RSS,

или контент туда и обратно

Почему формата HTML недостаточно?

Общеизвестно, что для написания веб-сайтов используется язык HTML, которым описывается их внешний вид, т. е. обеспечивается визуализация. Этот формат был разработан для решения задач отображения содержания на каждом конкретном ресурсе, поэтому не всегда удобен для автоматической обработки информации, в том числе и организации поиска. В результате сеть WWW оказалась ориентирована, прежде всего, на представление информации на отдельных сайтах и очень слабо приспособлена для автоматизированного обобщения информации, ее классификации и аналитической обработки.

Очень часто возникает необходимость обмена информацией, например, между несколькими сайтами, при этом всегда возникает вопрос о технологии однотипного представления их содержания. Если такая технология не используется, то изменение HTML-оформления одного сайта приведет к необходимости одновременной модификации программного обеспечения на всех сайтах, которые принимают от него информацию. Приблизительно такая же ситуация возникает при необходимости импортировать информацию на один ресурс с нескольких других, предположим, тематически близких. Изменения оформления на каждом из сайтов-экспортеров информации будет каждый раз требовать модификации соответствующего программного кода на сайте-импортере.

Все это обусловило необходимость использования унифицированного представления данных. Потребовался некоторый стандарт представления информации на сайтах, обеспечивающий однотипный обмен данными в такой сложной системе, как Интернет. Сегодня в качестве такого унифицированного формата все чаще используется формат RSS (рис. 1), базирующийся на компонентах Семантического веб XML и RDF.

Рис. 1. Официальная иконка RSS

Две компоненты из Семантического веб

Одним из первых проектов консорциума W3C, призванных решить задачи унификации обмена данными в Сети, стал Семантический веб. В его основу была положена идея, состоящая в том, что серверы должны уметь не только визуализировать информацию, но и использовать ее. Таким образом, различные программы разных производителей могли эффективно работать с данными из Сети. Дело оставалось за малым – создать правила формирования блоков информации, которые смог бы понять не только человек, но и компьютер. В рамках проекта Семантического веб были разработаны спецификации XML, предусматривающие разделение содержания, представления и смыслового значения. Коротко остановимся на двух компонентах Семантического веб, являющихся основами RSS-технологии.

XML представляет собой метаязык, то есть язык, на базе которого можно определять новые языки. В отличие от HTML, XML предназначен для представления информации в «чистом» виде, предполагая структурную, а не оформительскую разметку данных. Вместе с тем XML, являясь необходимой частью решения задачи обмена информационным наполнением сайтов, сам по себе не может дать ничего того, что необходимо для инфраструктуры обработки данных. Дело в том, что формально теги XML оторваны от определения их смыслового наполнения.

Параллельно с XML консорциумом W3C была начата разработка стандарта схемы описания источников (Resource Description Framework, или RDF), модели, в которой данные о ресурсах представляются в виде, пригодном для машинной обработки. Ресурсом в RDF может быть любая сущность – как информационная (например, веб-сайт или изображение), так и неинформационная (например, планета или табуретка). Утверждение, высказываемое о ресурсе, имеет вид «субъект – предикат – объект» и называется триплетом. Утверждение «веб-сайт с современным дизайном» в RDF-терминологии можно представить следующим образом: субъект – «веб-сайт», предикат – «имеет дизайн», объект – «современный». Множество RDF-утверждений образует ориентированный граф, в котором вершинами являются субъекты и объекты, а ребра помечены предикатами.

Назначение RSS

На основе XML и RDF был разработан формат RSS, специально предназначенный для легкого и быстрого обмена контентом между сайтами – организации информационной коммуникации между серверами. RSS – это 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.

Изначально RSS создавался компанией Netscape для своего портала Netcenter как одно из первых XML-приложений, но затем быстро завоевал популярность и стал достаточно широко использоваться. Вскоре эта технология уже использовалась для трансляции контента на многих новостных сайтах, в том числе таких, как BBC, CNET, CNN, Disney, Forbes, Wired, Red Herring, Slashdot, ZDNet и многих других.

Аббревиатура RSS предполагает неоднозначные, но понятийно близкие трактовки. В разных версиях аббревиатура RSS имела разные расшифровки:

- Rich Site Summary (RSS 0.9x) – обогащенная 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 простой сбор сводной информации.

Подразумевается, что речь идет о простом способе обобщения и распределения информационного наполне-

ния (синдикации) сайтов. Обычно с помощью RSS 2.0 дается краткое описание новой информации, появившейся на сайте, и ссылка на ее полную версию. Интернет-ресурс в формате RSS называется RSS-каналом, RSS-лентой или RSS-фидом.

Благодаря RSS администраторы сайтов новостей, блогов, форумов и других часто обновляемых web-ресурсов получили простой и унифицированный метод подачи информации о происходящих событиях. Сегодня RSS рассматривается, в первую очередь, как формат, предназначенный для публикации и обеспечения экспорта новостей на новостных сайтах. После того как информация преобразована в формат RSS, любая программа, ориентированная на данный формат, может загружать сведения об обновлениях веб-сайтов и далее, в зависимости от результата, выполнять определенные действия, например, автоматически обновлять список актуальных информационных сообщений.

Многие современные браузеры, почтовые клиенты и интернет-пейджеры умеют работать с RSS-лентами, среди них Safari, Maxthon, Miranda, Mozilla Firefox, Mozilla Thunderbird, Opera, Opera mini, Windows Internet Explorer (начиная с 7-й версии), Google Chrome. 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 Bloglines.

Семейство RSS

Итак, RSS – это формат данных и технический стандарт, который обеспечивает интегрированный доступ к новостной информации, представленной на сайтах, специально созданный для обмена их контентом. RSS имеет несколько независимых версий. Развитие данного формата началось с версии 0.90, разработанной компанией Netscape на основе RDF. Но так как он считался очень сложным, Netscape представила его упрощенную версию – 0.91, которую в 2000 г. передала компании UserLand Software. Затем был предложен формат RSS 1.0, также основанный на стандартах XML и RDF. Расширения формата предлагалось делать через модули расширений, описываемые в своих пространствах имен. Компания же UserLand решила развить ветку 0.9. Дейв Вайнер, работающий в компании «UserLand Software», опубликовал спецификацию RSS 0.92, которая является развитием версии 0.91 и ориентируется на тех пользователей, которым RDF-описание показалось излишне сложным. Вайнер смог популяризовать свою разработку и придумал свою расшифровку аббревиатуры – Really Simple Syndication (очень простое приобретение информации).

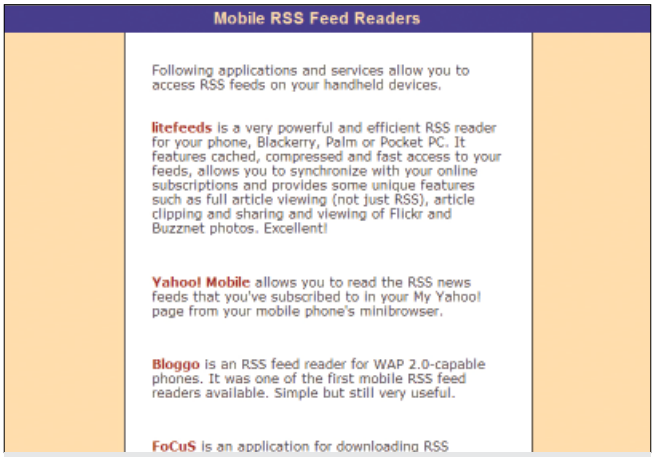


Рис. 2. Фрагмент списка RSS-агрегаторов для мобильных устройств

Дальнейшим развитием этой ветки стал формат RSS 2.0, который тоже поддерживает расширения с помощью модулей, лежащих в своих пространствах имен.

Дэйва Вайнера обвиняли в коренном изменении сути протокола RSS и в захвате номера версии 2.0. В знак протеста один из соавторов RSS 1.0, Аарон Шварц, сразу после выхода RSS 2.0 опубликовал спецификацию RSS 3.0 (http://rss30.com/), однако ее можно считать пародией или протестом, так как она не укладывается в XML-формат.

Фрагмент RSS-канала в формате 2.0 (http://ru.wikipedia.org/wiki/RSS):

```
<?xml version=»1.0»?>
<rss version=»2.0»>
 <channel>
  <title>Liftoff News</title>
  <link>http://liftoff.msfc.nasa.gov/</link>
  <description>Liftoff to Space Exploration.</description>
  <language>en-us</language>
  <pubDate>Tue, 10 Jun 2003 04:00:00 GMT</pubDate>

  <lastBuildDate>Tue, 10 Jun 2003 09:41:01 GMT</lastBuildDate>
  <docs>http://blogs.law.harvard.edu/tech/rss</docs>
  <generator>Weblog Editor 2.0</generator>
  <managingEditor>editor@example.com</managingEditor>
```

Рис. 3. Скриншот веб-сайта программы LiteFeeds (www.litefeeds.com)

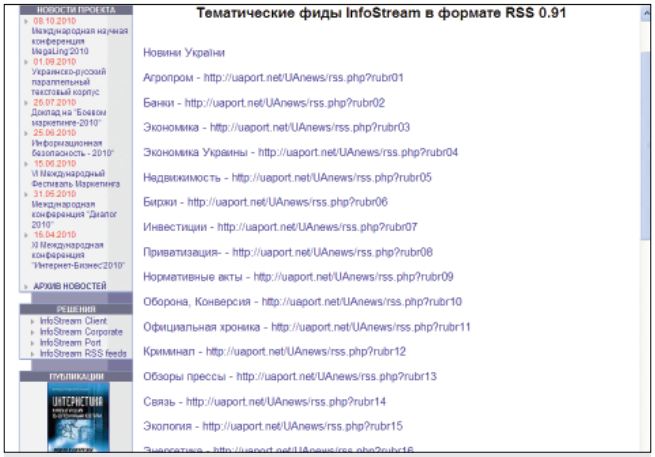


Рис. 4. Список стандартных запросов к UAport (www.infostream.ua/rss1)

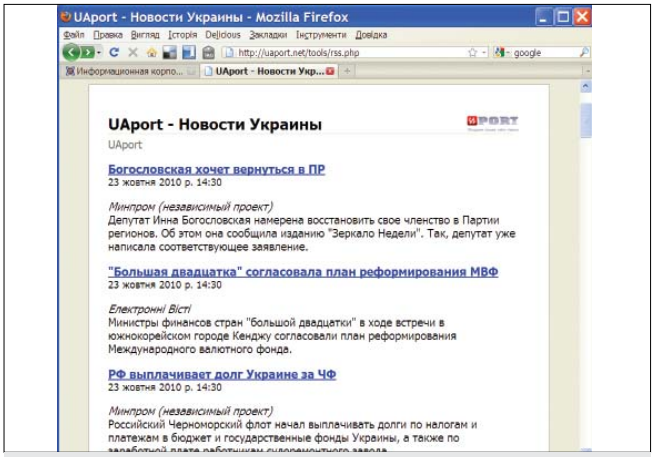


Рис. 5. Фрагмент RSS-фида, сформированного по стандартному запросу к UAport

```
    <webMaster>webmaster@example.com</webMaster>

    <item>
     <title>Star City</title>
     <link>http://liftoff.msfc.nasa.gov/news/2003/news-starcity.asp</link>
      <description>How do Americans get ready to work with Russians
aboard the
          International Space Station? They take a crash course in culture,
language
       and protocol at Russia's Star City.</description>
     <pubDate>Tue, 03 Jun 2003 09:39:21 GMT</pubDate>
     <guid>http://liftoff.msfc.nasa.gov/2003/06/03.html#item573</guid>
    </item>
    . . .
    <item>
     <title>Astronauts' Dirty Laundry</title>
     <link>http://liftoff.msfc.nasa.gov/news/2003/news-laundry.asp</link>
     <description>Compared to earlier spacecraft, the International Space
      Station has many luxuries, but laundry facilities are not one of them.
      Instead, astronauts have other options.</description>
     <pubDate>Tue, 20 May 2003 08:56:02 GMT</pubDate>
     <guid>http://liftoff.msfc.nasa.gov/2003/05/20.html#item570</guid>
    </item>
   </channel>
  </rss>
```

Все версии RSS отличаются друг от друга, но объединяет их то, что они ориентированы на один тип информации и содержат одинаковые базовые поля. Основной блок информации (channel), состоящий из названия (title), ссылки (link), данных о языке новостей (language) и логотипа (image). Затем идет список самих новостей, где в каждом пункте (item) указывается заголовок (title), краткое описание (description) и ссылка на новость (link). Пример фрагмента RSS-канала см. во врезке. Вместе с тем из-за существования различных версий формата RSS-каналов программы их обработки должны уметь работать со всеми вариантами, что создает некоторые трудности при их разработке. Так существуют проблемы с различными форматами представления дат и метаданных, таких как частота обновления.

RSS – агрегаторы

Пользователи могут получить доступ к данным в формате RSS с помощью специальных программ, называемых RSS-агрегаторами. То есть RSS-агрегатор – это клиентская программа или веб-приложение для автоматического сбора сообщений из источников, экспортирующих в формат RSS, например заголовков новостей, блогов, подкастов. Современная программа-агрегатор позволяет группировать публикации, т. е. обеспечивает возможность одновременно отслеживать появление новостей на всех RSS-каналах, без посещения каждого ресурса в отдельности.

Все RSS-агрегаторы можно условно разделить на те, ко-

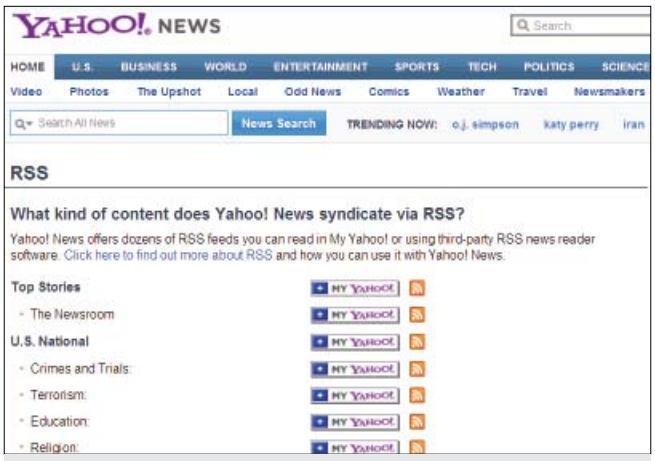


Рис. 6. Страница Yahoo! с «готовыми» RSS-фидами

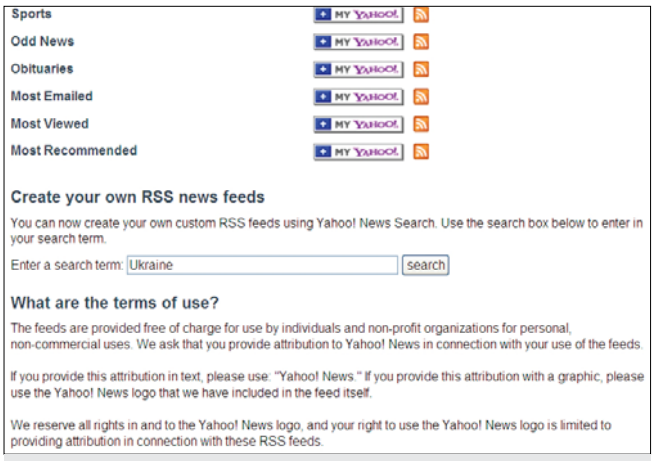


Рис. 7. Форма для ввода запросов в системе Yahoo! для подписки на RSS

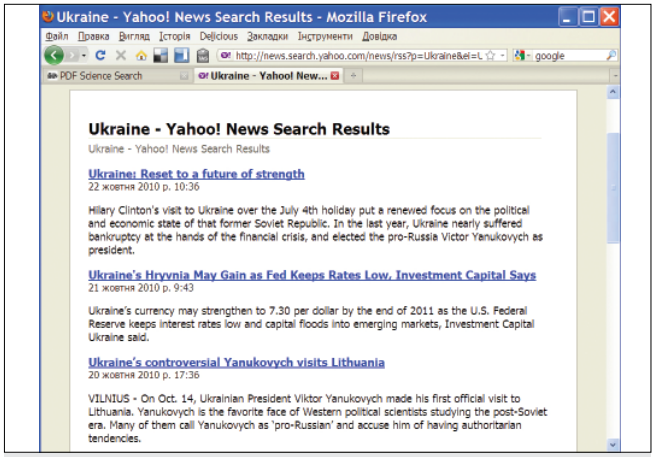


Рис. 8. Фрагмент RSS-фида Yahoo! по запросу «Ukraine»

Рис. 9. Фрагмент страницы с результатами поиска на сайте Google News (http://news.google.com/) с гиперссылкой на RSS-представление

торые доступны через веб-интерфейс, и устанавливаемые на компьютер. Задачи их одинаковы – получение обновлений из интересующих пользователя RSS-источников. Исчерпывающий список RSS-агрегаторов можно найти, например в Википедии (http://ru.wikipedia.org/wiki/Список_RSS-агрегаторов), приведем лишь некоторые из них.

Доступными через веб-интерфейс являются агрегаторы – веб-приложения, расположенные на серверах в Интернете, т.е. к ним можно получать доступ с любого компьютера, подключенного к Интернету. Среди доступных через веб-интерфейс можно назвать, например, FriendFeed (http://friendfeed.com/), Google Reader (http://reader.google.com/), RSS2Email (http://rss²email.infogami.com/, с возможностью получать RSS-новости на почту), RSSLenta.ru (http://rsslenta.ru/, веб-агрегатор с возможностью внедрения любых RSS на веб-страницу в виде настраиваемой ленты).

Устанавливаемый на компьютер агрегатор – это отдельная программа или встроенный в браузер, в почтовый клиент или даже в операционную систему модуль. Из устанавливаемых на компьютер под управлением ОС типа Windows можно назвать такие программы (стандартные веб-браузеры уже были названы), как Abilon (http://www.americantowns.com/wa/albion), FeedDemon (http://ru.wikipedia.org/wiki/FeedDemon), FeedReader (http://ru.wikipedia.org/wiki/FeedReader), RSS Reader (http://www.rssreader.com/), RSSBandit (http://rssbandit.org), Syndirella (http://www.yole.ru/projects/syndirella) и многие другие подобные программы.

Для ОС Linux существуют такие программы, как Liferea (http://liferea.sourceforge.net/), Akregator (http://akregator.sf.net/), Bottom Feeder (http://www.cincomsmalltalk.com/BottomFeeder/), Syndigator (http://syndigator.sourceforge.net/) и K.R.S.S. (http://krss.sourceforge.net/).

Существует более пятидесяти программ для чтения RSS-фидов под Mac OS X, из которых большая часть распространяется по свободным лицензиям, среди них NetNewsWire Lite (http://netnewswireapp.com/), NewsFire (http://www.newsfirerss.com/), NewsLife (http://www.thinkmac.co.uk/newslife/), Vienna (http://www.vienna-rss.org/) и др.

Благодаря своей простоте и стандартизированности RSS стал одним из самых популярных форматов для работы с новостной информацией на мобильных устройствах. Для работы с этим форматом существуют многочисленные программы, устанавливаемые на мобильных телефонах, коммуникаторах, КПК. По адресу http://www.feed-readers.com/mobile-rss-readers. htm (рис. 2) приведен список из 12 самых популярных программ чтения RSS-фидов для мобильных устройств. На рис. 3 представлен скриншот сайта одной из таких программ LiteFeeds.

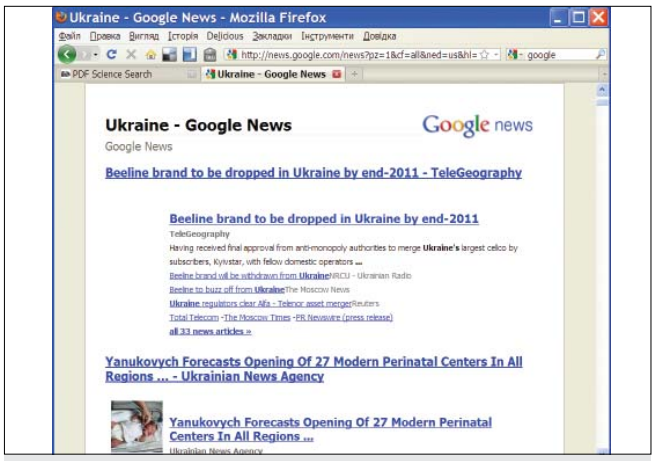


Рис. 10. Фрагмент RSS-фида Google News по запросу «Ukraine»

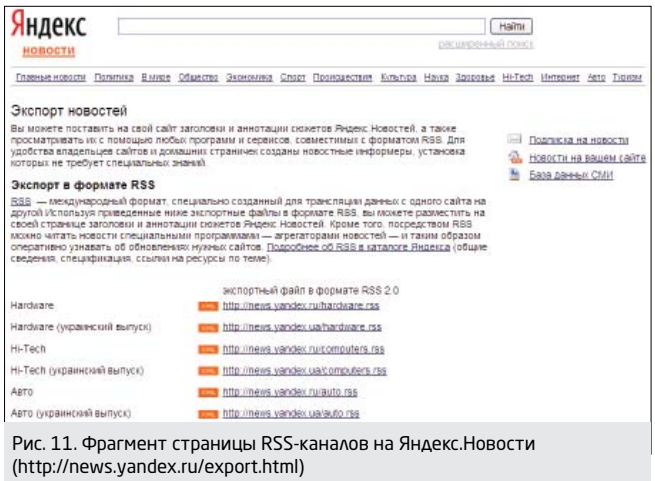
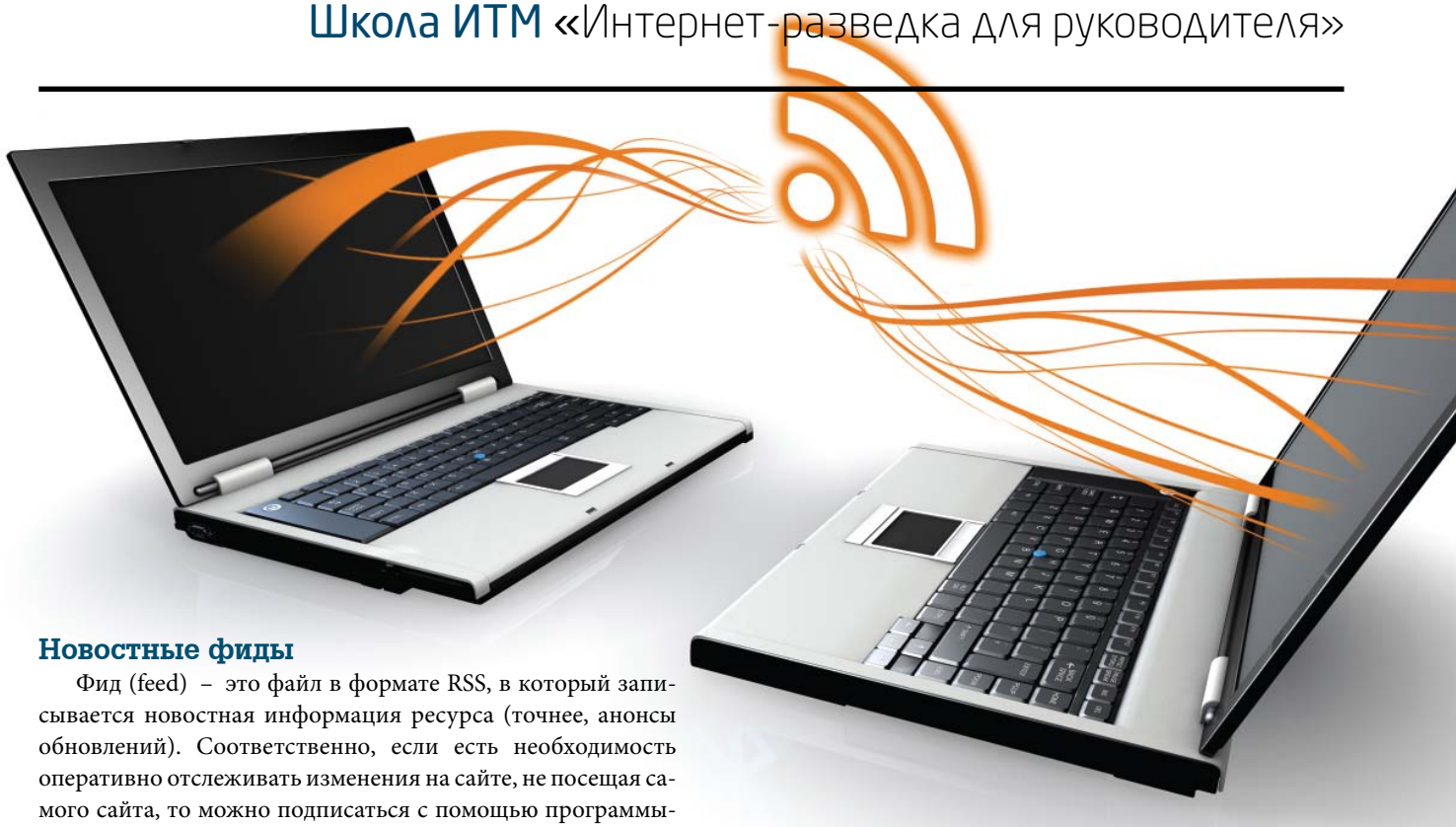


Рис. 11. Фрагмент страницы RSS-каналов на Яндекс.Новости (http://news.yandex.ru/export.html)



Новостные фиды

Фид (feed) – это файл в формате RSS, в который записывается новостная информация ресурса (точнее, анонсы обновлений). Соответственно, если есть необходимость оперативно отслеживать изменения на сайте, не посещая самого сайта, то можно подписаться с помощью программы-агрегатора на фид.

Например, у пользователей веб-портала UAport (http://uaport.net) имеется возможность получить интегрированный доступ к потоку украинских и российских новостных сообщений из Интернета с помощью RSS-шлюза с системой InfoStream. Последняя предоставляет интегрированный доступ к информации, получаемой более чем с 4000 web-сайтов и сгруппированной по тематикам, языкам, странам, источникам. Объем данных, обрабатываемых в рамках технологии InfoStream, сегодня превышает 80000 сообщений в сутки. RSS-каналы UAport могут генерироваться системой по запросам пользователей. Список стандартных запросов к UAport приведен на рис. 4. В качестве примера новостного фида формата RSS 0.91 на рис. 5 показан динамический RSS-фид из этого списка.

Использование RSS-технологий допускает своеобразный замкнутый цикл – информация публикуется на информационных веб-сайтах, на основании чего формируются RSS-каналы. RSS-каналы, с одной стороны, считываются пользователями, а с другой, экспортируются в виде лент новостей на веб-сайты (к счастью для пользователей, первичная информация и экспортируемые новости чаще всего публикуются на различных местах веб-страниц).

RSS на поисковых системах

Сегодня практически все ведущие поисковые системы используют RSS-технологии для представления результатов поиска новостей. Это дает возможность пользователям подписываться на новостные RSS-фиды по своим постоянным запросам, оперативно экспортировать результаты поиска на мобильные устройства.

Так на поисковой системе Yahoo! наряду с возможностью подписки на RSS-фиды (рис. 6, http://news.yahoo.com/rss) существует возможность подписки на результаты поиска (рис. 7). На рис. 8 приведен фрагмент результатов подписки в системе новостей Yahoo! по запросу «Ukraine».

Самая популярная российская поисковая система Яндекс, вернее ее раздел Яндекс-Новости, также использует RSS-формат для представления результатов поиска по стандартным запросам. На рис. 11 приведен фрагмент страницы для подписки на RSS-фиды Яндекса.

Заключение

Ускоряющийся темп жизни требует оперативного получения полной и актуальной информации, которая должна учитываться при принятии управленческих, маркетинговых, производственных решений.

Современная RSS-технология интегрированного доступа к данным обеспечивает пользователям Интернета простой и оперативный доступ к оперативной информации, размещаемой на информационных сайтах. Можно утверждать, что RSS является одной из первых реально работающих технологий Семантического веб.

Вместе с тем ориентация исключительно на RSS-ресурсы при работе с новостным контентом Интернета связана с опасностью понижения информационной полноты, ведь не все веб-ресурсы до сих пор поддерживают собственные RSS-фиды, для некоторых ресурсов создание RSS-фидов сопряжено с ручной работой администраторов, которые не всегда ответственно относятся к их поддержке. То есть ориентация на стандартизированные средства RSS, конечно же, ускоряет работу с информацией, позволяет оперативно ее импортировать/экспортировать, однако на данном этапе пока еще ведет к некоторой потере качества.

Итак, перспективность и популярность RSS как стандарта обусловлена, прежде всего, его доступностью и простотой. Поэтому практически все ведущие мировые информационные сайты, блоги, форумы используют RSS в качестве инструмента оперативного представления новой информации.

Дмитрий ЛАНДЭ,
зам. директора ИЦ «ЭЛВИСТИ»