Модель генерации документов на основе семантических сетей для создания web-ориентированных информационных систем♣

© Федотов А.М., Гуськов А.Е.

Институт вычислительных технологий СО РАН
{fedotov, guskov}@ict.nsc.ru

Аннотация

В статье рассматриваются способы обмена информацией в среде Web. Особое внимание уделяется семантическим сетям, предоставляющим фундамент для максимально полного и формализованного описания целевой информации. Предложена модель генерации документов, основанная на формировании внутреннего представления документов посредством семантической сети и его последующем приведении к требуемому формату. Описывается технология SMART для разработки информационных систем на основе предложенной модели, приводятся краткие описания реально функционирующих информационных систем. Рассматриваются некоторые вопросы создания качественных информационно-поисковых сервисов на основе семантических сетей.

1 Введение

Являясь одним из самых значительных достижений XX века, Интернет постоянно предлагает технологически новые, более эффективные решения самых различных задач, так или иначе связанных с обменом информацией. В данном исследовании рассматривается одна из таких задач: разработка информационных систем (ИС), которые, среди прочего, реализуют функции хранения информации (как правило, в виде наборов данных) и ее предоставления запрашивающим клиентам – программным агентам, осуществляющим доступ к услугам ИС. При этом возникает проблема разных требований клиентов к формату предоставления информации, что осложняет организацию взаимодействия между различными ИС. Также может быть затруднено и осуществление эффективного поиска, поскольку документы, предназначенные для прочтения человеком, и документы, пригодные для семантического анализа поисковыми агентами, должны быть опубликованы в разных форматах, которые соответствуют требованиям клиентов.

Поскольку одну и ту же информацию можно представить и передать различными способами в зависимости от возможностей ее отправителя и потребностей получателя, то при разработке ИС большую роль играют средства, используемые для организации обмена информацией с клиентами. На практике архитектура ИС основывается на схеме отображения данных из внутреннего хранилища, обычно управляемого СУБД, в конечный документ. При этом внимание акцентируется на том, *как* составить документ, а не на том, *что* является его содержанием. Вследствие этого, ИС оказываются спроектированными в расчете на определенный тип клиентов, удовлетворяющий специфичным требованиям; обслуживание других клиентов производится не эффективно.

Также в данной архитектуре трудно указать уровень, на котором может быть организовано информационное пространство, являющееся основным источником содержательного наполнения информационных ресурсов. Традиционно, вместо него определяется уровень баз данных, на котором посредством функций СУБД выполняются поисковые запросы (например, так устроены ИС, разработанные на языке *PHP*, либо на основе протокола *Z39.50* [10] или *LDAP*). Но идеология СУБД предполагает выполнение операций только со структурами данных. Поэтому данный подход не может реализовать полноценный информационный поиск, предоставляющий среду для создания гибких поисковых запросов, включающих не только структурные отношения, но и отношения более высокого семантического порядка, такие как “часть-целое” или “синоним”.

Результатом данного исследования является технология публикации информации, основанная на ее представлении в виде семантических сетей. Основная идея заключается в выделении в архитектуре ИС информационного уровня, который

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

обеспечивает унифицированный источник информации, и презентационного уровня, который определяет правила создания конечных документов. Предложенная технология позволяет оптимизировать процесс создания документов, имеющих одинаковое содержание, но предназначенных для разных клиентов, благодаря чему уменьшаются затраты на разработку ИС в целом. Другой результат исследования – построение прототипа системы поиска информации в семантической сети (онтологии [2]), предоставляющей этот сервис на более качественном уровне по сравнению с традиционными способами поиска.

Среди схожих работ можно выделить технологию публикации документов *Apache Cocoon* [9], в которой реализована схожая внутренняя архитектура. Принципиальным отличием предложенной авторами технологии является ориентация на создание единого информационного пространства, снабженного функциональностью для интеграции с другими подобными ИС.

2 Обоснование задачи

Авторами были рассмотрены обобщенные понятия информации и информационного обмена, лежащие в основе любой ИС. Выделены три основные способа представления информации для ее передачи в виде *электронного документа* – ресурса, который выступает в роли носителя информации и обладает внутренней структурой и описанием:

- Документы, содержащие тексты на естественном языке, которые обычно требуют наличие специализированного программного обеспечения для отображения в читаемом виде.
- Документы, хранящие данные согласно определенным схемам, применяются для представления структурированной информации посредством отношений типа “атрибут-значение”.
- Документы, представляющие информацию посредством семантических сетей, применяются для ее полного и формализованного описания, доступного для анализа программным алгоритмам.

На основе сравнительного анализа этих подходов, можно выделить наиболее характерные области их применения. При этом особый интерес представляют семантические сети – структуры данных, состоящие из узлов, соответствующих понятиям, и связей, указывающих на отношения между узлами. Семантические сети являются, своего рода, расширением концепции схем данных, в которой определены средства формального описания предметной области. Наиболее характерными примерами являются расширяемый язык разметки XML (eXtensible Markup Language), используемый преимущественно при работе со схемами данных, и язык описания ресурсов RDF (Resource Description Framework) [5], который в настоящее время является наиболее популярной реализацией семантической сети. Показано, что применение схем данных целесообразно только в случае, если для участников информационного обмена существует возможность согласовать используемую схему, т.е. определить и зафиксировать семантическую интерпретацию всех структурных элементов передаваемых документов. В децентрализованной распределенной среде, какой является Web, это не всегда возможно; в таких случаях для достижения большей эффективности взаимодействия программных агентов и, как следствие, обеспечения качества информационных услуг, следует применять технологии, основанные на использовании семантических сетей.

В качестве иллюстрации обоснованности этого утверждения можно рассмотреть наиболее популярный информационный сервис – поиск. В соответствии с указанными способами публикации информации выделяются три модели поиска документов: контекстный, атрибутный и семантический. Авторами были рассмотрены их свойства и показано, что качество поиска напрямую зависит от того, насколько формализовано была описана публикуемая информация.

Исследования показали, что используемые до настоящего времени технологии не предназначены для оперирования информацией в терминах семантических сетей, тогда как этот подход может предоставить качественно новые решения разных задач в области информационного обеспечения.

3 Модель генерации документов

Определим модель предметной области (МПО) как помеченный ориентированный граф, состоящий из двух частей: понятийной и содержательной (рис. 1). Понятийная часть определяет концепты предметной области и отношения между ними (например, концепты *Person*, *Publication* и отношения *is a*, *is author of*). Элементы содержательной МПО соответствуют реально существующим объектам предметной области, которые связанны определенными отношениями с концептами понятийной модели (например, элемент, соответствующий реальной персоне, связан отношением *is a* с концептом *Person*, а также отношением *is author of* с элементом, соответствующим реальной публикации).

Рис. 1. Структура модели предметной области

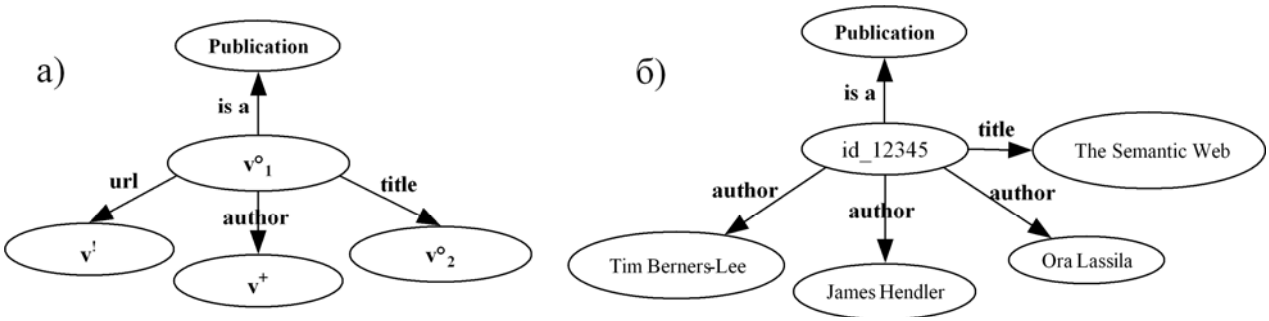

Рис. 2. Примеры: а) информационной структуры; б) наполнение информационной структуры

Для формального определения электронного документа определим термин *семантическая информация* как любой подграф МПО. Данное определение продиктовано следующим соображением: прием (и анализ) информации всегда связано с изменениями представлений ее получателя, т.е. с изменениями его модели предметной области. Таким образом, информацию можно отождествить с разностью между графами МПО после и до момента ее получения.

Определим *информационную структуру* – граф специального вида, который накладывает определенные ограничения на семантическую информацию. Эти ограничения можно разделить на два типа. Одни образуют подграф понятийной МПО и указывают, какие концепты и отношения должна содержать семантическая информация, т.е. задают, своего рода, ее каркас. Другие ограничения декларируют, каким образом и с какими отношениями остальные элементы содержательной МПО встраиваются в понятийную МПО. Семантическая информация, удовлетворяющая этим ограничениям, называется *наполнением информационной структуры*.

Пример информационной структуры и ее наполнения изображен на рис. 2, где вершины с метками v обозначают элементы, вместо которых должны быть подставлены данные, специфичные для конкретного документа. Вершина с меткой v_i^0 допускает подстановку вместо себя равно одного элемента, с меткой v^+ – одного и более элементов, с меткой $v^!$ – нуля или одного элемента; при подстановке каждого нового элемента все отношения с другими элементами сохраняются согласно исходной структуре.

Теперь определим *электронный документ в формате* φ, как ресурс, имеющий информационную структуру, ее наполнение и стиль, где *стиль* есть функция, определяющая правила преобразования семантической информации в последовательность символов. В данном определении можно выделить две компоненты: содержательную и презентационную. Содержательная компонента является композицией информационной структуры и ее наполнения, и описывает то, что отражает документ – информацию. Презентационной компонентой является стиль документа, который описывает то, *как* информация должна быть представлена клиенту. Можно показать, что данное определение документа применимо на практике для публикации информацию любым из трех рассматриваемых способов.

Одно из фундаментальных предположений, сделанных в исследовании, состоит в том, что наиболее хорошо зарекомендовавший себя подход к созданию ИС является подход на основе понятия *коллекции* – множества документов, имеющих одинаковую структуру и одну и ту же тематическую направленность. Таким образом, рассмотрение ИС сводится к классу *коллекционных ИС* – систем, где множество публикуемых документов состоит из объединения обозримого числа коллекций.

Одним из основных результатов исследования является разработка оригинальной модели ИС. Ее основные конструктивные отличия состоят в следующей схеме динамического создания документов (рис. 3):

1. Каждый клиентский запрос представляет собой набор параметров, по которому система однозначно определяет коллекцию, к которой принадлежит требуемый документ, процесс генерации содержания этого документа и процесс его публикации в требуемом формате.
2. Для каждой коллекций определен исходный *информационный шаблон* – описание информационной структуры документов коллекции. К информационному шаблону последовательно применяется ряд *трансформеров* – функций наполнения шаблона содержанием.
3. Результатом применения набора трансформеров к информационному шаблону является *внутреннее представление документа* (ВПД), которому взаимнооднозначно соответствует семантическая информация документа и из которого может быть получен сам документ.
4. Для каждой коллекций определен набор стилей; полученное ВПД согласно выбранному стилю преобразуется в документ конечного формата, запрошенного пользователем.

По определению, трансформер – суть функция, преобразующая информационный шаблон. Каждый

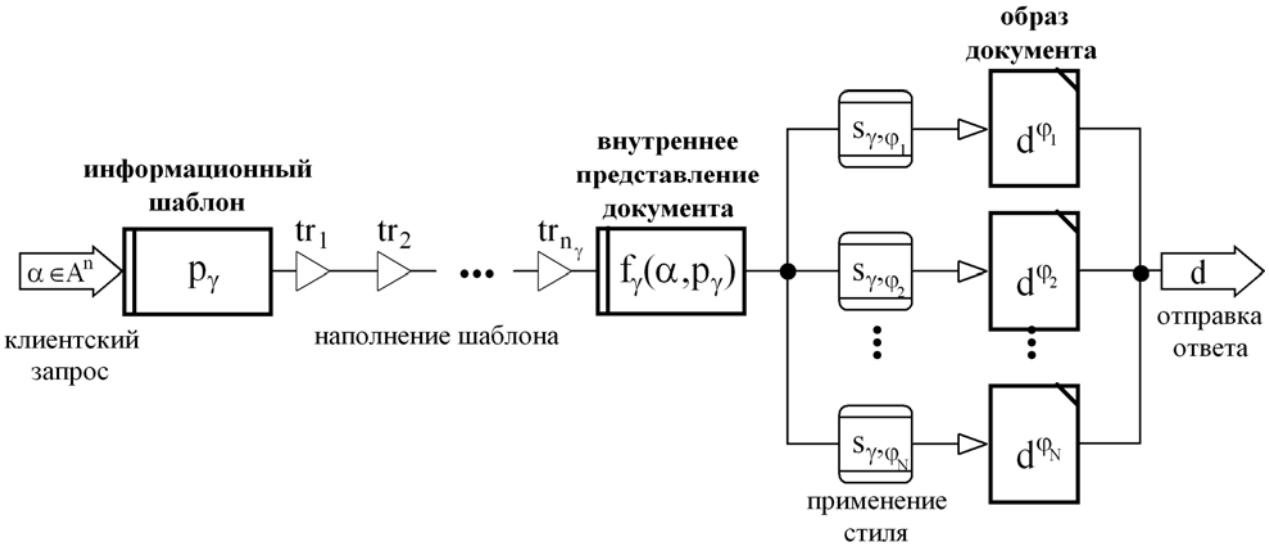

Рис. 3. Модель генерации документов

трансформер функционирует независимо, хотя и может при этом изменять любые части шаблона, в том числе те, которые были изменены другими трансформерами. Таким образом, трансформер реализует логически завершенное действие над информационным шаблоном.

Ключевым аспектом предложенной модели является язык внутреннего представления документов. В связи с этим был рассмотрена спецификация RDF [6] и показано, что данный язык является вполне адекватным средством для решения задачи описания ВПД.

4 Разработка ИС на основе предложенной модели

Для оценки практической применимости предложенной модели на ее основе была разработана система управления web-приложениями *SMART* (System for Managing Application based on RDF Technology) [8,12]. Функционально система представляет собой web-сервер, реализованный на основе технологии *Java Servlets* и библиотеки *Jena* [4] для работы с RDF, который при поступлении клиентского запроса инициирует процесс генерации документа на основе предложенной модели, по завершении которого полученный документ отсылается запросившему его пользователю.

При этом под информационной системой понимается система, публикующая информацию (или данные), содержащуюся в локальном хранилище под управлением СУБД, в виде электронных документов и функционирующая в рамках технологии “клиент-сервер”. При разработке технологии использовались некоторые парадигмы, которые к настоящему времени считаются общепринятыми концепциями, применяемыми при создании ИС (точнее, эти парадигмы были приняты во внимание еще на этапе разработки модели генерации документов). В частности, *концепция трехслойной архитектуры* приложений предполагает выделение слоя данных, слоя бизнес-логики и слоя презентации. Для реализации функциональности компонент слоя данных и слоя бизнес-логики обычно применяются *императивные языки программирования* с алгоритмической основой (PHP, Java, C#), для компонент презентационного слоя могут использоваться описания на *декларативных языках* (языки на основе XML) [11].

Декларативное программирование целесообразно использовать в случаях, когда алгоритм является менее значимым, чем результат его выполнения, и, более того, представляется целесообразным избежать явной записи алгоритма. В случаях, когда функциональность слоя бизнес-логики сводится к построению запросов к СУБД и подстановке полученных данных в некоторый шаблон, также представляется целесообразным применять декларативное программирование для описания исходных шаблонов и правил их наполнения (что и было отражено в модели генерации документов).

На практике разработка ИС в рамках технологии SMART предполагает определение следующих *ресурсов*:

1. *Репозитарий коллекций* хранит структурные описания коллекций документов на языке RDF, каждое из которых является реализацией информационного шаблона.
2. *Репозитарий трансформеров*. Трансформеры предназначены для наполнения структурного описания коллекции содержательной информацией, в результате чего будет получено ВПД, описанное на языке RDF. В отличие от других ресурсов, трансформеры описываются на императивном языке (Java), при этом, благодаря их универсальности и

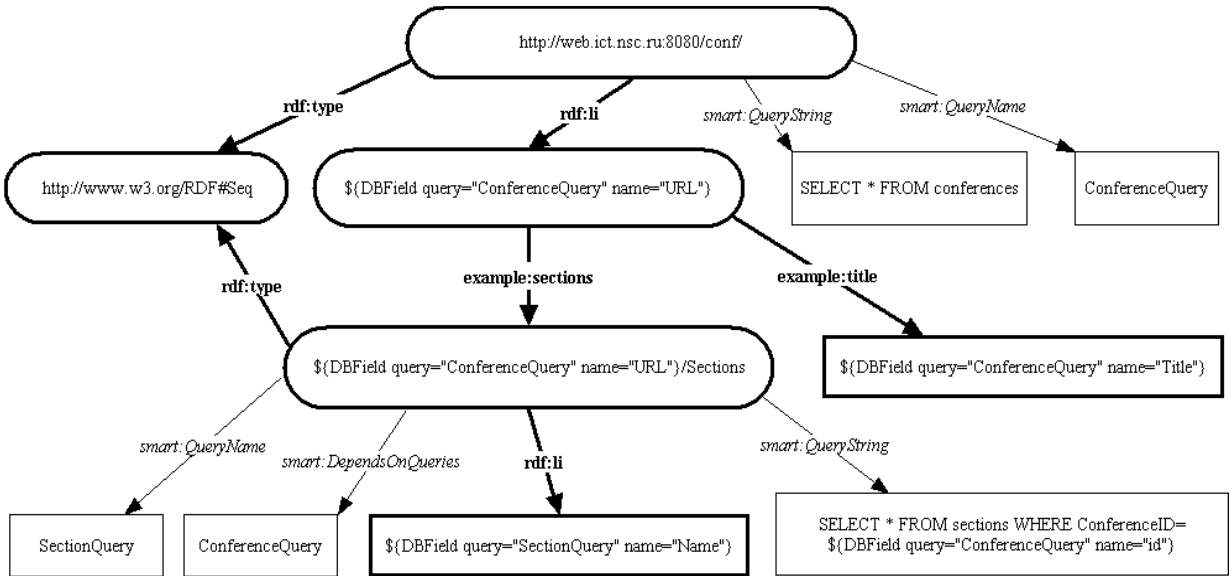

Рис. 4. Исходная RDF-модель информационного шаблона коллекции

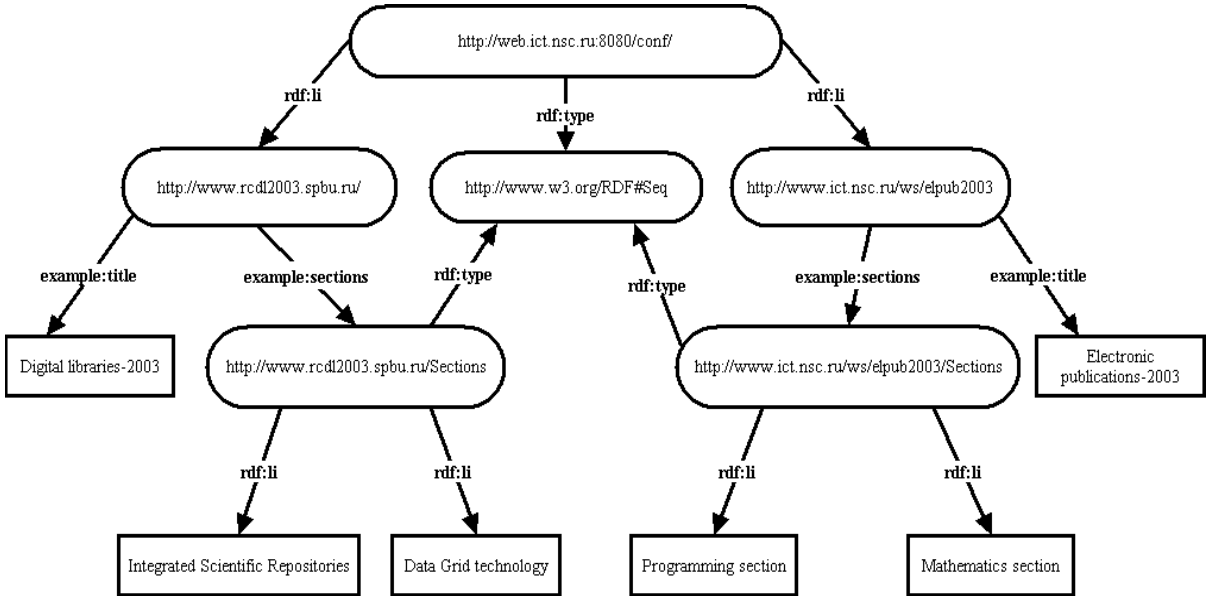

Рис. 5. RDF-модель, описывающая внутреннее представление документа

абстрагированности от предметной области, большинство из них могут использоваться в разных ИС. Система SMART предоставляет несколько стандартных трансформеров для работы с параметрами клиентского запроса, базами данных, словарями и текстовыми данными, к которым могут быть добавлены трансформеры специфичные для разрабатываемой ИС.

3. *Репозитарий стилей*. Стили описывают правила отображения ВПД в документ запрошенного формата. Каждой коллекции соответствует несколько стилей, по одному для каждого формата. Каждый стиль представляет собой XSLT-преобразование (eXtensible Stylesheet Language Transformation), язык стилевого преобразования XML-документов} XML-сериализации RDF-модели ВПД.
4. *Модуль конфигурирования ИС* описывает параметры, которые необходимы системе SMART для определения последовательности действий, в результате которых будет создан запрошенный документ. В частности, конфигурация определяет, какой набор трансформеров соответствует каждой коллекции, и какие стили следует применять для получения документа требуемого формата.

Система SMART берет на себя управление этими ресурсами и обеспечивает функционирование ИС. При этом для генерации ВПД трансформеры могут использовать данные, полученные из

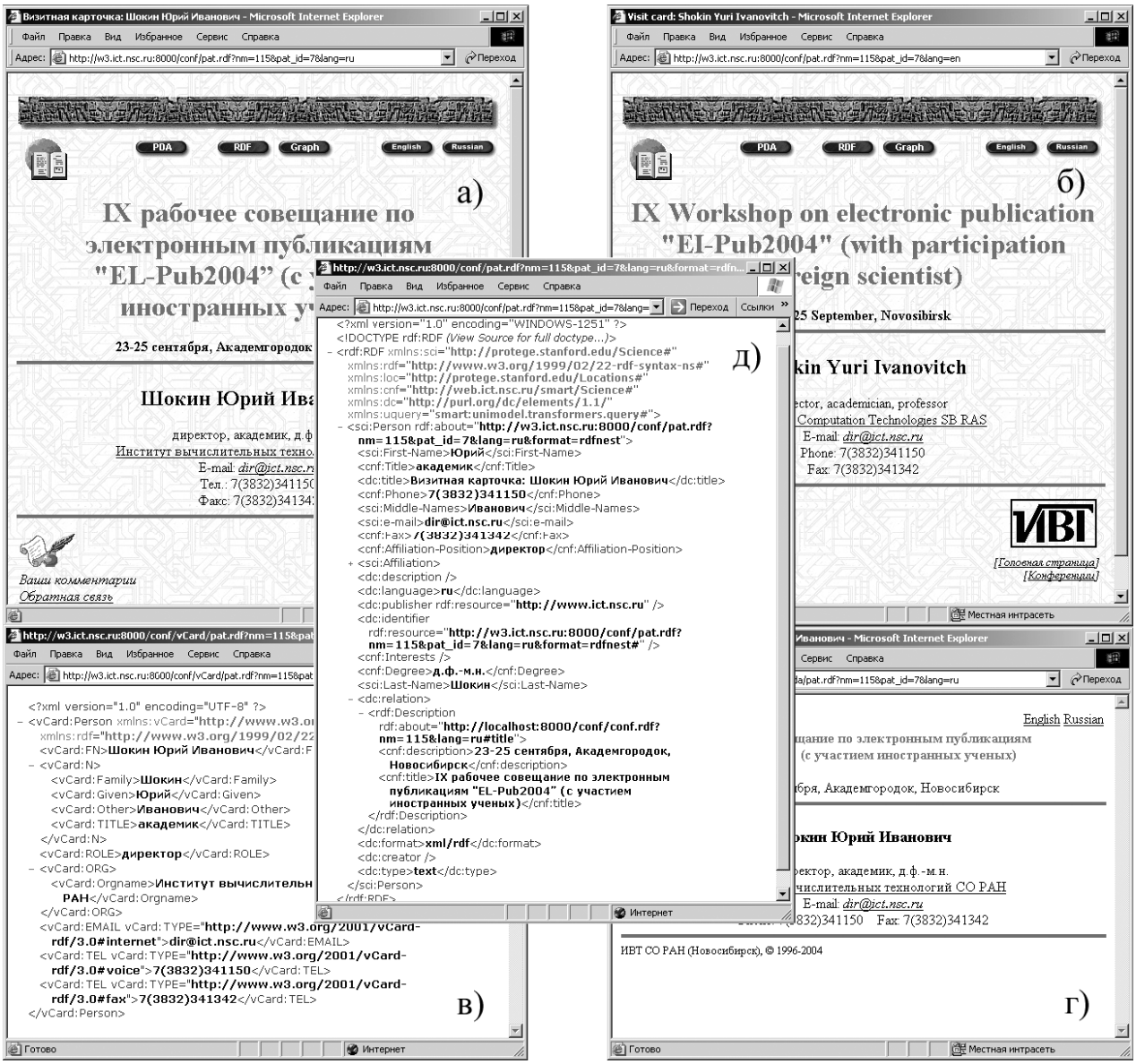

Рис. 6. Возможные конечные представления информации на примере визитной карточки: а) на русском языке (HTML); б) на английском языке (HTML); в) в формате vCard; г) для пользователей КПК; д) исходное RDF-описание документа

различных, внешних по отношению к системе SMART ресурсов, таких как базы данных, файлы, соединения с удаленными объектами и др.

Рассмотрим подробнее упрощенный пример получения ВПД, которое состоит в применении единственного трансформера обращений к БД. На рис. 4 представлен исходный информационный шаблон – RDF-модель коллекции. Жирными линиями выделены подграфы, соответствующие будущему содержанию документа, невыделенные – техническая информация, необходимая для функционирования трансформера и удаляемая на заключительной стадии формирования документа. Данная RDF-модель содержит две инструкции, говорящие трансформеру о выполнении двух SQL-запросов: *ConferenceQuery* и *SectionQuery*. Второй запрос зависит от первого, о чем свидетельствует свойство *smart:DependsOnQueries*, поэтому сначала будет выполнен запрос *ConferenceQuery*. Далее выполняется запрос *SectionQuery*, после чего будет сгенерировано ВПД, представленное в виде RDF-модели на рис. 5.

Следует отметить, что в результате выполнения SQL-запросов может быть получено несколько кортежей результатов (как это и случилось в запросе *ConferenceQuery)*. В таких ситуациях исходная RDF-модель тиражируется необходимое число раз (в нашем примере, 2 раза), и для каждого экземпляра происходит подстановка результатов только из одного кортежа. После этого все экземпляры объединяются, а продублированные элементы удаляются.

Одной из областей применения технологии SMART является преобразование данных из реляционных хранилищ в семантические сети. Подобная задача продиктована тем, что в

определенных случаях в качестве средства обмена информацией оптимально использовать семантические сети, тогда как большинство существующих хранилищ построено на основе реляционных СУБД, поскольку они предоставляют удобные средства для выборки отдельных групп данных, их изменения и редактирования структуры.

На технологической платформе SMART была разработана и внедрена информационно-вычислительная система *Атлас "Атмосферные аэрозоли Сибири"*, расположенная по адресу *http://web.ict.nsc.ru/aerosol*. Атлас предназначен для решения различных задач в области сбора и публикации сведений об атмосферных аэрозолях, в том числе функции математической обработки данных о химическом составе аэрозолей: статистические характеристики, факторный, кластерный и дискриминантный анализ временный рядов и другие функции.

В процессе создания Атласа была подтверждена гипотеза о том, что язык RDF плохо применим для представления больших массивов однородных числовых данных: получаемые документы имеют размер, многократно превышающий количество содержащейся в них информации, что, в частности, приводит к дополнительным затратам ресурсов при их обработке. Поэтому, несмотря на то, что это не помешало выполнить поставленную задачу и в полном объеме реализовать требуемую функциональность Атласа, следует заключить, что технология SMART не является оптимальной для создания систем, ориентированных на вычислительные процессы.

Другим практически значимым результатом является разработка *ИС "Конференции"*, предназначенной для поддержки проведения научных конференций, в рамках которой реализованы все необходимые организаторам средства, включая подготовку и публикацию материалов на web-сайте и в печатных изданиях. Системой поддерживается русскоязычный и англоязычный web-интерфейсы, для каждого из которых могут быть получены документы в HTML-формате, HTML-документы с минимальным числом графических объектов для пользователей карманных персональных компьютеров (КПК) и несколько различных RDF-представлений, включая графическое. Так, все документы на рис. 6 были получены из одного информационного шаблона; все документы, кроме 6.б), имеют одинаковое ВПД, которое представлено в XML-формате на рис. 6.д).

В общем случае можно показать, что технологию *SMART* целесообразно использовать для информационных систем с несколькими форматами представлений документов или для явно выраженного представления информации посредством семантических сетей. При этом время на разработку и поддержку подобных ИС в худшем случае сравнимо со временем, затраченным при использовании других технологических платформ.

5 Семантический поиск

Система SMART включает в себя средства для создания информационного пространства ИС, представленного в виде семантической сети на языке RDF. Уникальность особенностью системы SMART среди подобных систем заключается в возможности осуществления семантического поиска в этом пространстве, т.е. поиска информации в семантической сети. Поисковым запросом в данном случае является фрагмент семантической сети специального вида, определяющий шаблон, которому должны удовлетворять результаты.

Реализация семантического поиска в большой степени зависит от реализации семантической сети. В частности, для семантических сетей, реализованных на основе технологии RDF, существует ряд языков описания поисковых запросов: *TRIPLE, RDQL, RQL, RDFStore, Inkling, SeRQL* (их сравнительные характеристики опубликованы в [3]). К основным недостаткам семантического поиска следует отнести отсутствие простых пользовательских интерфейсов для описания поисковых запросов, что является существенным препятствием для его распространения и массового использования.

В качестве языка построения поисковых запросов был использован язык *RDQL* [7], разработанный в рамках проекта Jena [4] и потому наиболее хорошо интегрируемый в систему SMART. Следует отметить, что хотя RDQL имеет не самые богатые возможности среди своих аналогов, главным аспектом при его выборе была простота интеграции в существующую архитектуру системы. В рамках ИС "Конференции" на платформе технологии SMART был реализован сервис, который осуществляет семантический поиск информации, содержащейся в документах конференций. Показано, что функциональность данного сервиса превышает возможности контекстного поиска и поиска по атрибутам элементов схемы данных.

Так, были реализованы интерфейсы для выполнения следующих часто используемых поисковых запросов:

- Найти все документы, в которых содержатся данные о человеке с указанной фамилией.
- Найти все публикации по вхождению строки в название или аннотацию.
- Найти все публикации по фамилии автора.
- Найти все ресурсы, у которых указанный атрибут схемы *Dublin Core* содержит ключевое слово.

Кроме того, показаны примеры построения более сложных семантических запросов:

- Поиск публикаций, авторы которых работают в Иркутске.
- Поиск участников конференций из серии *El-Pub* (Электронные публикации), работающих в Красноярске.

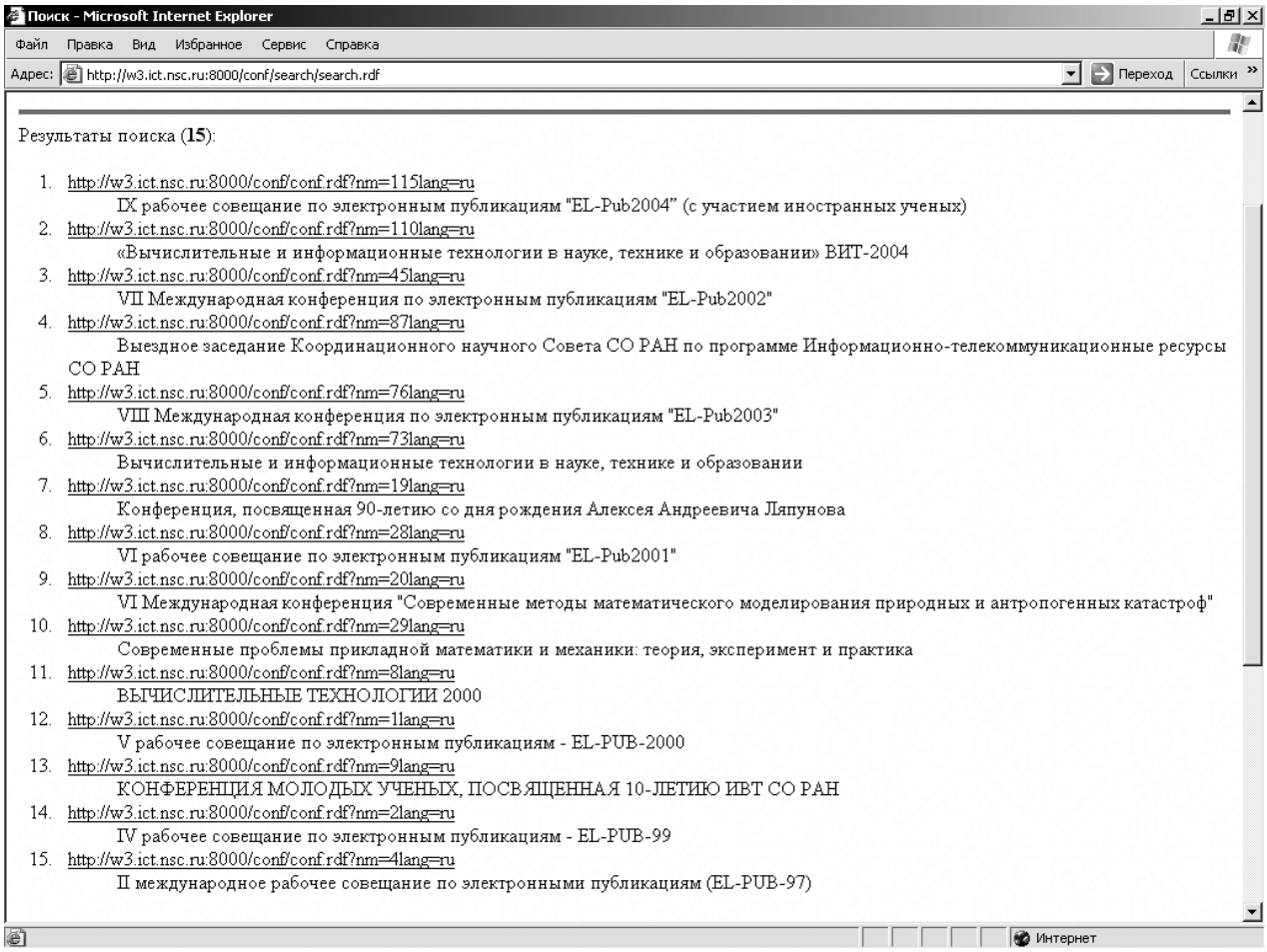

Рис. 7. Результаты поиска конференций, тематика которых относится к информационным технологиям и на которых был заявлен участник по фамилии *Шокин*

- Поиск конференций, тематика которых относится к информационным технологиям и на которых был заявлен участник по фамилии *Шокин* (рис. 7).

Разработанный метод осуществления семантического поиска в рамках отдельной ИС допускает масштабирование и применение в распределенной среде Web. Поэтому в перспективе полученные результаты могут быть использованы при создании информационно-поисковых сервисов, функционирующих в рамках идеологии *Semantic Web* [1].

6 Заключение

Предложенная авторами оригинальная технология SMART была успешно применена при создании ИС для научно-исследовательской и научно-организационной деятельности. Показано, что применение технологии целесообразно в случае разработки web-ориентированных ИС, предоставляющих информацию для нескольких типов пользователей, в частности, web-браузеров и узко специализированных программных агентов проекта Semantic Web.

Литература

[1] Berners-Lee T., Hendler J., Ora Lassila. The Semantic Web // Scientific American, May 2001.

[2] Gruber, T.R. A translation approach to portable ontology specifications. // Knowledge acquisition, 5(2), 199-220. http://ksl-web.stanford.edu/knowledge-sharing/papers/README.html

[3] Haase P., Broekstra J., Eberhart A., Volz R. A Comparison of RDF Query Languages // Proceedings of the Third International Semantic Web Conference, Hiroshima, Japan, 2004. http://www.aifb.uni-karlsruhe.de/WBS/pha/rdf-query/rdfquery.pdf

[4] Jena – A Semantic Web Framework for Java. http://jena.sourceforge.net/

[5] Manola F., Miller E., McBride B. RDF Primer, W3C Recommendation, 10 февраля 2004 http://www.w3.org/TR/rdf-primer/

[6] Ora Lassila, Swick R. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[7] Seaborne A. RDQL – RDF Data Query Language, part of the Jena RDF Toolkit, HPLabs Semantic Web activity. http://hpl.hp.com/semweb/
[8] SMART: System for Managing Applications based on RDF Technology. http://web.ict.nsc.ru/smart
[9] The Apache Cocoon Project. http://cocoon.apache.org
[10] Жижимов О.Л. Введение в Z39.50: изд. 4-е доп. и перераб. – Новосибирск: Изд-во НГОНБ, 2003. – 263 с.
[11] Пратт Т., Зелковец М. Языки программирования: разработка и реализация. 4-е изд. СПб.: Питер, 2002.
[12] Шрайбман В.Б., Гуськов А.Е. Разработка информационных систем на основе RDF-технологии // Труды XLI Международной научной студенческой конференции “Студент и научно-технический прогресс”, Новосиб. гос. ун-т. Новосибирск, 2003 г., Ч. 1. – С. 143-150.

Document generation model based on the semantic nets for creating web-oriented information systems

Fedotov A.M, Guskov A.E.

In this paper the ways of information interchanging in the Web environment are considered. Additional attention is paid to semantic nets concept, which gives the foundation for complete and formalized information description. The document generation model is proposed; it is based on the forming of internal document representation, expressed by semantic nets, and its following transforming to the document in required format. The technology named SMART for developing information systems, which is based on the proposed model, is described. Some questions related to the creation of high-quality information-retrieval services using semantic nets are considered.

♣ Работа выполнена при частичной поддержке Президентской программы НШ 2314.2003.19