Семантические модели и наборы данных в проектировании связанных данных в учебном процессе

А.А. Сытник, Н.С. Вагарина, Н.И. Мельникова
Саратовский государственный технический университет имени Гагарина Ю.А.
as@sstu.ru, v-n-s@yandex.ru. MelnikovaNI@gmail.com

Аннотация

Технологии связанных данных используются для представления наборов данных и их увязки с другими опубликованными наборами данных. Первой задачей при проектировании связанных данных является поиск и отбор подходящих семантических моделей, изучение состава терминов в выбранных моделях и особенностей их использования. В работе рассмотрены современные семантических модели, такие как тезаурусы, таксономии, словари, онтологии, а также наборы связанных данных под открытой лицензией. Предложены подходы к проектированию учебных связных данных на основе повторного использования и доработки существующих семантических моделей. Повторное использование и доработка имеющихся семантических моделей позволит эффективно осуществить функциональное расширение разработанных учебных программных средств RDF-редактора Make Sense и SPARQL-тренажера. Акцент на повторном использовании и доработке существующих семантических моделей влечёт необходимость создания действенных программных средств для учебного процесса в части проверки лицензий, под которыми были созданы семантические модели и наборы данных, верификации семантических моделей для их последующей интеграции.

Ключевые слова: веб данных; связанные данные; семантические модели; тезаурусы; таксономии; онтологии; словари

1.Введение

Концепция семантического веба (Semantic Web) предложена консорциумом W3C в качестве новой модели развития веб сервиса. Нынешний веб нацелен исключительно на предоставление информации людям, и его документы могут эффективно читаться только ими. Главная идея семантического веба состоит в представлении информации в форме, доступной для компьютерной обработки. Такой подход позволяет распространить принципы веба с документов на

данные, т.е. осуществить переход от «веба документов» к «вебу данных». Текущее проектирование сетевых информационных ресурсов также следует осуществлять с использованием семантических моделей [1]. Развитие технологий семантического веба вызвало потребность в преподавании новых дисциплин в подготовке специалистов по направлению «Информатика и вычислительная техника». В настоящее время используются разрозненные программные средства, использующие различные форматы, для первоначального обучения технологиям семантического веба. Стандартный набор включает несколько программных средств, например, RDF парсер Redland Raptor (http://librdf.org/parse/), RDF конвертер EasyRDF (http://www.easyrdf.org/converter) и RDF валидатор (http://www.w3.org/RDF/Validator/). В связи с этим возникла необходимость в программных средствах, пригодных для первоначального знакомства с технологиями семантического веба. Для учебных целей в СГТУ имени Гагарина Ю.А. разработаны RDF-редактор MakeSense и SPARQL-тренажер. Эти программные средства в настоящее время используются в преподавании дисциплин «Основы технологий семантического веба» и «Структуры и стандарты связанных данных» в подготовке бакалавра по направлению «Информатика и вычислительная техника» в СГТУ имени Гагарина Ю.А.

Разработанные RDF-редактор MakeSense и SPARQL тренажёр позволили существенно расширить спектр программных средств для обучения технологиям семантического веба. Для совершенствования учебного процесса необходимо дельнейшее расширение учебных программных средств. В связи с этим представляется актуальным представить в ёмком и лаконичном виде базовые подходы к порядку работы с семантическими моделями и связанными данными для дальнейшей разработки необходимых учебных программных средств в области технологий семантического веба.

2. Подходы к описанию семантических моделей

Первым значимым проектом в области технологий семантического веба стали так называемые связанные данные (linked data). Связанные данные получают всё большее распространение в образовании [2]. Технологии связанных данных используются для представления наборов данных и их увязки с другими опубликованными наборами данных таким образом, чтобы приложения могли использовать данные из множества различных источников. Сам термин «связанные данные» подсказывает, что всякий создаваемый набор данных должен быть связан с ранее созданными данными и семантическими моделями. Основным содержанием данного подхода является использование семантических моделей – тезаурусов, таксономий, онтологий, словарей. Поэтому краеугольным камнем в технологиях связанных данных является использование ранее созданных семантических моделей и наборов данных. Их отбор для последующего связывания является важнейшим этапом жизненного цикла проектирования связанных данных. Особое значение это имеет для учебного процесса, поскольку именно в процессе обучения закладывается культура использования существующих информационных активов.

Для описания семантических моделей используются специальные языки (RDF, OWL), для доступа к данным используются языки запросов (SPARQL). Для проектирования онтологий используются редакторы онтологий, например, Protégé (свободно распространяемый с открытым кодом редактор онтологий (http://protege.stanford.edu/) и системы логического вывода, например, Pellet (OWL-резонёр с открытым кодом).

Описание сущностей в наборах данных и их соотношений друг с другом выполняется с помощью семантических моделей, таких как тезаурусы, таксономии, словари, онтологии, выраженных в SKOS (Simple Knowledge Organization System) [3], RDFS (RDF Vocabulary Description Language или RDF-схемы) [4] или OWL (Web Ontology Language) [5]. Ввиду того, что технологии семантического веба находятся в состоянии интенсивного развития, то можно наблюдать некоторую нечёткость терминологии. Обобщённым названием семантических моделей является термин «словарь». Исторически сложилось, что для наиболее часто используемых словарей устоялось собственное обозначение, будь то онтология, словарь и пр. Подробное описание известных семантических моделей выполнено в рамках проекта «Linked Open Vocabularies» (LOV) (http://lov.okfn.org/dataset/lov/).

SKOS предназначен для представления концептуальных иерархий, часто называемых таксономиями или тезаурусами. Согласно данным проекта LOV SKOS использующийся в 152 наборах облака связанных открытых данных, является одним важнейших хабов для связывания данных (http://lov.okfn.org/dataset/lov/vocabs/skos). Известным проектом, реализованным с использованием SKOS, является проект связанных открытых данных газеты Нью-Йорк Таймс http://data.nytimes.com/).

Для описания концептуальных моделей в терминах классов и их свойств используются языки RDFS и OWL. RDFS это язык для описания лёгких онтологий в RDF, которые собственно и называются словарями. RDFS словари, которые также называются RDF схемы, состоят из классов и свойств. Примитивы RDFS определены в двух разных пространствах имён: пространство имён http://www.w3.org/2000/01/rdf-schema# связано (по соглашению) с префиксом rdfs:; пространство имён http://www.w3.org/1999/02/22-rdf-syntax-ns# имён связано (по соглашению) с префиксом rdf:. Основными классами в RDFS является класс ресурсов rdfs: Class и классе всех свойств rdf: Property. RDF ресурс объявляется как класс, с указанием его в качестве экземпляра rdfs:Class с помощью предиката rdf:type. RDFS также предоставляет примитивы для описания отношений между классами и между свойствами. Например, rdfs:subClassOf используется для того чтобы сделать утверждение, что все экземпляры одного класса также являются экземплярами другого класса.

В совокупности SKOS, RDFS и OWL предоставляют широкий спектр выразительности для построения семантических моделей. SKOS широко используется для представления тезаурусов, таксономий и предметных иерархий. RDFS и OWL используются в тех случаях, когда должны быть представлены отношения между терминами (например, что все студенты являются также людьми). Семантические модели RDFS и OWL в сочетании с подходящими поисковыми машинами позволяют выявлять неявные отношения в наборах данных. В этом состоит принципиальное отличие баз данных от

наборов связанных данных. Выразительной мощности RDFS часто бывает достаточно для построения семантической модели. Но для установления классов эквивалентности для URI из различных пространств часто необходимо использовать некоторые примитивы из OWL, например, sameAs. Использование этой примитивы позволяет сделать утверждение, что некоторые различные URI определяют на самом деле один и тот же ресурс.

Связанные данные всё ещё остаются новой областью деятельности, поэтому представляет интерес каталогизации открытых семантических моделей и связанных данных, для их последующего исследования в учебном процессе.

3. Методология поиска, отбора и использования семантических моделей и данных

В учебном процессе проектирование связанных данных и эффективное использование созданных программных средств выдвигает задачу отбора семантических моделей, которые бы позволили эффективно разрабатывать новые семантические модели, создавать наборы данных с использованием этих моделей и наращивать состав и функциональность программных средств для учебного процесса. Такой подход позволит вовлечь в учебный процесс, например, разработанные учебные онтологии, обеспечить преемственность процесса обучения технологиям семантического веба с наращиванием методического материала и пр.

Вполне вероятно, что для рассматриваемой предметной области семантические модели уже созданы и необходимо корректно их использовать [6]. Для учебных целей вполне логичным представляется сделать упор на повторное использование и доработку существующих, а не на разработку новых семантических моделей. Если найденная модель имеет недостаточное количество терминов для описания нового набора данных, то рекомендуется расширять её, а не создавать новую. Однако прежде чем создать новый термин, также необходимо убедиться, что подобный термина отсутствует в модели. И только после твёрдого убеждения в том, что необходимый термин отсутствует, его необходимо создать. Если в семантической модели существует схожий термин, но имеется потребность в его уточнении, то следует использовать такие инструменты как подклассы и подсвойства. Только в случае если нет подходящих терминов, которые можно уточнить с помощью подклассов и подсвойств, следует вводить новый термин. Если подходящие термины существуют в моделях, то они должны быть повторно использованы. Повторное использование существующих терминов весьма желательно, поскольку это увеличивает вероятность того, что данные могут быть использованы приложениями, работающими с известными словарями, без предварительной обработки данных или модификации приложений.

Ввиду отсутствия единого каталога семантических моделей в качестве отправной точки для их поиска можно использовать, например, репозитарий RDF схем *SchemaWeb* (http://www.w3.org/wiki/SchemaWeb), каталог SKOS тезаурусов (http://www.w3.org/2001/sw/wiki/SKOS/Datasets), данные проекта LOV, семантические поисковые системы, например, *Swoogle (*http://swoogle.umbc.edu/*)* и др. [7].

Существует ряд семантических моделей, которые употребляются при создании практически любого набора связанных данных, среди которых следует отметить: словарь *Dublin Core Metadata Initiative (DCMI) Metadata Terms* определяет общие атрибуты метаданных, такие как title, creator, date и subject; схема *The Creative Commons (CC) schema* определяет термины для описания лицензий авторского права; онтология *The Bibliographic Ontology (BIBO)* предоставляет концепты и свойства для описания цитирования и библиографических ссылок (т.е. цитаты, книги, статьи и т.д.); словарь *Friend-of-a-Friend (FOAF)* определяет термины для описания людей, их деятельности и их отношений с другими людьми; онтология *GeoNames* для представления геопространственной информации.

В случаях, когда существующие семантические модели не подходят для описания конкретного набора данных, новые термины должны быть разработаны в новой семантической модели с применением особенностей RDFS и OWL и следуя определённым принципам: тщательное документирование каждого нового термина; определение новых терминов в подконтрольных пространствах имён; применение принципов связанных данных одинаково строго как в словарях, так и в наборах данных; ограниченное использование онтологических аксиом.

Для управления процессом развития словарей существует ряд признанных инструментов, например, *Neologism* (http://neologism.deri.ie/), являющейся платформой для публикации словарей для веба данных на принципах связанных данных.

Поиск же наборов связанных данных также не является простой задачей. Один из первых зарубежных проектов в этой области был запущен в 2007 году, в рамках которого идентифицированы существующие наборы данных под открытой лицензией, затем преобразованы в RDF в соответствии с принципами связанных данных и опубликованы в вебе. Возникло понятие облака связанных открытых данных (Linked Open Data Cloud, LOD Cloud). Диаграммы LOD Cloud в динамике по годам представлены в Интернете (http://lod-cloud.net/). Целесообразно использовать веба облако связанных данных в учебном процессе.

RDF-редактор MakeSense, ориентированный на учебный процесс позволяет в простой и наглядной форме создавать и редактировать информацию в формате RDF/RDFS, изучать и сравнивать различные сериализации RDF, строить соответствующие RDF-графы, изучать принципы языка запросов SPARQL на создаваемых данных [8]. Облако связанных открытых данных является очень ценным методическим материалом. Разработанный SPARQL тренажёр позволяет выполнять запросы, как к локальным данным, так и к внешним точкам доступа [9]. Следует отметить, что при разработке SPARQL-тренажёра были учтены достоинства и недостатки существующих аналогов, в результате получился максимально приспособленный для использования в учебных целях программный продукт. SPARQL-тренажёр позволяет самостоятельно осваивать язык запросов SPARQL на данных облака связанных данных и творчески использовать разработанное методическое обеспечения [10].

В совокупности RDF-редактор MakeSense и SPARQL-тренажер позволяют не только читать на и создавать новые данные с использованием как данных из облака связанных данных, так и данных, создаваемые в процессе изучения по технологиям семантического веба. В результате в учебном процессе возникают новые наборы данных, имеющие сложное лицензионное происхождение, когда сопрягаемые данные имеют различные лицензии. В связи с этим важнейшей задачей в развитии указанных программных средств является включение механизма проверки лицензий, под которыми были созданы семантические модели и наборы данных, верификация семантических моделей для их последующей интеграции и др.

4. Заключение

Несмотря на впечатляющие успехи в развитии технологий семантического веба, они являются пока весьма разрозненными. Создание учебных программных средств в области семантического веба является по-прежнему является актуальным. Расширение их функциональности возможно при условии формулирования некоторых принципов в вопросах проектирования данных. Важнейшим аспектом проектирования связанных данных является повторное использование семантических моделей и наборов данных. Предложенная методология поиска и отбора словарей для проектирования учебных связанных данных с опорой на повторное использование и доработку существующих моделей является вполне допустимой, но не единственной. При проектировании общеупотребимых данных такой способ является естественным, что и отражает дух учебного процесса. Если же речь идёт о создании предметно-ориентированных моделей, то способы повторного использования, доработки и создания новых семантических моделей могут находиться совершенно в другой пропорции. Повторное использование семантических моделей в серьёзных проектах требует особой тщательности в верификации моделей, в том числе использования специальных программных средств для их визулизации и усовершенствования. Также для различных предметных областей представляет интерес автоматического построения и верификации семантических моделей [11]. Ввиду этого в разработке предметно-ориентированных семантических моделей упор может быть сделан не на повторное использование, а на разработку собственных моделей, отражающих тонкости некоторой предметной области. Развитие учебного процесса в области технологий связанных данных неизбежно потребует дальнейшего рассмотрения и систематизации способов продуктивного использования существующих семантических моделей и наборов связанных данных в проектировании связанных данных.

Литература

[1] Sytnik A.A., Papshev S.V. Semantic Segmentation of Hypertext on the Basis of Automata Model // International Journal of Computing Anticipatory Systems. 2014. T. 28. C. 109–115.

[2] Мельникова Н.И., Филина Е.В. Наборы связанных данных в зарубежном образовании // Информационно-коммуникационные технологии в науке, производстве и образовании. Саратов: Саратовский государственный технический университет, 2014. С. 48–50. URL: http://elibrary.ru/item.asp?id=22596945 (дата обращения: 08.04.2014).

[3] Miles A., Bechhofer S. SKOS Simple Knowledge Organization System Reference. URL: http://www.w3.org/TR/skos-reference/ (дата обращения: 08.04.2014).

[4] Brickley D., Guha R. RDF Schema 1.1. URL: http://www.w3.org/TR/rdf-schema/. (дата обращения: 08.04.2014).

[5] McGuinness D., Harmelen F. van. OWL Web Ontology Language Overview. URL: http://www.w3.org/TR/2004/REC-owl-features-20040210/ (дата обращения: 08.04.2014).

[6] Вагарина Н.С. Использование словарей RDF в контексте связанных данных // Проблемы управления в социально-экономических и технических системах. Саратов: Саратовский государственный технический университет, 2014. С. 46–49. URL: http://elibrary.ru/item.asp?id=22597423 (дата обращения: 08.04.2014).

[7] Мельникова Н.И. Поиск и извлечение связанных данных // Проблемы управления в социально-экономических и технических системах. Саратов: Саратовский государственный технический университет, 2014. С. 44–46. URL: http://elibrary.ru/item.asp?id=22597422 (дата обращения: 08.04.2014).

[8] Апсаликов М.Ю., Вагарина Н.С. О Создании редактора RDF-данных для использования в учебном процессе // Информационно-коммуникационные технологии в науке, производстве и образовании. Саратов: Саратовский государственный технический университет, 2014. С. 68–70. URL: http://elibrary.ru/item.asp?id=22596876 (дата обращения: 08.04.2014).

[9] Попов С.М. Точки доступа SPARQL для обработки информации в формате RDF // Математические методы в технике и технологиях (ММТТ-27) / под ред. А.А. Большакова. Тамбов: Тамбовск. гос. техн. ун-т, 2014. С. 40–42.

[10] Сытник А.А., Вагарина Н.С., Мельникова Н.И. Введение в язык запросов к семантическим данным SPARQL. Саратов: Саратовский государственный технический университет, 2014. 57 с.

[11] Романов С.В., Шульга Т.Э. Разработка программного обеспечения автоматического построения онтологий LSP-шаблонами для коллекций текстовых документов // Математические методы в технике и технологиях - ММТТ-27. Саратов: Саратовский государственный технический университет, 2014. С. 20–21.

Semantic models and data sets in the designing of linked data for curriculum

A. Sytnik, N. Vagarina, N. Melnikova
Yuri Gagarin State Technical University of Saratov

The aim of linked data is representing of data sets as interconnected data with another data sets. The first task of the designing of linked data is search and selection of appropriate vocabularies to study the set of terms in selected models and features of use. The article deals discusses the modern semantic models such as thesauri, taxonomies, vocabularies, ontologies and linked data sets under open source license. An approach to modelling and publishing training of linked data are made based on the reuse and refinement of existing semantic models. Reusing and refinement of existing semantic models will effectively improve the functional enhancement of training software such as RDF-editor MakeSense and SPARQL-simulator. The focus on reuse and refinement of existing semantic models implies the need to create effective software tools to the curricula. This software should have some functional features: check out the verification of licenses, under which were created semantic models and data sets; verification of semantic models for their subsequent integration.

Keywords: web of data, linked data, semantic models, thesaurus, taxonomies, ontologies, vocabularies