К ВОПРОСУ СОЗДАНИЯ СИСТЕМЫ ПОДДЕРЖКИ РАБОТЫ С НАУЧНЫМИ ПУБЛИКАЦИЯМИ

В.В. Костин

Вычислительный Центр им. А.А. Дородницына РАН
e-mail: kosvic11@mail.ru

Аннотация

В статье рассматриваются требования к системе для работы с научными трудами. Приводится обзор существующих семантических моделей, описывающих научные труды и научно-исследовательский процесс, а также моделей, которые могут использоваться для данных целей. Выделены онтологии, сущности которых, на взгляд автора, можно использовать для реализации различных аспектов описанной системы.

1. Электронные семантические библиотеки

В настоящее время современные поисковые средства предоставляют возможности производить быстрый и содержательный поиск по большим объёмам данных. И, несмотря на то, что поисковые системы не могут производить поиск по бумажным носителям, поиск стал весьма эффективным, потому что достаточно большое количество работ уже конвертировано в электронную форму. При этом поиск обычно проводится преимущественно по каким-либо ключевым словам. Семантическая же составляющая документов остаётся доступной преимущественно только для человека. Для увеличения доступности семантической информации и машинам в последнеие 10-15 лет активно разрабатываются семантические технологии, специальные языки для описания семантики RDF, RDFS, OQL, средства запросов к семантическим данным SPARQL, создаются проекты для обмена семантическими данными, как Linked Open Data.

В качестве результата этой деятельности появляется такое понятие, как электронные библиотеки. Электронные библиотеки представляют собой специализированные информационные системы, которые выполняют управление коллекциями электронных ресурсов (например, таких как текстовые документы, изображения, мультимедиа файлы) с целью повышения эффективности использования содержащихся в них знаний некоторыми сообществами пользователей. Под семантическими электронными библиотеками (СЭБ) понимаются электронные библиотеки, использующие семантические технологии для организации всех процессов своей работы, таких как описание ресурсов, ведение каталогов, описание профилей пользователей, поиск и рекомендация ресурсов пользователям и т. п. [1] Таким образом, в электронных библиотеках хранится информация о работах в виде метаданных, позволяющих осуществлять различные операции над трудами, такие как анализ близости, кластеризация текстов.

В связи с пополнением электронных библиотек особо актуальными вопросами становятся выделение сущностей в тексте с одной стороны и формирование и валидация связей между ними с другой.

На сегодняшний день существует ряд проектов, реализующих электронные библиотеки. В данной статье рассматриваются различные методологии, которые можно использовать при создании электронной библиотеки.

2. Система работы с научными трудами

В ходе работы с научными трудами была обнаружена потребность в системе, облегчающей данную работу. В ходе образовательной работы вместе с коллегами были выработаны и формализованы требования к системе.

В желаемой версии системы труд должен представлятоь собой текст, в котором выделены именованные сущности, он разбит на логические части, такие как («аннотация», «вступление», «основная часть», «заключение» и т.п.). У труда в системе должны определяться ряд связей и параметров, которые будут описывать как некоторые фактологические моменты (авторство, место и время публикации, формальные параметры наподобие ISBN), так и семантические моменты (корреляция с той или иной областью науки, научная значимость работы – при помощи таких параметров как индекс цитирования). В числе семантических параметров научного труда должны быть связи, определяющие степень соответствия труда различным научным разделам, определяющие отношения между трудами (близости, связанности и т.п.). Данные соотношения должны как определяться системой автоматически в качестве результатов семантрического анализа, так и задаваться пользователями.

Ключевым требованием к системе определена возможность удобной работы с научным текстом. При работе в системе пользователь должен иметь возможность просматривать труды, выделять из них наиболее интересные, добавлять их в свою электронную библиотеку. Также у пользователя должна быть возможность выделять имеющие для него интерес фрагменты в текстов. На основе этих данных формируется онтология его интересов. У пользователя должны быть возможность изменять собственную онтологию интересов. Для этого представляется необходимым возможность удобно перемещаться по ссылкам, которые привязаны к именованным сущностям текста (по ссылкам из списка литературы; по ключевым терминам работы; по определённым ранее терминам в работе; по терминам, описанным в других работах). Важной представляется возможность просматривать близкие работы, которые могут быть определены схожестью наборов ключевых терминов, в качестве результата семантического анализа, либо заданные и верифицированные сообществом пользователей. При семантическом анализе близость текстов должна оценивается с учётом онтологии интересов пользователя.

Система должна предоставлять удобный семантический поиск. При этом результаты поиска должны предоставляться после выполнения ряда итераций. В качестве первого шага – полнотекстовый или атрибутный поиск. На основе поулченных результатов поиска производится семантический анализ, формируется семантическая составляющая запроса в качестве весовых значений, определённых параметров. В качестве третьей итерации представляется внесение изменений в полученный семантический объект на основе онтологии интересов пользователя. Полученный результат применяется для произведения семантического поискового запроса, результаты которого и выводятся пользователю. Также важна возможность и другого поиска – на основе выбранных научных трудов, при котором

семантическая составляющая поискового запроса будет выделяться из выбранных пользователем трудов.

Важной частью системы представляется организация обмена информации между пользователями. Необходимо предоставить пользователями возможность обмениваться между собой текстовыми сообщениями, ссылками на элементы системы (труды, отзывы, рецензии, выделенные отрывки трудов, отдельные именованные сущности). Актуальной представляется реализация возможности рецензирования трудов и возможности оценки рецензий сообществом пользователей.

Полезной видится возможность вносить в систему свои ещё не опубликованные труды, работать с черновиками статьи, вести систему контроля версий статьи, реализовать возможность пользователю предоставлять доступ к своему черновику другим пользователям для оценки и рецензирования работы на различных этапах её формаирования.

3. Сравнение классов онтологий, описывающих научные труды и научно-исследовательскую деятельность

Для формирования онтологии для реализации описанной в предыдущей части системы были использованы результаты работы [2]. В ней были проведен анализ семантических моделей, которые либо описывают научные труды и научно-исследовательскую деятельность, либо напрямую не описывающие научные публикации или научную деятельность. Из итоговой таблицы были исключены онтологии Dublin Core, PRISM, CIDOC CRM и SWAN, потому что они напрямую не описывают научные труды или научно-исследовательскую деятельность. Также в сравнении классов не участвовали онтологии PSO (из-за отсутствия классов кроме Thing), PWO и C4O (из-за того, что они описывают сильно отличающиеся от остальных онтологий области).

Таким образом, в итоговую сравнительную таблицу попали онтологии FRBR [3], CERIF [4], SKOS [5], BIBO [6], PROV-O [7], а также онтологии FaBiO, CiTo, BiRo, PRO семейства SPAR [8]. Ниже приведены сравнительные таблицы сущностей онтологий. (Таб. 1)

Понятия	FRBR	CERIF	SKOS	SPAR				BIBO	PROV-O
				FaBiO	CiTO	BiRO	PRO		
Научная работа	+								+
Работа	+								
Текст	+								
Понятие	+		+	SKOS					
Начинание	+					FRBR			
Событие	+	+						++	
Образ	+								
Данное	+								

Изображение	+							FOAF	
Динамическое изображение	+								
Объект	+								
Классический труд	+	+		FRBR		FRBR			
Юридический труд	+			FRBR		FRBR			
Литературный труд	+			FRBR		FRBR			
Представление	+			FRBR					
Выражение	+			FRBR		FRBR			
Коллекция			+			+			
Схема			+	SKOS					
Упорядоченная коллекция			+						
Ситуация					+		+		
Субъект				+	+	+			
Библиотечный список						+			
Библиотечная ссылка						+			
Библиографичес-кая запись						+			
Библиографичес-кая коллекция						+			
Документ							FOAF	FOAF++	
Проект		+					FOAF		
Продукт		+							
Патент		+							
Измерение		+							
Содержание	+								+
Юридическое лицо	+							FOAF	
Персона	+	+						FOAF	

Агент							FOAF	FOAF	
Место	+								
Ответственность				+		+			
Цитирование					++				
Учётная запись							FOAF		
Роль							+		
Временной интервал							+		
Описание временного интервала							+		
Среда		+							
Финансирование		+							
Адрес		+							
Награда		+							
Организационная единица		+							
Действие									+

Таб. 1. Сравнение классов онтологий, описывающих научные труды и научно-исследовательскую деятельность.

В данной таблице указано, какие сущности в какой онтологии присутствуют. В каких-то онтологиях несколько сущностей относятся к одному классу – в этом случае ячейки соответствующих сущностей объединены. Какие-то онтологии заимствуют классы из других онтологий – в этом случае в ячейке расположение наименование той онтологии, из которой они заимствованы (например, класс «агент» заимствован онтологиями PRO и BIBO из онтологии FOAF). Двойным плюсом отмечается случай, когда в онтологии данному понятию соответствует несколько классов.

На основе представленной сравнительной таблицы можно сделать ряд выводов:

1. Для описания и классификации литературных трудов стоит использовать онтологию FRBR, в случае необходимости более подробного описания – онтологию BiRo.
2. Для описания аспектов, связанных с персонами, связями между ними, организации системы сообщений наподобие социальной сети наиболее подходящей видится онтология FOAF.
3. Для описания финансовой и формальной стороны научно-исследовательской деятельности наиболее целесообразным видится использование онтологии CERIF.
4. Для описания цитирования наиболее актуальным видится использование онтологии CiTo.

4. Заключение

В статье приведены требования для системы работы с научными публикациями. Также проанализирован обзор онтологий, описывающих научные труды и научно-исследовательскую деятельность, которые можно использовать для описания описанной модели.

Литература

[1] Хоай, Ле, А. Ф. Тузовский. "Семантическое аннотирование документов в электронных библиотеках." Известия Томского политехнического университета 322.5 (2013).

[2] В. В. Костин, "Обзор семантических моделей, описывающих научные публикации и научно-исследовательскую деятельность." Электронные библиотеки: перспективные методы и технологии, электронные коллекции. 13-16 октября(2014).

[3] Functional Requirements for Bibliographic Records, Final Report / IFLA Study Group on the Functional Requirements for Bibliographic Records. – München: K.G. Saur, 1998. (UBCIM Publications, New Series; v. 19)

[4] EuroCRIS | Research Information | CERIF http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1

[5] Miles, Alistair, et al. "SKOS core: simple knowledge organisation for the web." International Conference on Dublin Core and Metadata Applications. 2005.

[6] Bibliographic Ontology http://bibliontology.com/

[7] Lebo, Timothy, et al. "Prov-o: The prov ontology." W3C Recommendation, 30th April (2013).

[8] SPAR – Semantic Publishing and Referencing. http://sempublishing.sourceforge.net/