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.

Sample Script

```
#!/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>

<item>
<title>First Post</title>
<link>http://example.com/news/articles/1.html</link>
<description>Natalie Portman, hot grits, etc., etc.</description>
</item>

</channel>
</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).

Real-World DB Example

```
#!/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);
```

Real-World DB Example

```
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",
    },
);
```

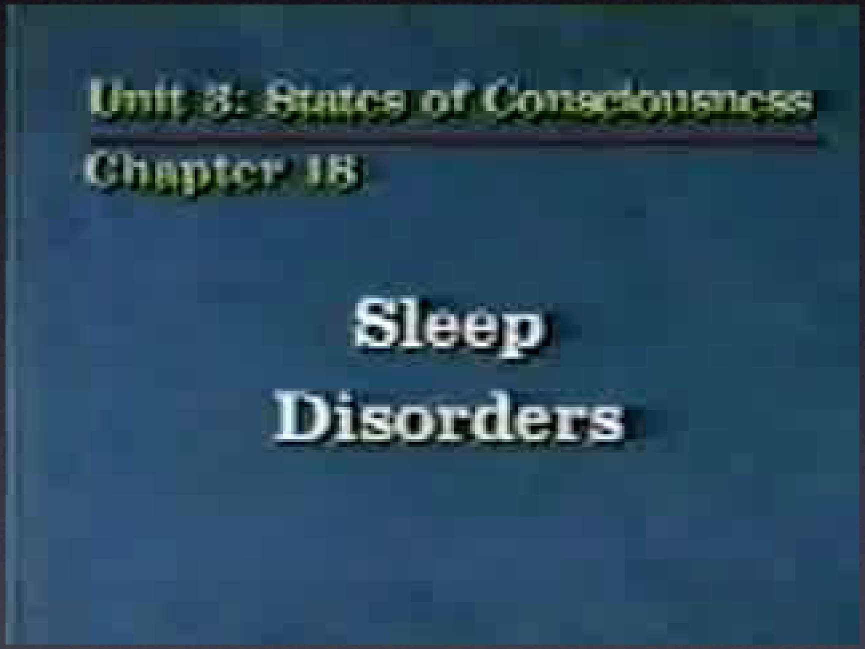
Real-World DB Example (cont.)

```
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  =>
          "http://www.comedyworx.com/players/news.php?action=read&story_id=$story_id",
        description => "$html",
        dc          => {
            date    => "$date",
            creator => "$poster",
        },
    );
}

$sth->finish;

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?



Unit 3: States of Consciousness

Chapter 18

Sleep

Disorders

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.

Peek Inside Parsed Feed

```
#!/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);
```

Peek Inside Parsed Feed

```
[minter@carlton ]$ perl parse.pl http://www.wral.com/news/topstory.rss
$VAR1 = bless( {
                  'channel' => {
                                 'http://www.ibsys.com/rss/' => {
                                                                'annotation' => 'This is a data file
meant to be read by an RSS reader. See http://www.wral.com/rss/index.html for more information.'
                                                              },
                                 'link' => 'http://www.wral.com/news/index.html?rss=ral&psp=news',
                                 'language' => 'en-us',
                                 'ttl' => '60',
                                 'copyright' => 'Copyright 2005, Internet Broadcasting Systems, Inc
Capitol Broadcasting Company',
                                 'category' => 'News',
                                 'title' => 'WRAL.com - Local News',
                                 'description' => 'Local News'
                               },
                  'version' => '2.0',
                  'items' => [
                               {
                                 'link' => 'http://www.wral.com/news/4394385/detail.html?
rss=ral&psp=news',
                                 'title' => 'Fuquay-Varina Yarn Maker To Cut Nearly Half Of
Workforce',
                                 'description' => 'The downturn in the textile industry has sent
another crushing blow to one Triangle town.'
                               },
[...........]
```

Application: Display WRAL current headlines

```
#!/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>\n";
foreach my $item (@{$rss->{items}})
{
  print qq|  <li><a href="$item->{link}">$item->{title}</a></li>\n|;
}
print "</ul>\n";
```

Headline Output

```
[minter@carlton minter]$ perl parse.pl http://www.wral.com/news/topstory.rss
<ul>
  <li><a href="http://www.wral.com/news/4394385/detail.html?rss=ral&psp=news">Fuquay-Varina Yarn Maker To Cut Nearly Half Of Workforce</
a></li>
  <li><a href="http://www.wral.com/news/4395536/detail.html?rss=ral&psp=news">Company Bringing Hundreds Of Jobs Begins Searching For
Workers</a></li>
  <li><a href="http://www.wral.com/news/4395372/detail.html?rss=ral&psp=news">Bikers Support Bill That Would Give Them The Choice To Wear
Helmets</a></li>
  <li><a href="http://www.wral.com/news/4393799/detail.html?rss=ral&psp=news">Animal Control Officers Remove 19 Pit Bulls From Durham
House</a></li>
  <li><a href="http://www.wral.com/news/4395776/detail.html?rss=ral&psp=news">Local Towns' Leaders Seek Support For Clean Water Bond</
a></li>
  <li><a href="http://www.wral.com/news/4395796/detail.html?rss=ral&psp=news">State Lawmaker Wants Casino Nights Legalized For Non-Profit
Groups</a></li>
  <li><a href="http://www.wral.com/news/4395725/detail.html?rss=ral&psp=news">Duke Student Admits To Making Fake IDs</a></li>
  <li><a href="http://www.wral.com/news/4393324/detail.html?rss=ral&psp=news">Crews Rescue Woman Trapped In Car After It Flips</a></li>
  <li><a href="http://www.wral.com/news/4395770/detail.html?rss=ral&psp=news">'Temporary' N.C. State Employees Seeking Back Benefits</a></
li>
  <li><a href="http://www.wral.com/news/4395741/detail.html?rss=ral&psp=news">Former CBC Employee Honored At Local YMCA</a></li>
  <li><a href="http://www.wral.com/news/4393596/detail.html?rss=ral&psp=news">Defense, Prosecution Squabble Over Psychiatrist In Grenade
Attack</a></li>
  <li><a href="http://www.wral.com/news/4394286/detail.html?rss=ral&psp=news">Police Investigate Bank Robbery In Fayetteville</a></li>
  <li><a href="http://www.wral.com/news/4392243/detail.html?rss=ral&psp=news">Public Invited To Discuss Future Plans For Dorothea Dix
Hospital</a></li>
</ul>
```

Conclusions

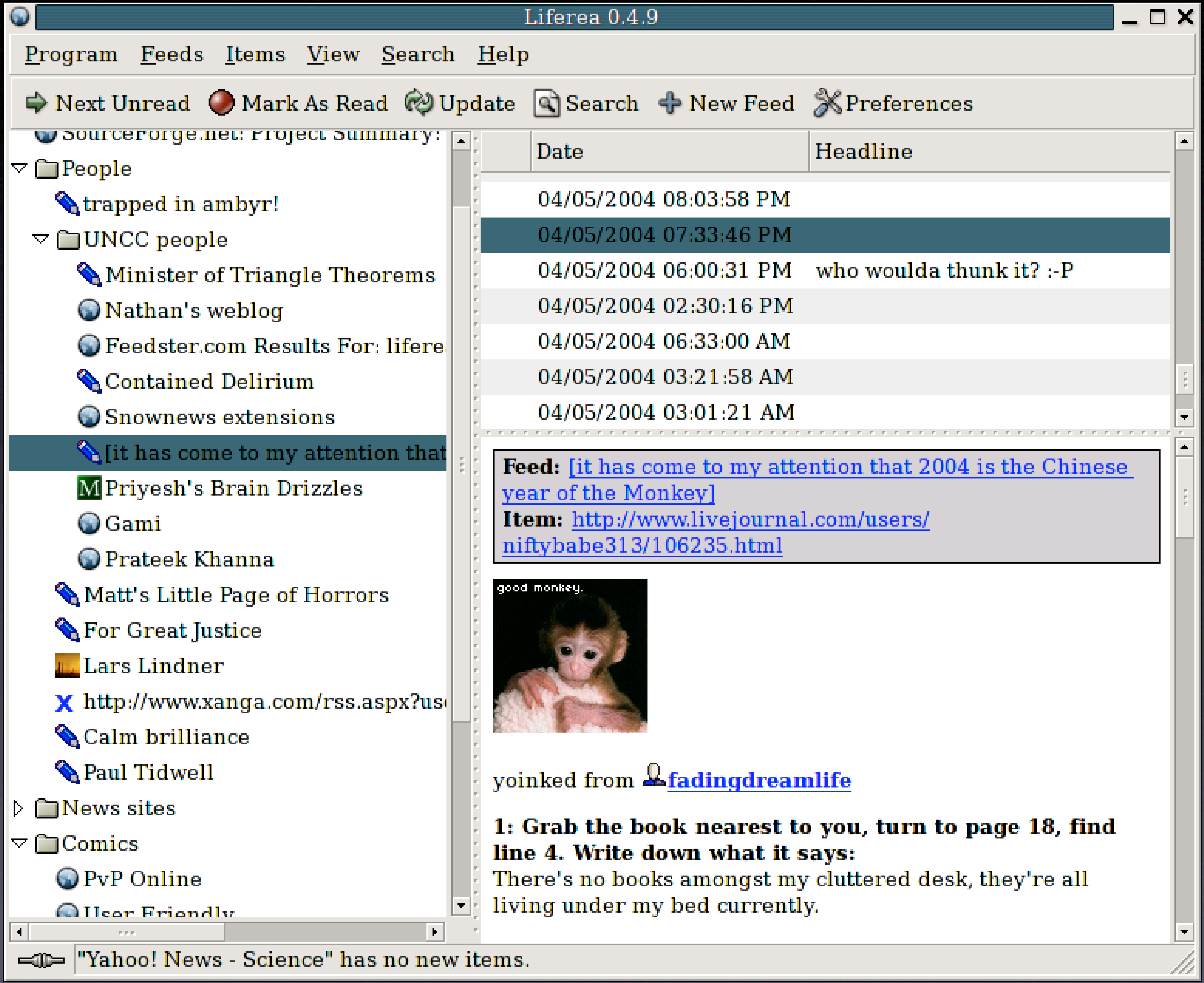
- The XML::RSS toolkit makes parsing feeds trivial.
- With more content being syndicated as RSS, the ability to do interesting things with the content increases.
- Other RSS parsing modules available on CPAN.

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/)



Date	Headline
04/05/2004 08:03:58 PM	
04/05/2004 07:33:46 PM	
04/05/2004 06:00:31 PM	who woulda thunk it? :-P
04/05/2004 02:30:16 PM	
04/05/2004 06:33:00 AM	
04/05/2004 03:21:58 AM	
04/05/2004 03:01:21 AM	

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

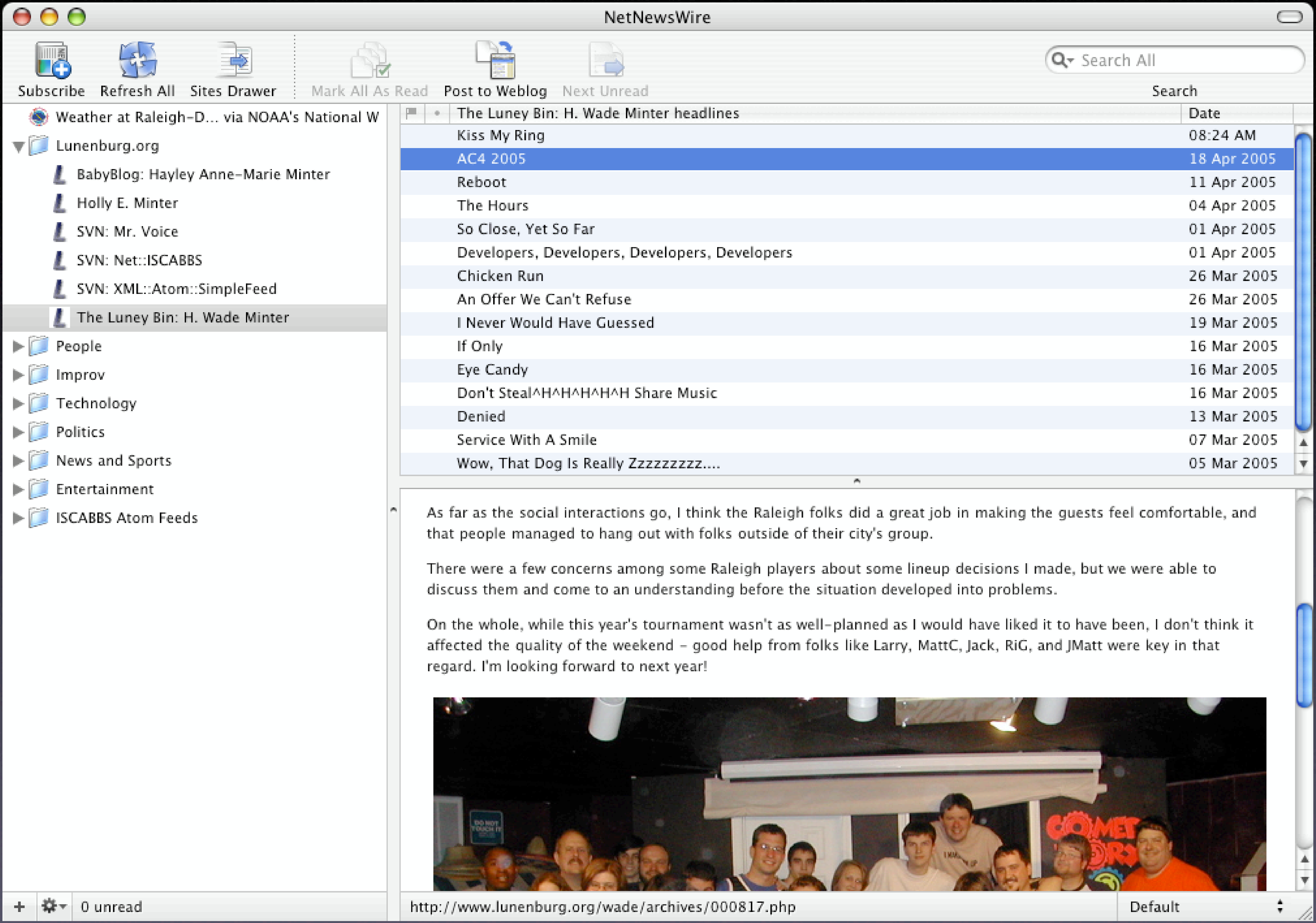
yoinked from fadingdreamlife

1: Grab the book nearest to you, turn to page 18, find line 4. Write down what it says:
There's no books amongst my cluttered desk, they're all living under my bed currently.

"Yahoo! News - Science" has no new items.

MacOS X

- NetNewsWire (http://www.ranchero.com/)
- Pulp Fiction (http://freshlysqueezedsoftware.com/products/pulpfiction/)
- NewsFire (http://www.newsfirerss.com/)



- Weather at Raleigh-D... via NOAA's National W
- Lunenburg.org
  - BabyBlog: Hayley Anne-Marie Minter
  - Holly E. Minter
  - SVN: Mr. Voice
  - SVN: Net::ISCABBS
  - SVN: XML::Atom::SimpleFeed
  - The Luney Bin: H. Wade Minter
- People
- Improv
- Technology
- Politics
- News and Sports
- Entertainment
- ISCABBS Atom Feeds

The Luney Bin: H. Wade Minter headlines	Date
Kiss My Ring	08:24 AM
AC4 2005	18 Apr 2005
Reboot	11 Apr 2005
The Hours	04 Apr 2005
So Close, Yet So Far	01 Apr 2005
Developers, Developers, Developers, Developers	01 Apr 2005
Chicken Run	26 Mar 2005
An Offer We Can't Refuse	26 Mar 2005
I Never Would Have Guessed	19 Mar 2005
If Only	16 Mar 2005
Eye Candy	16 Mar 2005
Don't Steal^H^H^H^H^H Share Music	16 Mar 2005
Denied	13 Mar 2005
Service With A Smile	07 Mar 2005
Wow, That Dog Is Really Zzzzzzzzz....	05 Mar 2005

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.

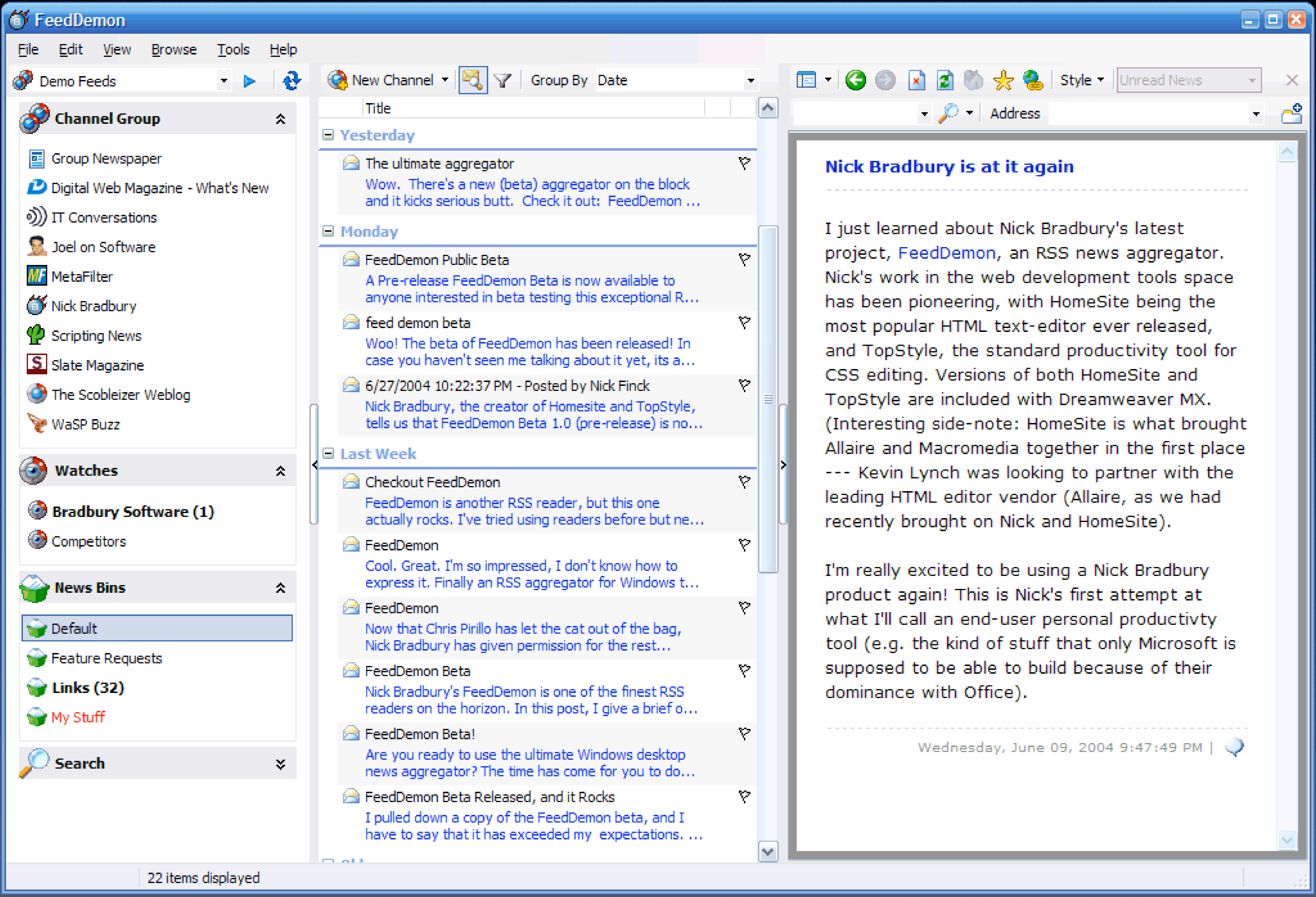
There were a few concerns among some Raleigh players about some lineup decisions I made, but we were able to discuss them and come to an understanding before the situation developed into problems.

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!

http://www.lunenburg.org/wade/archives/000817.php

Windows

- FeedReader (http://www.feedreader.com/)
- FeedDemon (http://www.bradsoft.com/feeddemon/)
- NewsGator (http://www.newsgator.com/)



Yesterday

The ultimate aggregator
Wow. There's a new (beta) aggregator on the block and it kicks serious butt. Check it out: FeedDemon ...

Monday

FeedDemon Public Beta
A Pre-release FeedDemon Beta is now available to anyone interested in beta testing this exceptional R...

feed demon beta
Woo! The beta of FeedDemon has been released! In case you haven't seen me talking about it yet, its a...

6/27/2004 10:22:37 PM - Posted by Nick Finck
Nick Bradbury, the creator of Homesite and TopStyle, tells us that FeedDemon Beta 1.0 (pre-release) is no...

Last Week

Checkout FeedDemon
FeedDemon is another RSS reader, but this one actually rocks. I've tried using readers before but ne...

FeedDemon
Cool. Great. I'm so impressed, I don't know how to express it. Finally an RSS aggregator for Windows t...

FeedDemon
Now that Chris Pirillo has let the cat out of the bag, Nick Bradbury has given permission for the rest...

FeedDemon Beta
Nick Bradbury's FeedDemon is one of the finest RSS readers on the horizon. In this post, I give a brief o...

FeedDemon Beta!
Are you ready to use the ultimate Windows desktop news aggregator? The time has come for you to do...

FeedDemon Beta Released, and it Rocks
I pulled down a copy of the FeedDemon beta, and I have to say that it has exceeded my expectations. ...

Nick Bradbury is at it again

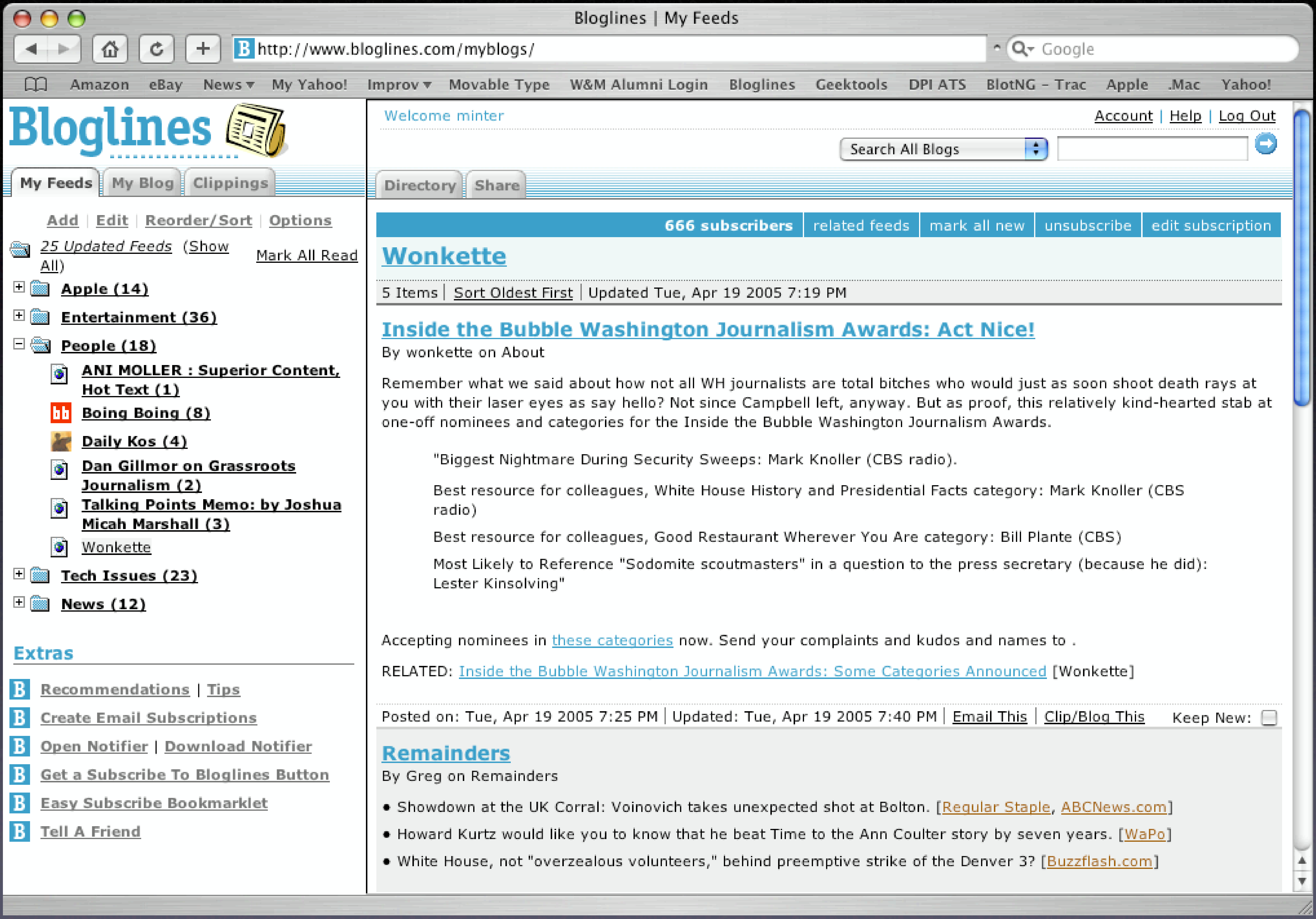
I just learned about Nick Bradbury's latest project, FeedDemon, an RSS news aggregator. Nick's work in the web development tools space has been pioneering, with HomeSite being the most popular HTML text-editor ever released, and TopStyle, the standard productivity tool for CSS editing. Versions of both 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 productivty 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.
- SAGE (Firefox extension) (https://addons.update.mozilla.org/extensions/moreinfo.php?application=firefox&id=77)
- RSSOwl (java) (http://www.rssowl.org/)
- BlogBridge (java) (http://www.blogbridge.com/)



Wonkette

5 Items | Sort Oldest First | Updated Tue, Apr 19 2005 7:19 PM

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]

Posted on: Tue, Apr 19 2005 7:25 PM | Updated: Tue, Apr 19 2005 7:40 PM | Email This | Clip/Blog This Keep New: □

Remainders

By Greg on Remainders

- Showdown at the UK Corral: Voinovich takes unexpected shot at Bolton. [Regular Staple, ABCNews.com]
- Howard Kurtz would like you to know that he beat Time to the Ann Coulter story by seven years. [WaPo]
- White House, not "overzealous volunteers," behind preemptive strike of the Denver 3? [Buzzflash.com]

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/
- BabyBlog: http://www.lunenburg.org/hayley/
- Atom project: XML::Atom::SimpleFeed (CPAN)