УДК 004.738.5

Анализ XML-подхода для описания метаданных и онтологий в Semantic Web

О.Н. Половикова

Алтайский государственный университет (Барнаул, Россия)

Analysis of XML-Based Approach to Description of Semantic Web Metadata and Ontologies

O.N. Polovikova

Altai State University (Barnaul, Russia)

Обозначен и проанализирован ряд проблем, связанных с использованием языка XML для описания метаданных ресурсов и онтологий. Исследование направлено на построение системы правил, регламентирующей использование определенной технологии для описания семантических конструкций. Такой выбор определяется свойствами языка декларирования и спецификой рассматриваемых web-ресурсов.

Анализируются следующие аспекты XML-подхода для описания онтологий: ограничения и возможности языка для построения онтологий различных типов; возможность расширения языка для адекватного отображения знаний некоторой предметной области; существование и использование технологий и методов автоматического построения метаданных и онтологий; возможность проверки на корректность структурированных в онтологии знаний.

Отмечено, что язык XML не разрабатывался как специализированное средство для построения семантических конструкций, но, несмотря на это, XML-система (данные, схема, набор преобразований) может быть использована для построения абстрагированных онтологий или онтологий программных агентов. В этом случае модель представления знаний описана деревом понятий.

Ключевые слова: язык XML, онтология, семантика ресурса, словарь терминов, XML-схема, XML-преобразования.

This paper states and investigates a number of problems related to the usage of XML for description of resource metadata and ontologies. The goal of the study is to elaborate a set of rules that regulate the employment of specific technology for descriptions of semantic structures.

Such choice is determined by properties of the declaration language and specifics of web-resources. In the paper, the following aspects of XML-based approach are analyzed: limitations and capabilities of the language for building ontologies of different types; capabilities to extend the language for adequate knowledge representation of a subject domain; existence and employment of technologies and methods for automatic construction of metadata and ontologies; capabilities of knowledge validation on ontologies.

It is stated that the XML language was not designed to be a special tool for building semantic structures. However, an XML-system (data, scheme, set of transforms) is well-placed to build ontologies or ontology-abstract software agents. In this case, the knowledge representation model can be described by a tree of concept.

Key words: XML, ontology , semantics resource, glossary, XML-scheme, XML-conversion

DOI 10.14258/izvasu(2014)1.2-19

Концепцию Семантического Web выдвинул Т. Бернерс-Ли на Международной конференции XML-2000, прошедшей в 2000 г. в Вашингтоне. Уже сегодня практически все известные компании уровня IBM, Adobe или Sun Microsystems, активно используют технологию Semantic Web в своих продуктах для решения задач управления данными [1]. Всего за несколько лет идеи проекта Semantic Web и предлагаемые конкретные стандарты для реализации этих идей получили широкое распространение среди пользователей-разработчиков глобальной сети. Это также подтверждается развитием и распространением базы методов и технологий для публикации новых ресурсов с потенциальной возможностью автоматически обрабатывать

их программными агентами. Кроме этого, обработке подвергаются и уже существующие документы и данные в глобальной сети с целью раскрытия их содержания системам интеллектуального поиска и анализа.

Несмотря на глобальное распространение концепции Semantic Web, а также множество конкретных практических результатов данного проекта, существуют проблемы методологического плана, решение которых определяют пути его дальнейшего развития. В рамках данного исследования обозначен и проанализирован ряд проблем, связанных с использованием языка XML для описания метаданных ресурсов и онтологий.

Именно онтология является ключевым звеном в организации всех этапов работы со знаниями в сети Internet: хранения, поиска, анализа, представления пользователям и программам-агентам, обмена ими и между приложениями. Онтология определяется как ключевая технология развития Semantic Web, а базовым этапом проектирования онтологии является выбор средства для ее спецификации. Поэтому тематика данного исследования является актуальной и востребованной.

Выбор XML в качестве объекта данного исследования обусловлен следующими факторами:

1. Доступность и простота. Для работы с XML-кодом не требуется дополнительного программного обеспечения. Любой человек может понять, что находится в XML-файле. XML поддерживается большинством популярных языков программирования и новыми версиями СУБД [2]. Создавать семантические конструкции с использованием данного языка может как начинающий разработчик, так и автор Web-ресурса. Кроме этого, XML-код не зависит от платформы и операционной системы.

2. Практически нет ограничений на типы данных. Существует возможность описывать структуру и содержание любых типов Web-ресурсов, в том числе и специфических.

3. XML-подход обеспечивает все уровни работы с содержанием. XML-файлы включают в себя содержание ресурса, средства для получения требуемой информации по определенным правилам, а также механизмы представления найденных конструкций на стороне пользователя-клиента.

Данное исследование поможет начинающим разработчикам Semantic Web сформировать свою систему правил для определения инструментария описания семантики ресурса, опираясь на свойства языка декларирования и специфику онтологии (модель знаний, тип онтологии, уровень детализации и другие факторы). От выбора языка будут зависеть свойства будущей онтологии, а также методы и технологии, которые будут использоваться для реализации функций по работе с ней (извлечение, добавление знаний, проверка на непротиворечивость и т.д.). Разные языки опираются на различные модели представления знаний (деревья, семантические сети, фреймы и т.д.), поэтому создаваемые на их базе семантические конструкции в зависимости от декларативной реализации подчиняются своему набору правил обработки и анализа.

Развитие проекта Semantic web осуществляется сразу по нескольким направлениям. Одним из приоритетных направлений является формирование технологической базы, которая реализует данный проект. В работе, посвященной исследованиям Semantic web, выделены следующие технологические составляющие [3]:

- расширяемый язык разметки eXtensible Markup Language, XML;
- система описания ресурсов — Resource Description Framework, RDF и его надстройка язык RDFS — язык описания словарей классов и свойств Web-ресурсов;
- язык онтологий — Web Ontology Language, OWL.

Правила и практические инструкции по использованию данных языков рассмотрены в работах [1–4].

Все представленные выше языки описания метаданных ресурсов и онтологий являются декларативными, так как позволяют описать, что представляют собой Web-ресурсы, т.е. каковы их свойства, и в частности семантика. Следует заметить, что сформированный перечень основных используемых декларативных языков является неполным. Анализ материалов исследований [5, 6] позволил сделать вывод, что разработчики сетевых ресурсов и инженеры знаний определяют также декларативные языки Prolog, Actor Prolog как необходимый инструментарий для реализации проекта Semantic Web. Это, прежде всего, связано с принципиальной особенностью prolog-системы автоматически находить решение поставленной перед ней задачи (если такое имеется) таким образом, как если бы эту задачу решал эксперт-человек. Actor Prolog в дополнение к классическому языку Prolog поддерживает объектно-ориентированную парадигму и обладает развитым аппаратом для разработки интеллектуальных агентов. Возможности языков Prolog, Actor Prolog являются востребованными в Semantic Web.

Прежде чем представлять результаты проведенного анализа, следует определить понятие «онтология». Одной из используемых трактовок данного понятия в работах по «Semantic Web является определение Тома Грабера (Tom Gruber) [7]: «Онтология — это формальное, точное описание (спецификация) согласованной концептуализации». Согласно данному определению под онтологией понимается абстрактная модель представления знаний, которая является машиночитаемой, описывает систему понятий некоторой области и поддерживается (согласовано с ними) определенным сообществом (группой людей).

В отличие от других определений, на которые также основываются разработчики Semantic Web, такая трактовка позволяет:

1. Выстраивать иерархию онтологий в зависимости от уровня формализации. Это способствует повышению релевантности семантического поиска, так как одним пользователям нужны обобщенные данные, другим — конкретные факты в рамках одного набора ресурсов. Кроме этого, иерархия онтологий способна решать задачи по сравнению и связыванию онтологий через более абстрактную модель знаний.

2. Рассматривать онтологии прикладных систем и программных агентов глобальной сети. Такое допущение позволяет реализовать одну их основных задач Semantic Web — предоставить возможность программным агентам понимать и взаимодействовать друг с другом. Интерфейс такого взаимодействия будет строиться на связях между их онтологиями.

С учетом представленного автором обоснования в рамках данного исследовании будем основываться на описанном выше определении онтологии. Следующим шагом определим ключевые моменты анализа XML-подхода для описания онтологий:

- ограничения использования и возможности языка для построения онтологий различных типов;
- возможность расширения языка, за счет новых деклараций для адекватного отображения знаний некоторой предметной области;
- существование и использование технологий и методов автоматического построения метаданных и онтологий;
- возможность проверки на корректность заложенных в онтологии знаний (существование методов и технологий).

По мнению автора, именно предлагаемый набор показателей поможет разработчику принять решение о целесообразности использования определенного декларативного языка для построения онтологии в конкретной прикладной области.

Прежде чем описывать возможности и ограничения языка XML, следует заметить, что данный язык не разрабатывался как средство для построения онтологий. Несмотря на тот вклад, который вносит данный язык в развитие информационных технологий, и в частности в Semantic Web, XML (как метаязык, а не его подъязыки) не обладает инструментарием для создания многоаспектных семантических структур с различного рода связями, ограничениями и механизмами построения нового знания. Но если рассматривать абстрагированные онтологии (иерархия по уровню формализации), модель знаний которых на некотором этапе семантического поиска или анализа можно описать деревом понятий (словарь), деревом категорий (иерархия типов), в этом случае XML может конкурировать с другими специальными языками. Кроме этого, XML выступает одним из средств построения единого универсального интерфейса к знаниям семантических конструкций.

В контексте возможностей языка XML для построения онтологии, ее верификации (различные проверки на соответствие и непротиворечивость), реализации web-сервисов по предоставлению доступа к ее знаниям (создание агентов), по мнению автора, необходимо анализировать не отдельный абстрактный язык разметки, а целостную XML-систему. XML-система включает в себя сочетание трех основных составляющих: XML-файл, в котором хранятся данные, XML-схема и необходимая совокупность *преобразований*.

XML-файл данных представляет собой текстовый документ (эта особенность определяет как преимущества, так и недостатки при работе с таким ресурсом), в котором представлены сами данные, а также задекларированы их структура и смысл.

XML-схема — это XML-файл, в котором определены правила для содержания данных. Схема обеспечивает необходимую структуру для хранимых в XML-файле данных, которая требуется для выполнения программными агентами своих задач. В XML-схема определяет следующие характеристики для данных XML-файла: какие понятия (термины) должны быть в него включены, как эти понятия связаны друг с другом, какие свойствами могут/должны обладать понятия, какие типы используются для спецификации терминов и др. Таким образом, XML-схема может использоваться для построения абстрагированных онтологий (без конкретной спецификации знаний).

XML-схема также может использоваться для описания семантики специфических web-ресурсов, таких как программный модуль, программный агент (см. свойства для термина «онтология»). Следует напомнить, что подобные семантические конструкции необходимы для взаимодействия программных агентов между собой, а также для реализации ими сервисов Semantic Web (поиск, анализ, построение нового знания и т.д.). Использование схем гарантируют корректность их значений для автора и других пользователей. Пока данные в XML-файле соответствуют правилам данной схемы, любая программа-агент может использовать эту схему для чтения, интерпретации и обработки данных.

Преобразование, или механизм, повторного использования данных (XSLT) позволяет на основе содержимого XML-файла создавать другие структуры и документы, не используя языки программирования. Такой механизм может быть использован, например, для получения определенной выборки из набора терминов нескольких XML-файлов, удовлетворяющих определенным условиям (фильтрация, сортировка). Можно также анализировать содержание множества XML-файлов данных на схожесть понятий словаря, на различие элементов одинаковых типов (категорий) и т.д.

Но такой механизм не позволяет построить логическую цепочку рассуждений для получения нового знания.

Полученные результаты работы *преобразования* могут быть зафиксированы не только в XML-файле, например, можно использовать rtf-формат или в формат любого диалекта языка XML. Таким образом, можно сделать вывод, что XML-система позволяет формировать новое знание в зависимости от потребностей пользователей, но ее возможности ограничены по сравнению с другими специализированными средствами для работы с онтологиями (например, язык OWL и Actor Prolog).

Важной особенностью языка является возможность для автора данных самостоятельно создавать XML-файл с данными и схему с требованиями к структуре данных и к содержанию смысловых элементов. Поскольку и данные, и схема оформляются в текстовом формате, дополнительное программное обеспечение можно не использовать.

Потенциал данного языка позволяет строить на его основе специализированные подъязыки для работы с семантическими конструкциями, обеспечивая необходимый уровень детализации любого типа модели знаний. Другими словами, разработчику онтологии можно не выбирать средство для ее реализации, сравнивая свойства созданных для этих целей языков, а самому сформировать синтаксис и семантику для нового языка с учетом специфики предметной среды и модели представления знаний. XML-схема такого специализированного диалекта позволит определить правила для его структуры и содержания.

XML — не просто расширяемый язык, это метаязык, на базе которого разработаны другие востребованные языки и стандарты:

- LOM (описание информационных ресурсов в области образования);
- WSDL (язык описания Web-сервисов);
- MathML (описание математических формул и специальных символов) и др.

Используя подъязыки XML, можно описывать данные произвольного типа. Таким образом, существует потенциальная возможность создавать онтологии специфических web-ресурсов, например, химических или математических трудов с формулами, диаграммами, рисунками; электро- и радиосхемы, нотные записи и т.д.

XML позволяет также осуществлять контроль за корректностью данных, хранящихся в документах, производить проверки иерархических соотношений внутри документа. На сегодняшний день существует два способа контроля правильности XML- документа: DTD-определения(Document Type Definition) и схемы данных (XML-схема). По сравнению с DTD схемы обладают более мощными средствами для определения сложных структур данных, обеспечивают более понятный способ описания грамматики языка, способны легко достраиваться и расширяться. Кроме этого, схемы описывают правила для XML-данных средствами самого языка [8].

Поскольку XML получил широкое распространение и используется для разработки большинства web-сервисов, соответственно, развиваются и автоматизированные средства работы с ним, среди которых необходимо выделить программы для проверки корректности XML-кода, генераторы XML-схем.

Как правило, функция проверки за корректностью хранящихся данных в XML-файле обеспечивается анализаторами (parsers), также поддерживается интегрированными средами разработки XML-кода, кроме этого, есть и отдельные программы-редакторы, которые выполняют указанные задачи. Программы для проверки корректности XML определяют соответствие данных схемам и синтаксис, формирует и выводит список нарушений.

Существуют и применяются на практике генераторы XML-схем, которые могут быть использованы для автоматизированного построения абстрагированных онтологий или онтологий программных агентов, модель представления знаний которых может быть описана деревом понятий. XML-схемы генерируются из данных с XML-кодом.

Таким образом, можно сделать вывод, что XML-система в рамках решения задач по работе с семантическими конструкциями может конкурировать с другими, более специализированными для этих целей языками.

На этапе проектирования и анализа разработчик онтологии должен решить принципиальную проблему — как описать, машиночитаемо представить и проверить на непротиворечивость систему понятия некоторой области. Несмотря на существование стандартизованных языков описания знаний, полнофункциональных систем автоматического выделения семантических конструкций для различных типов ресурсов, а также реальных примеров программных агентов, решающих задачи Semantic Web, на сегодняшний день нет универсальной системы правил, регламентирующих использование той или иной технологии для описания метаданных и онтологии.

Такая система правил должна предлагать способ описания необходимых семантических конструкций в зависимости от системы разнородных условий, которые и будут определять выбор языка для декларирования знаний (расставлять приоритеты между разными технологическими средствами). Соответствие возможностей языка задачам, для решения которых разрабатываются онтология и ее программные агенты, как раз и определяет адекватность такого выбора. Мы представили ключевые моменты анализа XML-системы, которые определяют применимость данного средства к спецификации определенного класса онтологий.

Библиографический список

1. Ландэ Д. Семантический Веб: от идеи к технологии [Электронный ресурс]. — URL: http://www.visti.net/~dwl/art/sw/index1.html.

2. Язык XML: назначение и область применения [Электронный ресурс]. — URL: http://nrd.pnpi.spb.ru/UseSoft/Journals/WebCreator/webc19/xml.htm.

3. Кудрявцев Д. Технологии применения онтологий [Электронный ресурс]. — URL: http://bigc.ru/theory/km/onto_technologies.php.

4. Добров Б.В. Онтологии и тезаурусы [Электронный ресурс]. — URL: http://www.intuit.ru/studies/courses/1078/270/info.

5. Морозов А.А. Об одном подходе к логическому программированию интеллектуальных агентов для поиска и распознавания информации в Интернет [Электронный ресурс]. — URL: http://jre.cplire.ru/iso/nov03/1/text.html

6. Морозов А.А., Обухов Ю.В. Акторный Пролог [Электронный ресурс]. — URL: http://www.cplire.ru/Lab144/aprolog.pdf, свободный.

7. Gruber T. Collective Knowledge Systems: Where the Social Web meets the Semantic Web // Journal of Web Semantics. — 2008. —V. 6, № 1.

8. Печерский А. Язык XML — практическое введение. Схемы данных. [Электронный ресурс]. — URL: http://www.ods.com.ua/win/rus/web-tech/xml/part5.phtml.