Обзор семантических моделей, описывающих научные публикации и научно-исследовательскую деятельность

© В.В. Костин
Федеральное государственное бюджетное учреждение науки
Вычислительный центр им. А. А. Дородницына Российской академии наук,
Москва
kosvic11@mail.ru

Аннотация

Данная статья содержит обзор моделей, описывающих научные публикации, научно-исследовательскую работу, а также онтологии, напрямую к ним не относящиеся, но содержащие элементы, которые можно использовать для их описания, и проводит сравнение между их сущностями.

1 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, 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. [18] 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 Онтологии

На сегодняшний день имеется ряд онтологий, описывающих научные труды частично или польностью (CERIF [4], SPAR [13], SKOS [11,12], FRBR [5], BIBO [1], PROV-O [10], ЕНИП [16]). Кроме того, существуют онтологии, напрямую с научными публикациями не связанные, которые, однако, можно использовать при описании научных публикаций (Dublin Core[3], PRISM [6], CIDOC CRM [2], JATS [8], SWAN [14]). Кроме того, для работы с научными трудами на основе вышеперечисленных онтологий создана онтология Соционет [17], учитывающая российскую специфику научных исследований. Далее будут рассмотрены каждая из этих инициатив и их применение.

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

2.1 Dublin Core

Онтология Dublin Core описывает различные аспекты, использующиеся в научной работе. В терминах онтологии описываются такие объекты, как Агент – нечто, что совершает или может соевршать действия (персона, организация, программный агент), Группа агентов (студенты, женщины), Библиографический ресурс, Формат файла, Частота, Подведомственность, Лицензия документа, Лингвистическая система, Место, Метод поступления (метод, которым элемент добавлен в коллекцию), Метод Обучения (метод, которым получены навыки, знания), Временной период, Физический способ (бумага, диск), Происхождение-авторство-версия, Физический ресурс (нечто физицеское), Политика (плна или курс, определяющий принятые решения власти), Происхождение, Права, Стандарт и несколько надклассов.

2.2 SWAN

SWAN (Semantic Web Applications in Neuroscience) – онтология сетевых приложений в нейромедицине. Схема главного модуля данной онтология представлена на рис. 1:

• Collections: определяет неупорядоченные и упорядоченные коллекции. Пример использования – список авторов;

• provenance, authoring and versioning (PAV): позволяет задать всю информацию по происхождению, авторству и версиям элементов, которая не описывается Dublin Core и Dublin Core Terms;

• discourse relationships: набор связей, которые используются для построения дискурса и, в частности, научного дискурса. Области определения и значения не заданы для возможности использования в других проектах;

• FOAF: онтология FOAF (Friend Of A Friend) описывает отношения и связи между людьми (написана на языке OWL-DL). Используется для описания персон, их отношений с другими персонами и организациями, группами и т.п.;

• agents: промежуточные средства, которые используются для интеграции FOAF в онтологию SWAN;

• SKOS: внешняя онтология для описания тезаурусов (описана ниже).

• qualifiers: используется для аннотирования сущностей на основе словарей SKOS;

• scientific discourse: вводит классы для описания научного дискурса;

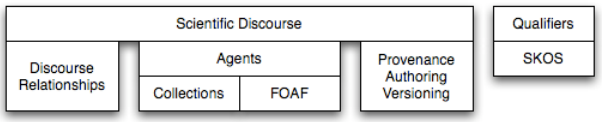

Рис. 1. Концептуальная схема онтологии SWAN

2.3 FRBR

Работа над данной онтологией по разным источникам была начата в 1992–1995 гг. В 1998 г. была выпущена первая спецификация.

Онтология создана, чтобы включить в себя все библиографические сущности и быть независимой от любых кодов катологизации.

Данная онтология может описывать объекты текстовых, музыкальных, картографических, аудивизуальных, графических и трехмерных материалов; она покрывает весь спектр физических носителей, описанных в библиографических записях – бумажные, магнитные, оптические и др.; она покрывает все форматы – книги, диски, кассеты, картриджи и т.д.; она отражает все модели записи информации – аналоговый, аккустический, электрический, цифровой, оптический и т.п.

Сущности онтологии относятся к 4 группам. К первой группе относятся сущности «работа» (work) – отличительное интеллектуальное или артистическое творение, «выражение» (expression) – интеллектуальная или артистическая реализация работы, «воплощение» (mainfestation) – физическое представление выражения работы и «экземпляр» (item) – отдельный экземпляр воплощения выражения работы. Вторую группу создают сущности «персона» (person), представляющие физическое лицо и «юридическое лицо» (corporate body). Связи между сущностями первых двух групп могут указывать на принадлежность, создание интеллектуального или артистического содержимого. Мощность таких отношений – «многие ко многим». К третьей группе относятся сущности «концепция» (concept) – абстрактное понятие или идея, «объект» (object) – нечто материальное, «событие» (event) – действие или происшествие и «место» (place). У сущностей третьей группы и сущности «работа» есть связь «тема», имеющая мощность «многие ко многим».

Отдельно в описании онтологии для первой группы выделяются связи содержимого, которые задаются между экземплярами одной сущности. Его примером могут служить связи «адаптация», «приложение», «имеет продолжение», «имеет аннотацию», «имеет пародию» и т.п. между экземплярами сущности «работа». Также в модели представлены отношения части целого.

2.4 SPAR

SPAR (Semantic Publishing and Referencing Ontologies) – это набор из 8 независимых онтологий для создания машиночитаемых метаданных в RDF для всех аспектов семантической публикации и перекрестных ссылок. Все онтологии SPAR могут использоваться вместе или независимо друг от друга. Все они написаны на языке OWL 2.0 DL.

Онтология FaBiO (the FRBR-aligned Bibliographic Ontology) содержит семантические описания уже опубликованных или потенциально публикуемых работ или сущностей, используемых для

определения ссылок. Сущности онтологии – это книги, журналы, газеты, периодические издания, а также их содержимое – произведения, статьи. Кроме того, в ней содержатся наборы данных, компьютерные алгоритмы, экспериментальные протоколы, формальные спецификации и словари, технические и коммерческие отчеты и схожие публикации, а также библиографии, списки ссылок, библиотечные каталоги и схожие коллекции.

Онтология CiTO (the Citation Typing Ontology) описывает цитирование. Цитирование может быть прямым (например, список литературы в статье), непрямым (ссылки на более ранние работы той же группы) и неточным (пародии, плагиат). Онтология также содержит классификацию трудов и их принадлежность.

Онтология BiRO (the Bibliographic Reference Ontology) структурирована в соответствии с моделью FRBR (модель функциональных требований для библиографических записей), чтобы задавать библиографические записи (в качестве подкласса rfbr:Work) и библиографические ссылки, а также их преобразование в библиографические коллекции и библиографические списки. Она позволяет задавать ссылки на отдельные труды, запись о которых может содержать не только имена авторов, но и библиографические параметры – ISBN и т.п.

Онтология C4O (the Citation Counting and Context Characterization Ontology – Онтология Подсчета Цитирования и Формализации Контекста) позволяет произвести формализацию библиографического цитирования в понятиях номеров и их контекста. Она предоставляет онтологические структуры для доступа к внутритекстовым ссылкам, чтобы ссылка могла указывать на конкретный раздел или фразу в тексте, а также количество цитирований, заданное в глобальных библиографических ресурсах, таких как Google Scholar, Scopus и т.п. на конкретную дату.

Онтология DoCO (the Document Components Ontology) предоставляет структурированный словарь компонентов документа – структурных (параграф, глава, раздел и т.п.) и риторических (вступление, обсуждение, список литературы, рисунок и т.п.), позволяя описать документ в RDF.

Онтология PRO (the Publishing Roles Ontology) создана для формализации ролей людей (автор, редактор, рецензент, издатель и т.п.) в процессе публикации. Она позволяет задать отношение работников и различных библиографических сущностей. Благодаря структуре онтологии количество ролей можно легко расширить, добавив экземпляры в класс pro:Role.

Онтология PSO (the Publishing Status Ontology) создана для формализации статусов публикации документа и других сущностей публикации на разных стадиях публикации (черновик, подтверждена, рецензируется, опубликована, катологизирована, переведена в архив и т.п.). Классы онтологии не заданы – только свойства, у которых, соответственно, не заданы области определения и значение.

Онтология PWO (the Publishing Workflow Ontology) создана для описания этапов рабочего процесса, сопутствующего публикации документа или других сущностей публикации (в процессе написания, рецензируется, дизайн страницы, сетевая публикация и т.п.).

2.5 CERIF

CERIF (Common European Research Information Format) – европейская онтология, разработанная для описания процесса научно-исследовательской деятельности. Она развивается на протяжении последних 25 лет и в данный момент является основной для ведения учета исследований в Европе.

Онтология разбита на несколько областей, описывающих какой-то аспект деятельности. В основе онтологии лежат три класса – «проект», «персона» и «организационная единица». Стоит обратить внимание на то, что собственно научный труд здесь описан не очень подробно. Большее внимание уделяется описанию процессов исследовательской деятельности и фиксированию хозяйственной стороны данного процесса – формальное фиксирование проведенных измерений, используемое и закупаемое оборудование, место проведения исследований, размер исследовательской группы, описание навыков каждого из сотрудников, финансирование и премирование, участие в подотчетных конференциях, публикации тудов и их цитирование, отслеживание формальных результатов исследований. Классы имеют большое количество связей для детального описания исследовательской деятельности.

Также в данной инициативе разработана спецификация CERIF XML, которая описывает обмен данными в XML виде.

2.6 BIBO

Онтология BIBO (Bibliographic Ontology) – онтология, описывающая библиографические сущности в RDF. Онтология включает в себя возможности цитирования, библиографической классификации документов. Она включает в себя сущности из онтологий Dublin Core и FOAF. В онтологии описываются организации и персоны, ресурсы, тезисы и агенты. Глубоко классифицированы документы (35 классов пяти уровней иерархии). Отдельно стоит отметить, что онтология BIBO написана на языке OWL Full, что существенно усложняет машинный вывод.

2.7 PROV-O

Онтология PROV-O (PROV Ontology) представляет собой переведенную на язык OWL2 модель данных PROV-DM. Модель PROV-DM создана для описания информации о сущностях, действиях или людях, занимающихся получением данных или чего-либо, что может формировать

суждения о качестве, надежности или достоверности этих данных.

Сущности онтологии позволяют задавать информацию о персоне, агенте или организации, их ролях и взаимодействии, данных, которые они предоставляют, цитировании, рецинзировании, производных материалах, валидации материалов и активности описываемых в онтологии лиц.

2.8 SKOS

SKOS (The Simple Knowledge Organization System) – модель данных для систем организации знаний, таких как тезаурусы, схемы классификации, таксономии и др. В рамках данного проекта написаны онтологии SKOS и SKOS-XL для описания данных в рамках рассматриваемой модели. Онтология SKOS написана на языке OWL Full. Данные онтологии задаются в виде троек RDF. Понятия (например, «лексический анализ» или «語彙分析») могут задаваться на любом естественном языке (а в только что описанном примере на русском или японском). В онтологии вводятся связи для указания на то, что один экземпляр является более широким/узким понятием по отношению к другому. Также имеются понятия для задания ассоциативных лексических связей между экземплярами, задания полного и краткого описания для указания на наличие экземпляра более высокого уровня (для эффективного определения верхнего уровня). Большинство схем классификации могут быть корректно заданы при использовании связей «более широкое», «более узкое» (skos:broader, skos:narrower) для классификации понятий, чем при помощи иерархии классов.

Также данная онтология предоставляет возможность для реализации массивов с узловыми метками (label nodes). Узловая метка необходима для построения группировки для упрощения просмотра иерархии терминов онтологии.

У онтологии существует расширение SKOS-XL для более точной классификации лексических сущностей.

Онтология является стандартом ISO 2788-1986.

2.9 PRISM

PRISM (The Publishing Requirements for Industry Standard Metadata) [7] – это спецификация, определяющая богатый набор терминов для описания печатных трудов. Термины метаданных PRISM могут использоваться как в XML, так и в RDF [6]. Главным плюсом онтологии PRISM является намного более богатый набор терминов для описания библиографических сущностей, чем Dublin Core. Главным недостатком PRISM является его гладкая структура. Такая структура приводит к отсутствию иерархической структуры и усложняет классификацию сущностей.

2.10 CIDOC CRM

CIDOC CRM ("Committee on Documentation" "Conceptual Reference Model") – это онтология, созданная для хранения информации об объектах культурного наследия. С 2006 года она является стандартом ISO 21127:2006. Модель позволяет описывать экспонаты и музейные коллекции, их характеристики и историю. Модель описывает как физические характеристики, так и временные, и состоит из примерно 100 классов.

2.11 Соционет

Соционет – это научно-образовательная социальная сеть, предоставляющая возможности семантического поиска и анализа научных трудов. В основе системы лежит онтология Соционет, которая заимствует сущности из онтологий CiTo и DoCo проекта SPAR, SKOS, CERIF и SWAN, дополненная некоторыми сущностями, которые характеризуют специфику российской научной среды. Важной частью онтологии является 6 словарей, представляющих таксономию семантических связей в системе:

1) Словарь свойств связей научного вывода (состоит из импортируемых связей онтологии CiTo).

2) Словарь свойств связей использования (состоит из импортируемых связей онтологии CiTo).

3) Словарь свойств связей мнений и оценок (состоит из импортируемых связей онтологий CiTo и SWAN).

4) Словарь свойств иерархических и ассоциативных связей между научными публикациями (состоит из импортируемых связей онтологий CiTo, SKOS, SWAN).

5) Словарь свойств связей между компонентами научных публикаций (состоит из импортируемых связей онтологии DoCo, дополненных внутри онтологии).

6) Словари свойств научно-организационных связей (состоит из импортируемых связей онтологии CERIF, дополненных внутри онтологии).

2.12 ЕНИП

В рамках российской инициативы по организации Единого Научного Информационного Пространства (ЕНИП) была разработана модель, в которой выделено четыре основных группы информационных сущностей: участники научной деятельности, научная деятельность, результаты научной деятельности, документы и публикации. К классу Документ относятся разного рода документы и публикации, как печатные, так и цифровые. Класс Публикация в данной модели является подклассом класса Документ. Данный абстрактный класс описывает метаинформацию об официально зарегистрированных печатных изданиях (публикациях). Публикации делятся на 3 группы: издания аналитического уровня, монографического уровня и сводного уровня. При описании

конкретных публикаций необходимо указывать конкретные неабстрактные классы, такие как Монография, Многотомное издание, Выпуск журнала и пр.

В основе лежит принцип разделенной системы из множества онтологий, позволяющих хранить данные о публикациях, вести сопровождение научных исследований, каталогизировать имеющуюся информацию.

3 Сравнение классов рассмотренных онтологий

При проведении сравнения было решено, что включать в него онтологии, напрямую не описывающие научные публикации или научную деятельность, не совсем корректно, поэтому из сравнения были исключены Dublin Core, PRISM, CIDOC CRM и SWAN. 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).

3.1 FRBR и CERIF

Онтологии FRBR и CERIF имеют лишь несколько точек пересечения. Это такие базовые понятия, как научный труд, которое более подробно классифицировано в онтологии FRBR, событие и персона. Столь малое пересечение объясняется тем, что онтология CERIF описывает научно-исследовательские работы с точки зрения финансирования и, следовательно, отчетности, в то время как онтология FRBR направлена на описание трудов, их представлений и различных воплощений (переводы, переиздания, публикация и проч.).

3.2 Цитирование в онтологиях CiTo, PROV-O

Если онтология PROV-O описывает получение информации, ее проверку и установление достоверности, то онтология CiTo детально описывает исключительно различные аспекты цитирования. Другие онтологии почти не описывают данный аспект – так, например, цитирование в онтологии CiTo описывается свойствами cito:’is cited as authority by’, cito:‘is cited as data source by’, cito:’is cited as evidence by’, cito:’is cited as metadata by’, cito:’is cited as potential solution by’, cito:’is cited as recommended reading’, cito:’is cited as related by’, cito:’is cited as source document by’, cito:’is cited for information by’, являющимися подсвойством свойства cito:’is cited by’, а в онтологии PROV-O только свойством prov:qualifiedQuotation.

3.3 FaBiO, SKOS и BIBO

Онтология SKOS описывает более узкий набор сущностей, детально структурируя понятия и их отношения. Сущности онтологии (и классы, и свойства) заимствует онтология FaBiO.

FaBiO дополняет классы «понятие» и «схема» и связанные с ними свойства из онтологии SKOS. Также импортируются классы трудов, представлений и выражений из онтологии FRBR. От этих классов наследуются подклассы, описывающие намного более детальную классификацию публикаций, изданий, их представлений и выражений. Также добавляется понятие ответственности и ответственных за тот или иной проект лиц.

Онтологии FaBiO и BIBO описывают одну и ту же область – библоиграфические сущности. При этом между ними есть ряд различий – в BIBO не используется модель данных FRBR, вследствие чего она менее подробна, чем FaBiO. Например, в BIBO есть класс bibo:AcademicArticle, который отображает одноврменно понятия академического труда и статьи из журнала, которые описываются классами fabio:ResearchPaper и fabio:JournalArticle. Если обобщить данные по онтологии целиком, то BIBO содержит 69 классов, 52 свойства объектов, 54 свойства данных и 14 экземпляров, в то время как FaBiO содержит 239 классов, 69 свойств объектов, 63 свойства данных и 15 экземпляров [15].

BIBO написана на языке OWL Full, а онтология FaBiO – на языке OWL DL. OWL Full в отличии от OWL DL, является неразрешимым в частных случаях, вследствие чего в онтологиях на этом языке нельзя проводить машинный вывод.

4 Заключение

В статье приведен обзор онтологий, описывающих научные труды и научно-исследовательскую деятельность. Рассмотрены онтологии, использующиеся для описания связей, образующихся между сущностями научно-исследовательской деятельности (исследователями, самими научными трудами, конференциями и т.п.). Проведено сравнение между онтологиями, выделены их схожие и импортированные части.

Литература

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

[2] Crofts N., Doerr M., Gill T., Stead S., Stiff M. (editors), Definition of the CIDOC Conceptual Reference Model, January 2008. Version 4.2.4.

[3] DCMI Home: Dublin Core® Metadata Initiative (DCMI). http://dublincore.org/

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

[5] 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).

[6] Hammond, T. (2008). RDF Site Summar. http://www.w3.org/TR/owl2-overview/y

1.0 Modules: PRISM. http://nurture.nature.com/rss/modules/mod_prism.html

[7] International Digital Enterprise Alliance (2009). Publishing Requirements for Industry Standard Metadata Specification Version 2.0. Alexandria, VA, USA: IDEAlliance. http://www.idealliance.org/specifications/prism

[8] Journal Article Tag Suite. http://jats.nlm.nih.gov/

[9] Lassila, Ora, and Ralph R. Swick. "Resource description framework (RDF) model and syntax specification". (1999).

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

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

[12] SKOS Simple Knowledge Organization System Reference. http://www.w3.org/TR/skos-reference/

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

[14] SWAN (Semantic Web Applications in Neuromedicine) – Scientific Discourse Relationships Ontology Specification. http://swan.mindinformatics.org/spec/1.2/discourserelationships.html

[15] Using the SPAR ontologies to publish bibliographic records http://semanticpublishing.wordpress.com/2013/03/01/lld5-using-spar-ontologies/

[16] Бездушный А.А., Бездушный А.Н., Серебряков В.А., Филиппов В.И. Интеграция метаданных Единого Научного Информационного Пространства РАН // Вычислительный центр РАН, г. Москва. – 2006. – 238 с.

[17] Паринов, С. И., Когаловский М. Р. Технология семантического структурирования контента научных электронных библиотек // Труды XIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL-2011. Воронеж, 2011.

[18] Хоай, Ле, Тузовский А.Ф. Семантическое аннотирование документов в электронных библиотеках // Известия Томского политехнического университета. 2013. Т. 322, № 5. С. 157–164.

Analysis of Semantic Ontologies That Describe Scientific Publications and Research Activities

Victor V. Kostin

This article contains analysis of ontologies that either describe scientific publications and research activities or contains elements (entities, classes or properties) that can be used for such purposes. The comparison of their entities is given.