

РАЗРАБОТКА СИСТЕМЫ ИНТЕГРАЦИИ ВЕБ-САЙТОВ НА ОСНОВЕ ТЕХНОЛОГИЙ SEMANTIC WEB

Быков Г. В.,
научный руководитель Бархатов А. В.
Сибирский федеральный университет
Институт математики

До настоящего времени всемирная компьютерная сеть наиболее быстро развивалась как носитель документов, используемых людьми, а не как хранилище информации, допускающей автоматическую обработку. Так, поисковые системы, получая запрос пользователя, производят сопоставление одной последовательности символов (запроса) с другими последовательностями (индексированные документы), не вникая в значение написанного. В итоге, вместо конкретного ответа на поставленный вопрос пользователю зачастую предлагается самостоятельно изучить значительное число Интернет-ресурсов, наличие на которых интересующей информации вовсе не гарантируется. В условиях стремительного роста Интернет, когда каждые полтора года объем данных в глобальной сети удваивается, заложенные на заре становления HTML принципы мета-тегов не справляются с возложенной на них задачей. Для повышения эффективности поиска в таких объемах материалов предлагается введение и использование новых механизмов, способных трансформировать сеть, понятную исключительно человеку, в сеть, понятную компьютерам. Одним из таких механизмов является Semantic Web (или «Семантическая сеть», «Семантическая паутина»).

Страницы в Семантической паутине, в отличие от своих HTML-предшественников, снабжены семантической разметкой (компьютерно-ориентированные данные, описания), никак не отображаемой пользователю при просмотре документов, однако позволяющей компьютеру «уловить общий смысл» написанного. Для того, чтобы позволить компьютеру интерпретировать информацию в Веб наравне с людьми, существуют:

- графовая модель описания ресурсов RDF (Resource Description Framework), с помощью которой можно создавать любые утверждения о любых ресурсах и публиковать эту информацию на сайтах в XML-формате. Каждое утверждение является тройкой субъект-отношение-объект (например, «Шекспир написал Гамлета», «Гамлет это пьеса»);
- язык описания онтологий OWL (Web Ontology Language), который, дополняя RDF, даёт возможность делать утверждения о принадлежности объектов более общим классам, свойствах бинарных отношений, существовании новых типов в терминах уже существующих;
- машины вывода (Semantic reasoners), позволяющие получать новые знания на основе уже имеющихся путём применения тех или иных правил логического вывода (в зависимости от структуры онтологии);
- язык запросов SPARQL, а также протокол передачи этих запросов и ответов на них (аналог SQL для реляционных баз данных).

Использование идей Semantic web позволяет снять часть работы по поиску и обработке информации с пользователя, передав эти функции специально разработанным приложениям. В данной работе рассматривается применение Семантических сетей в узком кольце ресурсов схожей тематики для интеграции данных и предоставления различных тематических сервисов.

Наглядным примером таких ресурсов может послужить кольцо сайтов СФУ о живой природе, включающее сайты о рыбах, птицах, красной книге Красноярского края (fish.krasu.ru, nature.krasu.ru, redbook.krasu.ru, birds.krasu.ru). Каждый из них придерживается определенной общей политики в описании разнородных животных: указание места в биологической систематике, латинского названия, года открытия, первооткрывателя, общей информации о полевых признаках, распространении, местах обитания. На основании этих данных можно предложить к реализации следующие сервисы:

- просмотр информации по данному животному (и его ближайших родственниках) на других сайтах кольца;
- единый интеграционный сайт (например, единая база животных);
- централизованный поиск информации (например, посмотреть все виды, которые были открыты в определенном году);
- поиск несоответствий и опечаток на сайтах кольца (один и тот же вид на разных сайтах имеет разный год открытия, первооткрывателя, принадлежит к разным более общим таксонам).

Для реализации вышеназванных сервисов были разработаны консольные приложения (Java SE 6.0 с применением библиотеки Jena), получающие информацию о представленных животных из баз данных и файлов ресурсов (*.htm, *.inc) сайтов кольца. Организованная в соответствии с разработанной OWL-онтологией информация о животных затем попадает в семантическое хранилище (база данных под управлением MySQL). Это хранилище используется веб-приложением (на Java Server Faces 2.0), которое:

- предоставляет пользователю единый веб-интерфейс для использования разработанных семантических сервисов;
- позволяет производить поиск информации на сайтах всего кольца по заранее заготовленным параметризованным или произвольным SPARQL-запросам;
- позволяет загружать на локальный компьютер всю используемую семантическим приложением информацию в формате RDF/XML.

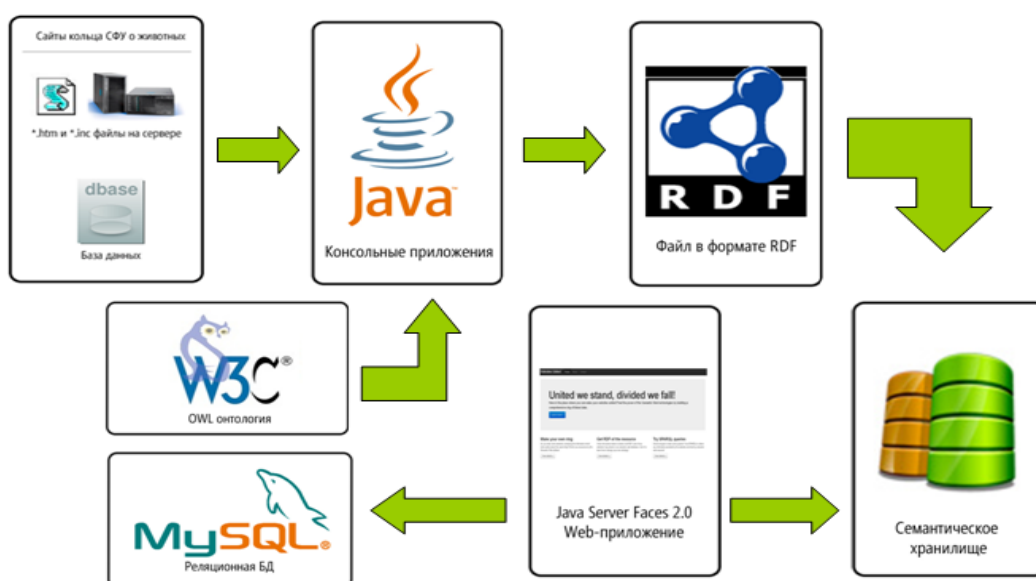
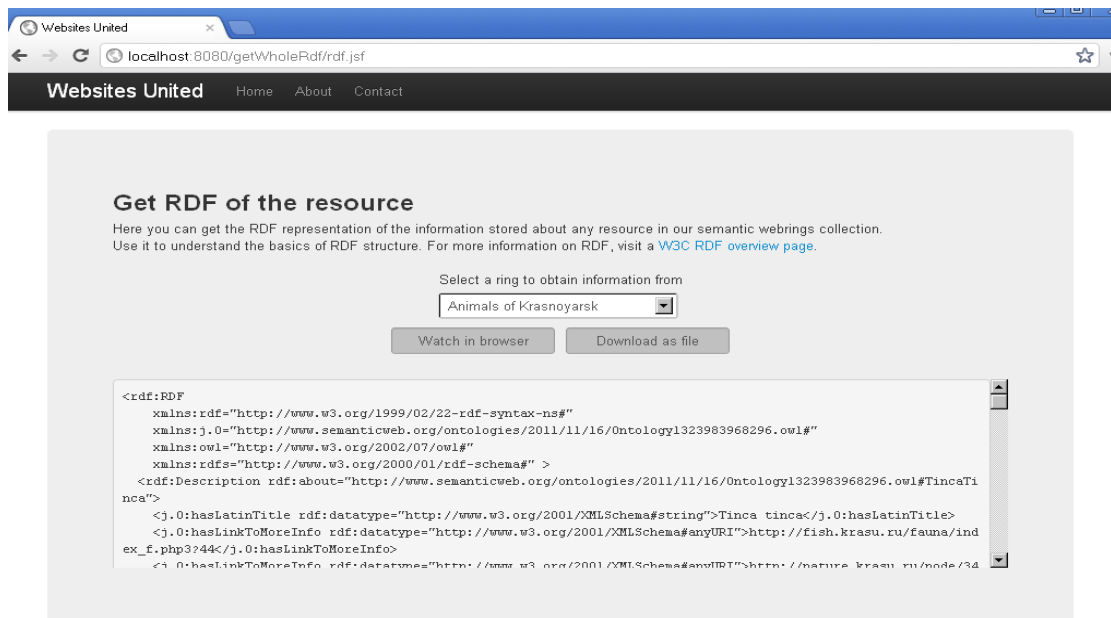


Рис. 1: Структура системы



Make your own ring

Do you have some websites containing the information which looks pretty about the same thing? Enrich you resources by the Semantic Web utilities!

[View details >](#)

Get RDF of the resource

Check the button below to obtain a full RDF code of any websites' ring stored in our semantic web database. Use it to learn how to design your own ontology!

[View details >](#)

Try SPARQL queries

Smart enough to make some queries? Use SPARQL to obtain any information provided by the websites enriched by semantic web reasoner!

[View details >](#)

Рис. 2: Получение RDF в форме на сайте

Развитием данной работы может послужить реализация авторизации и аутентификации на веб-сайте, предоставление веб-интерфейса для регистрации и подключения в работу новых колец сайтов, реализация javascript-виджетов, выполняющих асинхронные запросы к семантическому хранилищу и возвращающих результат без необходимости покидать текущую страницу, которые можно предложить администраторам колец для вставки на свои ресурсы.

Разрабатываемая система позволит производить интеграцию веб-ресурсов различных видов на основе технологий Semantic Web. Так, это могут быть сайты институтов, гостиниц, банков, авиакомпаний, предприятий. На данный момент подобные подходы в Интернет практически не применяются или применяются только в ограниченном кругу ресурсов, принадлежащих единому владельцу. Использование технологий Semantic Web в перспективе повысит интерпретируемость веб-ресурсов и сделает их доступными для различного рода интеллектуальной автоматической обработки.