Коммюнике онтологического саммита 2014

Прикладные онтологии в семантической сети и больших данных

Semantic Web and Big Data Meets Applied Ontology[1]

OntologySummit2014 Communiqué

Lead Editors: Michael Gruninger, Leo Obrst

CoEditors: Ken Baclawski, Mike Bennett, Dan Brickley, Gary BergCross, Pascal Hitzler, Krzysztof Janowicz, Christine Kapp, Oliver Kutz, Christoph Lange, Anatoly Levenchuk, Francesca Quattri, Alan Rector, Todd Schneider, Simon Spero, Anne Thessen, Marcela Vegetti, Amanda Vizedom, Andrea Westerinen, Matthew West, Peter Yim.

Beth Di Giulian, Bobbin Teegarden, Corey Leong, David Blevins, Deborah Nichols, Dennis Wisnosky, Ed Bernot, Elizabeth Florescu, Gail Hodge, Gilberto Fragoso, Hans Polzer, Mark Underwood, Michael Barnett, Michael Uschold, Patrick Cassidy, Pavithra Kenjige, Quentin Reul, Ravi Sharma, Shima Dastgheib, Stephane Fellah, Steve Ray, Terry Longstreth, James Solderitsch, Frank Olken, Naicong Li, Ali Hashemi, Matthew Kaufman, Katherine Goodier, Barry Smith, Malek Ben Salem, Uri Shani, Maria Poveda Villalon, Laura Daniele, Carlos Toro, Dagobert Soergel, Michael Riben, Marcia Zeng, Doug Holmes, Khalil Ben Mohamed, John Yanosy, Ranjith Kanimozhi, John Bateman, Nikolay Borgest, Oscar Corcho, Mara Abel, Torsten Hahmann, Adam Goldstein, Frank Loebe, Nathalie Aussenac Gilles, Oliver Kutz, Hanmin Jung, Michael Fitzmaurice, Mike Dean, John Mc Clure, Sunday Ojo, Jose Maria Garcia, Mitch Kokar, Megan Katsumi, Deborah Mac Pherson, Jens Ortmann, Jack Ring, Harold Boley, Cong Wang, Jie Zheng, Henson Graves, Rex Brooks, Ollie Faison, John Sowa, Vojtech Svatek, Yuh Jong Hu, Chih Hong Sun, Fabian Neuhaus, Christos Louis, Cyrus Nourani, Mohsen Sadighi Moshkenani, Howard Webb, Pauline Kra, Julita Bermejo Alonso, Leo Meerman, Dickson Lukose, Sameera Abar, Nancy Wiegand, Stefano Borgo, Nicola Guarino, Elisa Kendall

Резюме

Роль, которую онтологии играют или могут играть в разработке и использовании семантических технологий, получила широкое признание в сообществах Семантической сети (Semantic Web) и Связанных данных (Linked Data). Однако уровень сотрудничества между этими сообществами и сообществом Прикладные онтологии (Applied Ontology) гораздо ниже, чем ожидалось. Поэтому и онтологии, и онтологические методы слабо используются в Больших данных (Big Data) и их приложениях.

Для осмысления сложившегося положения и расширения сотрудничества, онтологический саммит (Ontology Summit 2014) собрал представителей Семантической сети, Связанных данных, Больших данных и Прикладных онтологий для поиска решений трех основных проблем, затрагивающих прикладные онтологии и эти сообщества [1, 2]:

1) роль онтологий [в этих сообществах];

2) текущее использование онтологий в этих сообществах;

3) разработка онтологий и семантическая интеграция.

Задача саммита состояла в том, чтобы идентифицировать и понять:

a) причины и проблемы (например, масштабируемость), которые препятствуют повторному использованию онтологий в Семантической сети и Связанных данных;

b) решения, которые могут уменьшить различия между онтологиями сетевыми (on line) и автономными (off line);

c) способы применения онтологий для преодоления технических «узких мест» в современных приложениях Семантической сети и Больших данных.

В течение четырех месяцев в рамках четырех секций прошло обсуждение практических вопросов с представителями сообществ Семантической сети, Связанных данных и Прикладных онтологий. Каждая секция фокусировалась на различных аспектах тематики саммита этого года:

(Секция A) исследование совместного доступа и повторно используемых онтологий;

(Секция B) инструменты, сервисы и методы для всестороннего и эффективного использования онтологий;

(Секция C) исследование технических «узких мест» и путей их предотвращения и преодоления;

(Секция D) рассмотрение проблемы разнообразия в Больших данных.

В дополнение к этим четырем секциям был проведен Hackathon[2], на котором были представлены шесть различных проектов. Все результаты доступны в соответствующих индивидуальных свободных репозиториях: библиотека онлайн-сообщества и онлайновый репозиторий онтологий.

[1] http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2014_Communique. *Перевод М.Д. Коровина.*

[2] **Хакатон** (англ. *hackathon*, от *hack* (хакер) и *marathon* — марафон) — мероприятие, во время которого специалисты из разных областей разработки программного обеспечения (программисты, дизайнеры, менеджеры) сообща работают над решением какой-либо проблемы. Сегодня хакатоны это просто марафоны программирования. Обычно хакатоны длятся от одного дня до недели. *Примеч.переводчика.*

Коммюнике онтологического саммита 2014 представляет собой отчет оригинальных результатов, в котором сделана попытка объединить подходы различных сообществ и достигнуть согласия относительно обозначенных проблем и выработки рекомендаций по их решению.

1. Введение, цель, мотивация

С момента появления семантической сети онтологии играли ключевые роли в разработке и развертывании новых семантических технологий. Однако, за эти годы уровень сотрудничества между сообществами Семантической сети и Прикладных онтологий оказался намного ниже ожидаемого. Что касается применения онтологии в Больших данных, то они получили незначительное распространение и влияние.

Онтологический саммит 2014 обеспечил возможность наведения мостов между сообществами Семантической сети, Связанных данных, Больших данных и Прикладных онтологий. С одной стороны, сообщества Семантической сети, Связанных данных и Больших данных выявляют широкий спектр серьезных проблем (проблемы производительности и масштабируемости и проблема разнообразия в Больших данных) и предлагают технологии (как инструменты автоматизации логического вывода), которые используют онтологии. Важной задачей в сети является понимание данных и информации, распределенной в сети. В отличие от сетевой задачи на небольших онтологиях используются локальное алгоритмы принятия решений, где единственные сетевые аспекты используют IRI[3] как символьные имена, и применяются правила вывода, основанные на открытом (или закрытом) представлении о мире. С другой стороны, сообщество Прикладных онтологий дает большое количество методов онтологического анализа и повторного использования онтологий.

По результатам саммита были выделены три основных направления исследований:

1. Как онтологии фактически используются в приложениях Семантической сети и Больших данных, и в чем состоит отличие от существующих применений онтологий?
2. Как могут сообщества Семантической сети и Больших данных использовать огромное количество уже разработанных и разрабатываемых онтологий?
3. До какой степени автоматизация и инструменты могут помочь преодолевать «узкие места» разработки онтологии?

2. Использование онтологий в Больших данных и Семантической сети

Семантические технологии, такие как онтологии и связанные рассуждения играют главную роль в семантической сети и все чаще и чаще применяются для того, чтобы обрабатывать и понимать информацию, выраженную в цифровом формате. Действительно, отделение точного знания от набора разнообразных (и связанных) данных является одной из основных тем Больших данных.

Проблема получения точного знания состоит в том, чтобы создать и использовать общий семантический контент, избегая множества понятий в различных онтологиях. Примерами такого контента являются целые или частичные онтологии, модули онтологий, онтологические шаблоны и архетипы, словари и общие, концептуальные теории, связанные с онтологиями и их адаптацией к проблемной области. Однако, обработка целого или даже частичного общего семантического контента через логическое объединение, сборку, расширение, специализацию, интеграцию, выравнивание и адаптацию длительное время представляла собой проблему. Достижения общности и повторного использования онтологии в установленные сроки и с управляемыми ресурсами остаются ключевыми компонентами для практической разработки взаимодействующих качественных онтологий.

Онтологии имеют широкий диапазон приложений, включая семантическую интеграцию, поддержку принятия решений, поиск, аннотацию и проектирование систем, как описано в структуре использования онтологии в коммюнике онтологического саммита 2011 [3]. Ключевой вопрос состоит в том, как приложения Больших данных и Семантической сети вписываются в эту структуру - какова роль онтологий в этих приложениях, и как используется семантический контент? Существуют также два общих вопроса, возникающих при решении проблем Семантической сети и Больших данных. Первым является характеристика онтологии, а именно, язык представления онтологии и компромисс, существующий между выразительностью этого языка и эффективностью обоснования с использованием онтологии на этом языке. Вторая особенность, которая характеризует проблемы, возникающие с Большими данными и Семантической сетью, появляется в новых направлениях, в которых онтологии используются в больших масштабах.

2.1 Как используются онтологии и как они могли бы использоваться?

В Больших данных семантическая интеграция решает проблему разнообразия, поскольку любое программное обеспечение, использующее множественные наборы данных, должно обеспечивать отсутствие семантических несоответствий. Онтологии могут также смягчить разнообразие в Больших данных путем помощи в аннотации данных и метаданных. Наборы данных будут отличаться по полноте метаданных, гранулярности и используемому словарю. Таким образом, онтологии могут частично уменьшить это разнообразие путем нормализации условий и обеспечения отсутствующих метаданных.

Новое многообещающее применение онтологий для аналитики данных – управление происхождением данных. К нему относятся любые преобразования, исследования или интерпретации, выполненные над данными. В настоящее время

[3] **IRI** (Internationalized Resource Identifier) — интернационализированный идентификатор ресурса. IRI — это короткая последовательность символов, идентифицирующая абстрактный или физический ресурс на любом языке мира. Идентификаторы IRI призваны в будущем заменить URI. *Примеч.переводчика.*

большинство проектов Больших данных обрабатывает происхождение данных лишь ситуационно, а не на системной основе. Онтологии, описывающие происхождение данных, действительно существуют, например, онтология PROV-O [4]. Разработка стандартных онтологий для часто используемых, но неформализованных процессных моделей, таких как цикл OODA [5] и фьюжн-модели JDL/DFIG [6] может оказать значительное влияние на анализ данных. Примером такой формализации является платформа KIDS [7]. Стандартные статистические онтологии обоснования — это еще одна потенциальная сфера успешного применения онтологий.

На глобальном уровне (например, в сети), существует слишком много доменов, для того, чтобы подробно описывать семантику каждого из них. Однако, существующий проект Schema.org занимается глобальной проблемой разработки общепринятого словаря, использующего в настоящее время более пяти миллионов интернет-доменов, и постепенно описывающего все более глубокую семантику. Внедрение онтологий в структуру Schema.org является сложной задачей, но имеет потенциал для получения существенных преимуществ.

Возможность внедрения онтологий в масштабах всей сети маловероятна. Пока проекты, такие как Уотсон (IBM) [8] ограничивают себя несколькими простыми таксономиями, другие большие совместные проекты приходят к соглашению об использовании ограниченного подмножества онтологий, как, например, разделы некоторых онтологий в молекулярной биологии, таких как Генная онтология [9] и другие открытые биологические и биомедицинские онтологии (OBO) [10]. Остро стоит вопрос целесообразности и выполнимости превращения полных наборов Больших данных в онтологии. Это видится нам выполнимой, но трудоемкой задачей. Ручное создание онтологий является трудоемким. Анализ данных, допускающий повторное использование семантического контента, страдает от потенциальной несогласованности, неполноты и лишних данных. Использование машинного обучения для того, чтобы создавать семантический контент из Больших данных может потребовать дальнейших исследований для того, чтобы реализовать формализацию онтологий на основе Больших данных.

Современные способы использования анализа данных текста, статистических, и других аналитических методов на больших данных для обнаружения корреляций и образов могут быть объединены с семантическим контентом, обеспечивающим некоторую семантическую интерпретацию полученных структур. В дальнейшем, связанный семантический контент будет полезен в обработке результатов, и затем поочередно может быть коррелирован и объединен в большие, постоянно растущие семантические структуры, обеспечивая многослойное богатство так называемого «глубокого обучения».

2.2 Роль выразительности

Понятие выразительности относится к логическим свойствам языка представления онтологии. Спектр онтологии характеризует диапазон различных языков от RDF, OWL и Формата обмена правилами (RIF) до Common Logic и модальной логики. Важнейшим вопросом и для пользователей онтологии, и для разработчиков, является выбор надлежащего языка и возможности эффективно на нем рассуждать. Фактически, многие ранние дебаты о природе онтологий (т.е. что такое онтология?) происходили от различных ожиданий того, что пользователи хотят от выразительности базового языка представления онтологии.

Существует общепринятое мнение, что различные применения онтологий требуют разных уровней выразительности. Для приложений онтологий, связанных с системами поддержки принятия решений, в которых запросы не известны во время проектирования, выразительность очень важна. С другой стороны, если запросы известны заранее, часто можно создать более строгую онтологию, которая ответит на эти известные запросы с лучшим качеством.

Многократная аксиоматизация онтологий на каждом из стандартных языков онтологии будет необходима для соответствия всем требованиям в домене. Разработчики онтологии в целом признают это условие, поэтому некоторые основополагающие онтологии, такие как Дескриптивная Онтология для Лингвистической и Познавательной Разработки (DOLCE) [11] и Основная Формальная Онтология (BFO) [12] имеют не только представления, основанные на логике первого порядка, но также и соответствующий более легкий вид представления на OWL с менее строгой (путем снижения выразительности) аксиоматизацией.

Выразительность языка представления онтологии тесно связана с требованиями для любой онтологии, предназначенной для использования в рамках некоторого приложения. RDF, родной язык связанных данных, находит широкое распространение в Больших данных из-за того, что при низком уровне требований к онтологичности позволяет создавать сложные описания. С другой стороны, много традиционных областей применения онтологий, таких как семантическая интеграция и поддержка принятия решений, требуют более выразительных языков, таких как RIF, Common Logic и логическое программирование.

Создание легких онтологий и словарей для семантической сети и приложений больших данных требует нового, быстрого и гибкого инжиниринга. Новый подход «Связанные открытые термины» (LOT) [13] начинается с повторного использования материалов, используя в своих интересах большое число словарей, уже существующих в сети. В случае, если термины, описывающие текущие условия, не обнаруживаются в существующих словарях, инженер по знаниям должен будет создать новые, при этом он должен стараться объединить их с существующими.

Разработчики Уотсона не создавали формальную онтологию мира, которой они бы попытались объединить формальные логические представления вопросов. Вместо этого они локально изучили онтологии по требованию, привлекая формальные, а также неформальные источники, с помощью различных методов обоснования. Во-первых, гипотезы сгенерированы. Во-вторых, получается доказательство для гипотез (подходы включают ключевое слово, соответствующее текстовым источникам на естественном языке). Проблема состоит в том, чтобы снять неоднозначность типов (например, «человек» по сравнению с «место») объектов и предикатов. Эта проблема может быть частично решена с помощью существующих онтологий и баз знаний, таких как YAGO [14].

Отход назад к легким подходам также произошел в области веб-сервисов. Обычно потребитель службы находит веб-сервис, который поставщик услуг зарегистрировал в центральном реестре, и затем взаимодействует с веб-сервисом для его

исполнения. Описания служб семантической сети, в дополнение к основному синтаксическому описанию WSDL, требуются для нахождения и сравнения поставщиков услуг, для согласования и выполнения служб, для их составления, ввода в действие и контроля, а также для того, чтобы обеспечивать взаимодействие неоднородных форматов данных, протоколов и процессов. Традиционно, семантика веб-сервисов описывается с помощью тяжелых онтологий, таких как язык моделирования веб-сервисов (WSMO) [15] или OWL-S [16] на основе выразительных языков онтологии. Эти службы, как предполагается, будут связываться тяжелыми сообщениями XML согласно SOAP. Поскольку подход моделирования «сначала семантика» не получил распространения на практике, и поскольку большинство веб-сервисов в настоящее время реализовано с помощью легких интерфейсов REST, новые подходы вместо этого продвигают более легкие семантические описания веб-сервисов: восходящая аннотация и связывающий подход под названием «Связанный сервис» («Linked Services»). Связанный сервис описан легкими онтологиями, главным образом, с помощью RDFS и нескольких конструкций OWL; например, Связанный объединенный язык описания службы (USDL) [17], реализации USDL с помощью связанных данных [18], которая обобщает Язык описания веб-сервисов (WSDL) [19].

2.3 Масштабируемость

Один аспект, в котором приложения и Больших данных, и Семантической сети отличаются от других приложений онтологий – это масштаб рассматриваемых проблем. Вместе с ограничениями производительности, масштабируемость оказывает глубокое влияние на то, как требуемые онтологии представляются и используются. Объединенные требования к размеру и скорости обработки требуют компромиссов между выразительностью языка онтологии и эффективностью механизма принятия решений для этого языка. Разработка крупномасштабных методов обоснования должна облегчить некоторые из этих проблем. Другой подход заключается в использовании гибридных методов, включающих семантический контент онтологии, не требующих явной аксиоматизации онтологии, используемой для принятия решений. Дальнейшее развитие этого подхода предполагает использование легких онтологий, поочередно соединенных с более сложными онтологиями, для включения по требованию (и дополнительно) более точного обоснования гранулированного семантического контента, т.е. помещение в прагматическую практику понятия модульного принципа онтологии.

Масштабируемость относится к использованию онтологий на наборах Больших данных, но также может быть применена к ситуациям, в которых сами онтологии являются слишком большими для стандартных систем обоснования. Даже редактирование и визуализация крупномасштабных онтологий ставит новые проблемы перед существующими инструментами онтологий.

2.4 Вопросы

- Какая комбинация разработки онтологии и методов обоснования будет использоваться для решения проблем Больших данных?
- Нужно ли пытаться представлять большие объемы знания с помощью онтологий? Могут ли хотя бы легкие онтологии масштабироваться к Большим данным? Или было бы достаточно, как предлагают в области биологии, использовать онтологии для того, чтобы аннотировать Большие данные терминами?

3. Совместно и повторно используемый семантический контент

Повторное использование семантического контента может быть определено как возможность включать контент из одного источника в другой, или просто использовать полезное содержание из некоторого источника. Повторное использование может совпадать с исходными намерениями разработчиков или может расширяться в неожиданных направлениях. Понятие семантического повторного использования подобно повторному использованию в разработке программного обеспечения. Оно требует, чтобы понятия, включая отношения, аксиомы и правила, предположения и выражения контента, соответствовали потребностям и могли быть включены в реализацию разработок пользователя. Повторное использование применяется с подобными целями во всех отраслях, связанных с проектированием: уменьшить трудоемкость разработки и её стоимость (путем минимизации количества труда), повысить привлекательность (увеличить доход от инвестиций) исходного контента и улучшить его качество. Так как распространение повторного использования предполагает, что ошибки идентифицируются и устраняются, возникает эффективный цикл, особенно когда использование разнообразно и все дефекты и изменения полностью задокументированы и объяснены.

3.1 Что ограничивает повторное использование онтологии?

С самого начала разработка допускающих повторное использование онтологий с обеспечением совместного доступа была одним из приоритетных направлений в области Прикладных онтологий. Много усилий ушло на разработку основополагающих (верхних) онтологий (таких как DOLCE и BFO) или создание широких моделей предметной области (таких как семантическая сеть для Земли и экологической технологии (SWEET) [20] как средства обеспечения повторного использования. Кроме того, мы в настоящее время видим быстрое развитие (иногда накладывающихся), описанных в экосистеме Связанных открытых словарей (LOV - Linked Open Vocabularies) [21]. Все же процент повторного использования существующих словарей и онтологий кажется нам довольно низким. В этом разделе мы исследуем несколько возможных причин возникновения этой ситуации и определим, создают ли они фундаментальные препятствия повторному использованию онтологий.

3.1.1 Несоответствия и непонимания

Одной из потенциальных причин небольшой частоты повторного использования онтологий является отсутствие подходящей онтологии, т.е. когда разработанные онтологии не удовлетворяют потребностям пользователя с новыми

приложениями. Определение, соответствует ли существующая онтология потребностям пользователя, приводит к обсуждению жизненного цикла онтологии – это тема онтологического саммита 2013[4], в котором онтологии рассматриваются как спроектированные артефакты в контексте разработки требований, анализа онтологии, проектирования, оценки и развертывания. В частности, пользователи должны понять, как требования для онтологии могут быть выработаны с помощью такого метода, как формирование вопросов компетентности. Существует много возможностей для повторного использования, но сначала должны быть определены область и вопросы её компетентности. Часто повторное использование не работает, потому что оно предпринято до определения требований, базовых понятий и предположений (управляющих созданием контента). В этом случае существует не реальное несоответствие, а непонимание - могут быть онтологии, которые пригодны для повторного использования, но пользователи не понимают, что существующие онтологии действительно фактически соответствуют их потребностям.

Существующие онтологии могут быть не предназначенными для повторного использования и могут быть реализованы способами, делающими повторное использование затруднительным (например, из-за несоответствия между фактической общностью/спецификой понятий и их метками и именами). То, что является подходящим для определенного приложения, может быть более или менее подходящим для способа, которым кто-то намеревается снова использовать эти понятия. Метки, в частности, могут вызвать недоразумения, так как разработчик онтологии, возможно, использовал очень общую метку для понятия, которое является специфическим для контекста другого приложения.

3.1.2 Поиск правильной онтологии

Возможно, подходящая онтология существует, но её тяжело найти. Где пользователи могут её найти? Исследования, такие как LOV и открытый репозитарий онтологий (OOR) [23] начинают находить подходы к решению этой проблемы. Конечно, необходимо нечто большее, чем простой реестр онтологий - должны также быть способы, позволяющие организовать и аннотировать онтологии надлежащими метаданными так, чтобы пользователи могли найти онтологии, соответствующие их требованиям (см. предыдущий раздел). В дополнение к понятиям, таким как происхождение (описанным в работе Словарь метаданных онтологии (OMV- Ontology Metadata Vocabulary) [24]), такие метаданные должны будут также включать более широкий диапазон функций. С точки зрения разработки, метаданные должны включать вопросы о компетентности, онтологические обязательства и проектные решения, использовавшиеся в разработке онтологии и существующих отображениях, а также связи с другими онтологиями. С точки зрения реализации, функции должны включать поддержку выводов, языки, правила и соответствие внешним стандартам, системам или приложениям, с которыми использовалась онтология. Наконец, с технической точки зрения, важно включать информацию об оценке, полученной онтологией. Таким образом, метаданные онтологии могут помочь выбрать требуемую онтологию из доступных в репозиториях.

Даже когда потенциальная онтология была найдена для повторного использования, возникает проблема оценки, проверки качества и доверия. Многократное использование онтологии просто потому, что в ней используется подходящий набор ключевых слов, несомненно, приведет к проблемам.

3.1.3 Онтология не подходит...

Допустим нужные онтологии существуют, но у них есть проблемы, препятствующие тому, чтобы они были снова использованы для определенных задач. В некоторых случаях существующие онтологии изначально не предназначены для повторного использования и могут быть реализованы способами, делающими повторное использование трудным, включая недостаточную семантическую явность и несоответствия между фактической общностью/спецификой понятий и их метками и именами.

Онтология может быть неполной, т.е. она может не удовлетворять всем требованиям для определенного приложения. Существующие онтологии обычно недостаточны для новой области или приложения и должны быть расширены или изменены. В этом отношении важно помнить роль вопросов о компетентности в выборе онтологии для повторного использования. Если пользователи в состоянии найти совпадение между своими вопросами о компетентности с ответами о компетентности, предоставляемыми существующими онтологиями, они могут лучше определить, как онтологии могут быть снова использованы или расширены для удовлетворения всех требований.

Наконец, онтология может быть на языке представления знаний, который не подходит пользователю, так что, даже если онтология удовлетворяет всем требованиям вопросов компетентности, она может не удовлетворить дополнительным требованиям, являющимся результатом использования онтологии в проектировании информационной системы в целом, её развертывании и работе. В этом случае важно понимать, что повторное использование онтологии может произойти с помощью универсальных языков представления знаний. Например, имея онтологию на выразительном языке, таком как Common Logic, мы можем определить менее выразительные версии или фрагменты онтологии на других языках представления, таких как RIF, OWL и RDF. Каждый из этих фрагментов может тогда быть снова использован более широким спектром приложений. В частности, приложения в Больших данных могут получить положительный результат от использования легких онтологий и методов. Эти представления на менее выразительных языках описания онтологий могут обеспечить локальные онтологические решения, при этом предоставляя преимущества соответствующей семантики при поддержке принятия решений в рамках их намеченного использования. Идея состоит в том, чтобы найти части онтологии для повторного использования, которые имеют соответствующую выразительность.

3.1.4 Модульный принцип

Во многих случаях пользователю требуется только часть онтологии, что приводит к появлению проблемы поддержки частичного повторного использования. Очевидный подход к решению этой проблемы – реализация модульного принципа,

[4] См. перевод коммюнике Онтологического саммита 2013 в журнале «Онтология проектирования» № 2(8), 2013

но модуляризация существующих онтологий сама по себе остается трудной проблемой. Разбиение, расширение, специализация, интеграция, выравнивание и адаптация маленьких модульных онтологий должны стать частью методологии разработки онтологии. Подходы, поддерживающие спецификацию отношений между модулями онтологии, такие как OntoIOp [25], направлены на решение этих проблем. Репозитории онтологий могут также быть в состоянии предоставить более выраженную поддержку для модуляризации онтологий по мере их загрузки. Разработка онтологии, редактирование и средства просмотра могут поддерживать модульный принцип лучшими эффективными представлениями и работать с наборами модулей онтологии.

3.1.5 Интеграция

Повторное использование обычно требует интеграции множественных онтологий, и проблема интеграции может быть столь же трудной, как и разработка новой онтологии. Ключом к решению является создание интеграционных модулей, объединяющих семантику повторно используемых компонентов.

Отображение онтологии играет ключевую роль в повторном использовании, когда существуют множество онтологий, которые потенциально могут быть снова использованы. Понимание, как различные онтологии в одной области (например, разные единицы измерения времени, величин или онтологии процессов) связаны друг с другом, является основной частью определения, может ли одна онтология быть интегрирована с другими, даже в случаях, когда терминология не совпадает.

Вопрос интеграции встает наиболее остро в проблеме разнообразия Больших данных, где онтологии могут решить эту проблему путем аннотации данных и метаданных. Наборы данных обычно отличаются по полноте их метаданных, гранулярности и используемой терминологии. Онтологии могут уменьшить часть этого разнообразия путем нормализации условий и обеспечения отсутствующих метаданных. Дополнительная проблема во многих приложениях Больших данных состоит в том, что терминология, использованная когда-то для одного набора данных, могла бы иметь различную интерпретацию другого набора данных, который, кажется, использует ту же терминологию, но в другое время. Для преодоления этой проблемы онтологии должны не только развиваться со временем, но также устанавливать соответствия предыдущих интерпретаций новым. Онтологии имеют потенциал для решения этой проблемы путем обеспечения стандартной модели, независимой от определенных представлений данных и терминологии, на которую могут быть отображены различные представления и терминология.

3.1.6 Разработка новой онтологии

Возможно, проще разработать новую онтологию для приложения, а не тратить время, чтобы найти возможные онтологии для повторного использования и лишь затем понять их достаточно хорошо, чтобы определить, удовлетворяют ли они требованиям пользователя. Если это действительно так, то важно будет создать новые среды разработки онтологии, лучше поддерживающие повторное использование проектов.

Использование шаблонов разработки онтологий является подходом, который может привести к непосредственно объединенному повторному использованию в методологии разработки онтологий. Путем явного получения допускающих повторное использование аспектов онтологии шаблон разработки позволяет разработчику эффективнее определять общности среди разрозненных компонентов.

Могут также быть ситуации, в которых слабые формы повторного использования являются более уместными. Например, в повторном использовании идей - условия, отношения или аксиомы определенной онтологии не используются повторно в явном виде, но они служат для принятия решений разработчиком новой онтологии относительно проектных решений. В этом подходе модификация онтологии становится методом проектирования онтологии.

Часто существуют барьеры и узкие места в использовании онтологий, с точки зрения повторного использования существующего контента или в разработке нового контента. Эти барьеры и узкие места могут проявиться из-за ряда факторов, включая:

- стоимость разработки и развертывания онтологий;
- непонимание задачи использования онтологий;
- своевременность способности обеспечить решения;
- неполное знание или навыки онтологического инжиниринга со стороны разработчиков онтологии;
- несоответствие между основными эксплуатационными характеристиками и намеченным доменным покрытием и обоснованием требований онтологий;
- использование несоответствующих инструментов на различных этапах жизненного цикла разработки онтологии;
- социологические, культурные и мотивационные проблемы среди заинтересованных лиц, разработчиков приложений, специалистов по проблемной области и онтологов.

В реальности, все вышеупомянутые факторы влияют на стоимость разработки и развертывания онтологий, и таким образом, повторное использование существующего семантического контента потенциально экономит деньги. Однако, на практике часто применяются одноразовые не онтологические решения, которые быстрее и дешевле, потому что повторное использование онтологии - значительно более дешевый метод разработки, позволяющий снизить затраты на обслуживание амортизируемых по многократным жизненным циклам приложений онтологии - не понято и не получило соответствующей поддержки сообществом.

3.1.7 Социальные факторы

Много онтологий, предназначенных для повторного использования, разработаны на английском языке, и предполагается, что все пользователи будут использовать английский язык; однако это не допустимое предположение для многих приложений. Несмотря на то, что это прагматично, идентификаторы должны быть на языке разработчика (так как это помогает в разработке и отладке), идентификаторы должны быть скрыты от конечных пользователей, которые должны быть в состоянии выбрать язык для меток, которые они видят. Когда намеченная семантика понятий в онтологии

определена в документации вместо того, чтобы быть формально полученной в аксиоматизированной онтологии, это может быть еще более проблематичным. В любом случае использование и словарей (терминов), и онтологий (понятия), объединенных вместе, позволяет специфичным для языка терминам быть отображенными на их логические понятия.

3.2 Где происходит повторное использование

Несмотря на суровую правду предыдущего раздела, у нас есть примеры успеха совместного использования и многократного применения словарей и онтологий. Например, рассмотрим проект Schema.org. Он представляет собой широко используемый (и расширяемый) словарь для описания веб-страниц. Понятия, содержавшиеся в Schema.org, полностью задокументированы, как и правила использования и расширения словаря. Кроме того, пользователи поддерживаются через блоги и дискуссионные группы. Подход, проявленный в разработке Schema.org, решает проблемы нахождения допускающего повторное использование контента, управления размером и сложностью контента, интегрирование различных частей и расширений, а так же поддержания качества и доверия. Все они являются важными проблемами, поднятыми в предыдущем разделе.

Другие примеры успешного повторного использования основываются на маленьких онтологиях и шаблонах разработки. Они могут быть в основном применимы или специфицированы к определенной области. Примеры обоих общих и частных структур могут быть найдены в Шаблонах разработки онтологии (ODP - Ontology Design Patterns) [26], в то время как проект OceanLink [27] (в рамках инициативы NSF Earth Cube) определяет проблемно-ориентированные структуры. Цель состоит в том, чтобы получить фундаментальные понятия, такие как наборы, списки, события, или, в случае OceanLink, траекторию круизного корабля.

Поскольку понятия распространены, они могут быть понятными и интегрироваться в разработку онтологий. Кроме того, они могут отображаться на данные в непересекающихся, разъединенных репозиториях и использоваться для интеграции их данных.

3.3 Лучшие практики

Что мы можем извлечь и из наших успехов, и из неудач? В следующем списке представлены некоторые лучшие практики.

- Грамотные возможности повторного использования следуют из знания требований проекта. Вопросы о компетентности должны использоваться, чтобы сформулировать и структурировать требования к онтологии как часть быстрого и гибкого подхода. Вопросы помогают изучать и структурировать в контексте области потенциального повторного использования контента.
- Планируйте формализацию. Повторно используйте контент на основе ваших потребностей, представляйте его в виде, отвечающем вашим целям, и затем подумайте, как его можно улучшить и снова использовать. Четко задокументируйте свои цели так, чтобы другие поняли, почему вы сделали данный выбор.
- Маленькие шаблоны разработки онтологии обеспечивают больше возможностей для повторного использования, потому что они имеют низкие барьеры для создания и потенциальной применимости, сфокусированы и согласованы. Они, вероятно, менее зависят от исходного контекста, в котором они были разработаны.
- Используйте модули «интегрирования» для слияния семантики вновь использованного индивидуального контента и шаблонов разработки.
- Рассмотрите повторное использование классов/понятий отдельно от свойств, от экземпляров и от аксиом. Путем разделения семантики (или для связанных данных, или онтологий) можно упростить повторное использование, ориентируясь на определенный контент, уменьшить объем необходимых преобразований и доработок.
- RDF обеспечивает основу для семантического расширения (например, OWL и RIF). Но RDF триплеты без этих расширений могут быть просто недоопределенными фрагментами данных. Они могут помочь в составлении словарей, но формализация на языках типа OWL может более формально определить и ограничить значение выражений. Это позволяет отвечать на запросы и осуществлять поддержку принятия решений.
- Метаданные лучшего качества (обеспечивающие определения, историю и любую доступную информацию о взаимосвязи) для онтологий и схем необходимы для упрощения повторного использования. Кроме того, важно отличать ограничения или понятия, которые являются категоричными (обязательными для получения семантики контента) в сравнении с теми, которые являются специфическими для области. Проблемно-ориентированное использование с практическими рекомендациями в обосновании приложений или аналитики данных также ценно. Исследования, проводимые в этой области, такие как Связанные открытые словари и несколько работ в рамках Хакатона онтологического саммита 2014, находятся в стадии реализации и должны поддерживаться.
- Необходимо улучшить управление онтологиями и схемами. Управление требует процесса, и этот процесс должен быть реализован в инструментах. Процесс должен включать открытое рассмотрение, комментирование, версионность и принятие версий сообществом.
- Явная спецификация фрагментов онтологии должна быть включена в методологии разработки в жизненном цикле онтологии.

4. Автоматизация и инструменты

Сеть данных (Semantic Web, Linked Data и Big Data) обеспечивает большие возможности для основанных на онтологии сервисах, но также и ставит проблемы перед инструментами для редактирования, использования и принятия решений с использованием онтологий, а также методов, выявляющих узкие места в разработке крупномасштабных онтологий. Разумно

начать с легких инструментов, но они не применимы для управления большими сложными онтологиями. Применение инструментов может помочь с обеспечением логической непротиворечивости, но существует еще много ошибок, которые могут быть сделаны вне логической непротиворечивости. Поддержка инструментов, которые могут идентифицировать и решить такие ошибки, находится все ещё в стадии зарождения.

4.1 Автоматизированное построение онтологии

Ключевой вопрос в использовании онтологий - автоматизированное построение онтологий. Это очень сложная задача, потому что она пытается получить и представлять семантику, которую люди получают из произвольных или иногда проблемно-ориентированных данных. Выявление и автоматизированное построение онтологий находятся все еще в младенчестве и требуют намного большего прогресса в машинном обучении (иногда называемым «глубокое обучение»), чем существует сегодня. Текущее положение с наиболее продвинутыми системами самообучения предполагает использование существующего машинного обучения, аналитической обработки текста и обработки естественного языка (а часто и всех трех одновременно) на аннотируемых или не аннотируемых данных для предоставления предполагаемых классов онтологии, отношений и свойств человеку, утверждающему кандидатов.

Несмотря на вышеупомянутое, выявление информации может значительно упроститься существующими онтологиями, так что в итоге неструктурированные данные становятся семантически аннотируемыми или индексированными и таким образом доступными для семантического поиска и навигации по онтологически описанным логическим триплетам связанных данных и семантической сети, которые могут быть добавлены для хранилищ триплетов, чтобы непосредственно упростить принятие решений по данным. В масштабе Интернета могут быть обеспечены навигация, поиск и логический вывод (например, через поиск свободного текста или запросы с использованием SPARQL) и агрегированная семантика. Автоматизация формулирования логических выводов (дедуктивного, индуктивного, абдуктивного и вероятностного) с помощью онтологий может быть разделена, распределена и распараллелена, но это может потребовать специальных инструментов (таких как реестры онтологий с сервисами и более специализированные аппаратные средства) и более длительных промежутков времени.

4.2 Инструменты для проектирования крупномасштабных онтологий

Инструменты, требуемые для проектирования крупномасштабных онтологий и поддержки семантического анализа больших данных, в масштабах Интернета от распределенных совместных инструментов разработки и обслуживания онтологии (примером которых является WebProtege), соединенных кластеров репозиториев онтологии (такие как Открытый репозиторий онтологии от онтологического саммита 2008 [28], BioPortal [29] и т.д.) и услуг, предоставляемых ими, к распределенной поддержке принятия решений. Вместе с провозглашением роста знания об онтологиях и семантических технологиях и их ценностном предложении, особенно к разработке, такие инструменты необходимы, чтобы помочь в преодолении распознанных барьеров и узких мест, описанных в предыдущих разделах.

4.2.1 Модульная архитектура онтологии

В последние годы появляется все больше онтологий модульной архитектуры и инструментов их поддержки, по крайней мере, в виде исследований и прототипов (например, Мастерская модульных онтологий (WoMO - Workshop on Modular Ontologies) [30]). Поскольку существуют потенциально разнообразные уровни гранулярности, необходимые для крупномасштабного использования онтологий, требуются инструменты и методы, поддерживающие модульный принцип и гранулярность.

4.2.2 Инструменты принятия решений в онтологии

Технологии семантической сети и связанных данных фокусируются на обеспечении семантического обогащения данных в Интернете и для этого используют онтологии разнообразными способами. Во многих случаях различные виды онтологий и правил рассуждения необходимы в пределах от классификационного обоснования до проверки непротиворечивости онтологий и триплетов, составляющих экземпляры базы знаний, к простому выводу (например, осуществляя утверждения транзитивности) и агрегации запроса SPARQL, к более сложному выводу, требующему точно выраженных правил для поддержки принятия решений и подобных приложений. Для более сложного обоснования часто необходимы гибридные инструменты обоснования, например, инструменты, поддерживающие логическое описание и обоснование логики первого порядка, и логическое, и вероятностное обоснование (например, мини-серия форума Ontolog «Онтология, правила и логическое программирование для выводов и приложений» [31]).

4.2.3 Онтология и словарь с инструментами настройки

Крупномасштабное использование онтологий для Интернета и больших данных также требует, использования инструментов, поддерживающих отображение онтологии и словаря, а так же их настройку. Как упомянуто ранее, пользователи и разработчики должны использовать свои собственные естественные языки и для разработки, и для использования онтологий. Во многих случаях те же онтологии должны будут отображаться на разнообразные словари (подход, представленный, например, в SKOS), возможно на разных естественных языках или для использования различными сообществами. Кроме того, разные онтологии или модули онтологий, должны быть отображены на другие онтологии, или иным образом выровнены, для обеспечения масштабируемой семантики. Необходимы инструменты и службы для поддержки словаря к онтологии и отображения от онтологии к онтологии (см.: семинар Согласование онтологии, (OM-2013) [32]).

4.2.4 Онтологическо-аналитические методы и гибридные инструменты

Особые требования к поддержке инструментариев выдвигают Большие данные, потому что многие аналитические инструменты, работающие над крупномасштабными данными, используют статистические и вероятностные аналитические

методы и массовое машинное обучение или гибридные алгоритмические методы (например, IBM Уотсон). Эти методы должны быть объединены с логическими и онтологическими методами разумным способом, чтобы «понять» большие данные и распространить это понимание между пользователями и приложениями, обеспечивающими поддержку принятия решений. Облачные технологии и Grid архитектура и инфраструктура часто требуются, чтобы находить существенные корреляции и образцы в Больших данных, которые могут использоваться, чтобы описывать и обогащать онтологию. Однако во многих случаях простая параллельная архитектура и вычислительные ресурсы недостаточны для того, чтобы объединить большие объемы данных со структурированными представлениями, то есть онтологиями. Таким образом, более специализированные аппаратные средства могут быть необходимы (например, Cray YarcData Urika [33]).

4.3 Вопросы

Среди вопросов, выдвинутых онтологическим саммитом 2014 относительно автоматизации и инструментов для онтологий, выделим следующие:

- Какие инструменты онтологии необходимы и когда они необходимы?
- Могут ли определение онтологии, разработка, интеграция и её повторное использование быть более автоматизированными?

5. Заключение и рекомендации

Гектор Левеск сделал приглашенный доклад в прошлом году на конференции IJCAI-13 в Пекине сообществу искусственного интеллекта (AI), и его заключительные слова могут быть полезны нашему сообществу [34]: «Мы должны избежать чрезмерного увлечения тем, что кажется нам самым многообещающим подходом. В нашей области, я полагаю, мы страдаем от того, что можно было бы назвать острым синдромом серебряной пули для AI; это усугубляется верой в серебряную пулю для AI, вместе с верой, что предыдущие верования о серебряных пулях были безнадежно наивны».

5.1 Рекомендации

1. Усилия идентифицировать значения онтологий в приложениях Больших данных имеют самый высокий приоритет, поскольку все еще существуют разрывы между сообществами Больших данных и Прикладных онтологий. Мы должны искать больше возможностей поощрять взаимодействие перекрестного сообщества.

2. Сообщество должно договориться и принять лучшие методы для допускающего повторное использование контента с обеспечением совместного доступа.

3. Разработчики Семантической сети и Больших данных должны идентифицировать функции онтологии, имеющие значение для них, т.е., те, которые они требуют в онтологии или которые они должны знать об онтологии при рассмотрении для повторного использования.

4. Разработчики онтологии и провайдеры должны рассмотреть вышеупомянутые функции и попытаться: (a) разрабатывать и/или осуществлять рефакторинг своих онтологий и методологий сообразно с этими потребностями, если это возможно, и (b) обеспечивать метаданные о своих онтологиях, указывающие на их состояние относительно этих потребностей.

5. Сообщество должно принять определение стандартных метаданных для повторного использования - документация предположений, требований, контекста, намерения, вариантов использования, истории. Репозитории онтологии и другие инструменты должны поддерживать эти метаданные и добавление более применимых метаданных будущими потенциальными пользователями и средствами анализа.

6. Должны быть разработаны инструменты для лучшей поддержки модульной разработки онтологий, их интеграции и повторного использования.

7. Более широкий массив функциональности должен быть добавлен к инструментам, включая поддержку разработки, публикации, поиска, понимания, визуализации, проверки, преобразования, интеграции онтологии в сети.

5.2 Острые проблемы

Мы можем также выделить несколько острых проблем, которые могут требовать совместных усилий и будущего сотрудничества среди трех сообществ Прикладные онтологии, Семантическая сеть (и Связанные данные) и Большие данные.

- Какие онтологии требуются приложениями Больших данных и Семантической сети?
- Каковы требования для инструментов, служб и методов, поддерживающих разработку онтологий в приложениях Больших данных и Семантической сети?
- Действительно ли масштабируемость является фундаментальной проблемой для использования онтологий в сети?
- Существенно ли отличаются проектирование и внедрение онтологий в сети от существующих практик?
- Каковы требования для инструментов, служб, методов, используемых для разработки и реализации семантического контента на Семантической сети и в приложениях Больших данных?
- Встречаемся ли мы с новыми узкими местами разработки онтологии в приложениях Больших данных и Семантической сети?
- Может проблема разнообразия в приложениях Больших данных быть адресована с помощью существующих методов для семантической интеграции, таких как визуализация онтологий?
- Какие наборы исходных данных могут использоваться для руководства будущей работой в интеграции онтологий?

Используемые источники

[1] Ontology Summit 2014 Recommended Readings and Ontology Repository. http://ontolog.cim3.net/OntologySummit/2014/readings.html.
[2] Ontology Summit 2014. http://ontolog.cim3.net/OntologySummit/2014/.
[3] Ontology Summit 2011: Making the Case for Ontology. http://ontolog.cim3.net/cgibin/wiki.pl?OntologySummit2011.
[4] PROVO, Provenance Working Group. http://www.w3.org/TR/provo-verview/.
[5] OODA Loop. http://en.wikipedia.org/wiki/OODA_loop.
[6] Joint Directors of Laboratories (JDL) / Data Fusion Information Group (DFIG). http://en.wikipedia.org/wiki/Data_fusion#The_JDL.2FDFIG_model.
[7] Chan, Eric. 2014. Enabling Enhanced OODA Loop with Modern Information Technology. Ontology Summit 2014 presentation. http://ontolog.cim3.net/cgibin/wiki.pl?ConferenceCall_2014_02_13#nid466S.
[8] IBM's Watson. http://www.ibm.com/smarterplanet/us/en/ibmwatson/.
[9] Gene Ontology. http://www.geneontology.org/.
[10] Open Biological and Biomedical Ontologies (OBO) Foundry. http://www.obofoundry.org/.
[11] Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE). http://www.loa.istc.cnr.it/old/DOLCE.html.
[12] Basic Formal Ontology (BFO). http://www.ifomis.org/bfo/. Also: http://ncorwiki.buffalo.edu/index.php/Basic_Formal_Ontology_2.0.
[13] Linked Open Terms (LOT). http://lot.linkeddata.es/.
[14] YAGO. http://www.mpiinf.mpg.de/yagonaga/yago/.
[15] Web Service Modeling Ontology (WSMO). http://www.wsmo.org/.
[16] OWLS: Semantic Markup for Web Services. http://www.w3.org/Submission/OWLS/.
[17] Linked Unified Service Description Language (Linked USDL). http://www.linkedusdl.org/.
[18] Unified Service Description Language (USDL). http://www.internetofservices.com/index.php?id=288&L=0.
[19] Web Services Description Language (WSDL). http://www.w3.org/TR/wsdl.
[20] Semantic Web for Earth and Environmental Technology (SWEET). http://sweet.jpl.nasa.gov/.
[21] Linked Open Vocabularies. http://lov.okfn.org/dataset/lov/.
[22] Ontology Summit 2013: Ontology Evaluation Across the Ontology Lifecycle. http://ontolog.cim3.net/OntologySummit/2013/.
[23] Open Ontology Repository (OOR). http://oor.net.
[24] Open Metadata Vocabulary (OMV). http://omv2.sourceforge.net/.
[25] Ontology Integration and Interoperability (OntoIOp). http://ontoiop.org.
[26] Ontology Design Patterns (ODP). http://OntologyDesignPatterns.org.
[27] EarthCube OceanLink. http://workspace.earthcube.org/oceanlink.
[28] Ontology Summit 2008: Toward an Open Ontology Repository: http://ontolog.cim3.net/cgibin/wiki.pl?OntologySummit2008).
[29] National Center for Biomedical Ontology (NCBO) BioPortal. http://www.bioontology.org/BioPortal.
[30] Workshop on Modular Ontologies (WoMO): http://womo2014.biolark.org/.
[31] Ontology, Rules, and Logic Programming for Reasoning and Applications, Ontolog Forum miniseries. http://ontolog.cim3.net/cgibin/wiki.pl?RulesReasoningLP.
[32] Workshop on Ontology Matching (OM2013): http://om2013.ontologymatching.org/.
[33] Cray YarcData Urika. http://www.cray.com/Products/BigData/uRiKA.aspx.
[34] Levesque, H. 2014. Artificial Intelligence, Volume 212, July 2014, Pages 27–35. http://dx.doi.org/10.1016/j.artint.2014.03.007.