УДК 004.891

В. В. Костин

Вычислительный центр им. А. А. Дородницына РАН
ул. Вавилова, 40, Москва, 119333, Россия

kosvic11@mail.ru

К ВОПРОСУ СОЗДАНИЯ СИСТЕМЫ ПОДДЕРЖКИ РАБОТЫ С НАУЧНЫМИ ПУБЛИКАЦИЯМИ

Рассматриваются требования к системе для работы с научными трудами. Описываются все аспекты такой системы с учетом использования современных семантических технологий. Приводится обзор существующих семантических моделей, описывающих научные труды и научно-исследовательский процесс, а также моделей, которые могут использоваться для данных целей.

Ключевые слова: семантические библиотеки, научные публикации, OWL онтологии, интеллектуальный анализ текста, семантический поиск, Linked Open Data.

Электронные семантические библиотеки

В настоящее время современные поисковые средства предоставляют возможности производить быстрый и содержательный поиск по большим объемам 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 остается 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, текстовых документов, изображений, мультимедиафайлов) 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 – 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.

Костин В. В. К вопросу создания поддержки работы с научными публикациями // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 4. С. 32–37.

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

Система работы с научными трудами

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

В желаемой версии системы труд разбит на логические части, такие как «аннотация», «вступление», «основная часть», «заключение» и т. п., в котором выделены именованные сущности, термины и структурные объекты наподобие формул. У труда в системе должен определяться ряд параметров – как формальные (авторство, место и время публикации, формальные параметры наподобие ISBN), так и семантические (использованные в научной работе термины, связи работы с другими работами в системе). Семантическая составляющая научного труда должна определяться в качестве характеристического вектора, который формируется на основе входящих в работу терминов, ссылок на другие работы и корректируется сообществом. Между трудами устанавливаются связи, указывающие на семантическую близость, принадлежность к одному циклу научных работ, следование одной работы из другой. Семантическая близость должна определяться в качестве значения функции, аргументами которой являются характеристические векторы, которое корректируется на основе связей, существующих между научными трудами.

Ключевым требованием к системе определена возможность эффективной работы с научным текстом. При работе в системе пользователь должен иметь возможность просматривать, изучать и анализировать научные работы, рекомендованные другими пользователями, привлекшие внимание или определенные системой как семантически близкие к рассматриваемому. Кроме того, у пользователя должна быть возможность формировать из них или их фрагментов «собственные» структуру и наполнение электронной библиотеки – индивидуальную книготеку. В дополнение к описанным функциям у пользователя должна быть возможность выделять и комментировать имеющие для него интерес фрагменты текстов. В зависимости от задачи пользователь должен иметь возможность задавать различные комментарии и отмечать различные тексты. Каждая из таких аннотаций будет являться «инфослоем» (информационным слоем) – своего рода проекцией для отдельной задачи. На основе описанных выше данных формируется онтология интересов пользователя. У пользователя должна быть возможность изменять собственную онтологию интересов, добавлять или убирать семантические связи, редактировать предпочитаемую научную область. Для улучшения эффективности работы с научными трудами необходима возможность удобного перемещения между работами – по ссылкам списка литературы; по ключевым словам, выражениям и тегам; по терминам и синтаксическим конструкциям, приведенным в научной работе.

Система должна предоставлять удобный адаптивный семантический поиск. При этом результаты поиска должны выдаваться после выполнения ряда итераций. В качестве первого шага – полнотекстовый или атрибутный поиск. На второй итерации производится семантический анализ результатов поиска (или набора заданных пользователем научных работ), формируется семантическая составляющая запроса, его характеристический вектор. Во время третьей итерации представляется внесение изменений в полученный семантический объект на основе онтологии интересов пользователя. Полученный результат применяется для произведения финальной итерации – 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 должна стать организация сообщества системы, обмена информации между пользователями. Необходимо предоставить пользователям возможность обмениваться между собой или в группах текстовыми сообщениями, комментариями, которые могут содержать стилевое или шрифтовое выделение, ссылками на объекты системы (труды,

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

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						+			

Окончание таблицы

Понятия	FRBR	CERIF	SKOS	SPAR				BIBO	PROV-O
				FaBiO	CiTO	BiRO	PRO		
Библиографическая запись						+			
Библиографическая коллекция						+			
Документ							FOAF	FOAF++	
Проект							FOAF		
Продукт									
Патент									
Измерение									
Содержание									
Юридическое лицо								FOAF	+
Персона								FOAF	
Агент							FOAF	FOAF	
Место									
Ответственность				+		+			
Цитирование					++				
Учетная запись							FOAF		
Роль							+		
Временной интервал							+		
Описание временного интервала							+		
Среда		+							
Финансирование		+							
Адрес		+							
Награда		+							
Организационная единица		+							
Действие									+

отзывы и рецензии, выделенные отрывки трудов, инфослои или выделенные заранее группы работ, отдельные термины и их конкретное использование в тексте).

Актуальной представляется реализация возможности рецензирования трудов, оценки рецензий сообществом пользователей, ведение совместного творчества, возможность вносить в систему свои еще не опубликованные труды. Это позволит вести публичное обсуждение, вырабатывать совместные идеи. Добавление еще не опубликованных трудов позволит выносить его на труд ограниченного пользователем круга лиц, работать с черновиками статьи, вести систему контроля версий статьи, автоматически размечать статью метаданными системы.

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 был 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 [1], SKOS [4], BIBO [2], PROV-O [5], а также онтологии FaBiO, CiTo, BiRo, PRO семейства SPAR [3]. Сущности онтологий сравниваются в 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 указывается 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 следует 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 подходит онтология FOAF.
3. Для описания финансовой и формальной стороны научно-исследовательской деятельности наиболее целесообразно использование онтологии CERIF.
4. Для описания цитирования наиболее актуально использование онтологии CiTo.

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 рассматриваемой модели.

Список литературы

1. *Хоай Л.*, *Тузовский А. Ф.* Семантическое аннотирование документов в электронных библиотеках // Изв. Том. политехн. ун-та. 2013. Т. 322, №. 5.

[1] EuroCRIS | Research Information | CERIF URL: http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1.
[2] Bibliographic Ontology. URL: http://bibliontology.com.
[3] SPAR – Semantic Publishing and Referencing. URL: http://sempublishing.sourceforge.net.

2. *Костин В. В.* Обзор семантических моделей, описывающих научные публикации и научно-исследовательскую деятельность // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. 2014.

3. Functional Requirements for Bibliographic Records, Final Report // IFLA Study Group on the Functional Requirements for Bibliographic Records. München: K.G. Saur, 1998.

4. *Miles A. et al.* SKOS core: simple knowledge organisation for the web // International Conference on Dublin Core and Metadata Applications. 2005. P. 3–10.

5. *Lebo T. et al.* Prov-o: The prov ontology // W3C Recommendation. 2013.

Материал поступил в редколлегию 15.12.2014

V. V. Kostin

Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS
40 Vavilov Str., Moscow, 119333, Russia Federation

kosvic11@mail.ru

ON A QUESTION OF CREATION OF A SUPPORT SYSTEM OF WORKING WITH ACADEMIC PAPERS

The article describes the demands for the support system that assists in working with scientific papers. All of these aspects are reviewed, with reference to nowadays semantic technologies. It also reviews the existing models that describe scientific papers and the process of scientific research and OWL ontologies that can be used for such purposes.

Keywords: semantic libraries, scientific papers, OWL ontologies, text mining, semantic search, Linked Open Data.

References

1. Hoaj L., Tuzovskij A. F. Semanticheskoe annotirovanie dokumentov v jelektronnyh bibliotekah. *Izv. Tom. politeh. un-ta*, 2013, vol. 322, no. 5.

2. Kostin V. V. Obzor semanticheskih modelej, opisyvajushhih nauchnye publikacii i nauchno-issledovatel'skuju dejatel'nost'. *Elektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii*. 2014.

3. Functional Requirements for Bibliographic Records, Final Report. *IFLA Study Group on the Functional Requirements for Bibliographic Records*. München, K. G. Saur, 1998.

4. *Miles A. et al.* SKOS core: simple knowledge organisation for the web. *International Conference on Dublin Core and Metadata Applications*, 2005, p. 3–10.

5. Lebo T. et al. Prov-o: The prov ontology. *W3C Recommendation*, 2013.