) и по стилю, предложенным в рассматриваемой статье. Чтобы их специфицировать, достаточно ввести в таксономию два контролируемых словаря:

- *Словарь мер* (или интенсивностей): высокая, средняя, низкая. Его следует использовать для характеристики значимости ссылки (в зависимости от места в тексте) и оценки частотности.
- *Словарь стилей* (характер контекста): прямая цитата, неконкретное упоминание источника, упоминание с пояснением, ссылка без контекста (для случая ссылки в списке литературы, не упоминаемой в тексте).

На основе приведенной классификации ссылок цитирования с помощью указанных контролируемых словарей могут генерироваться новые наукометрические показатели, например, следующие: *количество ссылок высокой* (а также средней/низкой) значимости на данную работу, *количество ссылок с высокой* (а также со средней/низкой интенсивностью), *количество ссылок с прямым цитированием* (а также с интерпретацией в контексте/с неконкретным контекстом/без контекста).

Необходимо упомянуть также онтологию ссылок цитирования C4O (the Citation Counting and Context Characterization Ontology) [12], представляющую собой составную часть модульного комплекса онтологий SPAR [13], некоторые элементы которых ранее уже были использованы в таксономии системы Соционет. Онтология C4O включает важные для нашей работы классы отношений между источниками из списков литературы и ссылками на них в текстах публикаций. Эти вопросы обсуждаются ниже в разд. 4.

Таксономический подход для описания семантики аннотаций используется и в системе Соционет. В этой системе поддерживается встроенная таксономия [1], используемая для классификации и тем самым для описания семантики связей между информационными объектами контента системы. Некоторые контролируемые словари, составляющие эту таксономию, используются и для семантического аннотирования. В частности, для этой цели можно использовать оценочный контролируемый словарь. Этот словарь может использоваться не только для аннотирования полного текста публикации и ее фрагментов, но также и ссылок на использованные источники в тексте публикации, а также в послестатейном списке литературы. Во всех указанных случаях, кроме последнего, аннотирование может осуществлять любой авторизованный пользователь системы, в последнем случае – только автор данной публикации. Оценочный контролируемый словарь включает, в частности, следующие классы: наилучшая, наиболее релевантная работа по обсуждаемой в ней теме;

новаторская работа (результат); интересная работа (результат); оценивается позитивно; оценивается негативно; основывается на заблуждении; возможно, является плагиатом.

Встроенная в систему Соционет таксономия может легко расширяться путем дополнения новых контролируемых словарей, позволяющих описывать новые аспекты семантики аннотаций. Обсуждается дополнение таксономии рядом новых словарей. Для аннотирования фрагментов авторефератов диссертаций и полных текстов диссертаций полезен словарь, позволяющий идентифицировать в текстах этих документов важные для их оценки оппонентами фрагменты, содержащие аргументацию соответствия диссертации требованиям ВАК. Словарь включает классы: *актуальность, новизна, достоверность, практическая ценность, теоретическая ценность.* Полезен также контролируемый словарь, позволяющий специфицировать *статус* аннотируемых фрагментов полного текста публикации: *аксиома*, *доказанное утверждение* (теорема), *цитата из используемого источника*, *фактография*, *результат исследования, постановка задачи.* Может быть также расширен оценочный словарь дополнительным включением в него следующих дополнительных классов: *актуальная тема исследования, актуальный результат, оригинальный результат, уже известный в науке результат, новый научный результат, фундаментальный результат, обоснованное утверждение, необоснованное утверждение, вода, раскавыченная цитата.*

Рассмотренные таксономии показывают, что их конкретные варианты следует использовать в соответствии с характером аннотируемых ресурсов и целями аннотатора.

3 Семантическое аннотирование в Соционет

В системе Соционет обеспечиваются возможности открытого семантического аннотирования. Важно отметить, что они реализуются с использованием тех же средств, которые уже имелись в системе для создания, поддержки и использования семантических связей между информационными объектами ее контента. Использовать возможности семантического аннотирования может зарегистрированный и авторизовавшийся пользователь, поскольку предусматривается фиксация авторства созданных аннотаций.

В Соционет поддерживаются информационные объекты – научные публикации, научные отчеты и научные произведения других видов, и семантические связи между ними [2]. Семантика связей определяется с помощью встроенной в систему таксономии, состоящей из нескольких контролируемых словарей. Эта таксономия подробно рассмотрена в работе [1] и кратко обсуждена вместе с некоторыми возможными ее расширениями в предыдущем разделе. Классы некоторых контролируемых словарей таксономии используются для описания семантики аннотаций. Это естественный подход, поскольку аннотации представляются в системе в виде семантических связей.

С точки зрения общей модели аннотаций, предложенной в [10], модель аннотаций, используемую в Соционет, можно назвать *комбинированной* – объект аннотации включает формальный и неформальный компоненты. Формальный компонент – это структурированные метаданные, указывающие один из классов подходящего контролируемого словаря встроенной в систему таксономии, определяющий семантику аннотации. Неформальный компонент, называемый в описании аннотации комментарием, – это неструктурированные метаданные, представленные в виде текста на естественном языке.

Субъектами аннотирования в Соционет могут быть полные тексты представленных в системе публикаций, фрагменты их абстрактов, а также фрагменты полных текстов. Кроме того, аннотироваться могут также и связи цитирования одних публикаций в других. Связи этого вида – это ссылки на источники из послестатейного списка литературы, а также сами библиографические описания использованных источников в этих списках.

Наряду со связями цитирования, выделяемыми пользователем-аннотатором в «ручном режиме», субъекты аннотирования такого рода могут порождаться в автоматическом режиме средствами анализа полных текстов публикаций, представленных в контенте системы в pdf-формате. Эта техника и ее возможности обсуждаются в следующем разделе.

Соционет является мультипользовательской системой, и поэтому для одного субъекта аннотирования может быть создано несколько аннотаций одним или разными пользователями системы. Аннотации представляются в Соционет в виде классифицированных связей «персона – субъект», и их описания включают идентификацию персоны-автора аннотации, идентификацию субъекта аннотации, класс выбранного аннотатором контролируемого словаря таксономии, а также текстовый комментарий.

Функциональные возможности системы Соционет позволяют использовать ее как платформу для виртуальной коммуникационной среды научного сообщества пользователей системы [9]. Эти возможности основаны на реакциях авторов публикаций, представленных в системе, на появлении семантических связей этих публикаций с публикациями других авторов либо оценочных связей, касающихся этих публикаций. Такая реакция состоит в создании новой связи профиля ее автора со связью, на появление которой он реагирует. Поскольку аннотации представляются в виде семантических связей, указанные возможности могут быть применены и к ним. Поэтому, хотя такая возможность пока еще не полностью реализована в

Соционет, создание аннотаций потенциально может быть вовлечено в возникающие в такой среде процессы коммуникаций, отображающие дискуссии относительно создаваемых аннотаций.

Формальные компоненты объектов аннотаций – структурированные метаданные – могут использоваться в критериях поиска аннотаций, интересующих пользователя классов, а также для генерации ряда новых наукометрических показателей наряду с другими, формируемыми сервисами системы. Для возможности генерации новых наукометрических показателей в описания создаваемых связей должны быть перенесены классификационные атрибуты цитирования. Должны быть также созданы в Соционет соответствующие сервисы, которые будут генерировать и показывать полученные показатели на странице метаданных (описателя) публикации, как это реализовано сегодня для других показателей в системе. Этими новыми показателями могут быть, например, следующие: *количество ссылок высокой* (а также средней/низкой) значимости на данную работу, *количество ссылок с высокой* (а также со средней/низкой интенсивностью), *количество ссылок с прямым цитированием* (а также с интерпретацией в контексте/с неконкретным контекстом/без контекста).

4 Генерация описаний ссылок цитирования и их визуализация в Соционет

Интересные перспективы для развития семантического аннотирования в системах, подобных Соционет, открывают новые подходы и технологии, создаваемые для поддержки анализа содержания цитирований. Общая концепция анализа содержания цитирований представлена в [14]. Описание создаваемых технологий, применение которых обсуждается в данной статье ниже, доступно в [4]. Основная новизна этих подходов связана с извлечением из научных публикаций более широкого по сравнению с традиционным подходом набора данных, связанных со ссылками цитирования, включая окружающий их контекст. Кроме того, создаются новые возможности визуализации этих данных, которые позволяют накладывать результаты анализа содержания цитирований поверх текста pdf-документов в виде программным образом генерируемых аннотаций.

Проект CitEcCyr (https://github.com/citeccyr), реализуемый с участием одного из авторов данной статьи в Российской академии народного хозяйства и государственной службы при Президенте РФ (РАНХиГС) с 2016 г., предусматривает разработку средств извлечения из русскоязычных научных публикаций, доступных в виде pdf-документов, расширенного набора сведений о цитированиях. На основе этих данных предполагается разработка новых наукометрических показателей, включая некоторые дополнительные данные о научной результативности. Предполагается учитывать количество ссылок в тексте публикации на источники из списка литературы, отделять источники без ссылок на них в тексте публикации. Кроме того, имеется в виду обрабатывать контекст вокруг ссылок для классификации содержания цитирований источников, а также ранжировать ссылки на источники по месту их в структуре статьи (например, ранг выше в разделе с результатами, ранг ниже в разделе обзор литературы) и др.

Источником публикаций для обработки средствами проекта является система Соционет. Первые результаты извлечения данных о цитированиях, полученные на основе публикаций архива НЭИКОН (https://socionet.ru/collection.xml?h=spz:neicon), доступны для ознакомления и тестирования по адресу http://no-xml.socionet.ru/~cyrcitec/citmap/spz/neicon/.

Средствами обсуждаемого проекта создаются новые данные о цитированиях. Рассмотрим их особенности, а также их визуализацию в Соционет на примере одной из научных публикаций гуманитарного профиля, доступной в виде pdf-документа по адресу http://nevolin.socionet.ru/files/2014_Nevolin_ rfbr.pdf.

На Рис. 1 приведен фрагмент этой публикации, в котором на экране компьютера ссылки на источники из списка литературы выделяются желтым цветом. К этим выделенным фрагментам текста публикации программным образом созданы аннотации. Кликая на выделенные цветом ссылки, пользователь получает различную дополнительную информацию.

Чтобы это стало возможным, создаваемые с помощью программного обеспечения проекта CitEcCyr данные о цитированиях преобразуются в соответствии с моделью данных веб аннотаций [15]. Затем эти данные интегрируются в среду системы Соционет в виде семантических связей, что является обычным для представления аннотаций в системе. Рассмотрим, какова общая схема получения этих данных о цитированиях и как они в данном случае используются.

На первом этапе выполняется конвертация бинарных pdf-документов в текстовый вид, который допускает анализ и извлечение необходимых данных о цитированиях. В проекте CitEcCyr разработана программа конвертации PDF-STREAM (https:// github.com/citeccyr/pdf-stream-cli), которая преобразует содержание pdf-документов в формат JSON. Пример данных, получаемых для указанной выше публикации, доступен по адресу http://no-xml.socionet.ru/citmap/convertedPDF/2014_Nevolin_ rfbr.json.

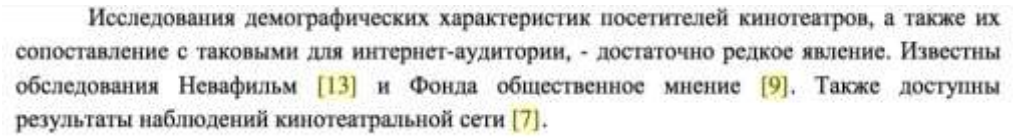
Исследования демографических характеристик посетителей кинотеатров, а также их сопоставление с таковыми для интернет-аудитории, - достаточно редкое явление. Известны обследования Невафильм [13] и Фонда общественное мнение [9]. Также доступны результаты наблюдений кинотеатральной сети [7].

Рисунок 1 Пример фрагмента публикации с выделенными аннотированными ссылками на цитируемые источники, одновременно служащими указателями на аннотации

На следующем этапе работает программа, также созданная в проекте CitEcCyr, которая создает XML-записи, содержащие, в том числе, сведения о ссылках на источники из списка литературы публикации. Для упомянутой выше публикации на основе ее JSON-версии ниже приведен пример XML-записи, содержащей извлеченные данные для двух ссылок (в тегах <intextref>) на один и тот же источник, который имеет в списке литературы порядковый номер 7.

```
<intextref>
	<Reference>7</Reference>
	<Exact>[7]</Exact>
	<Start>6125</Start>
	<End>6128</End>
	<Prefix>сом –имеются данные, что реклама и продажи в баре составляют, соответственно,
20-25% и 20-30% выручки кинотеатров</Prefix>
	<Suffix>.Итак, характеристики аудитории представляют коммерческую ценность для отрасли
и научный интерес для исследователей, н</Suffix>
</intextref>

<intextref>
	<Reference>7</Reference>
	<Exact>[7]</Exact>
	<Start>10119</Start>
	<End>10122</End>
	<Prefix>обследования Невафильм[13] и Фонда общественное мнение[9].
Также доступны результаты наблюдений кинотеатральной сети</Prefix>
	<Suffix>. Согласно данным Невафильм (см. Таблицу 2), профили аудиторий-посетителей
кинотеатров и интернет-пользователей, -з</Suffix>
</intextref>
```

Эти данные включают:

- номер источника в списке литературы, тег <Reference>, в примерах выше он содержит номер 7;
- вид ссылки на соответствующий источник, тег <Exact>, в примерах это - [7];
- текстовые координаты ссылки в тегах <Start> и <End>, которые содержат порядковые номера от начала текста документа первого и последнего символа строчки, содержащейся в теге <Exact>;
- контекст вокруг ссылки в тегах <Prefix> и <Suffix>, который в данном случае включает по 200 символов слева и справа от ссылки, содержащейся в теге <Exact>.

В частности, второй блок данных в теге <intextref> из приведенной выше XML-записи отображен на Рис. 1 как аннотация к ссылке на источник номер 7.

Кроме этого, из JSON-версии pdf-документов извлекаются данные о содержании списка литературы публикаций, которые иллюстрируются следующей XML-записью:

```
<reference>
	num="7"
	start="20952"
	end="21140"
	author="Гладких Михайлина"
	title="Кронверк Синема сколько стоит билет в кино"
	year="2011"
	handle="spz:cyberleninka:33099:16516633">
	<from_pdf>Гладких И.В., Михайлина А.П. «Кронверк Синема»: сколько стоит билет в кино?
(учебный кейс) / Вестник Санкт Петербургского университета. Серия 8: Менеджмент. 2011. No3.
с.145 159.</from_pdf>
</reference>
```

Эти данные содержат:

- атрибут num – номер источника в списке литературы, в приведенном выше примере он - номер 7;
- атрибуты start и end – текстовые координаты данных источника, которые содержат порядковые номера от начала текста документа первого и последнего символа строчки, содержащейся в теге <from_pdf>;
- атрибуты author, title и year, выделенные из данных тега <from_pdf> и используемые для поиска в Соционет публикации, которая указана в данных этого источника;

- атрибут handle – содержит уникальный код публикации, соответствующей данным этого источника, если она есть в Соционет;
- тег <from_pdf> – содержит «сырые» данные источника, которые извлечены из JSON-версии публикации.

Полный набор данных о содержании цитирований для научной статьи, к которому относятся приведенные выше примеры, доступен по адресу http://no-xml.socionet.ru/citmap/outputs/repec:rus:pgfhxz:wp9.xml.

Описанные выше данные о содержании цитирований допускают различные варианты их использования в научных информационных системах, подобных Соционет. Поскольку данные о ссылках на цитируемые источники включают их текстовые координаты, то возможно программное создание аннотаций, которые при просмотре соответствующих публикаций в Соционет выглядят визуально привязанными к тексту ссылок на источники. На Рис. 2 приведен пример визуализации данных о цитированиях в виде аннотаций ссылок на цитируемые источники, выделяемых цветом на экране компьютера. В частности, для ссылки на 7-й источник раскрыта ее аннотация (справа), которая в текущей версии содержит данные о соответствующем источнике и, если есть, ссылку на него в Соционет, а также статистику о количестве цитирований данного источника в этой публикации.

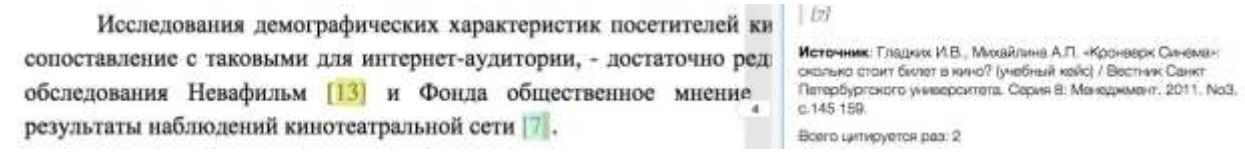

Рисунок 2 Пример программно-сгенерированной аннотации для ссылки на источник

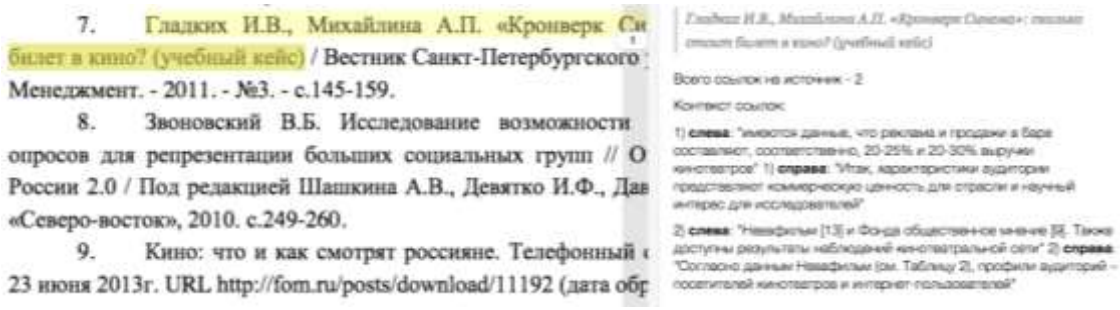

Рисунок 3 Пример программно сгенерированной аннотации для библиографического описания источника

В будущем планируется также приводить статистику обо всех цитированиях данного источника в контенте Соционет.

Похожим образом могут быть построены аннотации поверх данных о библиографических описаниях источников в списке литературы публикации. На Рис. 3 приведен пример аннотации в списке литературы, которая «наложена» поверх публикации с номером 7. Справа в текстовом блоке видно содержание этой аннотации.

Аннотация на Рис. 3 содержит сведения об общем количестве упоминаний (цитирований) данного источника в тексте публикации, а также контекст (по 200 символов справа и слева) для каждого такого случая.

Рассмотренные технологии аннотирования позволяют в нужных местах текста публикаций компактно предоставлять пользователям Соционет различную дополнительную информацию. Эта информация, как это представлено выше, может содержать обобщенные данные о содержании цитирований. Уже имеющиеся в Соционет сервисы для авторов публикаций для «ручного» семантического «раскрашивания» связей, в данном случае, позволяют им как уточнять программно-сгенерированную семантику аннотаций, так и добавлять к аннотациям новые семантические атрибуты.

5 Заключение

Современные научные информационные системы, к числу которых относится и система Соционет, начинают предлагать своим пользователям различные возможности для семантического аннотирования контента. Сравнительно новыми возможностями является доступное в Соционет «ручное» аннотирование полных текстов научных статей, представленных в виде pdf-документов, и их фрагментов. В дополнение к этому в Соционет разрабатываются средства программной генерации аннотаций для ссылок цитирования, которые являются важным элементом научных публикаций и академической культуры. Данный подход позволяет через аннотации, привязанные к определенным фрагментам pdf-документов, показать читателю разнообразную наукометрическую информацию, включая сводные сведения о том, сколько раз цитируются источники из списка литературы в данной публикации, а также и во всех других публикациях, имеющихся в системе Соционет.

Благодарности

Реализация методов аннотирования в системе Соционет выполнена в рамках работ по гранту РФФИ, проект 15-07-01294. Разработка подхода для извлечения данных о содержании цитирований, в том числе, для целей суперкомпьютерного моделирования взаимодействий между агентами и со средой научного сообщества, были получены С.И. Париновым в рамках работ по гранту РНФ, проект 14-18-01968.

Литература

[1] Когаловский, М.Р., Паринов, С.И.: Таксономия семантических связей информационных объектов контента научной электронной библиотеки. НТИ. Серия 2. Информационные процессы и системы, 9, сс. 15-23 (2015)

[2] Паринов, С.И., Когаловский, М.Р.: Технология семантического структурирования контента научных электронных библиотек. RCDL 2011, pp. 197-206 (2011)

[3] Annotation. Wikipedia. https://en.wikipedia.org/wiki/Annotation

[4] Barrueco, J.M., Krichel, T., Parinov, S., Lypunov,V., Medvedeva, O., Sergeeva, V.: Towards Open Data for Citation Content Analysis. Submitted to DAMDID/RCDL-2017

[5] Bennet, P.N., Gabrilovich, E., Kamps, J., Karlgren, J.: Sixth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'13). CICM'13, pp. 2543-2544 (2013)

[6] DBpedia. Википедия. https://ru.wikipedia.org/wiki/DBpedia

[7] Gagnon, M., Zouaq, A., Jean-Louis, L.: Can we use Linked Data Semantic Annotators for the Extraction of Domain-Relevant Expression. WWW 2013 Companion, pp. 1249-1246 (2013)

[8] Hernández-Alvarez, M., Gómez Soriano, J.M., Martinez-Barco, P.: Citation Function, Polarity and Influence Classification. doi: 10.1017/S1351324916000346 (2017)

[9] Kogalovsky, M.R., Parinov, S.I.: Scholarly Communications in a Semantically Enrichable Research Information System with Embedded Taxonomy of Scientific Relationships. In: Klinov, P. and Mouromtsev, D. (eds.): Knowledge Engineering and Semantic Web. 6th Int. Conf. KESW 2015. The Communications in Computer and Information Science series, 518. Springer, pp. 87-101 (2015)

[10] Oren, E., Hinnerk Moller, K., Scerri, S., Handschuh, S., Sintek, M. What are Semantic Annotations? (2006) http://www.siegfried-handschuh.net/ pub/2006/whatissemannot2006.pdf

[11] Parinov, S., Lyapunov, V., Puzyrev, R., Kogalovsky, M.: Semantically Enrichable Research Information System SocioNet. In: Klinov, P. and Mouromtsev, D. (eds.): Knowledge Engineering and Semantic Web. 6th Int. Conf. KESW 2015. The Communications in Computer and Information Science series, 518. Springer, pp. 147-157 (2015)

[12] Shotton, D.: C40, the Citation Counting and Context Characterization Ontology. Version 1.1.1, 11/05/2013. http://purl.org/spar/c4o

[13] SPAR Ontologies. Describing Publishing Domain. http://purl.org/spar/

[14] Zhang, G., Ding, Y., Milojević, S.: Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content. arXiv:1211.6321 (2012)

[15] Web Annotation Data Model. W3C Recommendation 23 February 2017. https://www.w3.org/TR/2017/REC-annotation-model-20170223/

[16] Web Annotation Protocol. W3C Recommendation 23 February 2017. http://www.w3.org/TR/annotation-protocol/

[17] Web Annotation Vocabulary. W3C Recommendation 23 February 2017. http://www.w3.org/TR/annotation-vocab