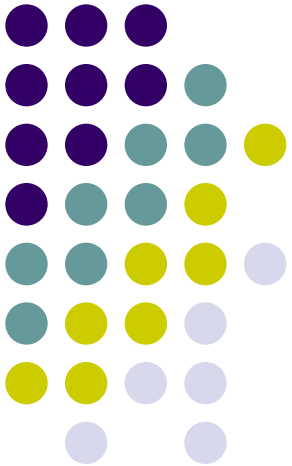
INFSCI 2480
Processing RSS Feeds

Yi-ling Lin

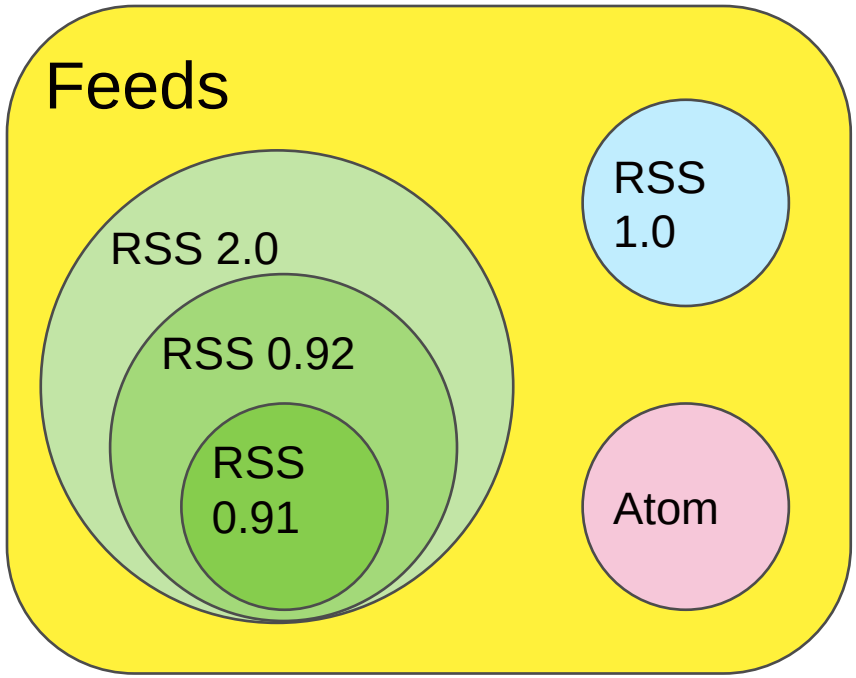


Feed? RSS? Atom?

- 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

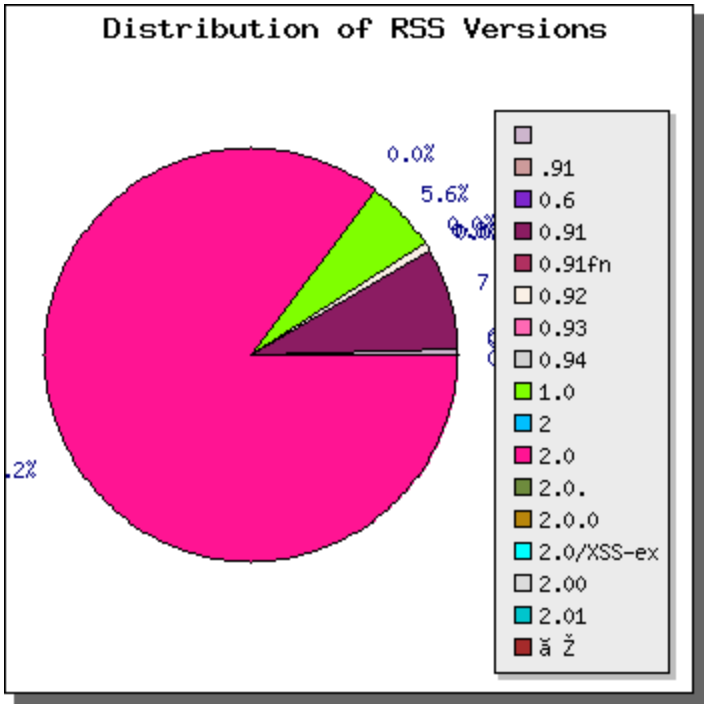
Why RSS(Feeds)?

- **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

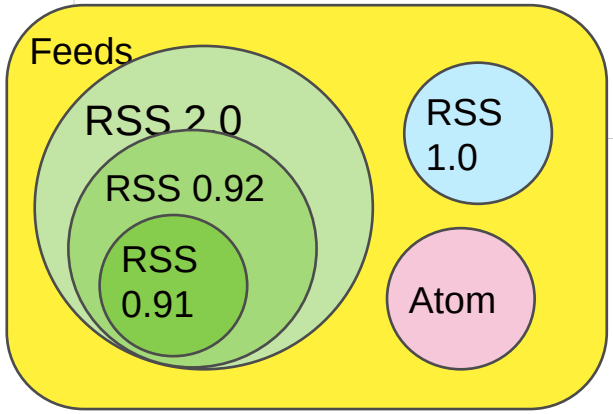
RSS Versions

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



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

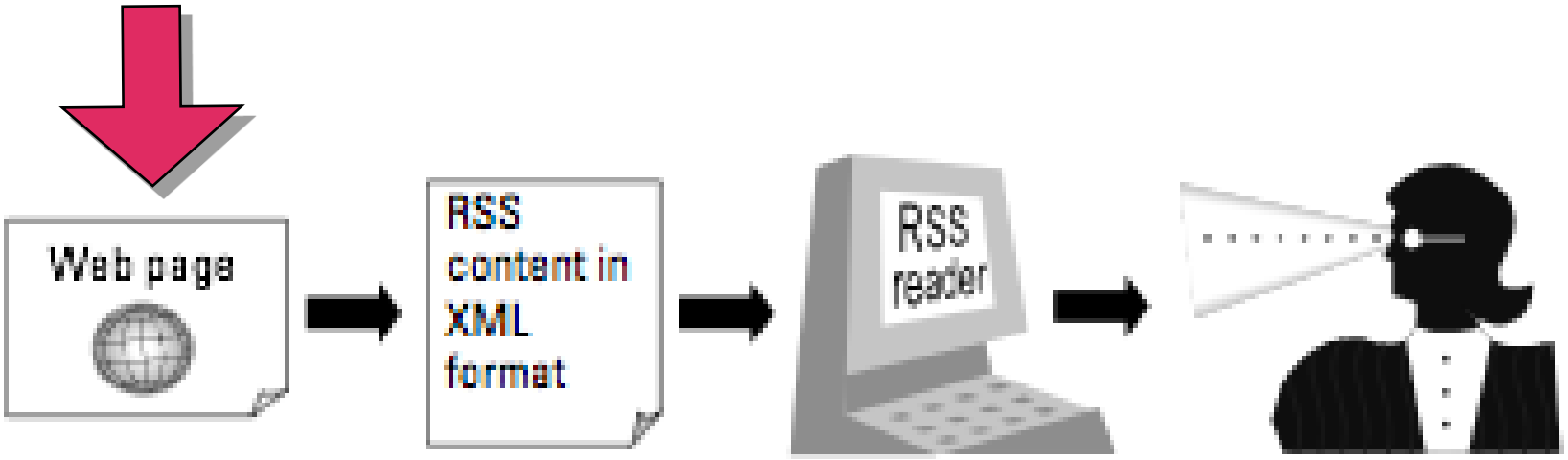
RSS2.0 VS. ATOM

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

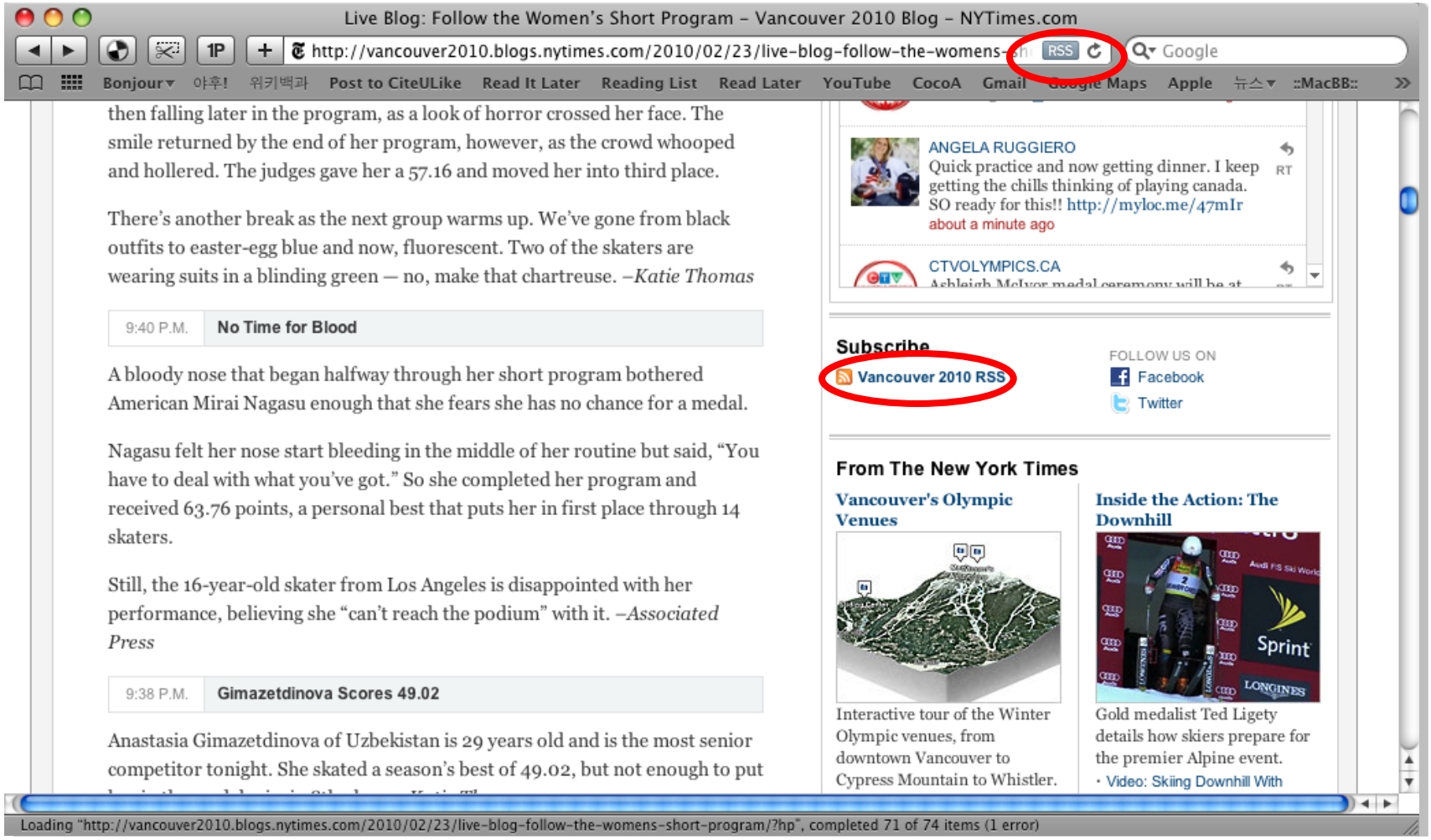
1. Improved control
2. Internalization options
3. More precise and standardized tag definitions
4. Ability to add features without enhancing the core structure.

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

Revealing RSS in Web pages



Revealing RSS in Web pages

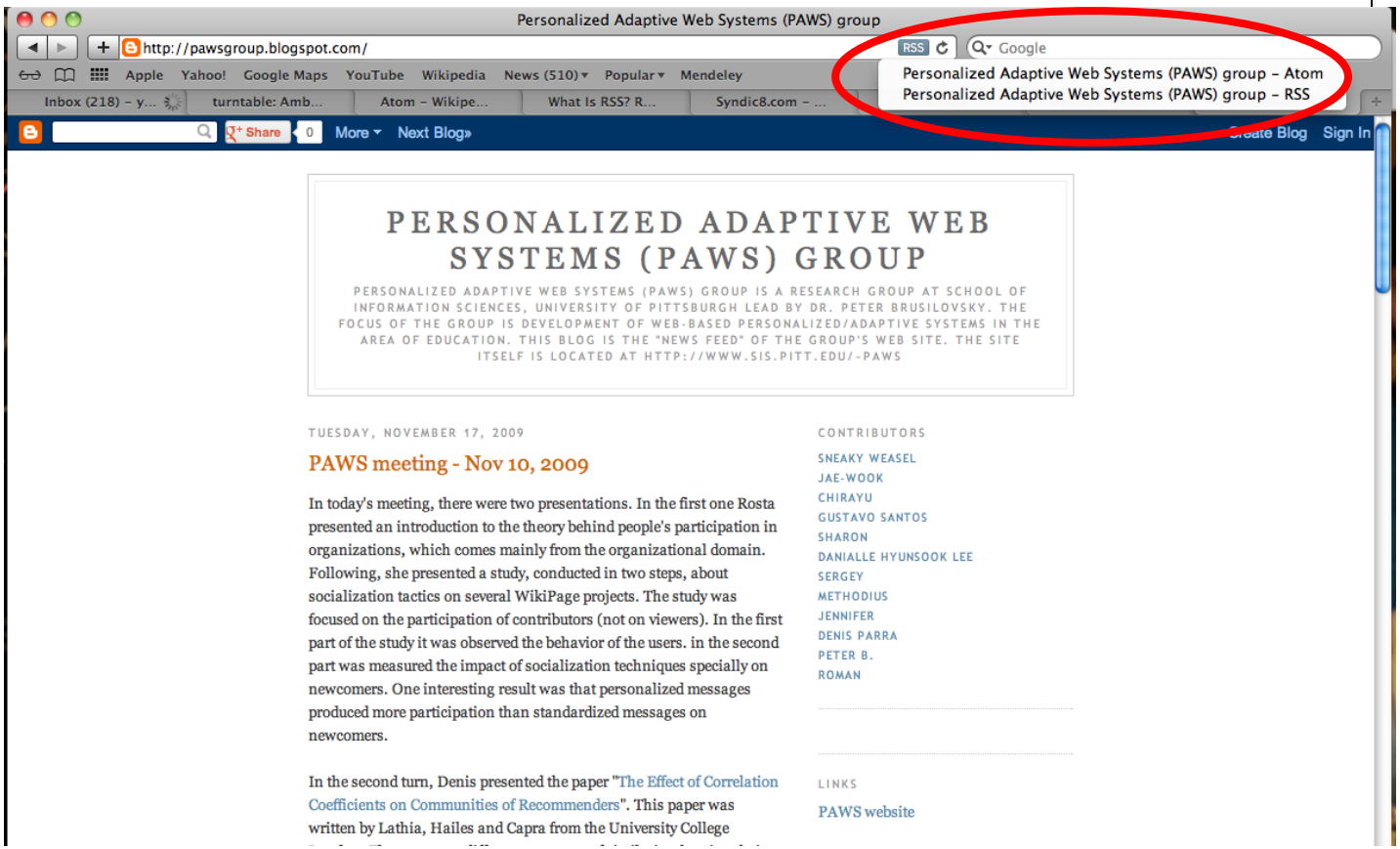


By visible links/icons

No actual standard

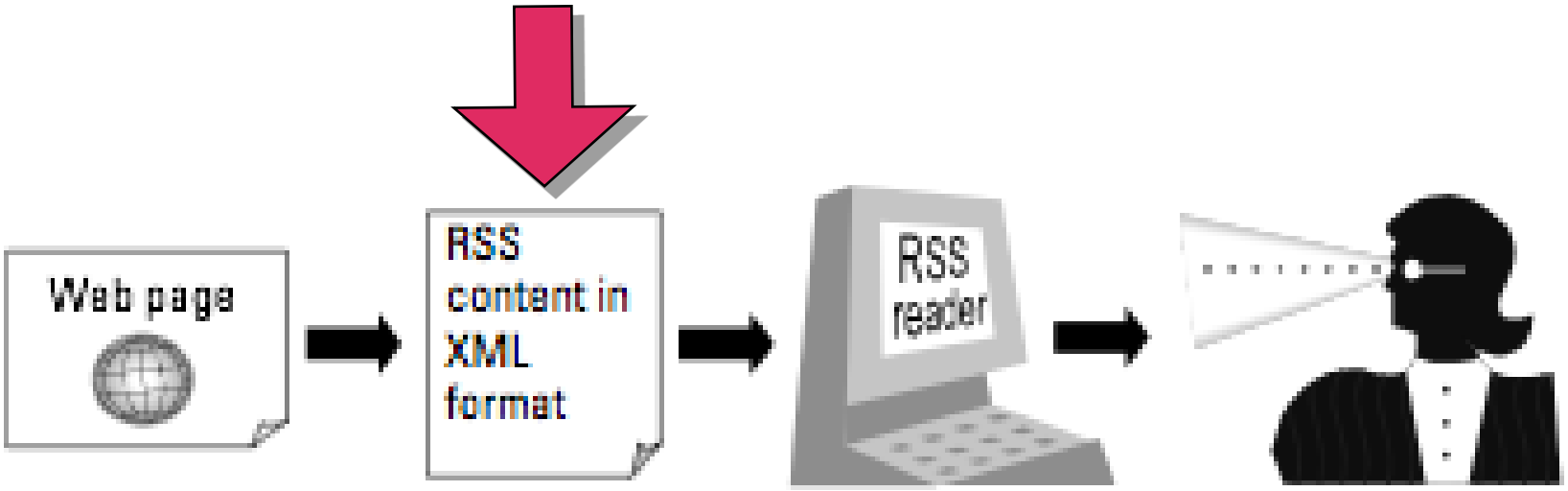
Browsers' convention, since Mozilla (mid 2005)

By hidden RSS Links



<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!

RSS 0.92

untitled

```
<?xml version="1.0"?>
<rss version="0.92">
<channel>
    <title>travelinlibrarian.info</title>
    <link>http://www.travelinlibrarian.info/</link>
    <description>The blog of Librarian, Trainer, and writer Michael P. Sauers</description>
    <lastBuildDate>Tue, 01 Feb 2005 13:23:02 GMT</lastBuildDate>
    <docs>http://backend.userland.com/rss092</docs>
    <managingEditor>msauers@travelinlibrarian.info (Michael Sauers)</managingEditor>
    <webMaster> msauers@travelinlibrarian.info (Michael Sauers)</webMaster>
    <image>
        <title>Michael in Lego</title>
        <url>http://travelinlinrarian.info/blog/lego.gif</url>
        <link>http://travelinlinrarian.info/</link>
        <width>155</width>
        <height>238</height>
        <description>Michael imagined as a Lego person. Create yours at http://www.reasonablyclever.com/mini/</description>
    </image>

<item>
    <title>Firefox 1.1 Delayed</title>
    <link>http://Weblogs.mozillazine.org/ben/archives/007434.html</link>
    <description>According to Ben Godger (lead Firefox engineer) version 1.1 of Firefox has been delayed and will not be released in March as originally scheduled.</description>
    <category domain="http://www.dmoz.org/">Computers: Software: Internet: Clients: WWW: Browsers: Firefox</category>
</item>

</channel>
</rss>
```

Line: 1 Column: 1 XML Soft Tabs: 4

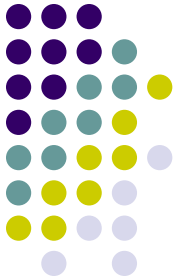
RSS 2.0

untitled

```
<?xml version="1.0"?>
<!-- RSS generated by Radio UserLand v8.0.8 on Sun, 30 Jan 2005 20:55:47 GMT
-->
Appendix: Feed Code Examples 253
<rss version="2.0">
<channel>
    <title>Library Web Chic</title>
    <link>http://www.librarywebchic.net/</link>
    <description></description>
    <language>en-us</language>
    <copyright>Copyright 2005 Karen Coombs</copyright>
    <lastBuildDate>Sun, 30 Jan 2005 20:55:47 GMT</lastBuildDate>
    <docs>http://backend.userland.com/rss</docs>
    <generator>Radio UserLand v8.0.8</generator>
    <managingEditor>kac@mailcity.com</managingEditor>
    <webMaster>kac@mailcity.com</webMaster>
    <ttl>60</ttl>
    <item>
        <title>Free E-books</title>
        <link>http://www.librarywebchic.net/2005/01/30.html#a196</link>
        <description>&lt;h4&gt;Free E-books&lt;/h4&gt; &lt;p&gt;This week I
spent some time adding free e-books collections to our OpenURL resolver. The
...
n trying to make as many free resources available to my users as
possible.&lt;br&gt; &lt;/p&gt;&lt;br&gt;</description>
        <guid>http://www.librarywebchic.net/2005/01/30.html#a196</guid>
        <pubDate>Sun, 30 Jan 2005 20:55:47 GMT</pubDate>
        <category>Ongoing Projects</category>
        <category>OpenURL</category>
    </item>
    <item>
        <title>Integrateable Standards compliant WYSIWYG Editor</title>
        <link>http://www.librarywebchic.net/2005/01/27.html#a195</link>
        <description>&lt;h4&gt;Integrateable Standards compliant WYSIWYG
Editor&lt;/h4&gt; &lt;p&gt;&lt;a
href=&quot;http://www.themaninblue.com&quot;&gt;The Man in
```

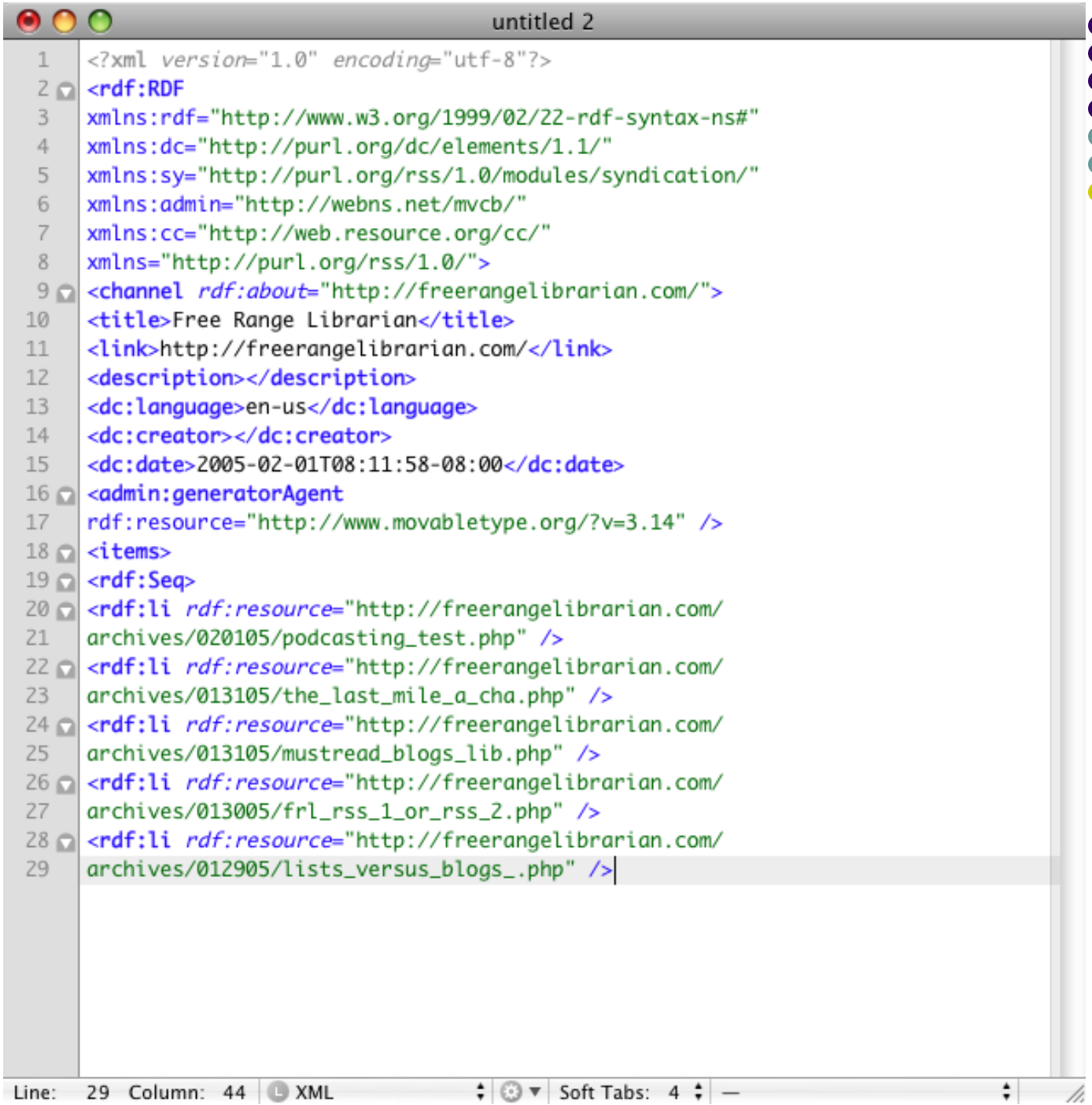
Line: 16 Column: 19 XML Soft Tabs: 4

Comparison of RSS versions

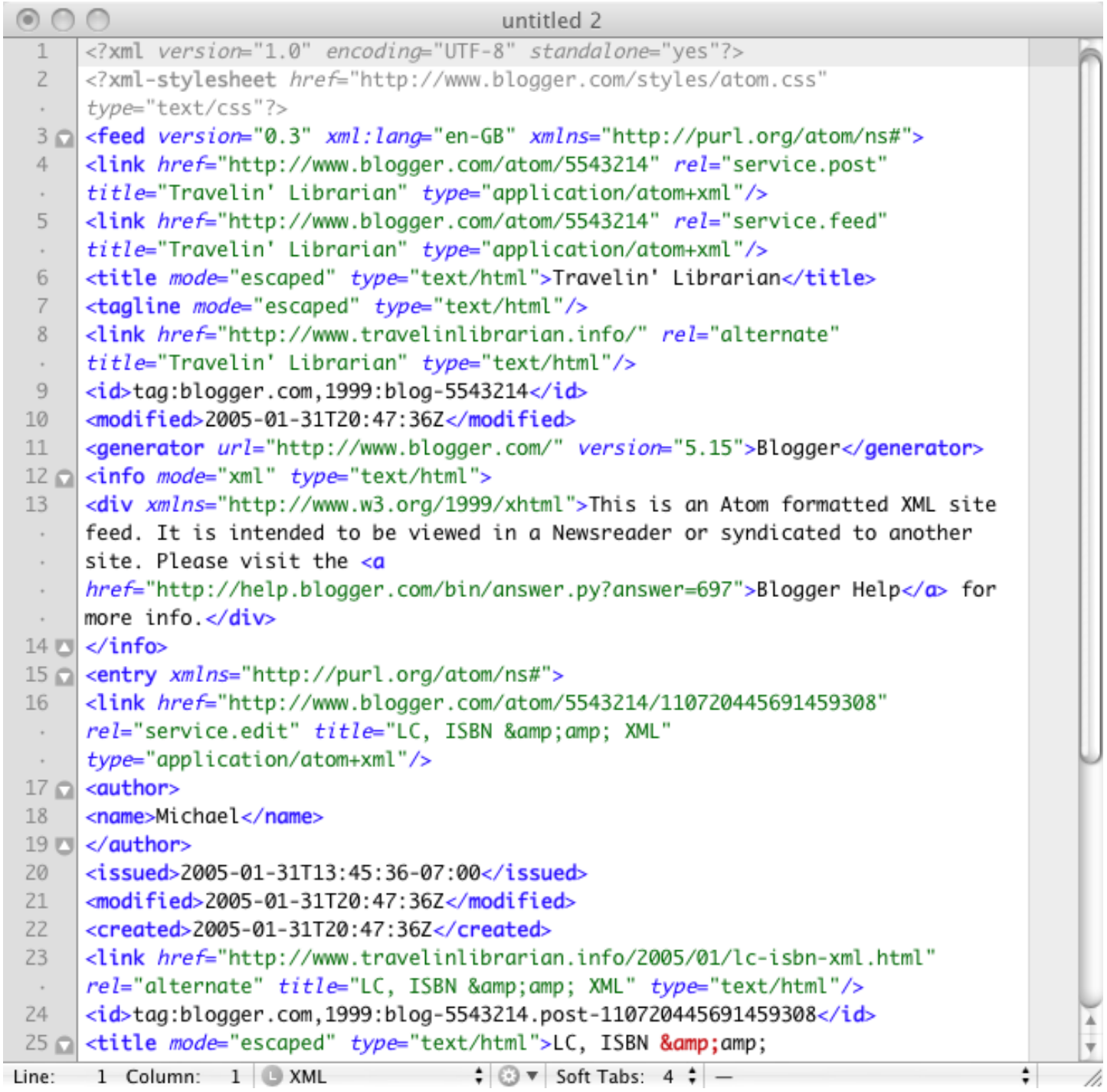


	RSS 0.91	RSS 0.92	RSS 2.0
Categories on channel or item	X	O	O
Elements on the channel : language, copyright, docs, lastBuildDate, managingEditor, pubDate, rating, skipDays, skipHours, generator, ttl	X	X	O
Item enclosures	X	O	O
Elements on items: authors, comments, pubDate	X	X	O
Item count limitation	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/



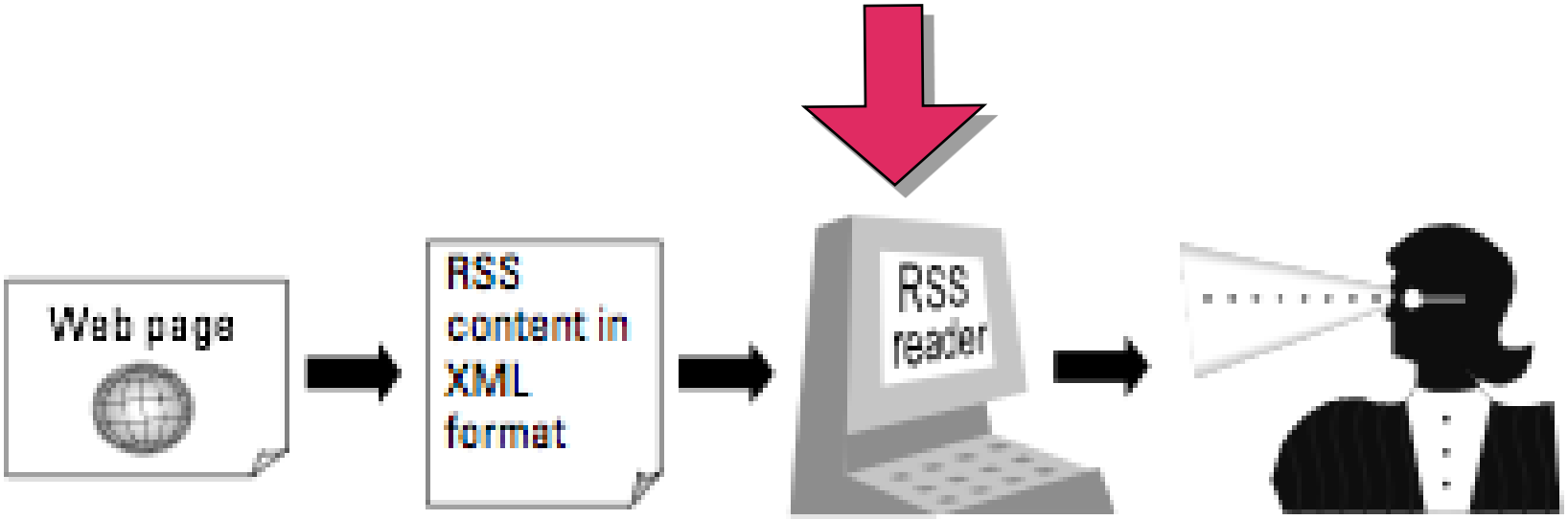
ATOM

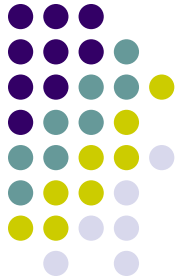

untitled 2
1 <?xml version="1.0" encoding="UTF-8" standalone="yes"?>
2 <?xml-stylesheet href="http://www.blogger.com/styles/atom.css" type="text/css"?>
3 <feed version="0.3" xml:lang="en-GB" xmlns="http://purl.org/atom/ns#">
4 <link href="http://www.blogger.com/atom/5543214" rel="service.post" title="Travelin' 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 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 Blogger Help for more info.
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 & 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 & XML" type="text/html"/>
24 <id>tag:blogger.com,1999:blog-5543214.post-110720445691459308</id>
25 <title mode="escaped" type="text/html">LC, ISBN &
Line: 1 Column: 1 XML Soft Tabs: 4

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

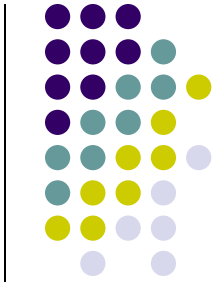




RSS Reader/Aggregator

- 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

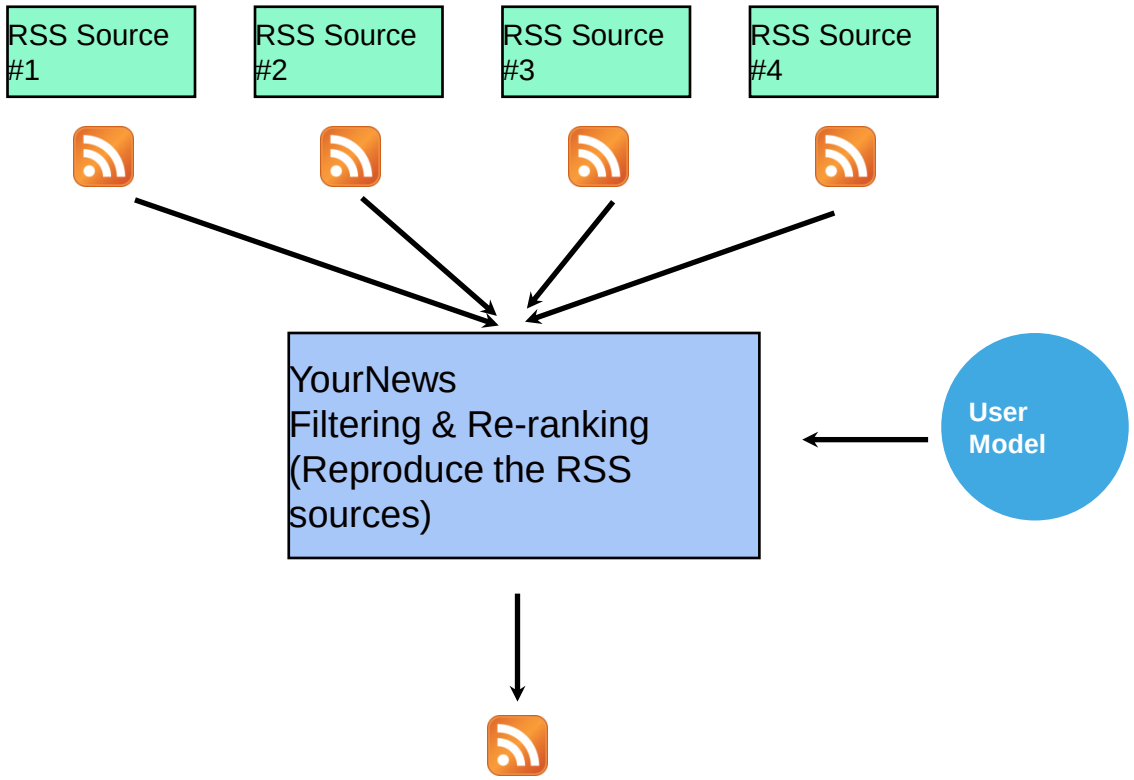
RSS Republishing Example : YourNews



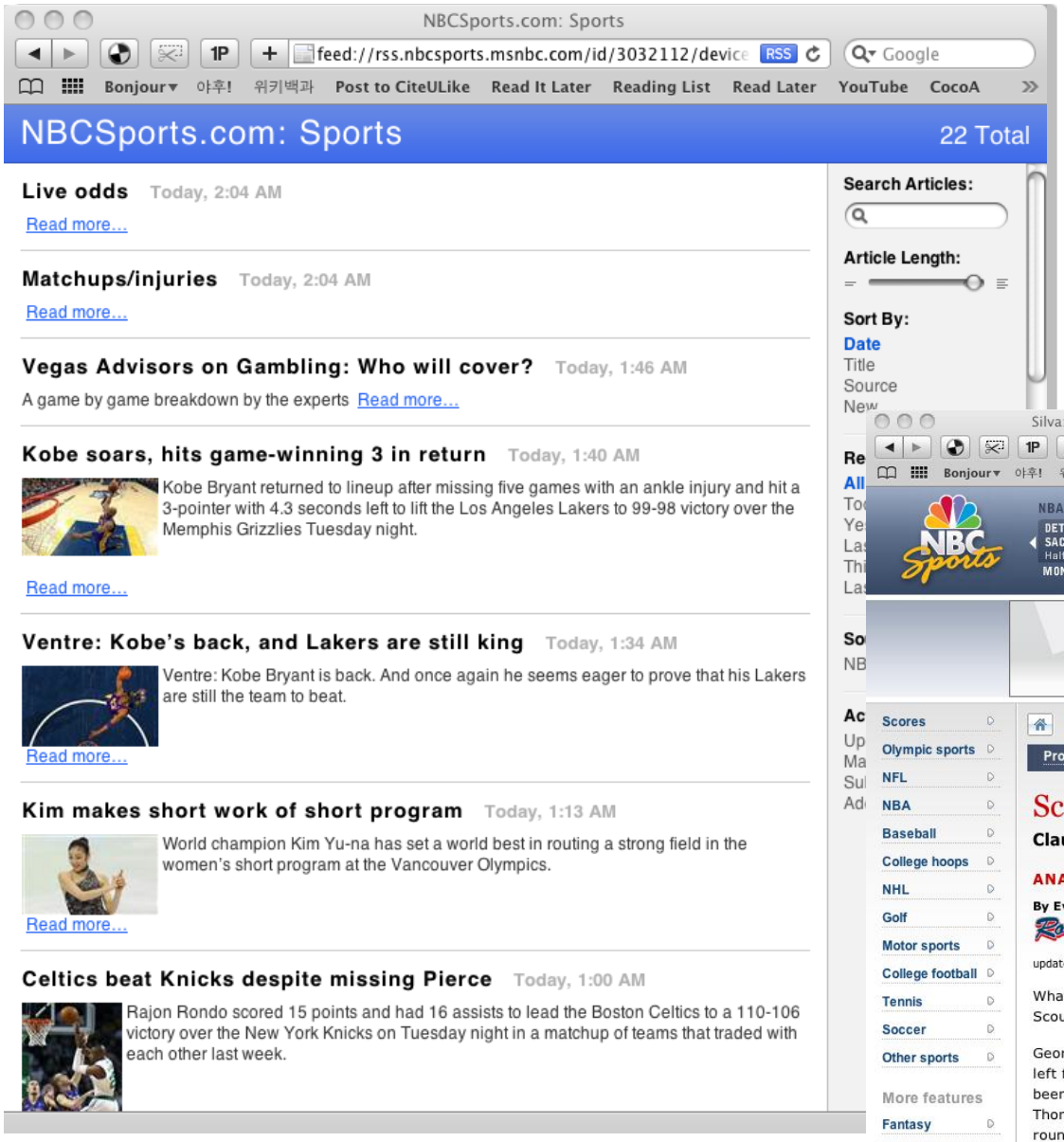
- Adaptive news filtering system
- Read/gather RSS feeds
- Reproduce news list (personalize) — filtering & re-ranking
- Re-publish it in a native interface or through RSS feeds again

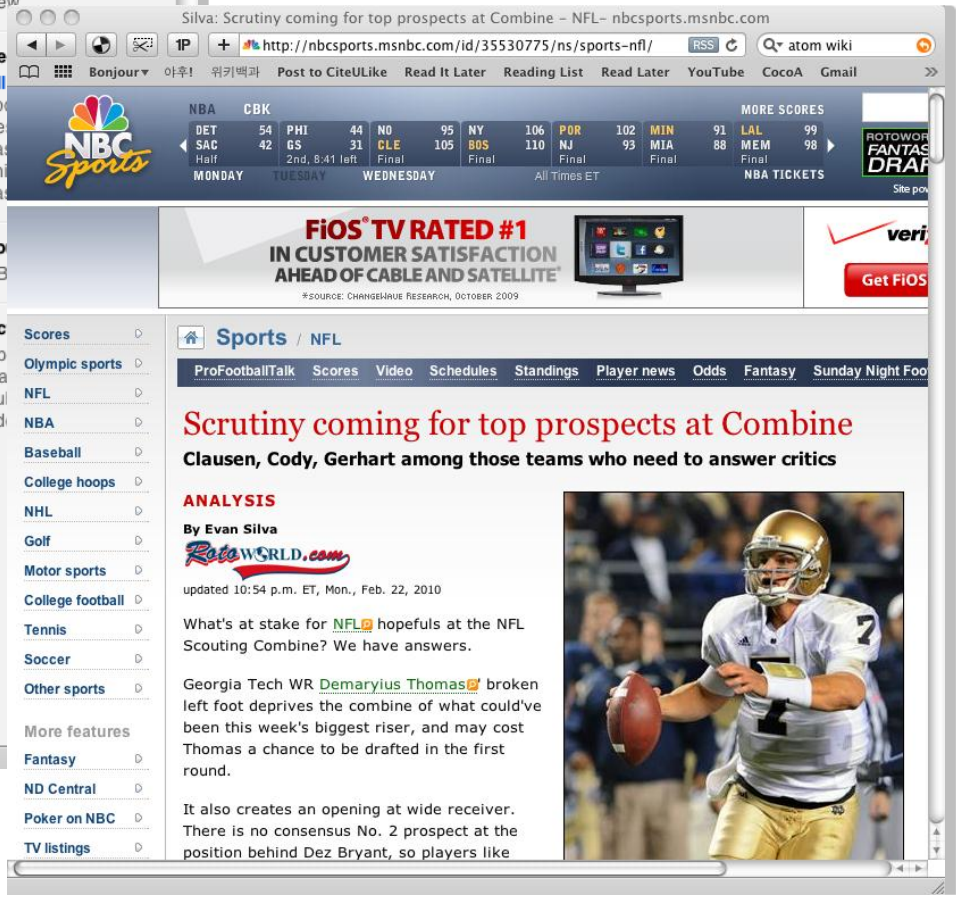
YourNews http://amber.exp.sis.pitt.edu/yournews/yournews.php

YourNews structure



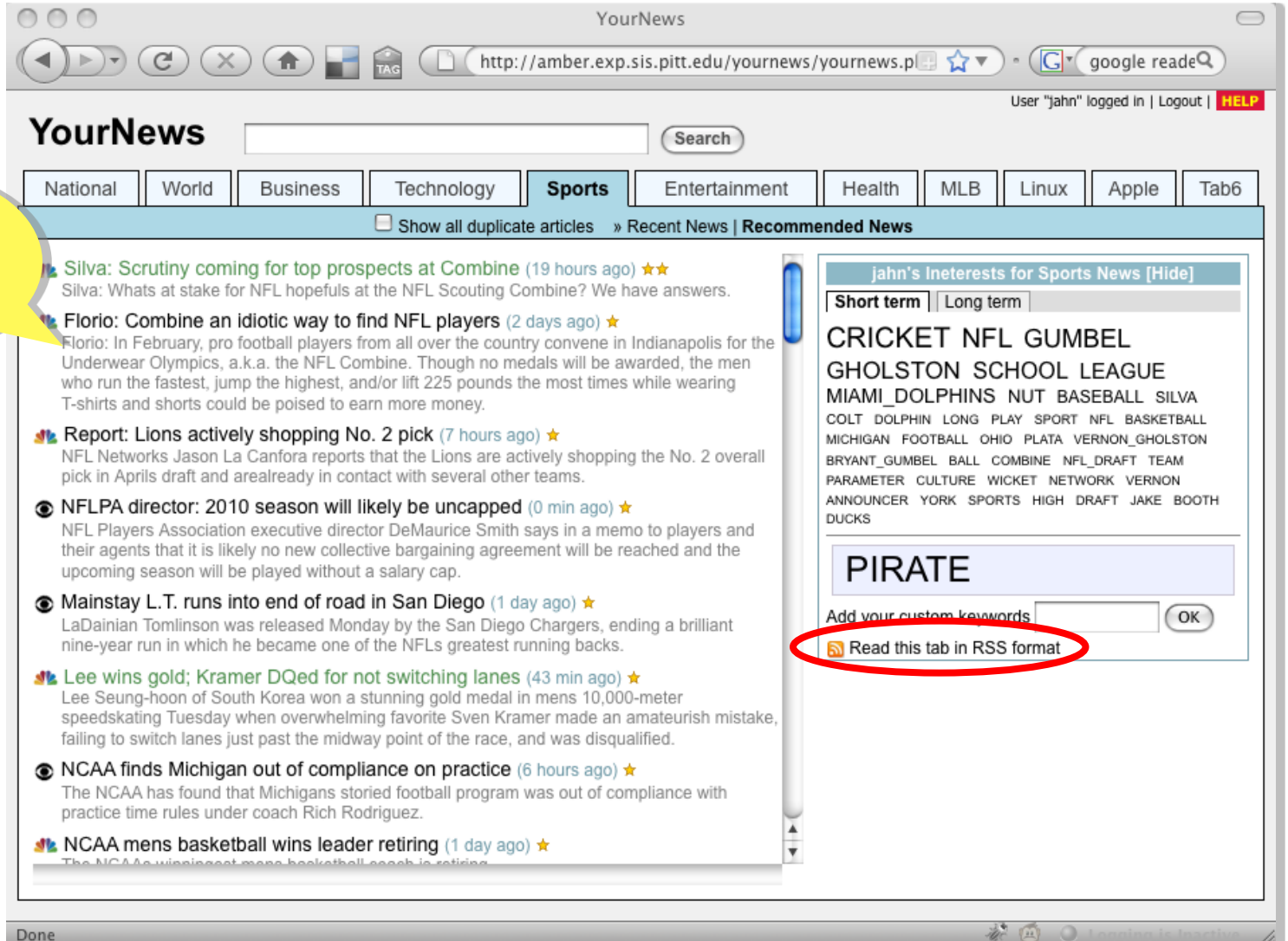
A new (personalized) RSS feed


NBCSports.com: Sports
feed://rss.nbcsports.msnbc.com/id/3032112/device
Google
Bonjour
Post to CiteULike
Read It Later
Reading List
Read Later
YouTube
CocoA
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
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.
Read more…
Ventre: Kobe's back, and Lakers are still king Today, 1:34 AM
Ventre: Kobe Bryant is back. And once again he seems eager to prove that his Lakers are still the team to beat.
Read more…
Kim makes short work of short program Today, 1:13 AM
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.
Read more…
Celtics beat Knicks despite missing Pierce Today, 1:00 AM
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.
Search Articles:
Article Length:
Sort By:
Date
Title
Source


Silva: Scrutiny coming for top prospects at Combine – NFL- nbcsports.msnbc.com
http://nbcsports.msnbc.com/id/35530775/ns/sports-nfl/
atom wiki
Bonjour
Post to CiteULike
Read It Later
Reading List
Read Later
YouTube
CocoA
Gmail
NBC Sports
NBA
CBK
MORE SCORES
DET 54 SAC 42 Half
PHI 44 GS 31 2nd, 8:41 left
NO 95 CLE 105 Final
NY 106 BOS 110 Final
POR 102 NJ 93 Final
MIN 91 MIA 88 Final
LAL 99 MEM 98 Final
MONDAY
TUESDAY
WEDNESDAY
All Times ET
NBA TICKETS
FiOS® TV RATED #1
IN CUSTOMER SATISFACTION
AHEAD OF CABLE AND SATELLITE*
*SOURCE: CHANGEWAVE RESEARCH, OCTOBER 2009
Get FiOS
Scores
Olympic sports
NFL
NBA
Baseball
College hoops
NHL
Golf
Motor sports
College football
Tennis
Soccer
Other sports
More features
Fantasy
ND Central
Poker on NBC
TV listings
Sports / NFL
ProFootballTalk
Scores
Video
Schedules
Standings
Player news
Odds
Fantasy
Sunday Night Foo
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

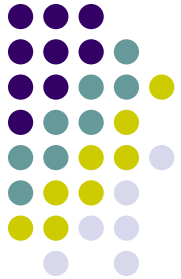
Reproduced (personalized) list

Personalized List: filtered & re-ranked



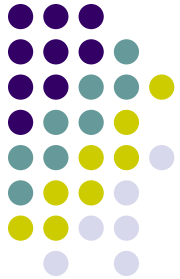
Parsing RSS Feeds

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

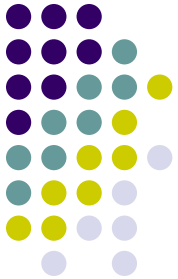


SAX and DOM

- 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

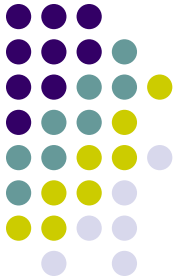


SAX Example



```
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

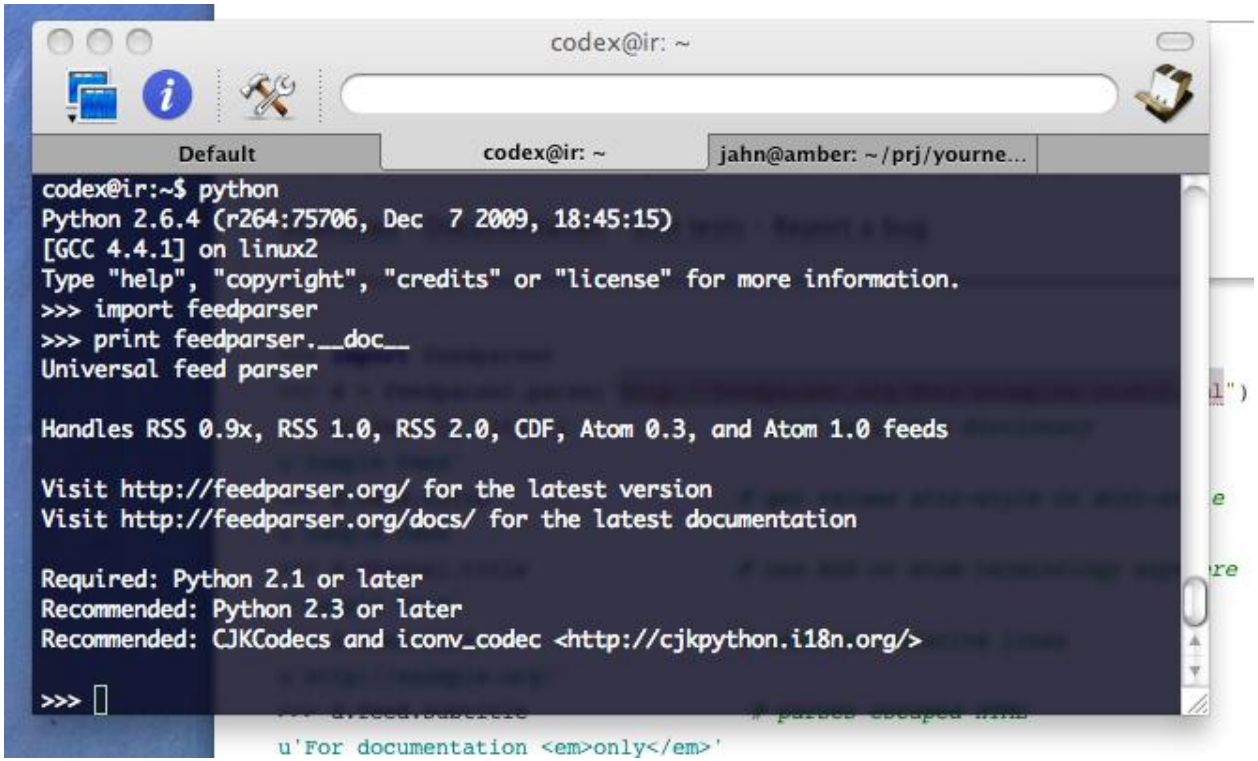


```
import xml.dom.minidom

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>"
```

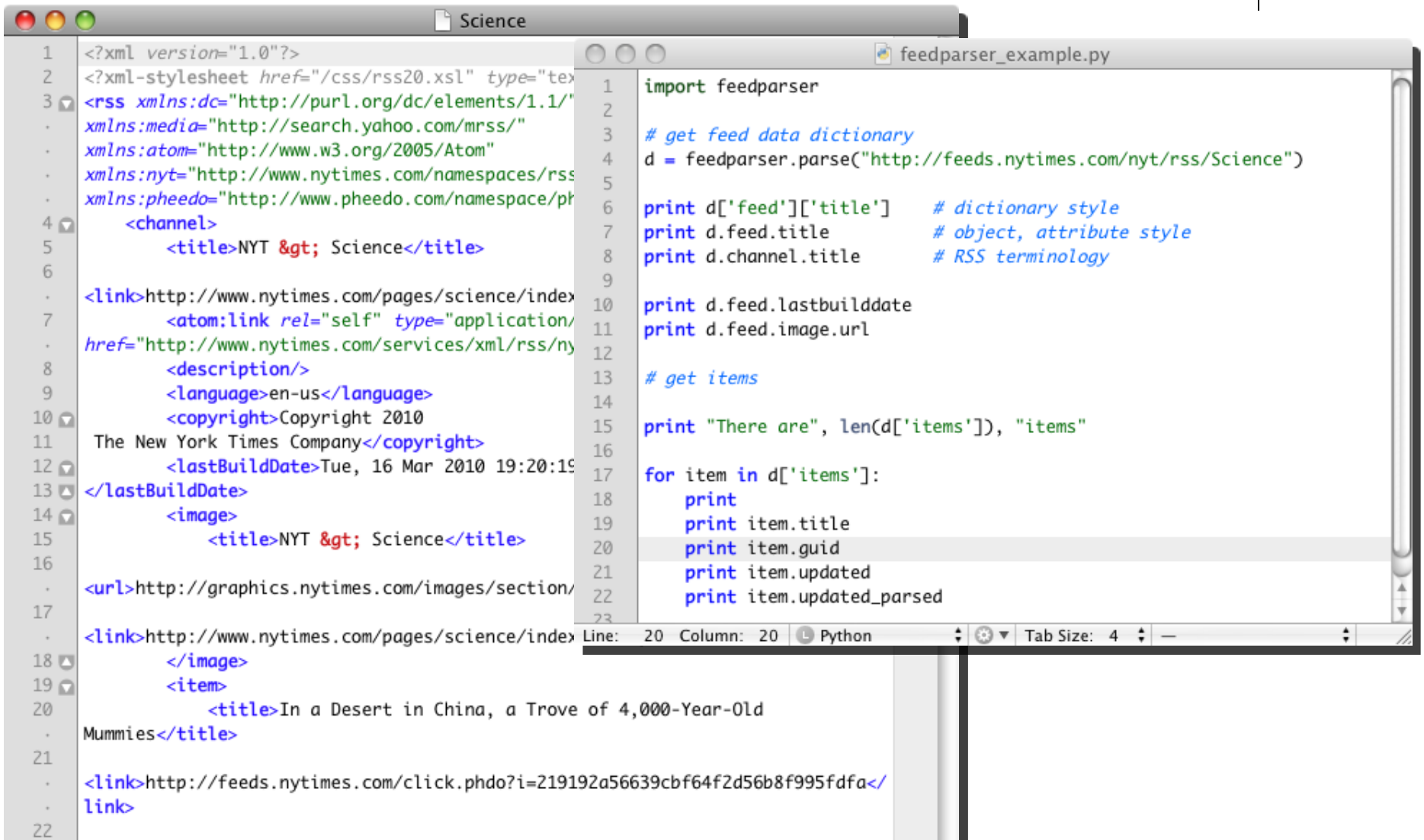
```
def handleSlides(slides):
    for slide in slides:
        handleSlide(slide)

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>



Universal Feedparser

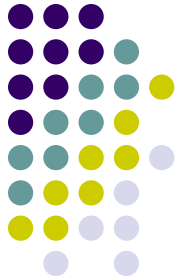


Core Attributes

- 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?