Новостной Интернет

Часть 1. Формат синдикации новостей — RSS

Для решения проблемы автоматизации обработки информации, ее обобщения и доставки целевым группам пользователей Сети создано несколько форматов описания данных. Самый распространенный из них - RSS (Really Simple Syndication или Rich Site Summary)

Сегодня Интернет — это огромное хранилище информации, интегрированный доступ к динамической составляющей которого — новостным ресурсам — затруднен. Разнообразие информации в Сети, в том числе и новостных сообщений, не может быть полезным на практике при отсутствии эффективного доступа. Так, по оценкам экспертов, около 79 % журналистов обращаются к Интернету в поисках новостей, и лишь 20 % находят ту информацию, которая им необходима.

Параллельно с визуальным вэбом

Язык HTML, основной формат представления информации в Интернете, описывает лишь внешний вид вэб-сайтов, обеспечивая прежде всего визуализацию данных. Он был разработан исключительно для отображения содержания сайтов, и не всегда удобен для автоматической обработки информации, в том числе и для организации поиска. То есть вся сеть Интернет ориентирована на показ пользователям отдельных сайтов и плохо приспособлена для автоматизированного сбора информации, ее классификации и аналитической обработки. Сегодня представление информации на разных сайтах настолько отличается по оформлению и расположению, что отбирать ее и обрабатывать можно только вручную.

Так, при необходимости обмена информацией между несколькими вэб-сайтами всегда возникает задача унифицированного представления контента. В противном случае изменение HTML-оформления одного сайта приведет к необходимости одновременной модификации программного обеспечения на всех

сайтах, которые принимают его информацию. Аналогичная ситуация возникает при необходимости импортировать информацию на один сайт с нескольких других. Изменения оформления на каждом из сайтов-источников информации будет всегда приводить к необходимости модификации соответствующего программного кода на целевом сайте.

Как видно, сегодня необходимо использование унифицированного формата данных на сайтах, стандарта, обеспечивающего однотипный обмен данными в Интернете. В качестве такого унифицированного формата все шире используется язык eXtensible Markup Language (XML) и его диалекты.

Семантический Web

Одним из первых проектов унификации обмена данными в Интернете стал Семантический Web. Основная идея проекта заключалась в такой организации данных, чтобы вэб-серверы могли их использовать, а не только визуализировать, чтобы программы разных производителей могли эффективно работать с вэб-контентом. Именно для Семантического Web были разработаны спецификации XML, предусматривающие разделение средств визуализации и смыслового содержания.

XML представляет собой метаязык, то есть язык, на базе которого можно определять новые языки. При этом он предназначен не только для организации обмена данными в Web, но и для распознавания семантики этих данных. В отличие от HTML, XML обеспечивает представление информации в чистом виде, предполагая ее структурную, а не оформительскую разметку.

Вместе с тем формально элементы разметки (теги) XML оторваны от определения их смыслового наполнения. Поэтому параллельно с XML была начата разработка стандарта схемы описания источников (Resource Description Framework, RDF) — языка формального описания содержимого вэб-сайтов в рамках единого стандарта. Спецификации RDF поддерживают теги, позволяющие определять любые понятия (например, тегами PRICE и INVOICE можно пользоваться для обозначения цены и счета, соответственно). Следует заметить, что данным в формате RDF присваиваются дескрипторы, которые могут определяться в отдельных файлах определения типов документов (Document Type Definitions, DTD). Сегодня практически в каждой отрасли знаний имеется свой, постоянно расширяющийся список DTD. На основе XML и RDF был создан формат RSS, специально предназначенный для организации информационной коммуникации как между людьми, так и между серверами.

ТЕЛЕКОМ-ИНФО

Спецификации отдельных версий формата RSS

RSS 0.90: *www.purplepages.ie/RSS/netscape/rss0.90.html*
RSS 0.91: *my.netscape.com/publish/formats/rss-spec-0.91.html*
RSS 0.92: *backend.userland.com/rss092*
RSS 0.93: *backend.userland.com/rss093*
RSS 1.0: *web.resource.org/rss/1.0/*
RSS 2.0: *backend.userland.com/rss/*

Синдикация новостной информации

Оптимальное решение, способное помочь ориентироваться в новостной информации Интернета, сегодня предоставляют информационные службы нового типа — системы синдикации новостей. Под синдикацией в данном случае понимается сбор информации в Сети и последующее распространение ее фрагментов в соответствии с потребностями пользователей. Кроме того, службы синдикации обеспечивают публикацию одних и тех же данных на различных сайтах (в том числе предназначенных для карманных компьютеров и мобильных телефонов).

Технология синдикации интернет-новостей включает в себя «обучение» программ сбора структуре выбранных источников (вэб-сайтов), непосредственное сканирование информации, ее приведение к общему формату (в последнее время — к XML), классификацию и доставку пользователям различными путями (e-mail, вэб, WAP, SMS и т. д.).

Форматы синдикации новостей

Для решения задачи синдикации новостей было создано несколько форматов описания данных на основе XML. Самый распространенный формат получил название RSS, что означает Really Simple Syndication, Rich Site Summary, хотя изначально он назывался RDF Site Summary. Смысл всех этих аббревиатур заключается в простом способе обобщения и распределения информационного наполнения вэб-сайтов — синдикации контента.

Изначально RSS создавался компанией Netscape для портала Netcenter как одно из первых XML-приложений, но затем стал использо-

ТЕЛЕКОМ-ИНФО

Самые популярные фиды

http://www.moreover.com/categories/ocs/ocsdirectory.rdf
http://10.am/extra/ocsdirectory.php
http://www.newsisfree.com/ocs/directory.xml
http://blogspace.com/rss/feeds/converted.ocs
http://www.groksoup.com/ocs/ocsdirectory.xml
http://theweb.startshere.net/channels.phtml?format=OCS
http://myrss.com/catalog/ocs04.rdf
http://www.syndic8.com/xml.php
NEWSru.com — *www.newsru.com/plain/rss/all.xml*
Газета.ru — Новости (RSS) — *www.gazeta.ru/export/gazeta_rss.xml*
Lenty.RU — *www.lenty.ru/export/bestnews.rss*
Подробности — *www.podrobnosti.com.ua/export/*
Lenta.ru — *lenta.ru/l/r/EX/import.rss*
Полит.ру — *www.polit.ru/rss/index.xml*
Портал «Юридическая Россия» — *law.edu.ru/rss/news.rss*
Водка он-лайн — *vodka.com.ua/export/rss.xml*
Портал «ПлейМобайл» — *playmobile.ru/news/rss*
3Dnews — *www.3dnews.ru/expnews/rss/newsrss.xml*

ваться на многих других сайтах. Сегодня практически все ведущие новостные сайты, «живые журналы», работающие в Интернете, используют RSS в качестве инструмента оперативного представления своих обновлений. Например, экспорт в RSS осуществляют крупнейшие порталы, включая CNN, BBC News, Amazon, CNet News, MSNBC, The Register, Wired и т. д.

RSS действительно обеспечивает согласованный способ резюмировать содержимое вэб-сайтов. Кроме того, его применение позволило администраторам новостных сайтов, онлайновых дневников — блогов, форумов и других часто обновляемых вэб-ресурсов, представить информацию в унифицированном виде.

Предполагается, что 2004 год станет «Годом RSS», то есть ожидается повсеместное широкое внедрение этого формата. Аналитики отмечают, что только в начале 2004 года интернет-пользователи по-настоящему открыли для себя все прелести технологии RSS. Сегодня для работы с новостями в формате RSS разрабатываются новые программы, сайты и поисковые системы, которые все более востребованы, в частности, пользователями карманных компьютеров.

Итак, RSS — это формат данных и технический стандарт, который обеспечивает интегрированный доступ к новостной информации, представленной на вэб-сайтах, специально созданный для обмена их контентом.

Развитие RSS началось с версии 0.90, разработанной компанией Netscape, но ее посчитали очень сложной, и Netscape создала упрощенную версию — 0.91, которую, после бума портальных технологий, передала компании UserLand Software. Это самый простой и доступный стандарт, который применяется сегодня в тех ситуациях, когда требуется несложный экспорт заголовков. Одновременно еще одна организация — RSS-DEV Working Group, создала свою версию RSS (1.0), близкую к исходной версии RSS 0.90 и максимально приближенную к стандарту RDF. RSS 1.0 предоставляет больше возможностей, чем все 0.9x, например, допускает расширение при помощи модулей. Компания же UserLand решила развить ветвь 0.9x и создала версии 0.92, потом 0.93, 0.94, которые позволяют представлять метаданные, и, наконец 2.0. При этом RSS 2.0 — не новая версия RSS 1.0, а логическое продолжение ветви 0.9x. В ней также добавлена поддержка модулей. В настоящее время существуют семь независимых версий RSS — RSS 0.90, 0.91, 0.92, 0.93, 0.94, 1.0, 2.0. Эти версии отличаются друг от друга, хотя все они ориентированы на один тип информации и содержат одинаковые базовые поля. При этом многие считают все версии, кроме 2.0, устаревшими и «отмененными», но это далеко не так: пока еще самой популярной является RSS 0.91. Что же касается версии 0.94, то ее спецификации не сохранилось даже на авторском сайте Userland. Так, по адресу *backend.userland.com/rss094* находится спецификация версии RSS 2.0. Адреса вэб-страниц, содержащие спецификации отдельных версий формата RSS, приведены во вставке на с. 55, вверху.

Во всех версиях RSS есть некоторые особенности, но объединяет их ориентация на один тип информации, вследствие чего они содержат общие базовые поля: основной блок данных (channel), который содержит атрибуты заглавия канала (title), ссылки (link), данные о языке сообщений (language) и логотип (image), после которых идет список самих сообщений, где в каждом пункте (item) указывается заголовок (title), краткое описание (description) и ссылка на новость (link). Кроме того, каждый RSS-файл начинается обязательными элементами xml и rss. Первый из этих элементов содержит атрибуты version (версия) и encoding (кодировка).

Среди множества необязательных элементов RSS можно назвать самые распространенные — язык (language), copyright, категория информации (category), дата и время публикации сообщения (pubDate), программа, которая использовалась для создания файла (generator), картинка, которую следует показывать наряду с текстовой информацией (image).

Кроме заголовка блока данных в формате RSS предусмотрено описание отдельных информационных элементов (item). Каждый элемент <item> — это отдельная статья или краткая аннотация и ссылка на полную версию статьи. Канал (channel) может содержать любое число элементов <item>, содержащих только два обязательных вложенных элемента — название (title) и описание (description). Кроме того, часто используются такие вложенные элементы, как ссылка на первоисточник (link), категория (category), комментарий (comments) и автор (author).

Помимо формата RSS, недавно появился формат Atom 3.0 (*www.mnot.net/drafts/draft-nottingham-atom-format-02.html*), пока окончательно не утвержденный, но используемый на крупнейшем поисковом портале Google, что предопределяет его популярность. Открытый стандарт Atom совершенствуется командой программистов из IBM, Google и других компаний. Как

ТЕЛЕКОМ-ИНФО

Новостные фиды

Аргументы и Факты — *www.aif.ru / info / rss.php?magazine=aif*
АвтоОБЗОР — *auto.obzor.ru / news / autonews.xml*
АвиаПорт.Ру — *www.aviaport.ru / news / yandex_export.xml*
Деловая Хроника — *www.chronicle.ru / l / r / EX / rsschannel.xml*
K2Kapital — *ad.k2kapital.com / cbp / mynetscape / mynews.news*
Linux.org.ru — *images.linux.org.ru / getrss.php3*
PalmQ Online — *www.palmq.net / backend.php*
СПОРТ сегодня — *www.sports.ru / sports_docs.xml*
TRAVEL.RU. Все о путешествиях — *www.travel.ru / inc / side / yandex.rdf*
АПК-Информ — *www.apk-inform.com / yandextr.php*
ФОНТАНКА.РУ — *www.fontanka.ru / _transmission_for_yandex.thtml*
IMA Press. Тема дня — *www.ima-press.ru / rss.php?newsblock=theme&limit=1*
Журнал «Итоги» — *www.itogi.ru / WebExport.nsf / Anons / itogi.xml*
Остров. Новости Донбасса — *www.ostro.org / yandex.php*
Полит.ру — *www.polit.ru / rss / index.xml?yandex_mode=1*
PRAVDA.Ru — *export.pravda.ru / yandex.txt*
PR NEWS (все пресс-релизы компаний) — *www.prnews.ru / yandex / business.asp*
Энциклопедия поисковых систем — *www.searchengines.ru / news / news.rdf*
Сетевой журнал — *www.setevoi.ru / weekly / export1.txt*

и RSS, Atom является подмножеством XML.

Дэйв Уинер, один из главных разработчиков RSS, недавно призвал разработчиков объединить свои усилия и создать единый формат, совместимый как с RSS, так и с Atom, чтобы слить конкурентные стандарты в единое целое. «Новый формат можно назвать RSS/Atom, — заявил Уинер, — Он бы имел всю функциональность, которую разработчики Atom обещают внедрить. Максимально авторитетный формат получил бы наиболее полную поддержку от всех разработчиков». Уинер предлагает, чтобы в RSS/Atom было как можно меньше отличий от RSS 2.0.

Еще один диалект XML — OPML (Outline Processor Markup Language), используется для описания совокупности RSS-фидов; его спецификация размещена по адресу *opml.scripting.com/spec*. С помощью OPML обеспечивается эффективный унифицированный обмен списками RSS-фидов. Так, для доступа ко всем новостям службы All Headline News пользователю достаточно указать адрес *www.allheadlinenews.com/feeds.opml* в соответствующем окне своей программы чтения RSS, поддерживающей OPML (например, FeedDemon). В списке доступных RSS-фидов сразу же окажутся более 100 каналов службы, таких как All Headline News — Accounting, All Headline News — Acupuncture, All Headline News — Adolescent Health, All Headline News — Adventure Sports, All Headline News — Advertising, All Headline News — Aerospace, и др.

Источники новостного контента

Основным применением RSS в настоящее время являются новостные фиды (feed). Фид — это файл в формате RSS, в который записывается новостной контент вэб-ресурса. Если есть необходимость оперативно отслеживать изменения на сайте, содержащем фид, то можно делать это, используя программу-агрегатор, не посещая самого сайта с помощью стандартных программ-браузеров. Адреса самых популярных в Интернете фидов приведены во вставке на с. 55, внизу.

Обширный список RSS-фидов русскоязычного сегмента Интернета находится по адресу *my.yandex.ru/rss.opml*. Наиболее интересные новостные фиды перечислены во вставке на с. 56.

На сегодня существует множество служб синдикации новостей, которые предоставляют доступ в тематические фиды, построенные на основе использования многочисленных источников. Такой фид, к примеру, доступен на портале UAport (*uaport.net*) и позволяет получить интегрированный доступ к потоку украинских и российских новостных сообщений, собираемому системой InfoStream. С помощью RSS-шлюза системой InfoStream предоставляется унифицированный доступ к информации более чем с 600 вэб-сайтов, сгруппированной по тематикам, языкам, странам, источникам. Объем этой информации сегодня превышает 20 тыс. сообщений в сутки. RSS-каналы UAport могут генерироваться системой по собственным запросам пользователей к поисковой системе.

Во второй части статьи речь пойдет о функциональности некоторых служб синдикации новостей, предоставляющих информацию в формате RSS, а также RSS-агрегаторах (в том числе и для мобильных устройств). ●

Дмитрий Ландэ, dwl@visti.net,
Александр Морозов, alex@visti.net