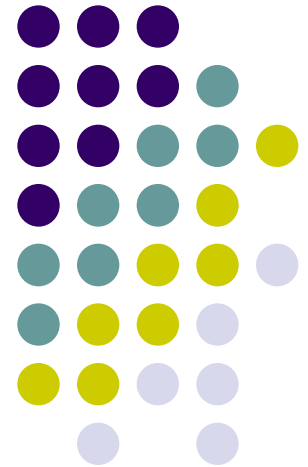


INFSCI 2480

Processing RSS Feeds

Yi-ling Lin





Feed? RSS? Atom?



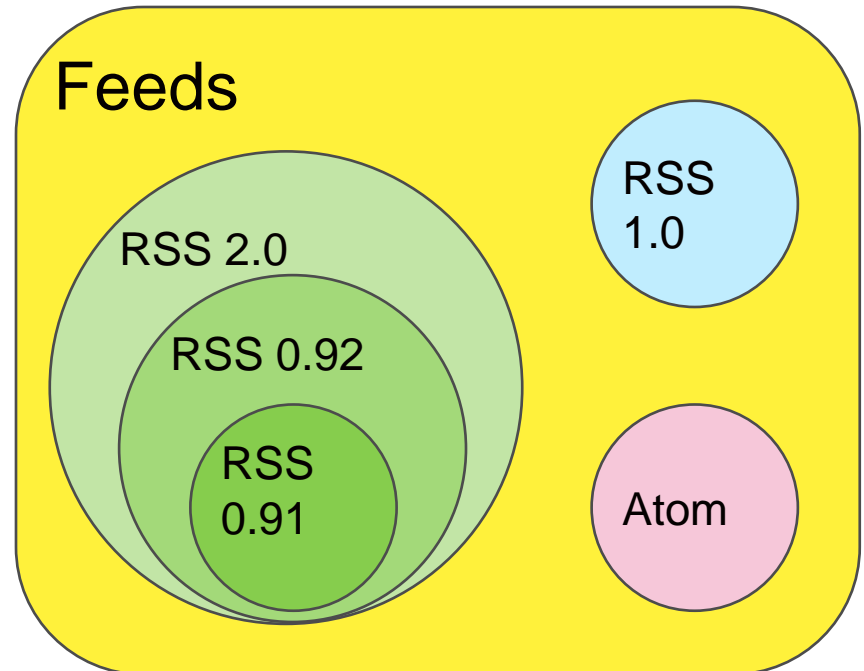
- RSS = Rich Site Summary
- RSS = RDF (Resource Description Framework) Site Summary
- RSS = Really Simple Syndicate
- ATOM

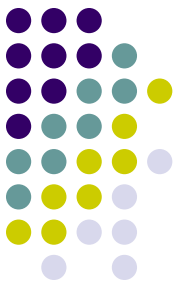
Feed



- Feed = “A document (often XML-based) which contains content items, often summaries of stories or weblog posts with web links to longer versions.

- Feed > RSS, Atom





Why RSS(Feeds)?

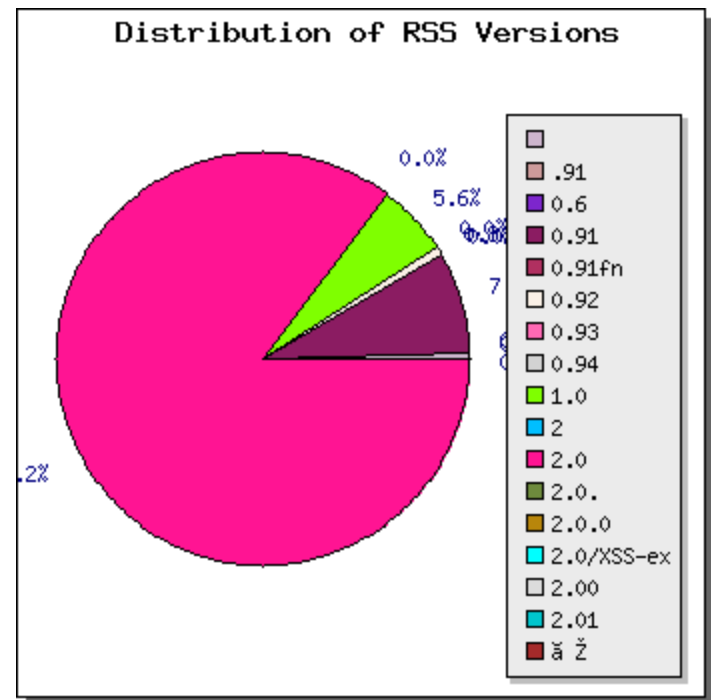
- **For publishers**
 - Syndicate content automatically
 - Simpler writing process
 - Easy republishing
- **For subscribers**
 - Easily stay informed
 - Save time
 - Ensure privacy
 - Easy to manage
 - Freedom from information overload

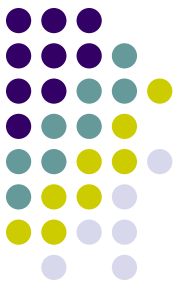
RSS is all about
publishing and
subscribing to
content



RSS Versions

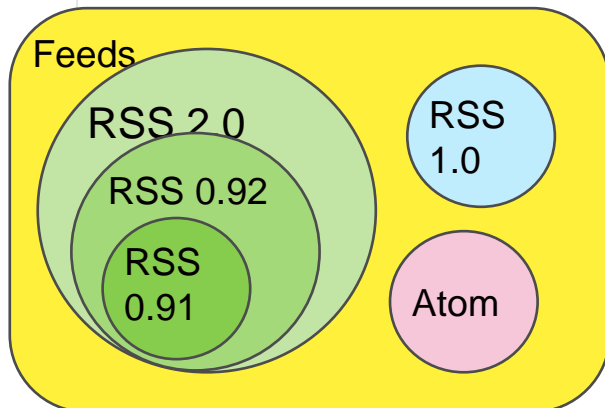
- Version distribution collected by an RSS search engine
- $2.0 > 1.0 > 0.91 > 0.92$
- <http://www.syndic8.com/stats.php?Section=rss#table>





Comparison of RSS versions

	RSS 0.91	RSS 0.92	RSS 2.0
Categories on channel or item	X	O	O
Elements on the channel : language, copyright, docs, lastBuildDate, managingEditor, pubDate, rating, skipDays, skipHours, generator, ttl	X	X	O
Item enclosures	X	O	O
Elements on items: authors, comments, pubDate	X	X	O
Item count limitation	15	X	X
Notes	Channel-level metadata only	Allows both channel and item metadata	Modularized



RSS2.0 VS. ATOM



RSS 2.0	Atom 1.0	Comments
rss	-	Vestigial in RSS
channel	feed	
title	title	
link	link	Atom defines an extensible family of rel values
description	subtitle	
language	-	Atom uses standard xml:lang attribute
copyright	rights	
webMaster	-	
managingEditor	author or contributor	
pubDate	published (in entry)	Atom has no feed-level equivalent
lastBuildDate (in channel)	updated	RSS has no item-level equivalent
category	category	
generator	generator	

Refer to <http://www.intertwingly.net/wiki/pie/Rss20AndAtom10Compared>



RSS2.0 VS. ATOM

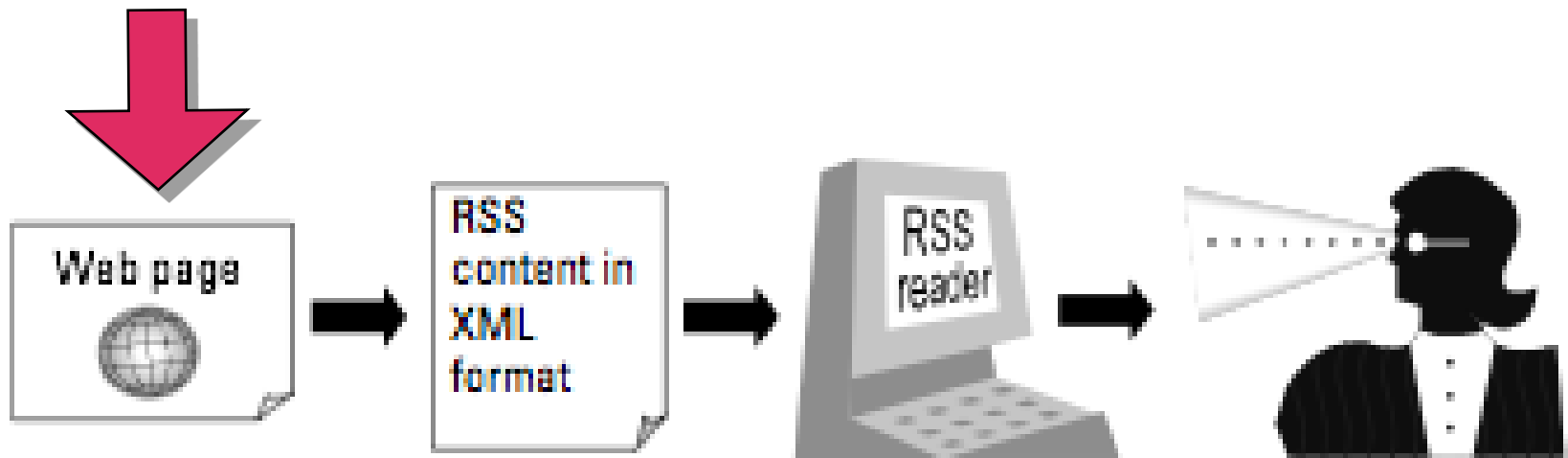
RSS 2.0	Atom 1.0	Comments
cloud	-	
ttl		<ttl> is problematic, prefer HTTP 1.1
image		ect ratio
-		
item		
author		
-		
description		ontent is
comments		
enclosure		Atom
guid	id	
source	-	rel="via" on <link> in Atom
-	source	Container for feed-level metadata to support aggregation

1. Improved control
2. Internalization options
3. More precise and standardized tag definitions
4. Ability to add features without enhancing the core structure.

Refer to <http://www.intertwingly.net/wiki/pie/Rss20AndAtom10Compared>



Revealing RSS in Web pages





Revealing RSS in Web pages

The screenshot shows a web browser window with the address bar containing the URL <http://vancouver2010.blogs.nytimes.com/2010/02/23/live-blog-follow-the-womens-short-program/>. The browser's address bar includes a search icon, a search box with the text "Google", and a red circle highlighting the "RSS" icon. Below the address bar, a navigation bar contains various links such as "Bonjour", "아후!", "위키백과", "Post to CiteULike", "Read It Later", "Reading List", "Read Later", "YouTube", "CocoA", "Gmail", "Google Maps", "Apple", "뉴스", and "MacBB:". The main content area features a live blog post with the following text:

then falling later in the program, as a look of horror crossed her face. The smile returned by the end of her program, however, as the crowd whooped and hollered. The judges gave her a 57.16 and moved her into third place.

There's another break as the next group warms up. We've gone from black outfits to easter-egg blue and now, fluorescent. Two of the skaters are wearing suits in a blinding green — no, make that chartreuse. —*Katie Thomas*

9:40 P.M. **No Time for Blood**

A bloody nose that began halfway through her short program bothered American Mirai Nagasu enough that she fears she has no chance for a medal.

Nagasu felt her nose start bleeding in the middle of her routine but said, "You have to deal with what you've got." So she completed her program and received 63.76 points, a personal best that puts her in first place through 14 skaters.

Still, the 16-year-old skater from Los Angeles is disappointed with her performance, believing she "can't reach the podium" with it. —*Associated Press*

9:38 P.M. **Gimazetdinova Scores 49.02**

Anastasia Gimazetdinova of Uzbekistan is 29 years old and is the most senior competitor tonight. She skated a season's best of 49.02, but not enough to put

On the right side of the page, there is a social media widget with a red circle around the "RSS" icon. Below it, a "Subscribe" section also has a red circle around the "Vancouver 2010 RSS" link. Further down, there are sections for "From The New York Times" and "Inside the Action: The Downhill".

At the bottom of the browser window, the status bar displays the message: "Loading 'http://vancouver2010.blogs.nytimes.com/2010/02/23/live-blog-follow-the-womens-short-program/?hp', completed 71 of 74 items (1 error)".



By visible links/icons



No actual standard

Browsers' convention, since Mozilla (mid 2005)



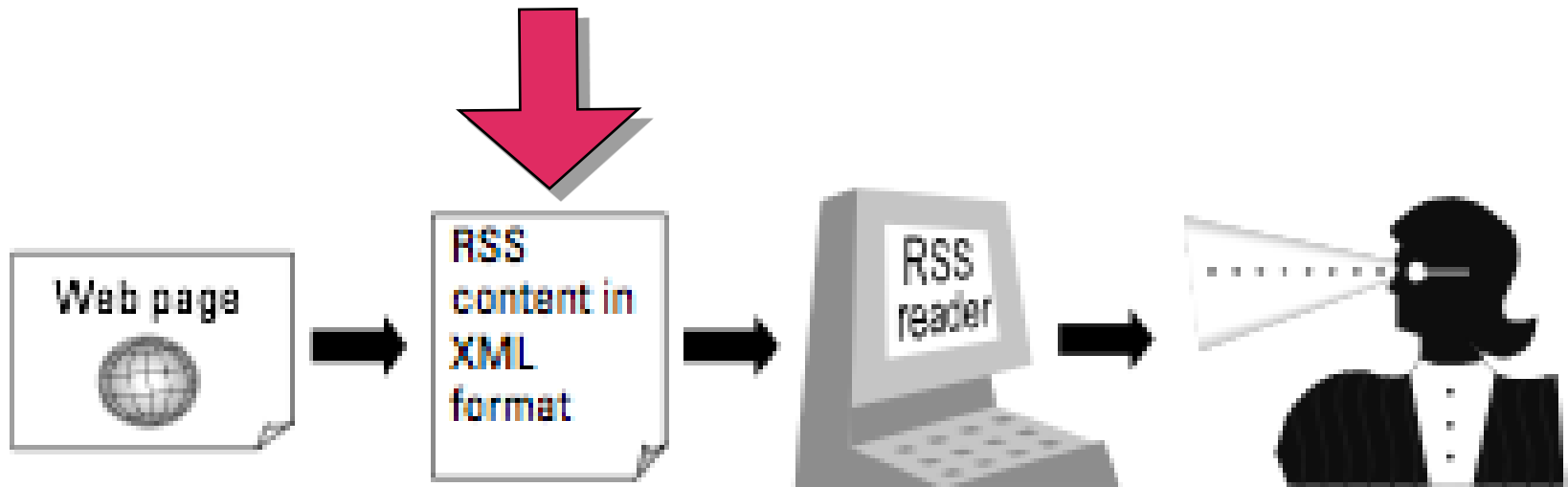
By hidden RSS Links



The screenshot shows a web browser window with the address bar containing `http://pawsgroup.blogspot.com/`. A red circle highlights the RSS icon and the text "Personalized Adaptive Web Systems (PAWS) group - Atom" and "Personalized Adaptive Web Systems (PAWS) group - RSS" in the browser's address bar. The page content includes a title "PERSONALIZED ADAPTIVE WEB SYSTEMS (PAWS) GROUP", a description of the group, a date "TUESDAY, NOVEMBER 17, 2009", a post title "PAWS meeting - Nov 10, 2009", and a list of contributors.

```
<link rel="alternate" type="application/atom+xml" title="Personalized Adaptive Web Systems (PAWS) group - Atom"
href="http://pawsgroup.blogspot.com/feeds/posts/default" />
<link rel="alternate" type="application/rss+xml" title="Personalized Adaptive Web Systems (PAWS) group - RSS"
href="http://pawsgroup.blogspot.com/feeds/posts/default?alt=rss" />
```

RSS content





RSS Content Structure

- RSS 0.90 to 2.0 family
- XML
- `<channel>` & `<item>` parts
Feed information (channel)
Each article content (item)
- Additional features with higher versions —
0.90 to 2.0
- **RSS 1.0 & Atom are in different formats!**

```
untitled
1 |<?xml version="1.0"?>
2 |<rss version="0.92">
3 |<channel>
4 |   <title>travelinlibrarian.info</title>
5 |   <link>http://www.travelinlibrarian.info/</link>
6 |   <description>The blog of Librarian, Trainer, and writer Michael P.
7 |   Sauers</description>
8 |   <lastBuildDate>Tue, 01 Feb 2005 13:23:02 GMT</lastBuildDate>
9 |   <docs>http://backend.userland.com/rss092</docs>
10 |  <managingEditor>msauers@travelinlibrarian.info (Michael
11 |  Sauers)</managingEditor>
12 |  <webMaster> msauers@travelinlibrarian.info (Michael Sauers)</webMaster>
13 |  <image>
14 |    <title>Michael in Lego</title>
15 |    <url>http://travelinlinrarian.info/blog/lego.gif</url>
16 |    <link>http://travelinlinrarian.info/</link>
17 |    <width>155</width>
18 |    <height>238</height>
19 |    <description>Michael imagined as a Lego person. Create yours at
20 |    http://www.reasonablyclever.com/mini/</description>
21 |  </image>
22 |</channel>
23 |<item>
24 |  <title>Firefox 1.1 Delayed</title>
25 |  <link>http://Weblogs.mozillazine.org/ben/archives/007434.html</link>
26 |  <description>According to Ben Godger (lead Firefox engineer) version 1.1
27 |  of Firefox has been delayed and will not be released in March as originally
28 |  scheduled.</description>
29 |  <category domain="http://www.dmoz.org/">Computers: Software: Internet:
30 |  Clients: WWW: Browsers: Firefox</category>
31 |</item>
32 |</rss>
```

Line: 1 Column: 1 XML Soft Tabs: 4

```
1 <?xml version="1.0"?>
2 <!-- RSS generated by Radio UserLand v8.0.8 on Sun, 30 Jan 2005 20:55:47 GMT
3 -->
4 Appendix: Feed Code Examples 253
5 <rss version="2.0">
6 <channel>
7   <title>Library Web Chic</title>
8   <link>http://www.librarywebchic.net/</link>
9   <description></description>
10  <language>en-us</language>
11  <copyright>Copyright 2005 Karen Coombs</copyright>
12  <lastBuildDate>Sun, 30 Jan 2005 20:55:47 GMT</lastBuildDate>
13  <docs>http://backend.userland.com/rss</docs>
14  <generator>Radio UserLand v8.0.8</generator>
15  <managingEditor>kac@mailcity.com</managingEditor>
16  <webMaster>kac@mailcity.com</webMaster>
17  <ttl>60</ttl>
18  <item>
19    <title>Free E-books</title>
20    <link>http://www.librarywebchic.net/2005/01/30.html#a196</link>
21    <description>&lt;h4&gt;Free E-books&lt;/h4&gt; &lt;p&gt;This week I
22    spent some time adding free e-books collections to our OpenURL resolver. The
23    ...
24    in trying to make as many free resources available to my users as
25    possible.&lt;br&gt; &lt;/p&gt;&lt;br&gt;</description>
26    <guid>http://www.librarywebchic.net/2005/01/30.html#a196</guid>
27    <pubDate>Sun, 30 Jan 2005 20:55:47 GMT</pubDate>
28    <category>Ongoing Projects</category>
29    <category>OpenURL</category>
30  </item>
31  <item>
32    <title>Integrateable Standards compliant WYSIWYG Editor</title>
33    <link>http://www.librarywebchic.net/2005/01/27.html#a195</link>
34    <description>&lt;h4&gt;Integrateable Standards compliant WYSIWYG
35    Editor&lt;/h4&gt; &lt;p&gt;&lt;a
36    href=&quot;http://www.themaninblue.com&quot;&gt;The Man in
```




Comparison of RSS versions

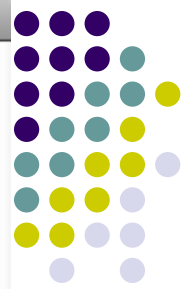
	RSS 0.91	RSS 0.92	RSS 2.0
Categories on channel or item	X	O	O
Elements on the channel : language, copyright, docs, lastBuildDate, managingEditor, pubDate, rating, skipDays, skipHours, generator, ttl	X	X	O
Item enclosures	X	O	O
Elements on items: authors, comments, pubDate	X	X	O
Item count limitation	15	X	X
Notes	Channel-level metadata only	Allows both channel and item metadata	Modularized

RSS 1.0

"uses RDF"

<http://www.w3.org/RDF/>

```
untitled 2
1 <?xml version="1.0" encoding="utf-8"?>
2 <rdf:RDF
3 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4 xmlns:dc="http://purl.org/dc/elements/1.1/"
5 xmlns:sy="http://purl.org/rss/1.0/modules/syndication/"
6 xmlns:admin="http://webns.net/mvcb/"
7 xmlns:cc="http://web.resource.org/cc/"
8 xmlns="http://purl.org/rss/1.0/">
9 <channel rdf:about="http://freerangelibrarian.com/">
10 <title>Free Range Librarian</title>
11 <link>http://freerangelibrarian.com/</link>
12 <description></description>
13 <dc:language>en-us</dc:language>
14 <dc:creator></dc:creator>
15 <dc:date>2005-02-01T08:11:58-08:00</dc:date>
16 <admin:generatorAgent
17 rdf:resource="http://www.movabletype.org/?v=3.14" />
18 <items>
19 <rdf:Seq>
20 <rdf:li rdf:resource="http://freerangelibrarian.com/
21 archives/020105/podcasting_test.php" />
22 <rdf:li rdf:resource="http://freerangelibrarian.com/
23 archives/013105/the_last_mile_a_cha.php" />
24 <rdf:li rdf:resource="http://freerangelibrarian.com/
25 archives/013105/mustread_blogs_lib.php" />
26 <rdf:li rdf:resource="http://freerangelibrarian.com/
27 archives/013005/fri_rss_1_or_rss_2.php" />
28 <rdf:li rdf:resource="http://freerangelibrarian.com/
29 archives/012905/lists_versus_blogs_.php" />
```



```
untitled 2
1 <?xml version="1.0" encoding="UTF-8" standalone="yes"?>
2 <?xml-stylesheet href="http://www.blogger.com/styles/atom.css"
· type="text/css"?>
3 <feed version="0.3" xml:lang="en-GB" xmlns="http://purl.org/atom/ns#">
4 <link href="http://www.blogger.com/atom/5543214" rel="service.post"
· title="Travelin' Librarian" type="application/atom+xml"/>
5 <link href="http://www.blogger.com/atom/5543214" rel="service.feed"
· title="Travelin' Librarian" type="application/atom+xml"/>
6 <title mode="escaped" type="text/html">Travelin' Librarian</title>
7 <tagline mode="escaped" type="text/html"/>
8 <link href="http://www.travelinlibrarian.info/" rel="alternate"
· title="Travelin' Librarian" type="text/html"/>
9 <id>tag:blogger.com,1999:blog-5543214</id>
10 <modified>2005-01-31T20:47:36Z</modified>
11 <generator url="http://www.blogger.com/" version="5.15">Blogger</generator>
12 <info mode="xml" type="text/html">
13 <div xmlns="http://www.w3.org/1999/xhtml">This is an Atom formatted XML site
· feed. It is intended to be viewed in a Newsreader or syndicated to another
· site. Please visit the <a
· href="http://help.blogger.com/bin/answer.py?answer=697">Blogger Help</a> for
· more info.</div>
14 </info>
15 <entry xmlns="http://purl.org/atom/ns#">
16 <link href="http://www.blogger.com/atom/5543214/110720445691459308"
· rel="service.edit" title="LC, ISBN &amp; XML"
· type="application/atom+xml"/>
17 <author>
18 <name>Michael</name>
19 </author>
20 <issued>2005-01-31T13:45:36-07:00</issued>
21 <modified>2005-01-31T20:47:36Z</modified>
22 <created>2005-01-31T20:47:36Z</created>
23 <link href="http://www.travelinlibrarian.info/2005/01/lc-isbn-xml.html"
· rel="alternate" title="LC, ISBN &amp; XML" type="text/html"/>
24 <id>tag:blogger.com,1999:blog-5543214.post-110720445691459308</id>
25 <title mode="escaped" type="text/html">LC, ISBN &amp;
```

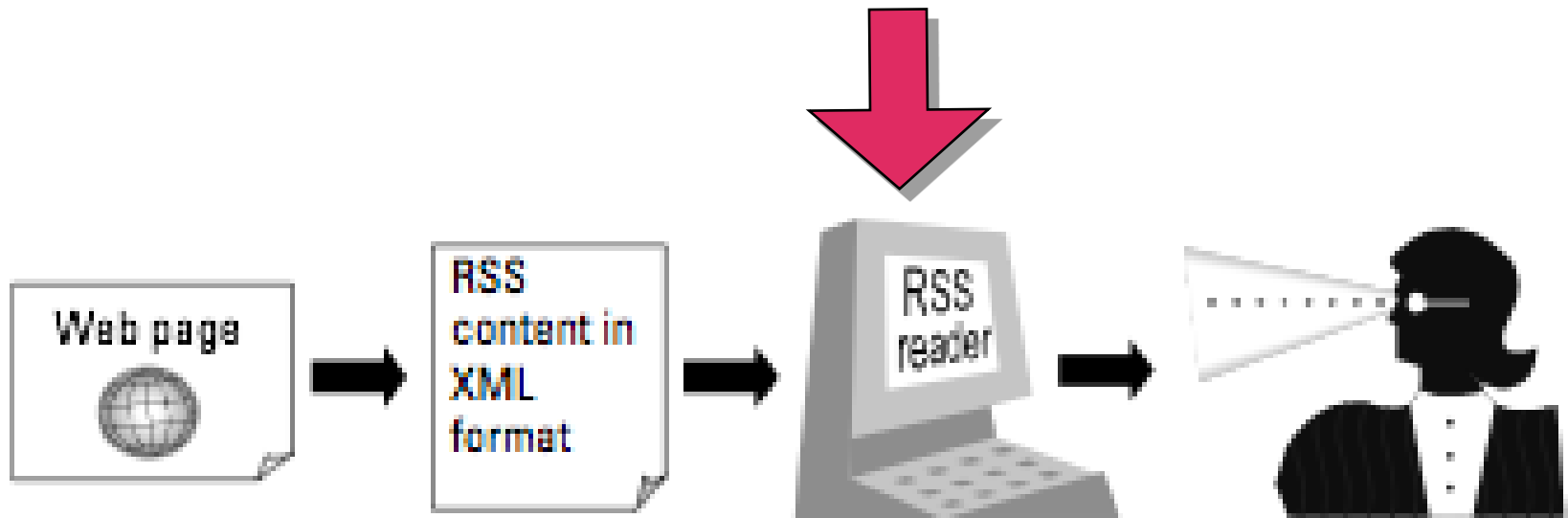


In more detail...

- Specifications
- RSS 0.91:
<http://www.rssboard.org/rss-0-9-1-netscape>
- RSS 2.0:
<http://cyber.law.harvard.edu/rss/rss.html>



RSS reader





RSS Reader/Aggregator

- Software for reading RSS feeds
- Parses RSS feeds (in XML) and displays (attaching style sheets)
- RSS *aggregator* — because it *aggregates* many sources of data in one place
- Desktop software
 - Dedicated, Web browsers, E-mail client
- Web service

<http://blogspace.com/rss/readers>

RSS Republishing Example : YourNews

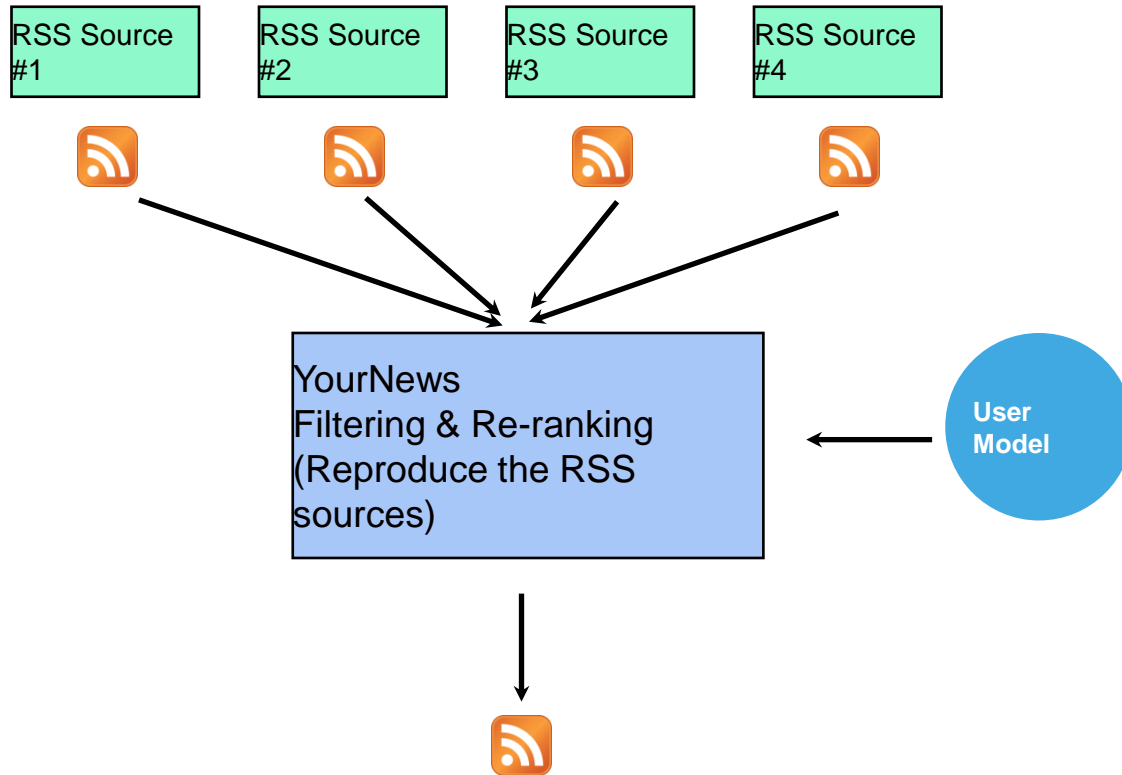


- Adaptive news filtering system
- Read/gather RSS feeds
- Reproduce news list (personalize) — filtering & re-ranking
- Re-publish it in a native interface or through RSS feeds again

YourNews <http://amber.exp.sis.pitt.edu/yournews/yournews.php>



YourNews structure



A new (personalized) RSS feed

NBCSports.com: Sports

22 Total

Live odds Today, 2:04 AM
[Read more...](#)

Matchups/injuries Today, 2:04 AM
[Read more...](#)

Vegas Advisors on Gambling: Who will cover? Today, 1:46 AM
 A game by game breakdown by the experts [Read more...](#)

Kobe soars, hits game-winning 3 in return Today, 1:40 AM

Ventre: Kobe's back, and Lakers are still king Today, 1:34 AM

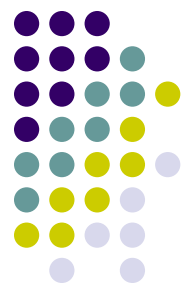
Kim makes short work of short program Today, 1:13 AM

Celtics beat Knicks despite missing Pierce Today, 1:00 AM

Search Articles:

Article Length:

Sort By:
 Date
 Title
 Source
 New



Kobe Bryant returned to lineup after missing five games with an ankle injury and hit a 3-pointer with 4.3 seconds left to lift the Los Angeles Lakers to 99-98 victory over the Memphis Grizzlies Tuesday night.

Ventre: Kobe Bryant is back. And once again he seems eager to prove that his Lakers are still the team to beat.

World champion Kim Yu-na has set a world best in routing a strong field in the women's short program at the Vancouver Olympics.

Rajon Rondo scored 15 points and had 16 assists to lead the Boston Celtics to a 110-106 victory over the New York Knicks on Tuesday night in a matchup of teams that traded with each other last week.

Silva: Scrutiny coming for top prospects at Combine - NFL - nbcsports.msnbc.com

atom wiki

NBC Sports

	NBA	CBK	MORE SCORES									
DET	54	PHI	44	95	NY	106	POR	102	MIN	91	LAL	99
SAC	42	GS	31	105	BOS	110	NJ	93	MIA	88	MEM	98
	Half		2nd, 8:41 left	Final	Final	Final	Final	Final	Final	Final	Final	Final
	MONDAY	TUESDAY	WEDNESDAY	All Times ET								

FIOS TV RATED #1
 IN CUSTOMER SATISFACTION
 AHEAD OF CABLE AND SATELLITE

Get FIOS

Sports / NFL

ProFootballTalk Scores Video Schedules Standings Player news Odds Fantasy Sunday Night Football

Scrutiny coming for top prospects at Combine
 Clausen, Cody, Gerhart among those teams who need to answer critics

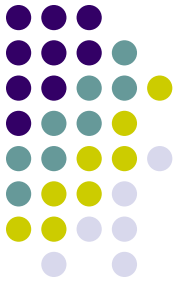
ANALYSIS
 By Evan Silva
Rotoworld.com
 updated 10:54 p.m. ET, Mon., Feb. 22, 2010

What's at stake for NFL hopefuls at the NFL Scouting Combine? We have answers.

Georgia Tech WR **Demaryius Thomas** broken left foot deprives the combine of what could've been this week's biggest riser, and may cost Thomas a chance to be drafted in the first round.

It also creates an opening at wide receiver. There is no consensus No. 2 prospect at the position behind Dez Bryant, so players like

Reproduced (personalized) list



YourNews [Search]

National World Business Technology **Sports** Entertainment Health MLB Linux Apple Tab6

Show all duplicate articles » Recent News | **Recommended News**

- Silva: Scrutiny coming for top prospects at Combine** (19 hours ago) ★★
Silva: Whats at stake for NFL hopefuls at the NFL Scouting Combine? We have answers.
- Florio: Combine an idiotic way to find NFL players** (2 days ago) ★
Florio: In February, pro football players from all over the country convene in Indianapolis for the Underwear Olympics, a.k.a. the NFL Combine. Though no medals will be awarded, the men who run the fastest, jump the highest, and/or lift 225 pounds the most times while wearing T-shirts and shorts could be poised to earn more money.
- Report: Lions actively shopping No. 2 pick** (7 hours ago) ★
NFL Networks Jason La Canfora reports that the Lions are actively shopping the No. 2 overall pick in April's draft and are already in contact with several other teams.
- NFLPA director: 2010 season will likely be uncapped** (0 min ago) ★
NFL Players Association executive director DeMaurice Smith says in a memo to players and their agents that it is likely no new collective bargaining agreement will be reached and the upcoming season will be played without a salary cap.
- Mainstay L.T. runs into end of road in San Diego** (1 day ago) ★
LaDainian Tomlinson was released Monday by the San Diego Chargers, ending a brilliant nine-year run in which he became one of the NFL's greatest running backs.
- Lee wins gold; Kramer DQed for not switching lanes** (43 min ago) ★
Lee Seung-hoon of South Korea won a stunning gold medal in mens 10,000-meter speedskating Tuesday when overwhelming favorite Sven Kramer made an amateurish mistake, failing to switch lanes just past the midway point of the race, and was disqualified.
- NCAA finds Michigan out of compliance on practice** (6 hours ago) ★
The NCAA has found that Michigans storied football program was out of compliance with practice time rules under coach Rich Rodriguez.
- NCAA mens basketball wins leader retiring** (1 day ago) ★
The NCAA mens basketball coach is retiring.

jahn's Interests for Sports News [Hide]

Short term | Long term

CRICKET NFL GUMBEL
GHOLSTON SCHOOL LEAGUE
MIAMI_DOLPHINS NUT BASEBALL SILVA
COLT DOLPHIN LONG PLAY SPORT NFL BASKETBALL
MICHIGAN FOOTBALL OHIO PLATA VERNON_GHOLSTON
BRYANT_GUMBEL BALL COMBINE NFL_DRAFT TEAM
PARAMETER CULTURE WICKET NETWORK VERNON
ANNOUNCER YORK SPORTS HIGH DRAFT JAKE BOOTH
DUCKS

PIRATE

Add your custom keywords

Read this tab in RSS format

Personalized List: filtered & re-ranked



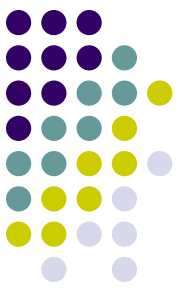
Parsing RSS Feeds

- Problem — extract texts from RSS structure
- They are **XML**
- Parsers
 - SAX
 - DOM
 - Out-of-box parser



SAX and DOM

- SAX (Simple API for XML) — serial access parser
 - Stream of XML data goes in
 - Event-driven parsing
- DOM (Document Object Model)
 - Use hierarchical structure for parsing



SAX Example

```
import xml.parsers.expat

# 3 handler functions
def start_element(name, attrs):
    print 'Start element:', name, attrs
def end_element(name):
    print 'End element:', name
def char_data(data):
    print 'Character data:', repr(data)

p = xml.parsers.expat.ParserCreate()

p.StartElementHandler = start_element
p.EndElementHandler = end_element
p.CharacterDataHandler = char_data

p.Parse("""<?xml version="1.0"?>
<parent id="top"><child1 name="paul">Text goes here</child1>
<child2 name="fred">More text</child2>
</parent>""", 1)
```

DOM Example



```
import xml.dom.minidom

document = """\
<slideshow>
<title>Demo slideshow</title>
<slide><title>Slide title</title>
<point>This is a demo</point>
<point>Of a program for processing slides</point>
</slide>

<slide><title>Another demo slide</title>
<point>It is important</point>
<point>To have more than</point>
<point>one slide</point>
</slide>
</slideshow>
"""

dom = xml.dom.minidom.parseString(document)

def getText(nodelist):
    rc = ""
    for node in nodelist:
        if node.nodeType == node.TEXT_NODE:
            rc = rc + node.data
    return rc

def handleSlideshow(slideshow):
    print "<html>"
    handleSlideshowTitle(slideshow.getElementsByTagName("title")[0])
    slides = slideshow.getElementsByTagName("slide")
    handleToc(slides)
    handleSlides(slides)
    print "</html>"

def handleSlides(slides):
    for slide in slides:
        handleSlide(slide)

def handleSlide(slide):
    handleSlideTitle(slide.getElementsByTagName("title")[0])
    handlePoints(slide.getElementsByTagName("point"))

def handleSlideshowTitle(title):
    print "<title>%s</title>" % getText(title.childNodes)

def handleSlideTitle(title):
    print "<h2>%s</h2>" % getText(title.childNodes)

def handlePoints(points):
    print "<ul>"
    for point in points:
        handlePoint(point)
    print "</ul>"

def handlePoint(point):
    print "<li>%s</li>" % getText(point.childNodes)

def handleToc(slides):
    for slide in slides:
        title = slide.getElementsByTagName("title")[0]
        print "<p>%s</p>" % getText(title.childNodes)

handleSlideshow(dom)
```



Ready-made Parser

- Universal Feed Parser
<<http://www.feedparser.org>>

A screenshot of a terminal window on a Linux system. The window title is 'codex@ir: ~'. The terminal shows the following output:

```
codex@ir:~$ python
Python 2.6.4 (r264:75706, Dec 7 2009, 18:45:15)
[GCC 4.4.1] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> import feedparser
>>> print feedparser.__doc__
Universal feed parser

Handles RSS 0.9x, RSS 1.0, RSS 2.0, CDF, Atom 0.3, and Atom 1.0 feeds

Visit http://feedparser.org/ for the latest version
Visit http://feedparser.org/docs/ for the latest documentation

Required: Python 2.1 or later
Recommended: Python 2.3 or later
Recommended: CJKCodecs and iconv_codec <http://cjkpython.i18n.org/>

>>> 
```



Universal Feedparser

```
Science
1 <?xml version="1.0"?>
2 <?xml-stylesheet href="/css/rss20.xsl" type="text/css"/>
3 <rss xmlns:dc="http://purl.org/dc/elements/1.1/"
4   xmlns:media="http://search.yahoo.com/mrss/"
5   xmlns:atom="http://www.w3.org/2005/Atom"
6   xmlns:nyt="http://www.nytimes.com/namespaces/rss"
7   xmlns:pheedo="http://www.pheedo.com/namespaces/pheedo"
8     <channel>
9       <title>NYT &gt; Science</title>
10
11 <link>http://www.nytimes.com/pages/science/index.html
12   <atom:link rel="self" type="application/rss+xml"
13     href="http://www.nytimes.com/services/xml/rss/nyt/science.rss"
14     <description/>
15     <language>en-us</language>
16     <copyright>Copyright 2010
17     The New York Times Company</copyright>
18     <lastBuildDate>Tue, 16 Mar 2010 19:20:15 GMT</lastBuildDate>
19     <image>
20       <title>NYT &gt; Science</title>
21       <url>http://graphics.nytimes.com/images/section/science/rss/nyt/science.jpg</url>
22     <link>http://www.nytimes.com/pages/science/index.html</link>
23     </image>
24     <item>
25       <title>In a Desert in China, a Trove of 4,000-Year-Old
26       Mummies</title>
27       <link>http://feeds.nytimes.com/click.phdo?i=219192a56639cbf64f2d56b8f995fdfa</link>
28     </item>
29   </channel>
30 </rss>
```

```
feedparser_example.py
1 import feedparser
2
3 # get feed data dictionary
4 d = feedparser.parse("http://feeds.nytimes.com/nyt/rss/Science")
5
6 print d['feed']['title'] # dictionary style
7 print d.feed.title # object, attribute style
8 print d.channel.title # RSS terminology
9
10 print d.feed.lastbuilddate
11 print d.feed.image.url
12
13 # get items
14
15 print "There are", len(d['items']), "items"
16
17 for item in d['items']:
18     print
19     print item.title
20     print item.guid
21     print item.updated
22     print item.updated_parsed
23
```

Line: 20 Column: 20 Python Tab Size: 4



Core Attributes

- Follows RSS/ATOM syntax normalization
- However, not always
 - updated
 - /atom10:feed/atom10:updated
 - /atom03:feed/atom03:modified
 - /rss/channel/pubDate
 - /rss/channel/dc:date
 - /rdf:RDF/rdf:channel/dc:date
 - /rdf:RDF/rdf:channel/dcterms:modified



Advanced features

- { Date parsing
- { HTML sanitization
- { Content normalization
- { Namespace handling
- { and more...

Questions?

