A FRAMEWORK FOR
AGGREGATING, ORGANIZING,
AND SEARCHING RSS FEEDS

by

Samir Bathla

A thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Computer Science

Pace University

2006

Approved by ___________________________________________________
Chairperson of Supervisory Committee

________________________________________________
________________________________________________
________________________________________________

Program Authorized
to Offer Degree ____________________________________________

Date _________________________________________________________
Recently, a new medium for publishing contents on the Internet has been gaining popularity. This new medium, called RSS, allows anyone to publish dynamic content in standard machine readable format from the internet. RSS has many advantages, however there are significant disadvantages arising mainly due to RSS being a single source publishing medium.

This thesis is an attempt to address these disadvantages and make the world of RSS a more useful environment for Web users. The web site developed as part of this project gathers content, stores it and then makes it accessible to and searchable by Web users.
<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>RSS File</td>
<td>5</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Main page showing categories and sources at the top navigation panel and Sources for the selected category “Top News” on the left link list</td>
<td>7</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Main page showing categories and sources in the top navigation panel and Categories for the NPR source on the left links bar</td>
<td>9</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Contents Page</td>
<td>11</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to Professor Narayan Murthy for his assistance in the preparation of this manuscript.
INTRODUCTION

Recently, many news and blog sites have begun providing news stories in a new medium. This new medium, called RSS, allows anyone to retrieve the latest content on demand. These sites are constantly changing and adding content. As an Internet user, RSS allows you to subscribe to these services. And unlike subscriptions to many printed newspaper and magazines, RSS feeds are free.

While RSS offers many advantages, there are also significant issues. How does a user find the most relevant stories across all these different sources? If one wishes to search by category, where is a master list provided? Finally, is there some easy way to find the sources that users trust most? This project is an attempt to address these issues, an attempt to present thousands of news stories to Web users in the most usable format possible.
RSS FEEDS

RSS is used by news websites, weblogs and any site where content keeps changing, for example a “deals” site. The abbreviation is generally used to refer to ‘Really Simple Syndication’ but there are other standards too:

- Really Simple Syndication (RSS 2.0)
- Rich Site Summary (RSS 0.91, RSS 1.0)
- RDF Site Summary (RSS 0.9 and 1.0)
- Real-time Simple Syndication (RSS 2.0)

RSS feeds are just a special kind of Web page, designed to be read by computers rather than people. It might help to think of them as the free, internet version of the old-fashioned ticker-tape news wire machines. Websites create an RSS document and save it with a .xml extension. Then they upload the file on Website. Creation of RSS is generally automated and is updated as soon as there is any new story. Many sites publish the date and time of the update in the title.

With RSS feeds, one can keep track of when sites from all over the internet have added new content. This enables one to get the latest headlines and articles (or even audio files, photographs or video) in one place, as soon as they are published, without having to remember to visit each site every day. It takes the hassle out of staying up-to-date, by showing the latest information of interest.

Not all Websites currently provide RSS, but it is growing rapidly in popularity and many large sites, including the Guardian, New York Times, and CNN provide it.

RSS SYNTAX

RSS is written in XML, so basic rules of XML apply here too. , for example, all elements must have a closing tag, elements are case sensitive, elements must be properly nested and attribute values must be quoted. RSS must also follow the syntax defined below.

The first line in the document - the XML declaration - defines the XML version and the character encoding used in the document. In this case the document conforms to the 1.0 specification of XML and uses the ISO-8859-1 (Latin-1/West European) character set.
The next line is the RSS declaration which identifies that this is an RSS document (in this case, RSS version 2.0).

The next line contains the <channel> element. This element is used to describe the RSS feed.

The <channel> element has three required child elements:

- <title> - Defines the title of the channel (e.g. BBC News | News Front Page | UK Edition)
- <link> - Defines the hyperlink to the channel (e.g. http://news.bbc.co.uk/go/rss/-/1/hi/default.stm)
- <description> - Describes the channel (e.g. Visit BBC News for up-to-the-minute news, breaking news, video, audio and feature stories. BBC News provides trusted World and UK news as well as local and regional perspectives. Also entertainment, business, science, technology and health news.)

Each <channel> element can have one or more <item> elements.

Each <item> element defines an article or "story" in the RSS feed.

The <item> element has three required child elements:

- <title> - Defines the title of the item (e.g. Blair defends embattled minister)
- <link> - Defines the hyperlink to the item (e.g. http://news.bbc.co.uk/go/rss/-/1/hi/uk_politics/4951730.stm)
- <description> - Describes the item (e.g. Tony Blair robustly defends his ministers and laughs off claims that he has suffered his own "Black Wednesday").

Finally, the two last lines close the <channel> and <rss> elements.

The following is an example of RSS file.
Figure 1 RSS File
STATEMENT OF PROBLEM

Today there are many Websites, blogs, and news channels that publish RSS feeds for the site’s content. As more and more Websites provide RSS feeds, it is getting extremely difficult to extract the information that is relevant and meaningful to the user. Some of the key challenges are:

- How to handle an ever increasing number of feeds?
- How to add the RSS feeds from a new source? Many people have favorite web sites and they would like to add content from new sources.
- Is there an easy way to find the sources that users trust most?
- How to categorize and arrange news items so that a user can get to them easily. If a user wishes to search by category, is a master list provided?
- Finally, can a user search the feeds using keywords?

METHODOLOGY

To deliver thousands of current news stories from hundreds of sources to our users, many different pieces must be brought together. The system needs to retrieve stories from the sources, store the stories to provide search across sources and then display the stories that fit the user’s needs. It should also provide a way to allow users to add their preferred links.

The framework developed in this thesis consists of modules that carry out the requisite functions, discussed in the previous paragraph, in an integrated manner. The key modules are – an aggregator that gathers the data from multiple RSS feeds and stores the contents in a database, a web-presentation module that retrieves the data and displays it according to default and user-specific preferences, and a customization and search module that provides user specific customization and search functionalities. These modules employ a wide array of technologies including PHP, HTML, XML, XSL, SQL, CRON and RDBMS.
The relational database used by the service is SQLite. This is a fast relational store that is closely integrated with PHP. It is used to store the users, user preferences, stories, and other relevant data.

All the server side scripting is carried out using PHP, a common scripting language. PHP provides rich functionality allowing one to fetch and parse content from the Web.

We now describe the system modules in detail.

**AGGREGATOR**

RSS feeds are available through URLs, therefore our system needs to store the URL for each source of news stories. Then the system can iterate through these URLs and retrieve the stories at any time. All news and blog sites publish these feeds at the RSS link. For example, the BBC site: http://news.bbc.co.uk/2/hi/help/3223484.stm has 25 RSS feed links. After storing the RSS-feed URLs, an automated retrieval (aggregation) of the news stories is performed using CRON scheduler of UNIX to run retrieval code. The scheduler can be set to update at any given interval, for this system, it updates once a day at midnight. Depending on the requirement, this time interval can be decreased to get stories more frequently.
After the stories are retrieved, they are parsed and stored in the relational database to provide searchable contents.

WEB PRESENTATION

Once stories are aggregated and stored in the database, a Web site can be created that allows users to enter search criteria. The system then searches the database and returns the stories that match those criteria. Users can also get stories by sources and categories. This enables the user to get their content quickly and easily. No knowledge of RSS or XML is required.

To show all the RSS links such that user easily finds what they are looking for, there is a Navigation panel on the top, see figure 2. In the Navigation panel, there are three rows – one each for region, category, and source. Top five regions, categories and sources are shown by default and at the end of the three rows for other categories and sources, more options are available, such as My Links, More Categories and More Sources. SQL query is used to retrieve and dynamically render sources or categories for the left link list when category or source is selected in the top navigation bar. The caption is set to the selected source or category. The caption is displayed in the blue bar under the top Navigation panel.

For example, to display all the categories of source NPR, select NPR in the top Navigation panel. The caption displays NPR. The left link list displays all the categories of NPR, for example - World News, Business, NPR Top News, US News, All Things Considered, Morning Edition, Talk of the Nation, Weekend Edition Saturday, Weekend Edition Sunday, News and Notes With Ed Gordon, Fresh Air With Terry Gross, World Café, The Motley Fool Radio Show, Day To Day, Art & Culture, health & Science, Opinion, People & Place, Politics & Society. A snapshot of this example is shown in Figure 1.

By default first the RSS is selected and the main content area shows all the stories from that link. Now the user can select any other link and customize the site.
Above the Left panel, there is a search box to search across the RSS links. There are also Quick Search keywords below the Left panel.

When a user visits the Web site for the first time, guest user links and keywords are displayed. On top there are links for user name, content and login. By clicking the login label, the user can enter their information to log into the system. Once the user successfully logs in, they can click on the contents label to add links and keywords, see figure 4.

To store users and their preferences, two tables are required in the database. There are two fields in the table for Users – User Name and Password. In one field, an email address is needed if the password retrieval feature is required. There are four fields in the Preferences table: Username, Title, URL and Use. This table is used for storing user links and keywords.
To store links and the associated stories, two tables are required in the database. There are five fields in the table for URLs: Source, Category, Title, Link and Priority. Priority is used to set the source order in the Left list panel. For example, in the figure below, for the top news category BBC, priority is set to 1, NPR is set to 2, and so on.

The News category has sub-categories, for example: Top News, World News, Regions like US, India and so on. Instead of saving News as a category and adding another field for sub-category, the sub-category itself is used as a category. They are also placed on the top Navigation panel in a separate line as they are the most commonly used links. There are five fields in the table for headlines also – Title, Link, GUID, PubDate, and Description. GUID are the unique IDs for each story. Before saving the story, a check is performed to see if the GUID exists or not. If it is present, it is deleted from the database and then saved again. This is required since the Title and Description of the story might have changed from the last time it was saved. It further helps avoid duplicate data in case the link is stored multiple times.

Two XSL files are used to format the Web contents. One is used for the top Navigation panel and the left link list, and the other for the news contents. When the left link is selected, all the news is returned in the RSS format. Using one of the XSL files, it is formatted and shown in the content area. When search is used, results are formatted in the RSS format and returned. This allows the same XSL file to be used to format it.
CUSTOMIZATION AND SEARCH

The ability to add or delete links by user, is very important as it ensures that all the links that are utilized by users are in the system. When the Web site is first opened, the user is considered a guest user. Returning users can subsequently login by clicking the login label at the top of the page. At the login screen, users are required to enter user name and password, and once authenticated they are redirected to the main page.

Once logged in, the user name is updated on the top of the Web page from ‘Guest’ to the User’s name. The user can then add and modify contents of My Links and Keywords by going to the contents page (see figure 3). After adding, deleting, arranging the links and keywords, the user can save the changes by clicking the Save button. Keywords are used for searching across RSS links. To access the added links, “My Links” is provided in the Navigation panel. Keywords are shown below the links on all the pages, as shortcuts for searching on those keywords.
To search the RSS feeds, a search text box and button is provided above the Left Panel. Users can type the text to search and select the Go button to search the stored stories of all RSS links. The result is returned in RSS format and displayed in the main content area. The same XSL used in formatting the other stories is used to format the returned result.

User’s saved keywords are shown below the Navigation left panel. Optionally, if users want to search any of the keywords listed under the Quick Search, they may simply click on the word to obtain the result.
SUMMARY

In the prototype implementation of the framework for aggregating, organizing, and searching RSS feeds, the basic functionality required is currently available and working on news sites. As RSS feeds are rapidly becoming a common internet publishing method, this aggregation framework will be a valuable contribution to the community. Displaying and arranging all available links on a single web page in a format that a user would easily find what they were looking for, was a challenging task. Running the automated task to fetch all the stories and store in the database presented some challenges.

The ability to customize was a very important objective to show how this feature can be extended to an almost unlimited number of links, so that users can get all desired links in one place. It also showed that by storing keywords, the search function can be used in different ways. In the future, APIs can be provided so that other Web sites can also search RSS feeds from this site. More Preferences can also be added to order source and category in the Navigation panel on the front page.

While it is a fully functional basic prototype, a multitude of enhancements can be made in the future. For instance, counters can be added in the source and category fields to keep track of how many times they were used. This can help in automatically showing the top five categories and sources in the navigation panel. Furthermore, in this project, one XSL is used to show the stories for a category and source and corresponding search results. Multiple XSL files can also be used to show search results, to give more control to the output of search results for example, displaying source name and time information of the story in the headline.
GLOSSARY

**RSS.** Really Simple Syndication (RSS) is a family of web feed formats, specified in XML and used for Web syndication.

**XML.** Extensible Markup Language (XML) is a W3C-recommended general-purpose markup language for creating special-purpose markup languages, capable of describing many different kinds of data.

**URL.** A Uniform Resource Locator (URL) is a string of characters conforming to a standardized format, which refers to a resource on the Internet (such as a document or an image) by its location.

**HTML.** HyperText Markup Language (HTML) is a markup language designed for the creation of web pages with hypertext and other information to be displayed in a web browser.

**PHP.** “PHP: Hypertext Preprocessor” is a scripted programming language that can be used to create websites. The initials actually come from the earliest version of the program, which was called "Personal Home Page" but in their current form, constitute a recursive acronym. It is an open-source, reflective programming language used mainly for developing server-side applications and dynamic web content.

**XSL.** The eXtensible Stylesheet Language (XSL) is a family of languages which allows one to describe how files encoded in the XML standard are to be formatted or transformed.

**SQL.** Structured Query Language (SQL) is the most popular computer language used to create, modify, retrieve and manipulate data from relational database management systems.

**CRON.** - Cron is the name of a program that enables Unix users to execute commands or scripts (groups of commands) automatically, at a specified time/date.

**RDBMS.** - A Relational Database Management System (RDBMS) is a database management system (DBMS) that is based on the relational model as introduced by Edgar F. Codd.
BIBLIOGRAPHY


Yank, Kevin *Build Your Own Database-Driven Website Using PHP & MySQL*, 2nd ed. SitePoint Pty. LTD., 2003.

http://en.wikipedia.org/wiki/Main_Page

http://news.bbc.co.uk/2/hi/help/3223484.stm

http://www.w3schools.com/
## Table – plusview: for all sources

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Type</th>
<th>Primary</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Text</td>
<td>Yes</td>
<td>CNN</td>
</tr>
<tr>
<td>Category</td>
<td>Text</td>
<td>Yes</td>
<td>Science</td>
</tr>
<tr>
<td>Title</td>
<td>Text</td>
<td></td>
<td>Science &amp; Space</td>
</tr>
<tr>
<td>Link</td>
<td>Text</td>
<td></td>
<td><a href="http://rss.cnn.com/rss/cnn_space.rss">http://rss.cnn.com/rss/cnn_space.rss</a></td>
</tr>
<tr>
<td>Priority</td>
<td>number</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

## Table – prefs: for customization

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Type</th>
<th>Primary</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Username</td>
<td>Text</td>
<td>Yes</td>
<td>Samirb</td>
</tr>
<tr>
<td>Title</td>
<td>Text</td>
<td></td>
<td>Ap Top News</td>
</tr>
<tr>
<td>url</td>
<td>Text</td>
<td></td>
<td><a href="http://hosted.ap.org/lineups">http://hosted.ap.org/lineups</a></td>
</tr>
<tr>
<td>use</td>
<td>Text</td>
<td></td>
<td>Ml</td>
</tr>
</tbody>
</table>

## Table – users: for authentication

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Type</th>
<th>Primary</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Username</td>
<td>Text</td>
<td>Yes</td>
<td>Samirb</td>
</tr>
<tr>
<td>Password</td>
<td>Text</td>
<td></td>
<td>Encoded password</td>
</tr>
</tbody>
</table>

## Table – headlines: for search

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Type</th>
<th>Primary</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Text</td>
<td></td>
<td>Home Secretary Charles Clarke must…</td>
</tr>
<tr>
<td>Description</td>
<td>Text</td>
<td></td>
<td>Opposition urge Clarke to resign</td>
</tr>
<tr>
<td>Guid</td>
<td>Text</td>
<td>Yes</td>
<td><a href="http://news.bbc.co.uk/go/rss">http://news.bbc.co.uk/go/rss</a>...</td>
</tr>
<tr>
<td>Link</td>
<td>Text</td>
<td></td>
<td><a href="http://news.bbc.co.uk/1/hi">http://news.bbc.co.uk/1/hi</a>...</td>
</tr>
<tr>
<td>Pubdate</td>
<td>Text</td>
<td></td>
<td>1146066393</td>
</tr>
</tbody>
</table>
First time when a user visits the site, main page is shown with guest as default user. On clicking login…
User is presented with login screen. If user is not a registered user then by clicking here, they can register to the site. On clicking “here”…
User is presented with registration screen where they can create a new user and new password.
On successful login user name is displayed on the top and user keywords under quick search. User can change keywords and links by going to the contents page.
On contents page user can see the existing links and keywords. Here they can add new links, arrange links and add or delete keywords. Once user preferences are saved, it is reflected on the main page.
On clicking My Links user saved links are shown same way as the other links.
On clicking keywords in quick search user can get all the news related to that keyword.
User can also type keywords in the search text box and press Go to get all the news related to that keyword.
APPENDIX C – UNIX CODE

Scheduled job code: Scheduled daily at 12:00 pm

Crontab -e

0 0 * * * /home/zequant/bin/store_headlines.sh > /tmp/store_headlines.log
2>&1

store_headlines.sh code:

#!/bin/bash

cd /www/zequant.com/subdomains/www/thesis

/usr/bin/php
APPENDIX D – AGGREGATOR CODE

StoreHeadlines.php – This file contains all the code required to store headlines.
APPENDIX E – WEB PRESENTATION CODE

plusview.php – Main page code in PHP. Echo is used to generate HTML.

page3.xsl – XSL file to format main page

rss.xsl – XSL file to format contents in the frame
APPENDIX F – CUSTOMIZATION AND SEARCH CODE

Login.php – PHP file for login screen.

Functions.php – File containing all the functions required for user creation and authentication.

Content.php – PHP file for showing contents page for customization.

Sort.js – Java Script file for arranging links in the contents page.

SearchHeadlines.php – PHP file to search keywords in the stored headlines.