Semantic Web как новая модель информационного пространства интернет

Ф.И. Андон, И.Ю. Гришанова, В.А. Резниченко

ИПС НАНУ

УКРПРОГ2008, *

Экспоненциальный рост количества информации

Исследования IDC: к 2008 году количество информации составит 5444 петабайт.

Количество информации, опубликованное в книгах – порядка 200 петабайт.

Проблема современности: информационное переполнение, обусловленное ростом объема информации, представленной в интернет

Интернет: сеть компьютеров, объединенных каналами и использующих протоколы (TCP/IP) для связи. Изобретатели: Vint Cerf и Robert Kahn

Веб: сеть сайтов, использующих гиперссылки для переходов от страницы к странице. Изобретатель: Tim Berners-Lee

Тенденция современного развития Интернет - это переход от документов, *читаемых компьютером* (machine readable) к документам, *понимаемым компьютером* (machine understandable).

Tim Berners-Lee «The Semantic Web». Scientific American, 17 мая 2001

«Semantic Web является абстрактным представлением данных во Всемирной Паутине, основанное на стандартах RDF и других имеющих распространение стандартах. Проект разрабатывается Консорциумом W3C в содружестве с большим количеством исследователей, ученых и промышленных партнеров»

(Semantic Web Homepage, W3C)

Семантический Веб это сеть информационных узлов, связанных друг с другом таким образом, что имеющаяся информация может легко обрабатываться компьютером.

Семантическая паутина — это надстройка над существующей Всемирной паутиной, которая призвана сделать размещённую в ней информацию более понятной для компьютеров. (Википедия)

Хроноло
гия

- **1994**. Создание **W3C**, которым разработаны стандарты:**HTML, URL, XML, HTTP, PNG, SVG, CSS**

- **1998**. Тим Бернерс Ли публикует план семантического веба (Semantic Web Roadmap)

- **1999**. W3C создает группы по проектированию семантического Веба, публикуется первая версия **RDF**

- **2000** Агентство Министерства обороны США (DARPA) начинает исследования по описанию онтологий (проект **DAML+OIL**)

- **2001**. В журнале Scientific American Тим Бернерс Ли публикует описание Семантического Веба

- **2004**. Выпущена новая версия **RDF** и схема **RDF-S**, представлен язык описания онтологий **OWL**

- **2006**. Представлена версия языка запросов **SPARQL** (candidate reccomendation)

Сценарий будущего

Человек – своему веб-агенту:
- закажи для меня эту книгу в ближайшей библиотеке
- Посмотри на расписание электричек и мое раписание и выбери билеты в театр, чтобы я мог успеть после работы
- Микроволновка, сходи на сайт призводителя продукта и загрузи оптимальные параметры подогрева
- найди мне рецепт пиццы и закажи в ближайшем магазине все необходимые ингредиенты, чтобы они были доставлены к моему приходу с работы. Подбери к этому вино. И, кстати, я не люблю «брют».

Неформальный план: Что нужно сделать, чтобы осуществить вышеприведенный сценарий будущего?

Наивный план.

1. Разработать МЕГА-язык, на котором можно описать все знания человечества и который был бы понятен компьютерам

2. Под угрозой смерти заставить ВСЕХ веб-мастеров снабжать сайты переводом на этот язык

3. Написать программы, работающие со знаниями на этом языке (обработка запросов, логический вывод)

(Юрий Лифшиц, Лекция «Семантический Веб», ПОМИ РАН СпбГУ ИТМО, 2006)

План Тима Бернерса Ли:

Нужно последовательно разработать:

1. Синтаксис для представления знаний, использующий ссылки на онтологии (сделано: RDF)

2. Язык описания онтологий (сделано: OWL)

3. Язык описания веб-сервисов (начато: WSDL, OWL-S, 2007 г – SML – Service Modelling Language)

4. Инструменты чтения/разработки документов семантического Веба (начато: Jena, Haystack, Protege)

5. Язык запросов к знаниям, записанным в RDF (начато: SPARQL)

6. Логический вывод знаний (не сделано)

7. Семантическая поисковая система (начато: SHOE)

8. Агенты семантического Веба (не сделано)

Технологии:

- Семантический поиск
- Вопросо-ответные системы
- Агенты в семантическом Вебе
- Объединение знаний (интеграция баз данных)
- Всепроникающие вычисления (ubiquitous computing)

Семантический Веб есть многодисциплинарная тема, охватывающая теории и методы из трех областей:
1. логика – формальные структуры и правила логического вывода
2. онтологии – описания типов сущностей, относящихся к предметной области
3. Теория моделей

Джон Сова

The Semantic Web is "…an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is the idea of having data on the Web defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across various applications… data can be shared and processed by automated tools as well as by people." World Wide Web Consortium

Semantic Web это есть расширение текущего Вэб, в котором информация дана с хорошо определенным значением, лучше позволяющая компьютерам и людям работать вместе. Это идея о том, чтобы иметь данные в Вэб, определенные и связанные между собой т.о., чтобы их можно было использовать для более эффективного исследования, автоматизации, интеграции и повторного использования через различные приложения…данные могут быть общедоступными и обрабатываемыми автоматическими средствами так же хорошо, как и людьми.

Тим Бернерс Ли, Decentralized Information Group AAAI, 18 July 2006

AI and SW

- SW is not AI and AI is not SW
- AI is a field; SW is a project
- SW owes a debt: used much from AI (имеет моральные обязательства)
- SW should be a great playground for AI
- AI projects should use SW to interoperate

What is AI anyway?
Lots of bits of CS including functional languages, machine learning, recognition, speech, etc., etc.?
An attempt to make smart machines?
An attempt to make machines which behave like people?
An attempt to be collectively more powerful?

What is SW?

- Data Interoperability across applications and organizations (for IT)
- A set of interoperable standards for knowledge exchange
- A architecture for interconnected communities and vocabularies

Semantic Web

Три уровня проекта Semantic Web

Сервисы Приложений

напр. сервис агентства путешествий

Базовый Сервис

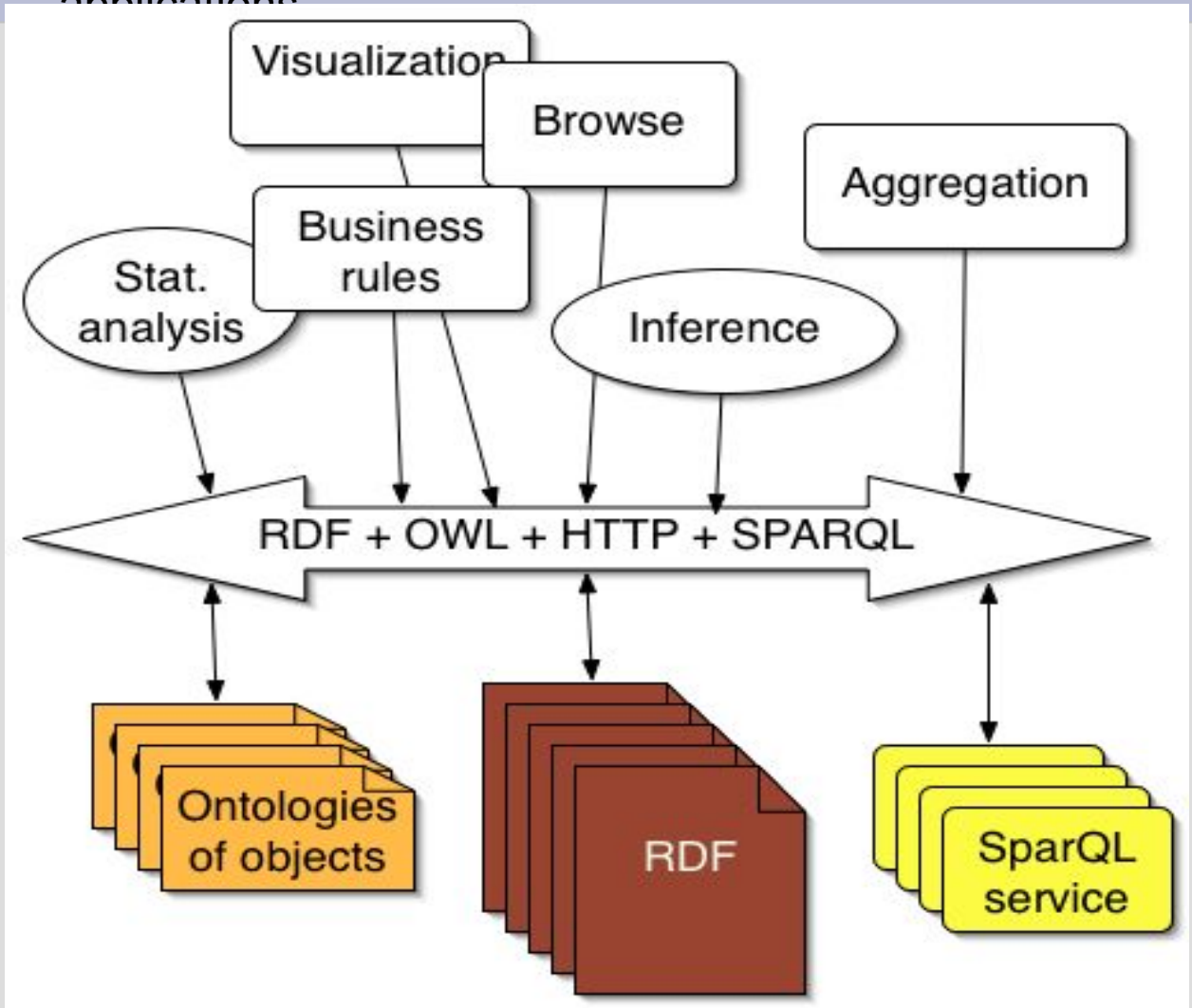
логический вывод и запросы по метаданным и онтологиям, разъяснение таких выводов (заключений), управление доверием (trust), агенты, поисковые системы, серверы онтологий.

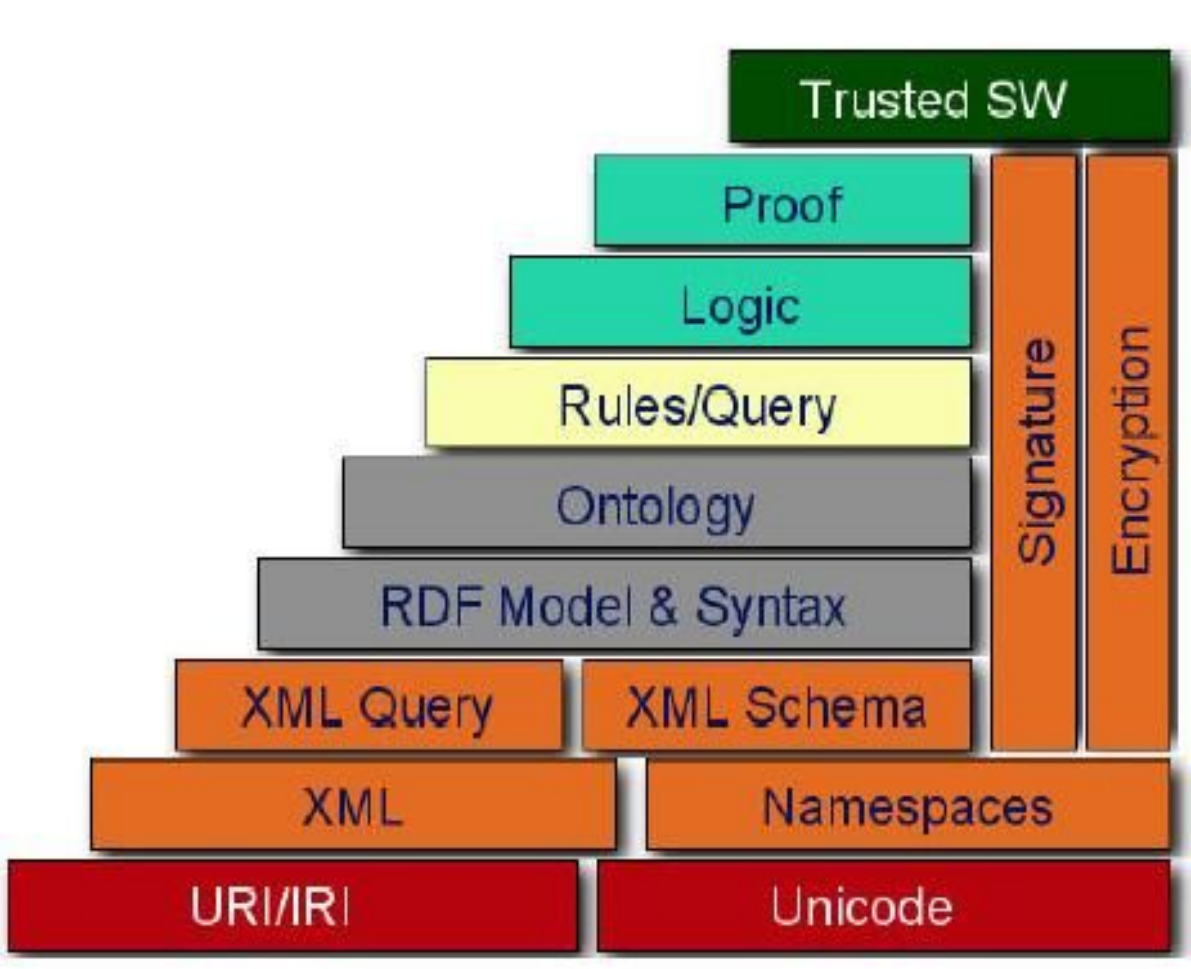
Базис

- уникальная глобальная *идентификация* ресурса;
- метаданные для декларирования фактов о ресурсах
- общий язык для выражения метаданных и знаний, реализованный при помощи онтологий для общедоступного понимания и общего словаря метаданных и правил для добавления новых метаданных и знаний.

Тим Бернерс Ли, Decentralized Information Group AAAI, 18 July 2006

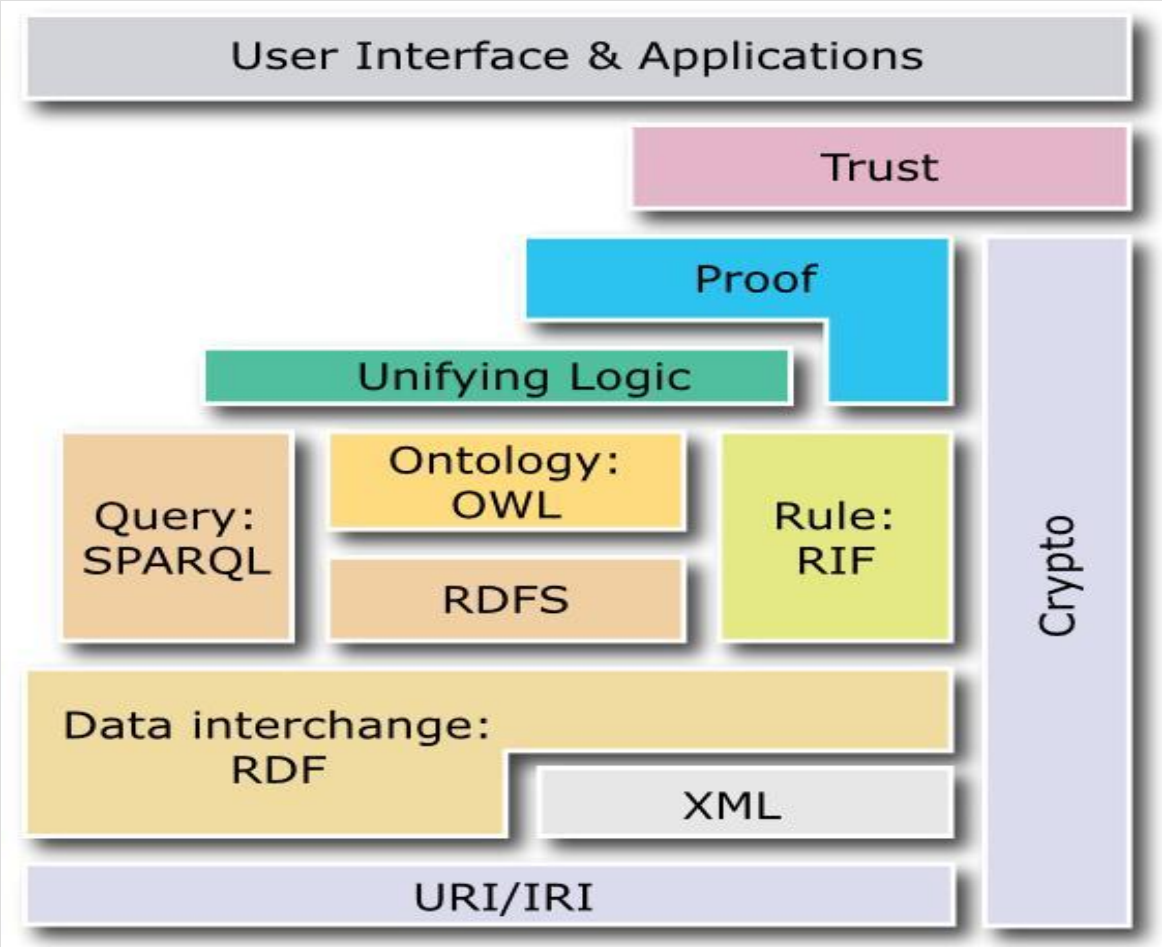
The RDF data bus
The RDF bus connects data sources and applications





Тим Бернерс-Ли, 2001

Тим Бернерс Ли, Decentralized Information Group AAAI, 18 July 2006



URI - Uniform Resource Identifier

W3C «SemanticWeb»:

“In fact, we can create a URI to refer to anything we want to talk about, including

- network-accessible things, such as an electronic document, an image, a service (e.g., "today's weather report for Los Angeles"), or a collection of other resources.
- things that are not network-accessible, such as human beings, corporations, and bound books in a library.
- abstract concepts that don't physically exist, like the concept of a "creator".

За синтаксисом URI следит комитет IETF, опубликованный им документ RFC 2396 является общей спецификацией URI. Консорциум W3C поддерживает список схем URI.

“A Uniform Resource Identifier (URI) is a compact string of characters for identifying an abstract or physical resource. “

(http://www.isi.edu/in-notes/rfc2396.txt,
Uniform Resource Identifiers (URI): Generic Syntax)

“URIs are **U**niform **R**esource **I**dentifiers, they are text strings that identify resources, or concepts. “

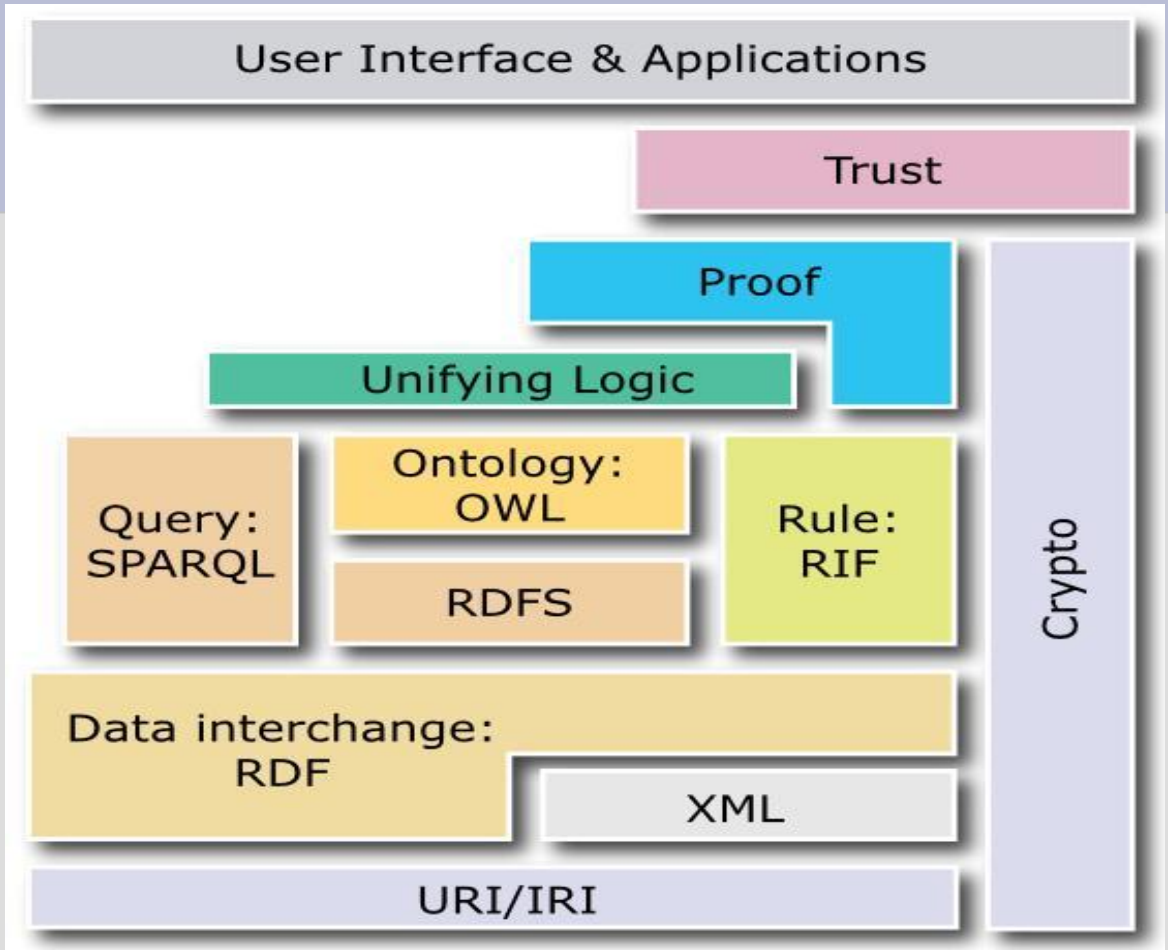
(W3C, Semantic Web for developers)

Internationalized Resource Identifiers (IRI)
интернационализированный идентификатор ресурса

“IRIs are a new protocol element, a complement to URIs. An IRI is a sequence of characters from the Universal Character Set (Unicode/ISO10646). There is a mapping from IRIs to URIs, which means that IRIs can be used instead of URIs where appropriate to identify resources. “

(W3C, 2005)

XML - eXtensible Markup Language



Тим Бернерс Ли, Decentralized Information Group AAAI, 18 July 2006

XML - eXtensible Markup Language

XML - текстовый формат для описания документов произвольной структуры

- разработан и принят в качестве стандарта в 1998 Консорциумом W3C для упрощения исполнения (реализации) и для обеспечения интероперабельности между SGML и HTML.

Он является подклассом языка SGML, однако более прост для понимания и обработки.

XML - eXtensible Markup Language

Функции XML:

- Представление синтаксиса для других языков разметки.
- Семантическая разметка Web-страниц. XML-представление может использоваться на Web-странице вместе с таблицей стилей XSL, которая определяет корректный вывод различных элементов.
- Единый формат обмена данных. XML-представление может передаваться между двумя приложениями как объект данных.

XML - eXtensible Markup Language

Тело документа XML состоит из элементов разметки и содержимого документа. Теги предназначены для определения элементов документа, атрибутов и других конструкций. Документ XML начинается с тега <?xml?>, в котором можно указывать кодовую страницу, версию языка и другую информацию.

```
<?xml version="1.0"?>
<list_of_items>
<item id="1"><first/>Первый</item>
<item id="2">Второй <sub_item>подпункт
1</sub_item></item>
<item id="3">Третий</item>
<item id="4"><last/>Последний</item>
</list_of_items>
```

```
I just got a new pet dog.
```

```
<sentence>
<person href="http://example.com/#me">
I</person> just got a new pet <animal>dog</animal>.
</sentence>
```

```
<dfgre><reghh bjhb="http://example.com/#me">I</reghh>
just got a new pet <yudis>dog</yudis>.</dfgre>
```

```
<my:sentence xmlns:my="http://example.org/xml/documents/">
   <my:person my:href="http://example.com/#me">I</my:person>
just got a new pet <my:animal>dog</my:animal>.
</my:sentence>
```

```
<sentence
xmlns="http://example.org/xml/documents/"
xmlns:c="http://animals.example.net/xmlns/">
<c:person c:href="http://example.com/#me">
I</c:person> just got a new pet <c:animal>dog</c:animal>.
</sentence>
```

XML - eXtensible Markup Language

Требования к XML-документу:
· Заголовок документа - <?xml> с необходимыми параметрами.
· В отличие от HTML, в XML каждый тег должен иметь закрывающий элемент.
· учитывается регистр символов.
· Атрибуты тегов заключаются в кавычки.
· Необходим строгий порядок следования открывающих и закрывающих тегов при использовании вложений.

В XML представлены следующие конструкции:

- Элементы данных.
- Атрибуты
- Директивы
- Специальные символы
- Комментарии
- CDATA

Атрибуты - это параметры, определяющие характеристики тех или иных элементов. Так же, как и в HTML, они задаются парой "название"="значение".
Директивы анализаторам обозначаются в XML тегами <? и ?>. Броузер использует эти инструкции для управления разбора документа.
Для отображения в тексте документа некоторых специальных символов, как, например, кавычки или угловой скобки, необходимо использовать их специальные символьные идентификаторы (&lt, &qout).
Комментарии определяются по аналогии с HTML как <!-- Текст комментария -->. Они игнорируются анализатором при разборе структуры документа.
Тегом CDATA определяется область документа, рассматриваемая анализатором, как просто текст.

Пример корректного XML- документа:

```
<?xml version="1.0"?>
<! DOCTYPE journal [
<!ELEMENT contacts (address, tel+, email?)>
<!ELEMENT address (street, appt)>
<!ELEMENT street PCDATA>
<!ELEMENT appt (PCDATA | EMPTY)*>
<!ELEMENT tel PCDATA>
<!ELEMENT email PCDATA>
]>
...
<contacts>
<address>
<street>Marks avenue</street>
<appt id="4">
</address>
<tel>12-12-12</tel>
<tel>46-23-62</tel>
<email>info@j.com</email>
</contacts>
```

XML Schema

Схемы данных (Schemas) являются альтернативным способом создания правил построения XML-документов. По сравнению с DTD, схемы обладают более мощными средствами для определения сложных структур данных, обеспечивают более понятный способ описания грамматики языка, способны легко модернизироваться и расширяться. Безусловным достоинством схем является также то, что они позволяют описывать правила для XML- документа средствами самого же XML.

XML Schema - Пример

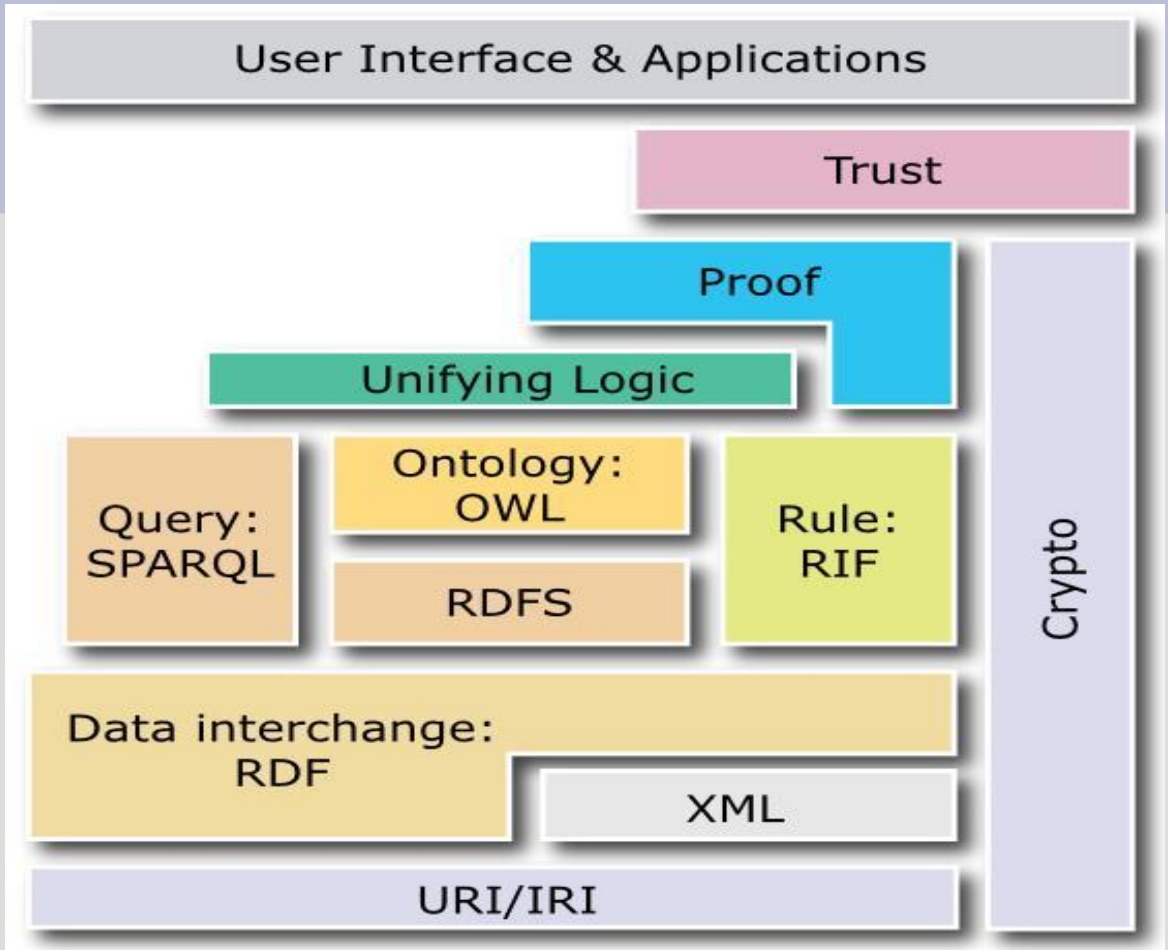
```
<schema id="OurSchema">
 <elementType id="#title">
<string/>
</elementType>
<elementType id="photo">
<element type="#title">
<attribute name="src"/>
</elementType>
 <elementType id="gallery">
 <element type="#photo">
 </elementType>
 </schema>
```

Корректно

```
<gallery>
<photo id="1"><title>My computer</title></photo>
<photo id="2"><title>My family</title></photo>
<photo id="3"><title>My dog</title></photo>
</gallery>
```

```
<gallery>
<photo id="1"/>
<photo index="2"><title>My family</title></photo>
<photo index="3"><title> My dog</title><dogname>Sharik</dogname></photo>
</gallery>
```

RDF - Resource Description Framework



Тим Бернерс Ли, Decentralized Information Group AAAI, 18 July 2006

Высказывания: Общая схема описания ресурсов

RDF - Resource Description Framework

RDF предназначен для описания предметной области ресурсов. Он был принят в 1999 году консорциумом W3C.
Первоначально назначение RDF заключалось в описании XML-ресурсов с различных точек зрения.
RDF представляет собой модель описания метаданных. Этот язык использует XML-синтаксис.
Модель Resource Description Framework имеет своей целью стандартизировать определение и использование метаданных, описывающих ресурсы Web.

RDF - Resource Description Framework

RDF включает две основные части:
1 RDF Syntax - собственно способ описания ресурсов - модель для описания объекта, рассматриваемого в качестве ресурса, как связей между ресурсами в терминах поименованных свойств и значений.
2 RDF Schema - RDFS - способ задания схем, по которым описывается ресурс - для задания структуры предметной области и аналогична диаграмме классов в UML.
На RDF можно описывать как структуру ресурса, так и связанную с ним предметную область (RDF Vocabulary Description Language).

RDF - Resource Description Framework

RDF описывает ресурсы в виде ориентированного размеченного графа – каждый ресурс может иметь свойства, которые в свою очередь также могут быть ресурсами или их коллекциями.

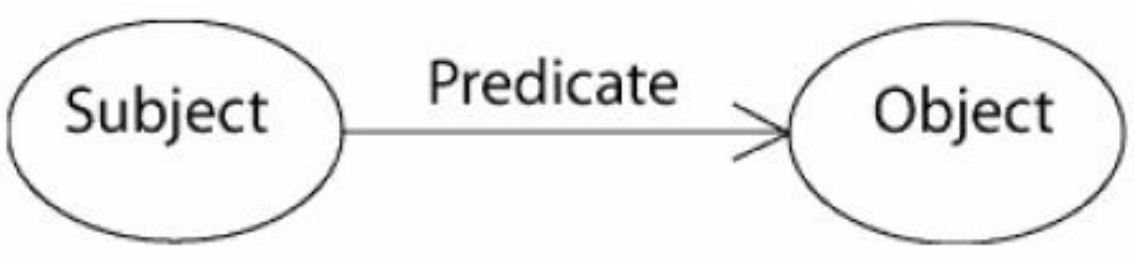
Базовый строительный блок в RDF — это тройка (триплет) «объект — атрибут — значение» A(O,V),

т.е. «объект O имеет атрибут A со значением V».

Эту связь можно также представить как ребро с меткой A, соединяющее два узла, O и V: [O]-A—>[V].

В литературе W3C: **тройка субъекта, предиката и объекта (S,P,O).**

Модель данных представляет собой набор триплетов и называется RDF-графом.

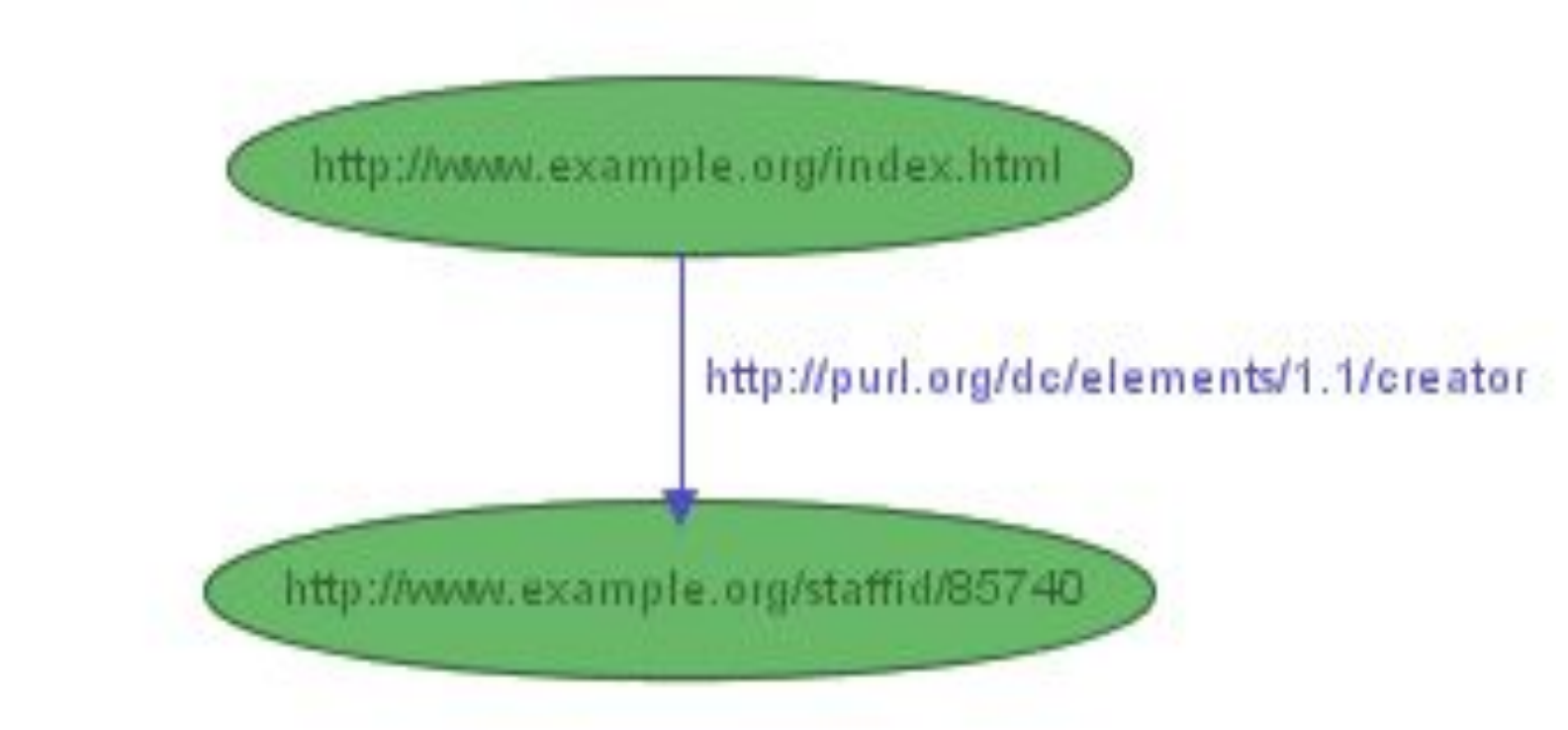


RDF - Resource Description Framework. Пример

http://www.example.org/index.html has a creator whose value is John Smith

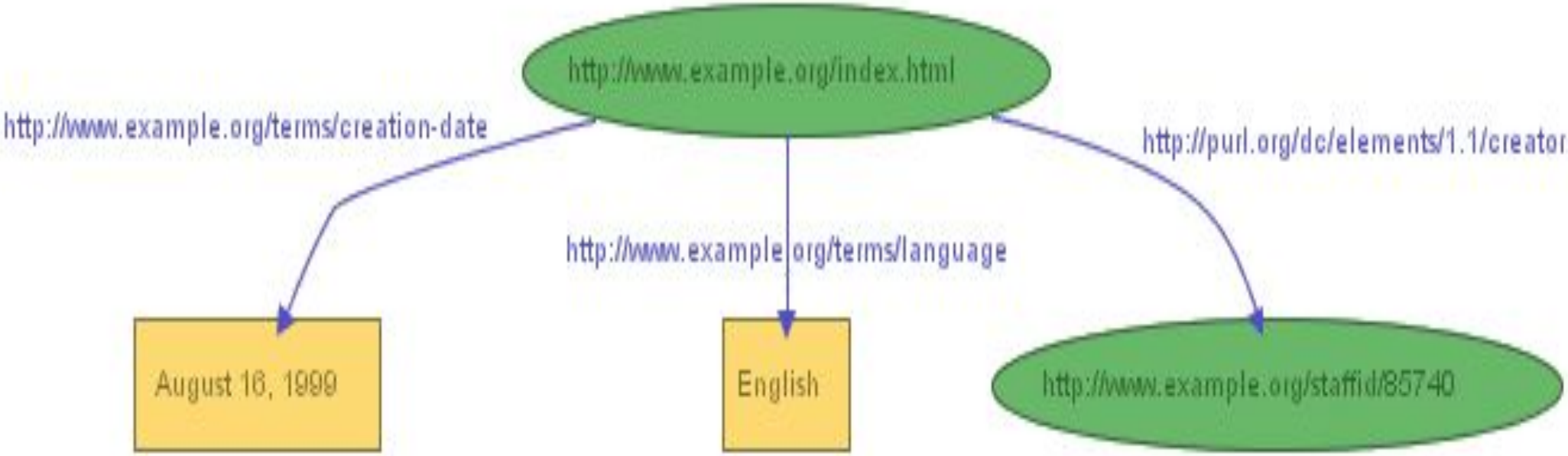
RDF-высказывание:

- a *subject* `http://www.example.org/index.html`
- a *predicate* `http://purl.org/dc/elements/1.1/creator`
- and an *object* `http://www.example.org/staffid/85740`



RDF - Resource Description Framework. Пример

http://www.example.org/index.html has a creation-date whose value is August 16, 1999
http://www.example.org/index.html has a language whose value is English



RDF - Resource Description Framework

RDF предоставляет возможность формулировать утверждения в виде, пригодном для обработки машиной.

- RDF является уточнением XML
- В начале RDF-документа идет список ссылок на онтологии (namespaces)
- Основная часть RDF-документа состоит из деревьев глубины 3
- Верхняя вершина – субъект, средний уровень – свойства, нижний уровень – объект
- Каждая вершина может быть задана строкой или ссылкой на объект из одного из namespaces
- Вершины могут иметь уточняющие квалификаторы

Метаданные

Метаданные – это данные, предназначенные для идентификации, описания или локализации информационных ресурсов независимо от физической природы ресурса

- Dublin Core Metadata Elements
- vCard
- RSS
- FOAF
- DOAP
- SIOC
- …

Набор элементов для создания метаданных

Dublin Core

Dublin Core Metadata Elements состоит из 15 базовых элементов, которые можно условно разбить на три группы:

1. Content - элементы, в основном относящиеся к содержанию ресурса; (Title, Subject, Description, Type, Source, Relation, Coverage)
2. Intellectual Property - элементы, в основном рассматриваемые с позиции интеллектуальной собственности;(Creator, Publisher, Contributor, Rights)
3. Instantiation - элементы, в основном относящиеся к данному экземпляру ресурса.(Date, Format, Identifier, Language)

Набор элементов для создания метаданных

Dublin Core

Пример

Subject (Тема) - выбор тезауруса для характеристики предметной области (рубрикатора тем). Для поиска можно выделить два тезауруса:

- Предметный тезаурус (subject tesaurus) содержит понятия предметной области и отражает содержание документа, он отвечает на вопрос "*о чём этот документ?*".
- Функциональный тезаурус (function thesaurus) отражает роль *документа в человеческой деятельности, бизнесе* и отвечает на вопрос: "*для чего этот документ?*".

Набор элементов для создания метаданных

Dublin Core

Пример

Type (Тип) Тип ресурса отражает *жанр*, категорию его содержания. Он включает следующие элементы:

- text
- image
- sound
- dataset
- software
- interactive
- event
- physical object

Набор элементов для создания метаданных

Dublin Core

Пример

Формат (*Format)* отражает среду, формат данных ресурса; материал, из которого состоит ресурс (если это физический объект), его физические размеры.

Если описываемый ресурс представлен в электронном виде, то его формат определяет способ работы с ним, программное обеспечение.

Формат электронного документа рекомендуется выбирать из списка MIME, если это возможно. Примеры электронных форматов:

- text/xml - текст в формате XML
- text/plain - текст без форматирования
- image/gif - рисунок в формате GIF

description of a Web site home page using Dublin Core properties

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
 <rdf:Description rdf:about="http://www.dlib.org">
    <dc:title>D-Lib Program - Research in Digital
Libraries</dc:title>
    <dc:description>The D-Lib program supports the community of
people with research interests in digital libraries and
electronic publishing.</dc:description>
    <dc:publisher>Corporation For National Research Initiatives
</dc:publisher>
      <dc:date>1995-01-07</dc:date>
      <dc:subject>
        <rdf:Bag>
          <rdf:li>Research; statistical methods</rdf:li>
          <rdf:li>Education, research, related topics</rdf:li>
          <rdf:li>Library use Studies</rdf:li>
        </rdf:Bag>
      </dc:subject>
      <dc:type>World Wide Web Home Page</dc:type>
      <dc:format>text/html</dc:format>
      <dc:language>en</dc:language>
    </rdf:Description>
</rdf:RDF>
```

Дополнительные спецификации Semantic Web

Спецификация FOAF (Друг моего друга)

Аббревиатура FOAF означает Friend of a Friend — то есть «друг друга» (в смысле «друг второго уровня»). Это — спецификация, основанная на RDF и OWL. Она позволяет описать отдельную персону с помощью ряда характеристик (имя-фамилия, пол, адрес электронной почты, участие в разных блогах и прочая…), а также создать из этих данных уникальный идентификатор. Проект FOAF позволяет установить связи между разными людьми и предоставить всю эту информацию в виде, легко читаемом и понимаемом машинами (это RDF и OWL в действии). Создать простейшую FOAF-страницу о себе можно, в принципе, и вручную; но для форматированного описания «толпы» френдов, информация о которых регулярно меняется, несомненно, понадобится автоматизация. Впрочем, готовые FOAF-потоки предоставляют многие блог-хостинги, например, LiveJournal.

FOAF позволяет определить и описать языком, понятным для компьютера, наше место в сети, а если свести FOAF-данные воедино, то и место Сети в нашей жизни. FOAF-данные легко могут передаваться по Интернету от одного сервиса к другому, и каждый из них, разобрав их на составляющие, выуживает оттуда интересующую его часть: если сообщить все тому же «Моему кругу» ваш ник в «ЖЖ», он обратится к Livejournal.com, добудет оттуда FOAF вашей записи, получит список френдов и немедленно скажет, кто из них уже зарегистрирован в «Круге», но еще не добавлен в список прямых контактов (1-й круг). Удобно, просто, технологично.

www.foaf-project.org.

Дополнительные спецификации Semantic Web

RSS — семейство XML-форматов, предназначенных для описания лент новостей, анонсов статей, изменений в блогах и т. п. Информация из различных источников, представленная в формате RSS, может быть собрана, обработана и представлена пользователю в удобном для него виде специальными программами-агрегаторами.

В разных версиях аббревиатура RSS имела разные расшифровки:

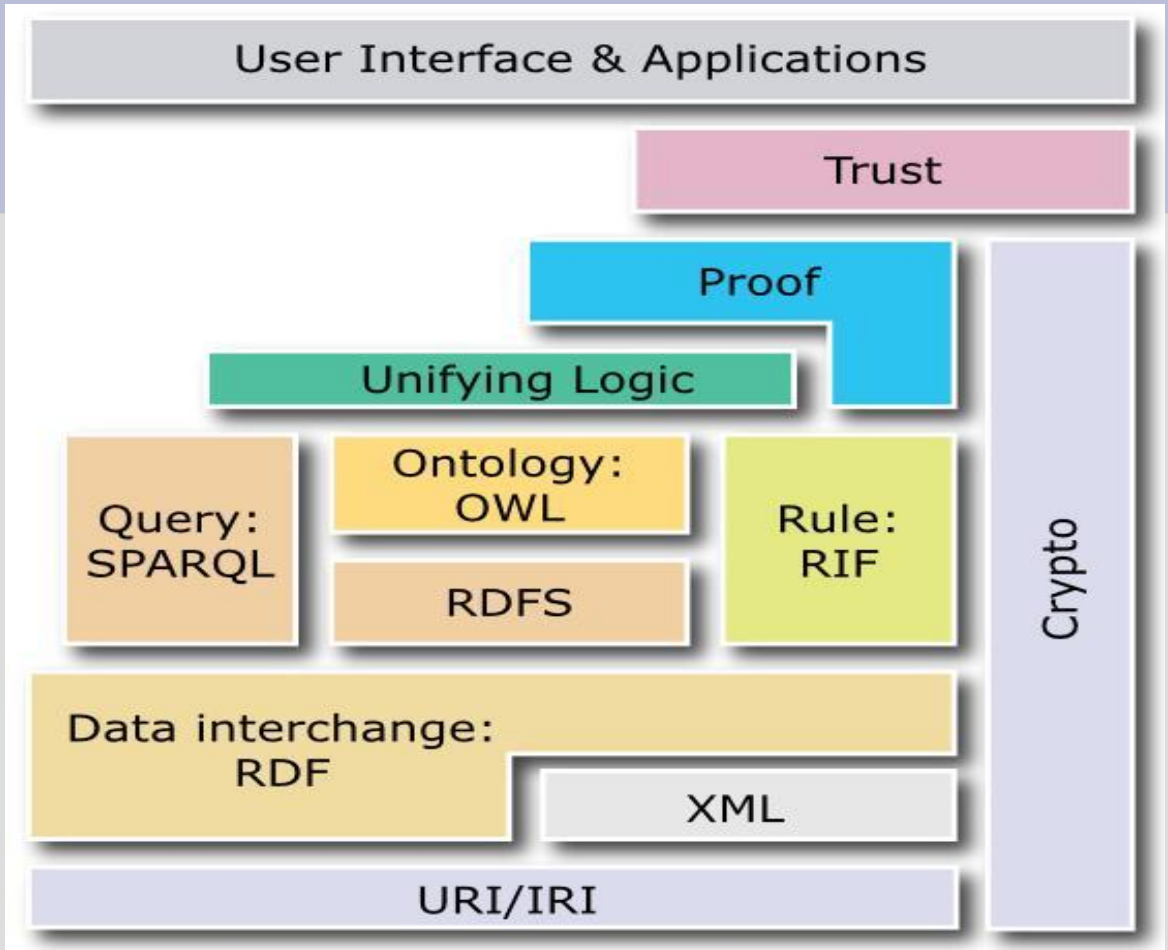
Rich Site Summary (RSS 0.9x) — *обогащённая сводка сайта*;

RDF Site Summary (RSS 0.9 и 1.0) — *сводка сайта с применением инфраструктуры описания ресурсов*;

Really Simple Syndication (RSS 2.x) — *очень простая синдикация информации*.

Обычно с помощью RSS 2.0 даётся краткое описание новой информации, появившейся на сайте, и ссылка на её полную версию. Интернет-ресурс в формате RSS называется RSS-каналом, RSS-лентой или RSS-фидом. Многие современные браузеры и почтовые клиенты умеют работать с RSS-лентами. Кроме того, существуют специализированные приложения (RSS-агрегаторы), собирающие и обрабатывающие информацию RSS-каналов. Также очень популярны веб-агрегаторы, представляющие собой сайты по сбору и отображению RSS-каналов, такие как Яндекс.Лента, Google.Reader и т.п.

Схемы и онтологии



Тим Бернерс Ли, Decentralized Information Group AAAI, 18 July 2006

Схемы и онтологии

Схема и онтология – это средства для описания смысла и связей между терминами.

«I'm not quite sure exactly what a schema is. In RDF terms it's a lot like an ontology but with XML it generally specifies the layout of a document.»

...

«According to the folks at Stanford an ontology is "a specification of a conceptualization" or a formal description of "the concepts and relationships that can exist for an agent or a community of agents".»

Aaron Swartz

RDF Vocabulary Description Language 1.0: RDF Schema

Схема RDF была разработана как простая модель типизации данных для RDF

RDF Schema определяет базовый словарь, предназначенный для описания RDF-словарей и принятые соглашения, которые могут быть использованы при создании приложений Semantic Web для поддержки более сложных словарей RDF описаний. Язык описания словаря RDF определяет классы и свойства, которые могут быть использованы для описания других классов и свойств.
RDF Schema определяется в терминах базовой информационной модели RDF – структура графа, описывающая ресурсы и свойства. Все словари RDF используют некоторую базовую структуру: они описывают классы ресурсов и типы связей между ресурсами.

Три наиболее важных понятия, которые дает нам rdf и схема rdf:
- это «Ресурс» (rdfs:Resource)
- «Класс» (rdfs:Class)
- и «Свойство» (rdfs:Property)

Эти понятия являются классами в том смысле, что этим классам могут принадлежать термины.
На языке схемы обычно задаются простые утверждения о разрешенных комбинациях.

```
rdfs:Resource rdf:type rdfs:Class .
rdfs:Class rdf:type rdfs:Class .
rdf:Property rdf:type rdfs:Class .
rdf:type rdf:type rdf:Property .
```

Дополнительные понятия:
- «Подкласс» (rdf:subClass)
- «Подсвойство» (rdf:subProperty)

Дополнительные понятия RDF-схемы:

- диапазоны (rdfs:**range**)
- домены (rdfs:**domain**)

Диапазоны и домены позволяют сказать, каким классам субъект и объект каждого свойства должен принадлежать

rdfs:domain всегда определяет, какому классу принадлежит субъект триплета, использующего это свойство, а rdfs:range всегда определяет, какому классу принадлежит объект триплета, использующего это свойство.

Пример. Высказывание: Свойство «:bookTitle» всегда должно применяться к книге и иметь литеральное значение:

```
:Book rdf:type rdfs:Class .
:bookTitle rdf:type rdf:Property .
:bookTitle rdfs:domain :Book .
:bookTitle rdfs:range rdfs:Literal .
:MyBook rdf:type :Book .
:MyBook :bookTitle "My Book" .
```

Необходимые свойства, не обеспеченные средствами RDF Schema:

- ***Ограничение мощности множества*** *(cardinality constraints)* на свойства, напр., что Человек имеет *только одного* биологического отца.
- Определение того, что представленное свойство (напр. `hasAncestor – имеет предка, прототип`) является ***транзитивным***, напр., что если A `hasAncestor` B, и B `hasAncestor` C, тогда A `hasAncestor` C.
- Определение того, что два различных класса, определенных в разных схемах, фактически представляют одну и ту же концепцию (понятие).
- Определение того, что два различных экземпляра (instances), определенные раздельно, фактически представляют одно и то же субъект.
- Способность определять новые классы в терминах комбинирования (напр. объединения и пересечения) других классов.

Онтологии

Онтологии, в общем виде определяемые как совместно используемые формальные концепции конкретных предметных областей, дают общее представление о темах, информацией по которым могут обмениваться люди и приложения.

Они позволяют концептуализировать домен фиксированием сущностей (entities) и связей в домене. Указанные, в каких связях участвует сущность, частично можно понять и ее значение (смысл), поскольку это позволяет нам видеть где данная сущность входит в отношения с остальным доменом.

DAML (DARPA Agent Markup Language)

Язык DAML (Aug 2000) был разработан как расширение XML и RDF. Последняя версия языка DAML+OIL обеспечивает богатый набор конструкций для создания онтологий и разметки информации таким образом, чтобы они были читаемы и понимаемы машиной.

DAML+OIL является семантическим языком разметки для Вэб-ресурсов. Он основывается на ранних стандартах W3C таких, как RDF и RDF Schema, и расширяет эти языки более полными примитивами моделирования. DAML+OIL (Mar 2001) расширяет DAML+OIL (Dec 2000) добавлением значений из типов данных XML Schema. DAML+OIL был разработан на основе первоначального языка DAML онтологий - DAML-ONT (Oct 2000) как попытка соединить многие из языковых компонент языка OIL (Ontology Inference Layer).

DAML+OIL

Онтология DAML+OIL (или база знаний, knowledge base) представляет собой коллекцию RDF – троек. Некоторые из этих троек релевантны DAML+OIL, а некоторые – нет. Эти семантики относятся только к тем тройкам, которые релевантны DAML+OIL и игнорируют остальные.

В языке OIL онтологии содержат определения свойств (slot-def) и определения классов (class-def). Определения Slot-def описывают отношения между двумя объектами. Class-def связывает имя класса с его описанием и содержит ряд компонентов

OWL - Ontology Web Language

Проект OWL (нач. 2002) расширяет возможности XML, RDF, и RDF Schema. Этот язык основан на DAML+OIL.

Основные существенные отличия от DAML+OIL

- Устранение ограничений;
- Способность напрямую указывать, что свойство может быть симметрично; и
- Устранение некоторых неиспользуемых конструкций DAML+OIL, особенно ограничения с дополнительными компонентами.

Существует также несколько маловажных различий, включающих в себя некоторые изменения имен некоторых конструкций, однако основная цель, преследуемая при создании OWL была максимально корректно сохранить имена DAML+OIL.

OWL - Ontology Web Language

Модификации OWL

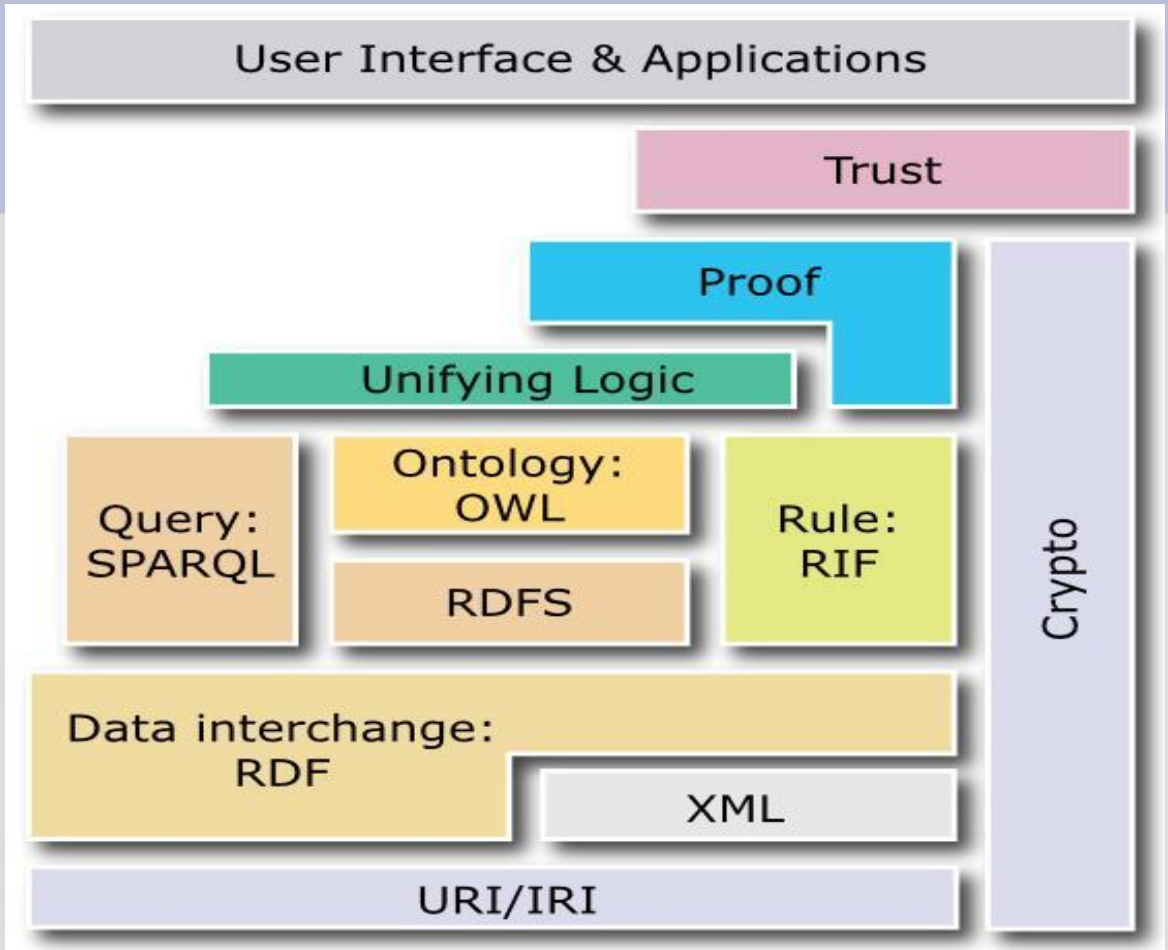
- OWL Lite
- OWL DL
- OWL Full

OWL - Ontology Web Language

Свойства OWL

- Использует синтаксис XML
- Имеет инструкции для представления дерева классов
- Имеет инструкции для указания принадлежности индивидов классам
- Имеет систему описания свойств: область определения, область значений
- Имеет возможность задавать характеристики свойств: симметричность, транзитивность, функциональность
- Имеет возможность указывать эквивалентность классов

Языки запросов к RDF-хранилищам



Тим Бернерс Ли, Decentralized Information Group AAAI, 18 July 2006

Языки запросов к RDF-хранилищам

В 2006 консорциум W3C предложил разработку языка запросов к RDF и OWL-хранилищам – **SPARQL**, который на данный момент находится в состоянии рекомендованного кандидата (candidate recommendation)
SPARQL является языком запросов, который базируется на паттернах графов.
Вид синтаксиса запроса (упрощенный) :

```
SELECT      <v_list>
FROM   <onologyURI>
WHERE       {     <template_list>.
            FILTER <filter_expr>
        }
```

где
v_list – список имен переменных,
onologyURI – ссылка на онтологию,
template_list – список шаблонов
filter_expr – ограничения на значения переменных.

SPARQL одновременно является языком запросов и протоколом доступа к данным

пример
Определен предикат haveDad (имеет отца).

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:som="http://somesite.ru/" >
<rdf:Description rdf:about="Alex">
<som:haveDad rdf:resource="Misha" />
</rdf:Description>
<rdf:Description rdf:about="Misha">
<som:haveDad rdf:resource="Dima" />
</rdf:Description>
<rdf:Description rdf:about="Dima">
<som:haveDad rdf:resource="Ivan" />
</rdf:Description>
</rdf:RDF>
```

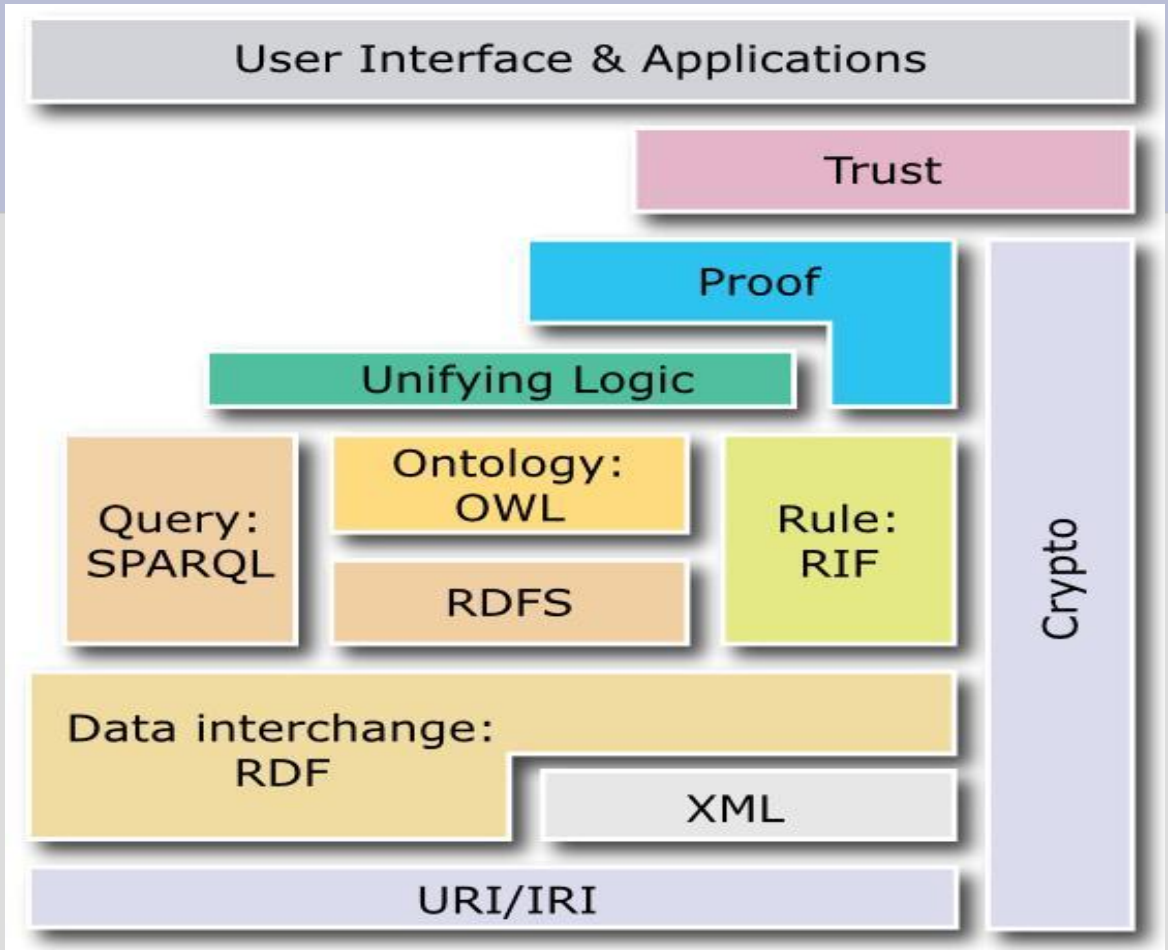
нам необходимо узнать деда Миши

```
PREFIX som: <http://somesite.ru/>
SELECT $x
WHERE {
<Misha> som:haveDad [ som:haveDad $x ] .
}
```

результат: дедом Миши является Иван

```
-----------------------------------------
| x                                     |
=========================================
| <file:///D:/bin/Arq-1.3/work/Ivan> |
-----------------------------------------
```

Логика



Тим Бернерс Ли, Decentralized Information Group AAAI, 18 July 2006

Логика

Принцип «логического вывода» очень прост:
возможность выводить новые данные из данных, которые уже известны.
В математическом смысле, выполнение запроса есть одна из форм логического вывода (например, возможность вывести из массы данных некоторый результат поиска).

Пример:
:**MyCar de:macht «160KW»**

английский обработчик этого не поймет. Добавим строчку:
de:macht daml:equivalentTo en:power

Объединение баз данных становится простой задачей фиксирования где-то в RDF соответствия, или несоответствия полей двух и более баз данных.

Логический вывод

Логический вывод - один из базовых принципов Семантического Веба, поскольку он позволяет нам очень легко создавать SW (Semantic Web)-приложения

Для того, чтобы Семантический Веб стал достаточно выразительным и смог помогать людям во всяких ситуациях, возникает необходимость построения мощного логического языка, который поддерживает логический вывод.

Проблемы логики предикатов подробно рассмотрены в монографии Джона Сова (John Sowa's) Математические предпосылки (логика предикатов) Mathematical Background (Predicate Logic)

Правила вывода новых фактов SWRL

словарь **SWRL (A Semantic Web Rule Language)**

- дополнение OWL языком **RuleML**.
В связи с этим появилась возможность использовать дизъюнкты Хорна (Horn-like rules) для явного указания способа вывода новых фактов из RDF-утверждений. Пока словарь SWRL находится в стадии стандартизации.

Хотя работы над этим уровнем Семантического Веба еще продолжаются, однако в нашем распоряжении уже есть достаточный набор средств для построения Семантического Веба:

утверждения (т.е. “И"), цитирование (материализация) в RDF,
классы, свойства, области и документирование в схеме RDF,
непересекающиеся классы, свойства однозначности и уникальности, типы данных, инверсии, эквивалентности, списки проч. из DAML+OIL

Правила вывода новых фактов SWRL

словарь **SWRL (A Semantic Web Rule Language)**

parent(?x,?y) & brother(?y,?z) → uncle(?x,?z):

```
<swrl:Imp rdf:ID="Def-hasUncle">
   <swrl:body>
     <swrl:AtomList>
       <rdf:first>
         <swrl:IndividualPropertyAtom>
           <swrl:propertyPredicate rdf:resource="#hasParent"/>
           <swrl:argument1 rdf:resource="#x"/>
           <swrl:argument2 rdf:resource="#y"/>
         </swrl:IndividualPropertyAtom>
       </rdf:first>
       …
      </swrl:AtomList>
   </swrl:head>
  </swrl:Imp>
```

Rule Interchange Format (RIF) - Формат обмена правилами

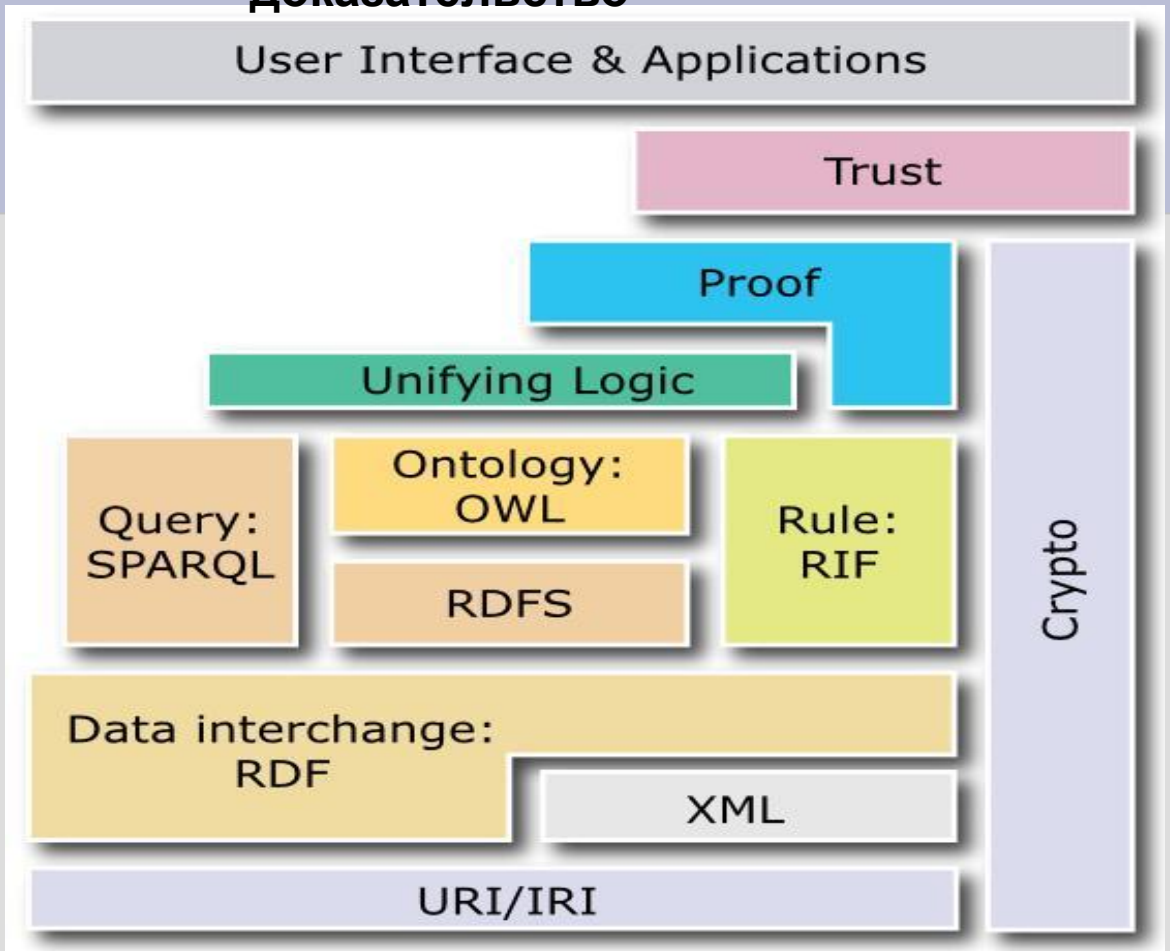
Целью данного разрабатываемого консорциумом W3C стандарта является определение формата, который позволит транслировать правила между разными языками правил и благодаря этому обеспечить обмен правилами между системами правил.
Системы, основанные на правилах, получили широкое распространение в информационных технологиях:
експертные системы, системы дедуктивных баз данных и т.п.

Разработки технологий Семантического Веба обеспечат новую среду использования таких систем. Консорциум W3C уделяет особое внимание этой области.

Спецификация RIF может рассматриваться как составная часть комплекса стандартов консорциума для Семантического Веба

В настоящее время обсуждается рабочий проект документа, который систематизирует случаи использования RIF и требования к этому языку.

Доверие и доказательство



Тим Бернерс Ли, Decentralized Information Group AAAI, 18 July 2006

Доверие и доказательство

В суровой действительности простейшим образом это звучит так: если один человек говорит, что x – что-то голубое, а другой утверждает, что x не голубое, то не следует из этого, что идея Семантического Веба развалится?

Естественно, нет, потому что

- приложения Семантического Веба зависят от контекста,
- и b) приложения в будущем будут, как правило, содержать механизмы проверки доказательства и цифровых подписей.

Направления фундаментальных исследований только формируются.

Агенты и сервисы

Веб-сервис – это программная система, предоставляющая некоторую услугу и обеспечивающая взаимодействие по сети.

Веб-сервис – это программная система, идентифицируемая строкой URI, чьи общедоступные интерфейсы определены на языке XML. Описание этой программной системы может быть найдено другими программными системами, которые могут взаимодействовать с ней согласно этому описанию посредством сообщений, основанных на XML, и передаваемых с помощью интернет-протоколов.

Википедия

Агенты и сервисы

Технология веб-сервисов базируется на следующих открытых XML-стандартах:

SOAP (Simple Object Access Protocol) – XML-протокол для удаленного вызова методов веб-сервиса;
UDDI (Universal Description, Discovery and Integration) – описывает модель данных для каталогизации и обнаружения услуг, предоставляемых веб-сервисами;
WSDL (Web Services Description Language) – язык описания интерфейсов веб-сервисов.

OWL-S – онтологический язык веб-сервисов

SML (Service Modelling Language) – язык моделирования сервисов (проект)

Программная поддержка технологии
Semantic Web

Дальнейшему развитию проекта Семантического Веба способствует наличие свободно распространяемых каркасов для разработки приложений Semantic Web:
Jena Framework (Java)
Drive RDF Parser (C#)

Уже существуют:
библиотеки для интерпретации стека языков RDF для всех популярных языков программирования (Jena, Redland, RDFLib);
редакторы онтологий (Protege);
системы размышлений над онтологиями (Racer, KAON, FACT);
семантические хранилища (Sesame, Kowari, YARS);
семантические браузеры (Simile, Piggy Bank, Gnowsis, Haystack);
поисковики семантических данных (Swoogle);
конвертеры из разных форматов представления данных в/из RDF/XML (Aperture, RDFizers, D2R);
прикладные программы (Bibster, FOAF Explorer).
Существующие коммерческие продукты: Adobe's XMP – инструментарий для создания метаописаний о файлах; Oracle's 10.2 Database – уже имеет встроенную поддержку модели RDF; Tucana's Knowledge Discovery Suite - платформа для интеграции информации приложений (Enterprise Information Integration , EII)

Заключение.
Основные вехи на пути к Semantic Web

Технология Семантического Веба в данное время уже успешно решает следующие задачи:

- независимость данных от приложений;
- семантическая интеграция данных;
- создание основы для широкого использования компьютерных агентов и сервисов.

Заключение.
Основные вехи на пути к Semantic Web

Широкое распространение Web стандартов (рекомендаций W3C)
- XML
- RDF/RDFS
- OWL
- SPARQL

Наличие свободно распространяемых каркасов для разработки Semantic Web приложений
- Jena Framework (Java)
- Drive RDF Parser (C#)

Массовая разработка и использование онтологий

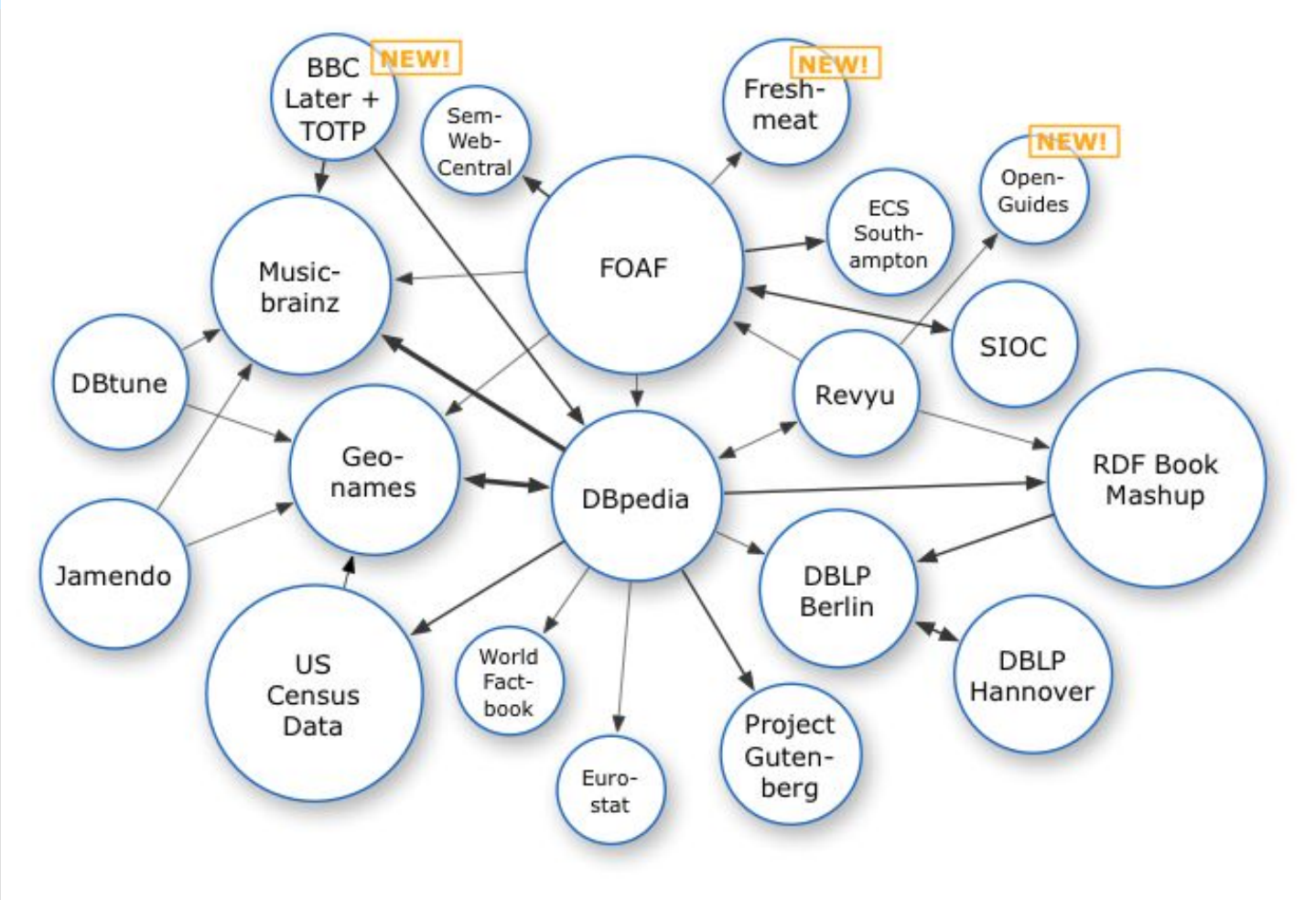
Заключение.
“Core” vocabularies

There are also a number “core vocabularies” (not necessarily OWL based)

- Dublin Core:
- FOAF: about people and their organizations
- DOAP: on the descriptions of software projects
- Music Ontology: on the description of CDs, music tracks, …
- SIOC: Semantically-Interlinked Online Communities
- vCard in RDF
- …

Open Data Community Project

"Expose" open datasets in RDF
Set RDF links among the data items for different datasets
Set up SPARQL endpoints to query the data
Over 2 billion triples served so far (August 2007)



Проект Semantic Web - Semantic Web Challenge

2003

…

2007 Sixth International Semantic Web Conference),
11-15 ноября 2007, Busan, Корея

Проект Semantic Web - Semantic Web Challenge

Определение приложения Семантической Сети (Минимум):

1. Приложение должно использовать информационные источники, которые:
   - географически распределены
   - имеют разнообразное право собственности (нет никакого контроля за их развитием).
   - являются гетерогенными (синтаксически, структурно, и семантически).
   - содержат данные реального мира, а не игрушечные примеры.
2. Приложение должно воспринимать открытый мир (это значит, что оно знает, что информация никогда не бывает полной).
3. Приложение должно использовать некоторое формальное описание значения данных.

Semantic Web Applications

– (Semantically) Closed SW Applications *(1st generation)*

• Typically in the domain of Enterprise Information Integration

– (Semantically) Open SW Applications *(2nd generation)*

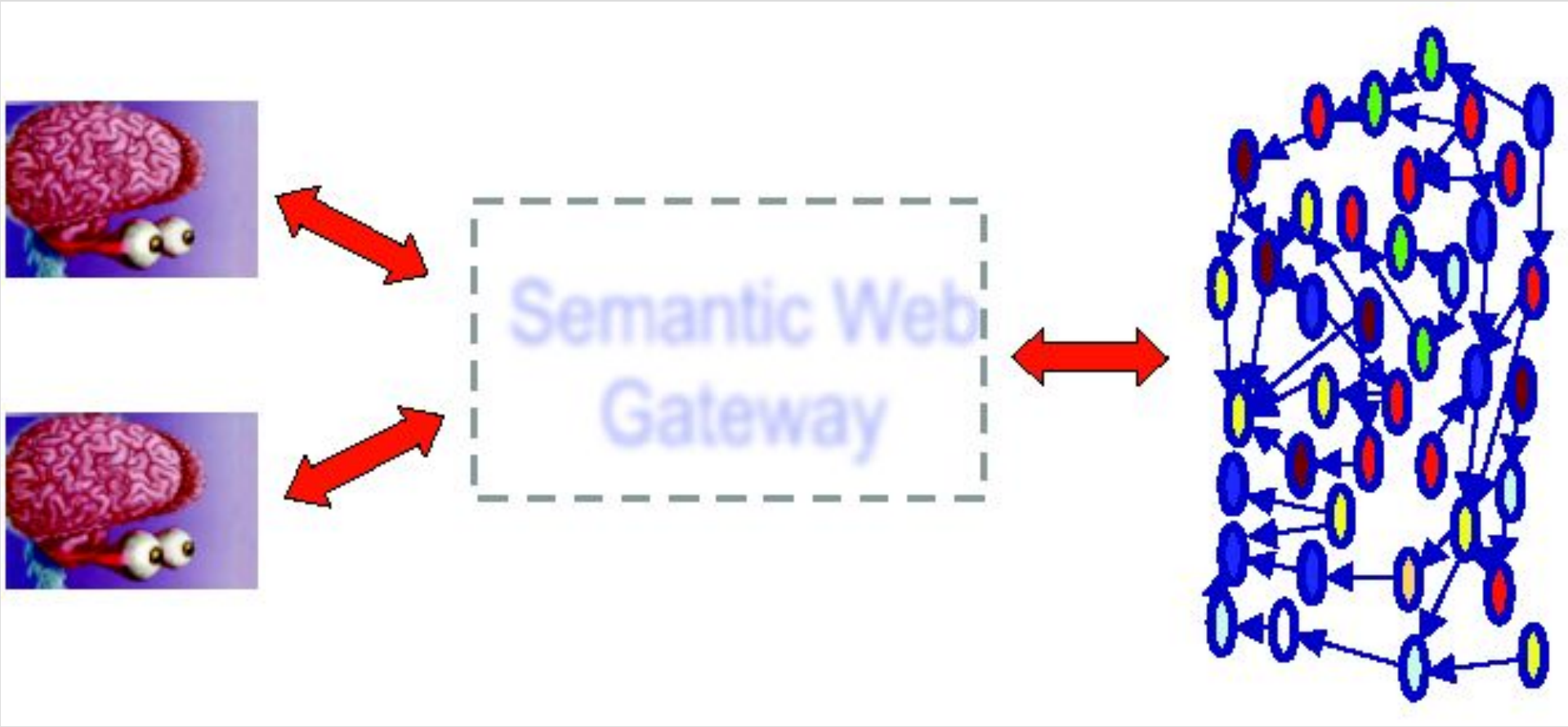
Prof. Enrico Motta, *PhD*
Knowledge Media Institute
The Open University Milton Keynes, UK

Semantically Open Semantic Web Applications

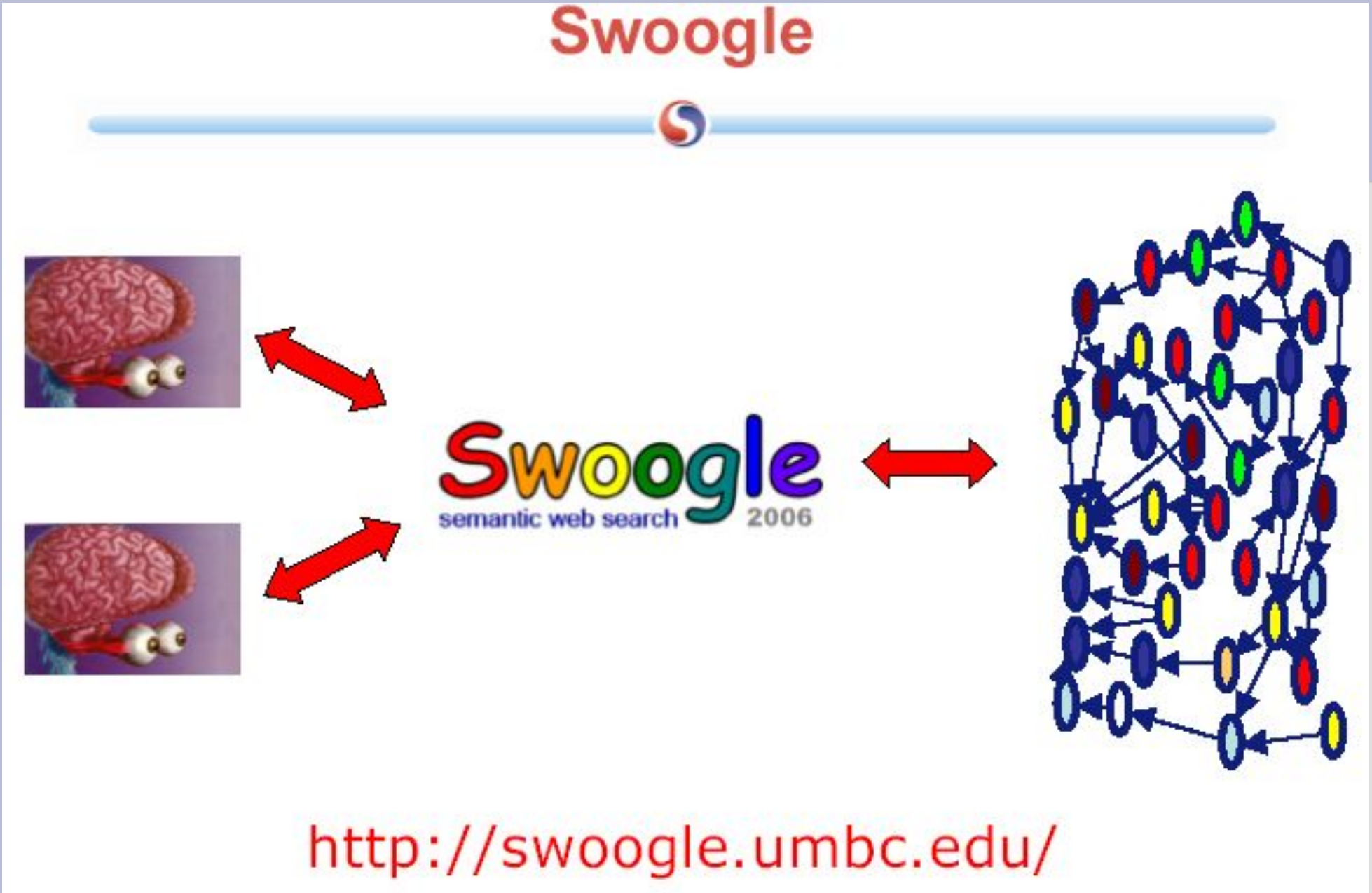
Должны быть способны использовать:
– множество онтологий
– быть открытыми для семантических ресурсов
– быть открытыми для работы с пользователем (user interaction)

В идеале, они должны уметь использовать не только данные SW, но и другие форматы данных, напр. фолксономии и т.п., следовательно должны иметь мощные механизмы по автоматическому извлечению информации.

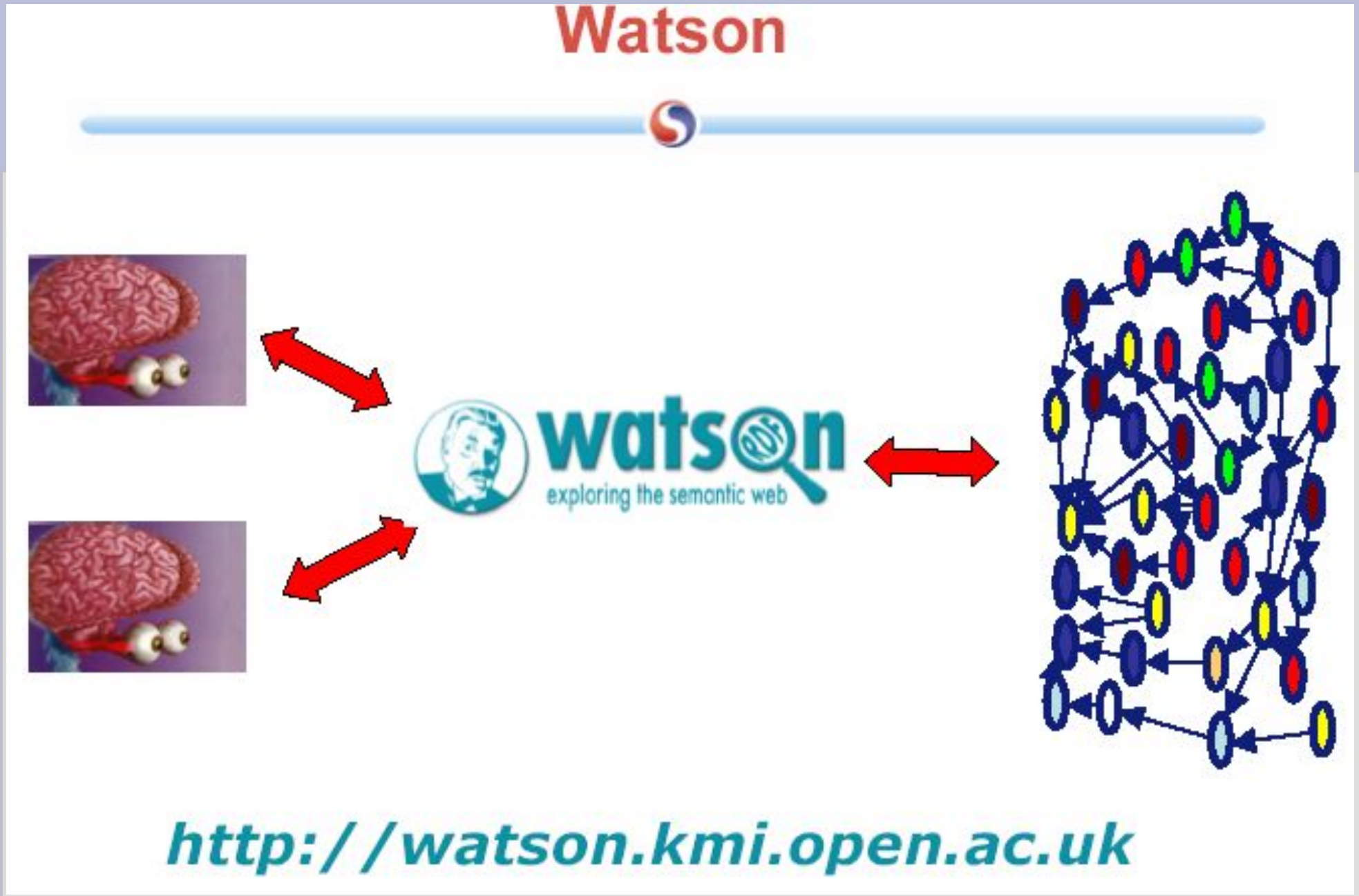
Architecture of OSW Apps



Проект Semantic Web - Semantic Web Challenge



Проект Semantic Web - Semantic Web Challenge



Проект Semantic Web - Semantic Web Challenge

watson RDF
exploring the semantic web

- Sophisticated quality control mechanism
  - Detects duplications
  - Fixes obvious syntax problems
- E.g., duplicated ontology IDs, namespaces, etc..
- Structures ontologies in a network
  - Using relations such as: *extends, inconsistentWith, duplicates*
- Provides efficient API
- Supports formal queries (SPARQL)
- Variety of ontology ranking mechanisms
- Modularization/Combination support
- Plug-ins for Protégé and NeOn Toolkit

Prof. Enrico Motta

Conclusions

- Many ‘classic’ applications of SW technology use ontologies to support the integration of distributed data sources
  - These applications are typically ‘semantically closed’
- As more and more semantic information becomes available on the SW, researchers are also looking at ‘semantically open’ applications, able to exploit large scale semantics to support intelligent problem solving
- This approach is being used in a number of other scenarios, including:
  - Semantic Web Browsing
  - Question Answering
  - Integration of Folksonomies with the SW

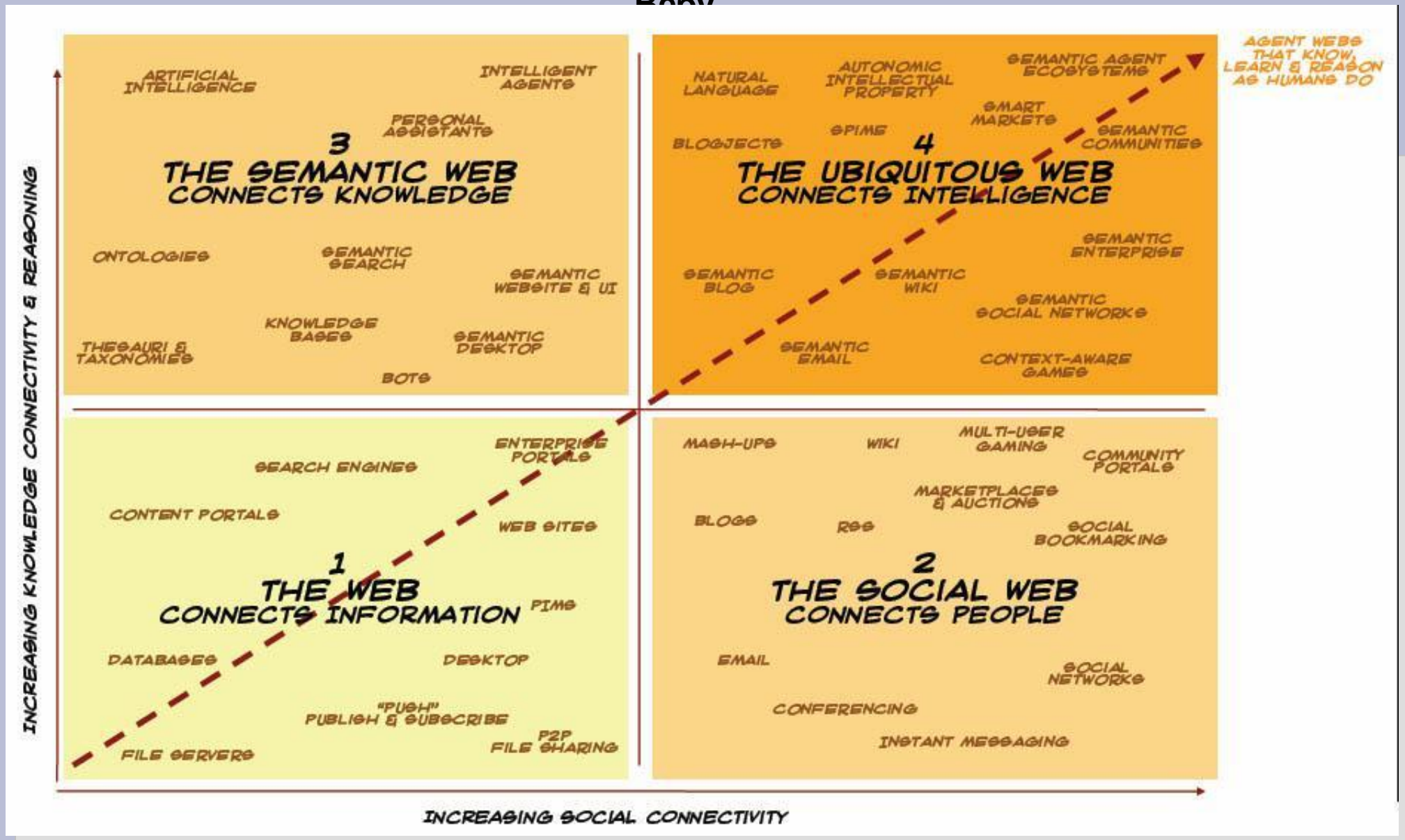
Выводы

- ✔ Обозначился резкий рост и возникновение компаний, использующих технологию Semantic Web (Radar Networks, MetaWeb, Joost, SiberLogic etc.)
- ✔ произошло вовлечение крупных поставщиков ПО – Adobe, Cisco, HP, Microsoft, Nokia, Oracle, Sun…)
- ✔ активно развиваются правительственные программы – в США, Объединенной Европе, Японии, Корее, Китае
- ✔ сильно вырос медико-фармацевтический рынок – создана специальная группа при консорциуме Health Care and Life Sciences Interest Group at W3C
- ✔ появилось много инструментов с открытым кодом – Kowari, RDFLib, Jena, Sesame, Protégé, SWOOP, Onto(xxx), Willbur

К 2012 году в 80% общедоступных веб-сайтов будет, в той или иной степени, использоваться семантический гипертекст для создания семантических веб-документов (с вероятностью 0.7), а в 15% общедоступных веб-сайтов будут использоваться развитые веб-онтологии для создания семантических баз данных (с вероятностью 0.6)

Прогноз Gartner Group

Путь движения к Семантическому Вебу



Выводы

Задачи и проблемы:

- ✔ Индексация и поиск информации
- ✔ Разработка и поддержка метаданных
- ✔ Разработка и поддержка методов аннотирования
- ✔ Представление Веб в виде большой интероперабельной БД – линковка баз
- ✔ Организация машинной добычи данных
- ✔ Обнаружение и предоставление веб-сервисов
- ✔ Исследования в области интеллектуальных агентов

Базовая литература

Тим Бернерс-Ли **«The Semantic Web»**
Тим Бернерс-Ли **«The Semantic Web Roadmap»**
Sean B.Palmer **«The Semantic Web: An Introduction»**
Aaron Swartz **«The Semantic Web In Breadth»**
Материалы конференций SemanticWebChallenge

Спасибо за внимание