Semantic Web – на пути к новому поколению информационных систем♣

© Марчук А.Г.

Институт систем информатики им. А.П.Ершова СО РАН, г. Новосибирск
mag@iis.nsk.su

Аннотация

В докладе рассматриваются концептуальные и прикладные аспекты реализации идей Semantic Web. Оптимизм автора относительно возможностей предложенной консорциумом World Wide Web системы стандартов основывается на изначально распределенной парадигме формирования информационного пространства, связывающей отдельные группы данных и описаний в содержательную сетевую конструкцию, позволяющую выполнять как типовые поисковые и манипуляционные действия, так и расширять действия в сторону логического вывода. Автор постарается показать, что современные технологии и методологии, в том числе идеи и наработки Semantic Web, приводят к созданию прикладных информационных систем нового поколения и, соответственно, к формированию нового поколения информационных технологий. Доклад также частично отражает текущее состояние исследований и разработок, проводимых в Институте систем информатики им. А.П.Ершова СО РАН.

1 Концепция

Инициатива Semantic Web [1] была сформулирована одним из законодателей в области современных информационных технологий – консорциумом W3C и, естественно, сразу привлекла к себе внимание. Справедливости ради надо отметить, что ряд идей и рекомендаций, предложенных еще в 1998 году и получивших развитие как RDF (Resource Description Framework), не привлекал заметного внимания специалистов до выхода в свет концептуальной статьи классика Web'а Тима Бернерс-Ли с соавторами с броским заголовком The Semantic Web [2]. Вслед за авторами статьи можно отметить, что нынешнее «море» информации в Интернете является «бессмысленным» с точки зрения программ и ориентировано, в основном, на использование этих ресурсов человеком, что заметно ограничивает развитие информационных систем и информационных технологий. Предложенная идея «подтащить» смысл данных к самим данным и при должном уровне формализации использовать смысл вместе с данными – действительно революционна. Более ранние построения, связанные со схемами данных и метаинформацией, решали лишь частные задачи и не порождали качественного продвижения, наметившегося сейчас.

Для формулирования концепции рассмотрим две модели:

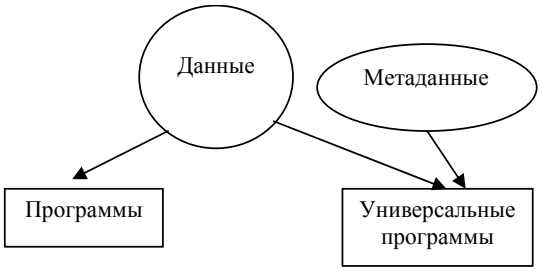

Рис. 1.

На рис. 1 изображены данные и программы, их обрабатывающие. Если рассмотреть традиционную пару «данные программы», нарисованную в левой части картинки, то можно задать вопрос: где находится смысл обрабатываемых данных? Ответ очевиден – смысл данных «знают» программы обработки, именно в их коде «зашито» понимание этих данных. В правой части рисунка появляются метаданные, т.е. данные о данных. В метаданные мы можем попытаться вложить смысл данных или часть смысла. Если метаданные обрабатывать вместе с данными, то программы смогут стать универсальными, что явно выглядит как прогрессивный фактор.

Труды 7ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005, Ярославль, Россия, 2005.

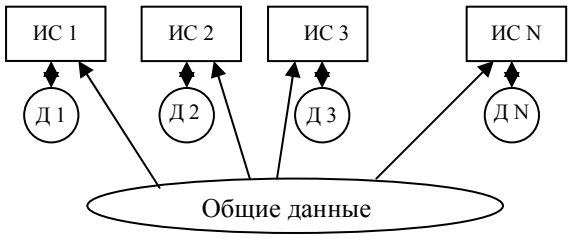

Рис. 2

На рис. 2 изображены различные информационные системы ИС1, ИС2,…, работающие каждая со своим набором данных. Однако, как легко видеть, практически каждая информационная система, сколь бы специализированной она ни была, в свою обработку вовлекает данные, являющиеся отражением традиционных сущностей реального мира, таких, как персоны, организации, события, адреса и т.д. Назовем эти данные неспецифическими, в противовес специфическим для каждой ИС «своим» данным, отражающим предметные особенности этой информационной системы. Правильным построением информационных систем в соответствии с рассматриваемой моделью является двухуровневое разбиение данных на общие (неспецифические) и локальные (специфические), как это изображено на рис. 2. Проблема заключается в том, что традиционные технологии не позволяют (точнее – не помогают) реализовывать такую архитектуру. Двухуровневость разбиения данных – достаточно условна. Понятно, что в развитых случаях обобществлению могут подвергаться и специфические данные, а в целом общее поле данных может иметь некоторую структуру.

Информационные системы нового поколения видятся как распределенные системы, опирающиеся на множественные базы «осмысленных» данных, содержащие неспецифические общие данные, неспецифические приватные данные, специфические для информационной системы общие и частные данные, спецификации модели мира, предметной области и задачи. Совместимость данных и описаний должна обеспечиваться общей методологией, едиными стандартами. Программный код таких систем должен быть универсальным, интерпретирующим произвольные спецификации и данные, или специализированным, сгенерированным в стиле смешанных вычислений из универсального решения путем фиксации нужного контекста.

Наиболее адекватной парадигмой для построения таких систем является подход Semantic Web, сформулированный консорциумом W3C, разрабатывающим также соответствующие стандарты, такие как RDF, RDFS, OWL. База данных при таком подходе формируется из (распределенного) множества семантических сетей, построенных по простым правилам, включающих в себя данные в виде высказываний, «склеивающихся» в единый граф через склейку по уникальным идентификаторам сущностей и разделенных по предмету и степени доверия через RDF-документы и пространства имен. Семантическими сетями описываются данные, метаданные, схемы данных, другие структурные построения, требующиеся для целостного описания информационной системы. Если технологическое решение, позволяющее объединять данные из разных источников, уже имеется (RDF), то единого подхода для отображения фактов реального мира (модели мира) пока не создано. Также непроработанным является вопрос спецификации функционирования информационной системы и интерфейсов.

2 Общая архитектура

Логически архитектура клиент-серверных систем нового поколения состоит из распределенной системы RDF-данных (документов), опубликованных в Internet/Intranet, интерфейсов к данным, позволяющим человеку и машинам использовать и изменять, в пределах своих полномочий, эти данные, а также универсальных и прикладных агентов (процессов, задач, демонов), порождающих предметную активность, решающих задачи обеспечения целостности данных и корректности функционирования программной среды.

Физически данные могут пребывать в разных формах существования, хранения и обработки. Например, возможной формой реализации больших систем данных может быть реализация через реляционную базу данных, помещенную в реляционную СУБД. Типовым средством взаимодействия распределенных программных компонентов предполагается технология Web-сервисов.

Потенциальная распределенность данных потребовала осмысления представлений об информационном пространстве. Была предложена и опробована логика публикации документов и их связывание через метаинформацию, позволяющие разделить идентификацию документов и их координаты.

Это позволило решить две важных задачи:

- определение координаты оригинала документа по его идентификатору в условиях распределенного реестра;
- использование копии документа вместо оригинала.

В целом подход описан в работах [3, 4].

Достоинством идей, заложенных в основу Semantic Web, является гибкость, комбинируемость и относительная простота базовых блоков, из которых может быть набрана та или иная информационная система. К таким блокам относятся: менеджер документов, универсальный

просмотрщик-редактор, система исполнения сложных запросов (в качестве формализма задания запросов пока используется XML-версия языка SPARQL).

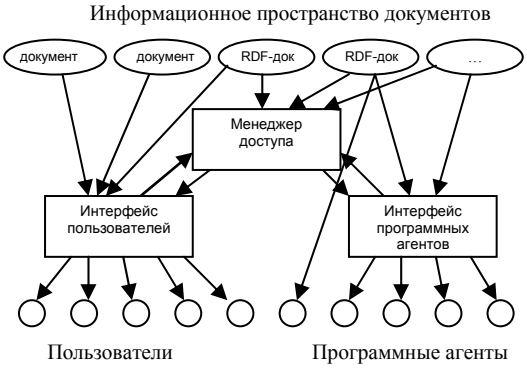

Рис. 3

Приведенный рисунок иллюстрирует архитектуру, построенную на распределенной системе опубликованных документов, в том числе – RDF и OWL-документов.

3 RDF

Вначале предполагалось использовать базовый стандарт RDF с обеспечением адаптации других видов структуризации данных к логике RDF и к выбранным политикам (пространства имен, изменчивость схем, синхронизация асинхронного доступа). Схему данных предполагалось описывать средствами RDFS. Сразу же мы отказались от попыток решения задачи «взаимопонимания» систем, смысл данных которых описан разными схемами данных или онтологиями. Ориентация была взята на описанный в этой работе подход по формированию двухуровневой модели мира с иерархией классов сущностей, при этом предметная специализация строится как совместимое расширение базовой модели.

В результате экспериментального проектирования информационных систем выяснилось, что RDF в целом является адекватным и удобным механизмом структуризации разнохарактерной информации. Однако есть и проблемы предлагаемой модели и ее технического воплощения. Во-первых, RDF-документы, если не предпринимать специальных и несколько искусственных мер, текстуально существенно больше по объему, чем аналогичные прямые XML-построения. И это при том, что XML не является образцом компактности записи данных, скорее – наоборот.

Во-вторых, графовая модель RDF-сети, являющаяся базовой в семантике RDF, не так удобна для работы, как хотелось бы. Сведение способов структуризации к базовому примитиву «субъект – предикат – объект» потребовало некоторых «жертв», целесообразность которых не является очевидной. Это видно, например, в появлении «непомеченных» узлов, структурированных литералов, сомнительных механизмов создания неупорядоченных и упорядоченных множеств и списков. С другой стороны, текстовое XML-представление, которое является достоинством подхода, недостаточно однозначно, что не всегда удобно. Например, не понятно, как после преобразования RDF-документов в граф и редактирования этого графа можно при записи семантической сети в файл восстановить вид документа, близкий к исходному, если в исходном документе были использованы разные виды сокращений.

В результате эксперименты с использованием RDF для информационных систем проводились при весьма сильных ограничениях. Часть из этих ограничений теперь уже снята, но многие остались. До сих пор созданные нами системы не способны адаптировать произвольные RDF-документы, использующие все возможности и сокращения, допущенные в стандарте.

4 Логика структуризации данных, модель мира, онтологии

Легко показать, что для порождения единого смыслового пространства сколь-нибудь содержательного толка недостаточно только средств структуризации, присущих RDF. Ситуация в чем-то аналогична системам программирования на базе языков высокого уровня. Язык действительно дает универсальное средство, в том числе – и определения библиотек (подпрограмм, функций, классов и др.). Однако совместимости библиотек, стихийно проектируемых разными командами, чрезвычайно трудно добиться, опыт показывает, что нужны стандартные библиотеки.

Одной из целей инициативы Semantic Web, если не главной, является описание смысла используемых данных и обработка формально определенного смысла вместе с данными. Такие определения смысла данных ныне принято называть онтологиями. Проблема заключается в том, что пока не существует однозначной процедуры формулирования смысла данных (онтологии), а значит, даже для одной и той же задачи он может быть определен по-разному, со сложным переплетением используемых понятий. Современная теория не дает надежного способа сочетания данных, определенных разными онтологиями. Поэтому наш подход предполагает наличие единственной базовой онтологии относительно общего поля неспецифической информации и ее расширений, необходимых для адекватного описания специфики предмета информационной системы.

Такую онтологию естественно назвать моделью мира, подчеркивая ее выделенную роль относительно моделей предметных областей. Проблема построения модели мира не так проста,

как может показаться. Свидетельством тому отчасти является тот факт, что метаинформационные построения типа Dublin Core, GILS и др. не порождают полноценную схему данных и, как правило, отражают лишь библиографический взгляд на описываемые ресурсы.

Последовательный переход от эксперимента к эксперименту показал возможность построения базовой онтологии неспецифической информации достаточно широкого класса, охватывающей такие понятия, как вещи, персоны, документы, организационные системы, географические образования, коллекции. В группе отношений присутствуют отношения родства, коммуникационные возможности, именование, датирование, отражение, ролевые отношения в организационных системах, отношения принадлежности. Как выяснилось, одним из наиболее важных в исторических базах данных является отношение датирования, позволяющее, с одной стороны, фиксировать темпоральность в данных и не «стирать» устаревшие данные, а помещать отметки конца системы, действия или отношения и, с другой стороны, неявно вводя в данные событийность и возможность генерации планов, протоколов и хроники. Впрочем, датирование важно и в других классах информационных систем, например, в системах делопроизводства.

5 Прикладные разработки

Возможности, которые дает концепция, еще только осмысливаются коллективом разработчиков. Однако уже сейчас новые решения закладываются в модернизацию ранее разработанных информационных систем, например, Электронного архива академика А.П.Ершова [5, 6], проводятся эксперименты с малыми информационными системами, например, переработке подверглась информационная система обеспечения работы кафедры программирования ММФ НГУ. Ведутся также большие новые проекты с использованием сформированного подхода и разработанного программного обеспечения. К таким системам относится проект «Хроника Сибирского отделения» и система делопроизводства академического института.

В качестве платформы используется программное обеспечение фирмы Microsoft. Для систем, построенных с использование реляционных таблиц, применяется SQL-server. При этом RDF формируется как экспорт для объединения информационных ресурсов.

В других системах RDF используется как для хранения документов, так и для организации обработки. Программное обеспечение реализовано в .NET и написано на языке C#. Существенная часть кода пока выполнена в виде XSLT-преобразований. В качестве базового формализма задания запросов к RDF-сети используется SPARQL.

Схемы данных для различных созданных экспериментальных и производственных систем пока логически недостаточно согласованы и выражены в разных формализмах. В экспериментах с RDF-моделями таким формализмом является специальная XML-структура, однако идет перевод систем на единое OWL-описание со специализированными расширениями.

Полученные результаты вселяют оптимизм в разработчиков и исследователей и, хотя существует еще значительное количество нерешенных проблем, есть уверенность, что получающаяся конструкция представляет собой качественное продвижение в технологии информационных систем.

Литература

[1] Semantic Web. http://www.w3.org/2001/sw/

[2] Tim Berners-Lee, James Hendler, Ora Lassila The Semantic Web, Scientific American, May 2001.

[3] А.Г.Марчук, А.Е.Осипов К вопросу об идентификации электронных документов и коллекций // Программирование, N 4, 2000

[4] А.Г.Марчук Распределенные электронные архивы, библиотеки и базы данных. Препринт 122, Институт систем информатики СО РАН, 2004.

[5] Антюфеев С.В и др. Проектирование, создание и наполнение электронного архива // Электронные библиотеки: перспективные методы и технологии, электронные коллекции / Труды четвертой Всерос. конф. Дубна, 2002. – Дубна: ОИЯИ, 2002. – Т. 2. – С. 189-196.

[6] Электронный архив академика А.П.Ершова. http://ershov.iis.nsk.su

Semantic Web – on the way to a new generation of information systems

Alexander Marchuk

The paper presents conceptual and applied aspects of implementation of Semantic Web approach. Proposed by World Wide Web Consortium, group of standards is based on initially distributed paradigm of information space formation. Separate pieces of data, metadata and definitions are united into the data net, which allows traditional search and transformational actions as well as logical deduction to be perfomed. The goal of paper is to show that modern technologies and methodologies, including Semantic Web ideas lead to a new generation of information systems and information technologies. The paper partially reflects current state of research and development in A.P. Ershov Institute of Informatics Systems, SB RAS.

♣ Работа поддержана грантом РФФИ № 03-07-90330в