Content Syndication with RSS and Perl

H. Wade Minter
HCS^H^H^H WebAssign
<minter@lunenburg.org>
http://www.lunenburg.org/
Raleigh Perl Mongers
Thursday, April 21, 2005
Topics Tonight

• What is content syndication/RSS?
• Generating feeds with Perl.
• Parsing feeds with Perl.
• Tools for using feeds.
What is RSS?

- Originally developed by Dave Winer at Userland in 1997.
- RSS 0.90 debuts for the my.netscape.com portal in 1999. Mostly XML.
- RSS 1.0 spec published in 2000. RDF, not related to any other RSS format.
- RSS 0.92 also shows up in 2000.
- RSS 2.0 spec designed in 2002, released under Creative Commons license in 2003.
What IS RSS?

• A way to provide content in a manner that is easily distributed, parsed, and read by client applications.

• A fast-growing way to publish, adopted by blogs and major media outlets alike.

• A framework for building some pretty cool applications.
What’s good about RSS?

• Popular.
• Very well-supported by content generators and consumers.
• Newer versions of the spec are under open community development (CC license) (vs. one-man kingdom).
What’s bad about RSS?

• Dave Winer :-)

• Spec designed by crack monkeys - each version is incompatible with the next (http://diveintomark.org/archives/2004/02/04/incompatible-rss).

• Limited feature set (by design).
Atom

- New content syndication format on the block.
- First known as Pie, then Echo, now Atom.
- Aiming to become an IETF standard.
- Version 0.3 the current “standard” version.
What’s good about Atom?

• Designed in large part by Triangle residents Mark Pilgrim and Sam Ruby, both of whom know their stuff.

• Fixes many mistakes in RSS, format much more robust.

• Is actually an entire content publishing API, instead of just a syndication format.

• Has support of industry heavyweights like Google.
What’s bad about Atom?

• Spec still in draft, process moving slowly.

• Much less support in toolkits and client software.

• Perl bindings for crap.

• May be too complicated for its own good.
Focus tonight on RSS

- Most common format.
- Perl toolkits very useful.
- Questions on overview?
Generating RSS with Perl

- Several Perl toolkits (XML::RSS, XML::RSS::SimpleGen, Apache::RSS, etc).
- We will focus on XML::RSS.
- Good documentation in the POD.
- Requires XML::Parser, which requires expat.
#!/usr/bin/perl

use warnings;
use strict;

use XML::RSS;

my $rss = new XML::RSS( version => '2.0' );
$rss->channel(
    title => 'Example Feed',
    link => 'http://example.com'
);

$rss->add_item(
    title => "First Post",
    link => "http://example.com/news/articles/1.html",
    description => "Natalie Portman, hot grits, etc., etc."
);

print $rss->as_string;
Sample Output

<?xml version="1.0" encoding="UTF-8"?>

<rss version="2.0" xmlns:blogChannel="http://backend.userland.com/blogChannelModule">
  <channel>
    <title>Example Feed</title>
    <link>http://example.com</link>
    <description></description>
  </channel>
  <item>
    <title>First Post</title>
    <link>http://example.com/news/articles/1.html</link>
    <description>Natalie Portman, hot grits, etc., etc.</description>
  </item>
</rss>
Notes

• Previous output was valid RSS 2.0 (use Feed Validator at http://www.feedvalidator.org/ to check).

• Easy interface (create the channel, add items to it, spit out the output).

• Many other options available (categories, dates/times, etc. - check POD).

• Can even save RSS as JavaScript print statements (XML::RSS::JavaScript).
#!/usr/bin/perl -w

use strict;
use warnings;

use DBI;
use XML::RSS;
use Date::Manip;
use DateTime;
use DateTime::Format::W3CDTF;
use HTML::FromText;

# How many stories do we want to show at once?
my $stories = 10;

use DBI;
my $dbh = DBI->connect( "dbi:Pg:dbname=$dbname", $dbuser, $dbpass );

my $query = "SELECT * FROM news ORDER BY date_modified DESC LIMIT $stories";
my $sth = $dbh->prepare($query);
$sth->execute;

my $dt  = DateTime->now;
my $f   = DateTime::Format::W3CDTF->new;
my $now = $f->format_datetime($dt);
my $rss = new XML::RSS( version => '1.0' );
$rss->channel(
    title => "ComedyWorx Player Headlines",
    link => "http://www.comedyworx.com/players/",
    description => "Headlines for ComedyWorx players",
    dc => {
        date => "$now",
        subject => "Improv",
        creator => 'webmaster@comedyworx.com',
        publisher => 'webmaster@comedyworx.com',
        rights => 'Copyright 2004, ComedyWorx',
        language => 'en-us',
    },
    syn => {
        updatePeriod => "hourly",
        updateFrequency => "1",
        updateBase => "1901-01-01T00:00+00:00",
    },
);
while ( my @table_row = $sth->fetchrow_array )
{
    my ( $story_id, $poster, $headline, $body, $date_modified ) = @table_row;
    my $epoch = UnixDate( ParseDateString($date_modified), "%s" );
    my $dt = DateTime->from_epoch( epoch => $epoch );
    $dt->set_time_zone( 'America/New_York' );
    my $f = DateTime::Format::W3CDTF->new;
    my $date = $f->format_datetime($dt);
    my $html = text2html( $body, lines => 1 );

    $rss->add_item( 
        title => "$headline",
        link => 
        description => "$html",
        dc => { 
            date => "$date",
            creator => "$poster",
        },
    );
}

print $rss->as_string;

$dbh->disconnect;
Conclusions

• If your content is in a database, it’s absurdly easy to provide RSS feeds.

• RSS feeds can be generated statically or dynamically.

• Links should be permanent.

• Questions on content generation?
Still Awake?
Parsing RSS Content

• Once you have RSS content available, what do you do with it?

• Perl toolkits can also parse RSS feeds and let you get to the data.

• Big advantage of XML - data is standardized and in machine-readable format.

• Once you have the data, you can do almost anything.
#!/usr/bin/perl

use warnings;
use strict;

use XML::RSS;
use LWP::Simple;
use Data::Dumper;

my $url = shift or die "Pass URL as argument";
my $content = get($url) or die "Couldn't get $url";
my $rss = new XML::RSS;
$rss->parse($content) or die "Couldn't parse content";
print Dumper($rss);
#!/usr/bin/perl

use warnings;
use strict;

use XML::RSS;
use LWP::Simple;

my $url = shift or die "Pass URL as argument";
my $content = get($url) or die "Couldn't get $url";
my $rss = new XML::RSS;
$rss->parse($content) or die "Couldn't parse content";

print "<ul>
";
foreach my $item (@{$rss->{items}}) {
  print qq|  <li><a href="$item->{link}">$item->{title}</a></li>
|;
}
print "</ul>";
Fuquay-Varina Yarn Maker To Cut Nearly Half Of Workforce

Company Bringing Hundreds Of Jobs Begins Searching For Workers

Bikers Support Bill That Would Give Them The Choice To Wear Helmets

Animal Control Officers Remove 19 Pit Bulls From Durham House

Local Towns' Leaders Seek Support For Clean Water Bond

State Lawmaker Wants Casino Nights Legalized For Non-Profit Groups

Duke Student Admits To Making Fake IDs

Crews Rescue Woman Trapped In Car After It Flips

'Temporary' N.C. State Employees Seeking Back Benefits

Former CBC Employee Honored At Local YMCA

Defense, Prosecution Squabble Over Psychiatrist In Grenade Attack

Police Investigate Bank Robbery In Fayetteville

Public Invited To Discuss Future Plans For Dorothea Dix Hospital
Conclusions

- The XML::RSS toolkit makes parsing feeds trivial.
- With more content being syndicated as RSS, the ability to do interesting things with the content increases.
- Other RSS parsing modules available on CPAN.
Feed Readers

• Standalone applications designed to parse RSS and Atom feeds.

• Generally let you categorize many feeds, set update times (not too frequently!), and keep track of what you’ve read.

• Newer ones offer synchronization.

• New ones showing up all the time.
Linux

- Straw (http://www.nongnu.org/straw/)
- LiFeRea (http://liferea.sourceforge.net/)
- Syndigator (http://syndigator.sourceforge.net/)
**[it has come to my attention that 2004 is the Chinese year of the Monkey]**

**Feed:** [it has come to my attention that 2004 is the Chinese year of the Monkey]

**Item:** http://www.livejournal.com/users/niftybabe313/106235.html

Yesterday I was surfing the web (I know. I can’t help myself.) and I came across a site that told me that 2004 is the Chinese year of the Monkey. This is interesting because I was born in the year of the Monkey (1968, if I remember correctly). It’s also interesting because I have found out that I am a very unlucky year of the Monkey. This is because I was born near the end of the year of the Monkey. The year of the Monkey is typically associated with good fortune and success, but it is not the case for me. I have always been the exception to the rule. I have had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me. I’ve had a lot of financial problems, and I’ve also had a lot of health problems. I’ve had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me. I’ve had a lot of financial problems, and I’ve also had a lot of health problems. I’ve had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me. I’ve had a lot of financial problems, and I’ve also had a lot of health problems. I’ve had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me. I’ve had a lot of financial problems, and I’ve also had a lot of health problems. I’ve had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me. I’ve had a lot of financial problems, and I’ve also had a lot of health problems. I’ve had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me. I’ve had a lot of financial problems, and I’ve also had a lot of health problems. I’ve had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me. I’ve had a lot of financial problems, and I’ve also had a lot of health problems. I’ve had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me. I’ve had a lot of financial problems, and I’ve also had a lot of health problems. I’ve had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me. I’ve had a lot of financial problems, and I’ve also had a lot of health problems. I’ve had a lot of bad luck, and I can’t seem to shake it off. It’s not that I’m looking for trouble, but it just seems to find me.
MacOS X

- NetNewsWire (http://www.ranchero.com/)
- Pulp Fiction (http://freshlysqueezedsoftware.com/products/pulpfiction/)
- NewsFire (http://www.newsfirerrss.com/)
As far as the social interactions go, I think the Raleigh folks did a great job in making the guests feel comfortable, and
that people managed to hang out with folks outside of their city's group.

There were a few concerns among some Raleigh players about some lineup decisions I made, but we were able to
discuss them and come to an understanding before the situation developed into problems.

On the whole, while this year's tournament wasn't as well-planned as I would have liked it to have been, I don't think it
affected the quality of the weekend – good help from folks like Larry, MattC, Jack, RiG, and jMatt were key in that
regard. I'm looking forward to next year!

Windows

- FeedReader (http://www.feedreader.com/)
- FeedDemon (http://www.bradsoft.com/feeddemon/)
- NewsGator (http://www.newsgator.com/)
Nick Bradbury is at it again

I just learned about Nick Bradbury's latest project, FeedDemon, an RSS news aggregator. Nick's work in the web development tools space has been pioneering, with HomeSite being the most popular HTML text-editor ever released, and TopStyle, the standard productivity tool for CSS editing. Versions of both HomeSite and TopStyle are included with Dreamweaver MX. (Interesting side-note: HomeSite is what brought Allaire and Macromedia together in the first place --- Kevin Lynch was looking to partner with the leading HTML editor vendor (Allaire, as we had recently brought on Nick and HomeSite).

I'm really excited to be using a Nick Bradbury product again! This is Nick's first attempt at what I'll call an end-user personal productivity tool (e.g. the kind of stuff that only Microsoft is supposed to be able to build because of their dominance with Office).

Wednesday, June 09, 2004 9:47:49 PM
Other

- Bloglines (web-based) (http://www.bloglines.com/)
- Firefox Hot Bookmarks.
- RSSOwl (java) (http://www.rssowl.org/)
- BlogBridge (java) (http://www.blogbridge.com/)
Inside the Bubble Washington Journalism Awards: Act Nice!

By wonkette on About

Remember what we said about how not all WH journalists are total bitches who would just as soon shoot death rays at you with their laser eyes as say hello? Not since Campbell left, anyway. But as proof, this relatively kind-hearted stab at one-off nominees and categories for the Inside the Bubble Washington Journalism Awards.

"Biggest Nightmare During Security Sweeps: Mark Knoller (CBS radio).

Best resource for colleagues, White House History and Presidential Facts category: Mark Knoller (CBS radio)

Best resource for colleagues, Good Restaurant Wherever You Are category: Bill Plante (CBS)

Most Likely to Reference "Sodomite scoutmasters" in a question to the press secretary (because he did): Lester Kinsolving"

Accepting nominees in these categories now. Send your complaints and kudos and names to .

RELATED: Inside the Bubble Washington Journalism Awards: Some Categories Announced [Wonkette]
Random Bits

- OPML - Used for transferring subscription lists.
- Podcasting - Syndicating more than just text.
- Subversion logs as RSS.
- ISCABBS as RSS.
Parting Shots

- RSS is amazing.
- Once you start aggregating content with RSS, it’s hard to stop reading.
- Perl makes it trivial to provide RSS feeds of your content.
- Thanks.
Syndicate Me

- Personal Blog: http://www.lunenburg.org/wade/
- Atom project: XML::Atom::SimpleFeed (CPAN)