

# News Website Refinement Using RSS Aggregator

<sup>1</sup>P.Nitin Chandra, <sup>2</sup>Dr.B.Sankara Babu

<sup>1</sup>*B.Tech Scholar, Department of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana – 500 090*

<sup>2</sup>*Professor, Department of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana – 500 090*

**Abstract-** Real Simple Syndication (RSS) is a type of web feed which allows users and applications to access updates to online content in a standardized, computer-readable format. In computing, a news aggregator, also termed a feed aggregator, feed reader, news reader, RSS reader or simply aggregator, is client software or a web application which aggregates syndicated web content such as online newspapers, blogs, podcasts, and video blogs (vlogs) in one location for easy viewing. Using this aggregator, it is possible to subscribe to various online content and maintain them in separate folders as an individual collection. Whenever there is an update in that specific subscribed online content, a notification is given out by the aggregator and that particular updated information can be viewed and accessed. We provide a website for getting the data from various websites and storing them in a database. We provide a User Interface to display this information to the user and provide links to the websites. Text Summarization is done such that the user can understand the required concise data thereby providing a faster and efficient way of knowledge to our users. This reduces the burden of the reader as the entire information does not have to be read. When any given word is highlighted, we provide a dictionary meaning to that word for comprehensive understanding. Bookmarking a specific content is also provided along with these features. Through this project, we provide a solution to our users to easily understand access and store their favorable data.

## I. INTRODUCTION

### REAL SIMPLE SYNDICATION

RSS is a family of Web feed formats used to publish often updated content such as blog entries, news headlines or podcasts. An RSS document, which is called a "feed," "web feed," or "channel," contains either a summary of content from an associated web site or the full text. RSS makes it possible for people to keep up with their favorite web sites without having to check them manually. This flow of content between websites and users is called "web syndication".

RSS content can be read using software called an "RSS reader," "feed reader" or an "aggregator." The user subscribes to a feed by entering the feed's link into the reader or by clicking an RSS icon in a browser that begins the subscription process. The reader checks the user's subscribed feeds often for new content, downloading any updates that it finds. This content is stored in a standard XML file that many different computers and programs

can understand. This file tells the programs how to show the content to the user.

## II. RSS AGGREGATOR

The news aggregator will automatically check the RSS feed for new content, allowing the content to be automatically passed from website to website or from website to user. This passing of content is called web syndication. Websites usually use RSS feeds to publish frequently updated information, such as blog entries, news headlines, or episodes of audio and video series. RSS is also used to distribute podcasts. An RSS document (called "feed", "web feed", or "channel") includes full or summarized text, and metadata, like publishing date and author's name.

A standard XML file format ensures compatibility with many different machines/programs. RSS feeds also benefit users who want to receive timely updates from favorite websites or to aggregate data from many sites.

Subscribing to a website RSS removes the need for the user to manually check the website for new content. Instead, their browser constantly monitors the site and informs the user of any updates. The browser can also be commanded to automatically download the new data for the user.

RSS feed data is presented to users using software called a news aggregator. This aggregator can be built into a website, installed on a desktop computer, or installed on a mobile device. Users subscribe to feeds either by entering a feed's URI into the reader or by clicking on the browser's feed icon. The RSS reader checks the user's feeds regularly for new information and can automatically download it, if that function is enabled. The reader also provides a user interface.

### HISTORY OF RSS

The RSS formats were preceded by several attempts at web syndication that did not achieve widespread popularity. The basic idea of restructuring information about websites goes back to as early as 1995, when Ramanathan V. Guha and others in Apple Computer's Advanced Technology Group developed the Meta Content Framework. RDF Site Summary, the first version of RSS, was created by Dan Libby and Ramanathan V. Guha at Netscape. It was released in March 1999 for use on the My.Netscape.Com portal. This version became known as RSS 0.9. In July 1999, Dan Libby of Netscape produced a new version, RSS 0.91, which simplified the format by removing RDF elements and

incorporating elements from Dave Winer's news syndication format. Libby also renamed the format from RDF to RSS **Rich Site Summary** and outlined further development of the format in a "futures document". This would be Netscape's last participation in RSS development for eight years. As RSS was being embraced by web publishers who wanted their feeds to be used on My.Netscape.Com and other early RSS portals, Netscape dropped RSS support from My.Netscape.Com in April 2001 during new owner AOL's restructuring of the company, also removing documentation and tools that supported the format. In September 2002, Winer released a major new version of the format, RSS 2.0, that redubbed its initials Really Simple Syndication. RSS 2.0 removed the typeattribute added in the RSS 0.94 draft and added support for namespaces. To preserve backward compatibility with RSS 0.92, namespace support applies only to other content included within an RSS 2.0 feed, not the RSS 2.0 elements themselves.<sup>[16]</sup> (Although other standards such as Atom attempt to correct this limitation, RSS feeds are not aggregated with other content often enough to shift the popularity from RSS to other formats having full namespace support.)

In July 2003, Winer and UserLand Software assigned the copyright of the RSS 2.0 specification to Harvard's Berkman Center for Internet & Society, where he had just begun a term as a visiting fellow.<sup>[18]</sup> At the same time, Winer launched the RSS Advisory Board with Brent Simmons and Jon Udell, a group whose purpose was to maintain and publish the specification and answer questions about the format. In January 2006, Rogers Cadenhead relaunched the RSS Advisory Board without Dave Winer's participation, with a stated desire to continue the development of the RSS format and resolve ambiguities. In June 2007, the board revised their version of the specification to confirm that namespaces may extend core elements with namespace attributes, as Microsoft has done in Internet Explorer 7. According to their view, a difference of interpretation left publishers unsure of whether this was permitted or forbidden.

### III. VARIANTS OF RSS

There are several different versions of RSS, falling into two major branches (RDF and 2.\*). The RDF (or RSS 1.\*) branch includes the following versions:

RSS 0.90 was the original Netscape RSS version. This RSS was called *RDF Site Summary*, but was based on an early working draft of the RDF standard, and was not compatible with the final RDF Recommendation.

RSS 1.0 is an open format by the RSS-DEV Working Group, again standing for *RDF Site Summary*. RSS 1.0 is an RDF format like RSS 0.90, but not fully compatible with it, since 1.0 is based on the final RDF 1.0 Recommendation.

RSS 1.1 is also an open format and is intended to update and replace RSS 1.0. The specification is an independent draft not supported or endorsed in any way by the RSS-Dev Working Group or any other organization.

The RSS 2.\* branch (initially UserLand, now Harvard) includes the following versions:

RSS 0.91 is the simplified RSS version released by Netscape, and also the version number of the simplified version originally championed by Dave Winer from Userland Software. The Netscape version was now called *Rich Site Summary*; this was no longer an RDF format, but was relatively easy to use.

RSS 0.92 through 0.94 are expansions of the RSS 0.91 format, which are mostly compatible with each other and with Winer's version of RSS 0.91, but are not compatible with RSS 0.90.

RSS 2.0.1 has the internal version number 2.0. RSS 2.0.1 was proclaimed to be "frozen", but still updated shortly after release without changing the version number. RSS now stood for *Really Simple Syndication*. The major change in this version is an explicit extension mechanism using XML namespaces.

Later versions in each branch are backward-compatible with earlier versions (aside from non-conformant RDF syntax in 0.90), and both versions include properly documented extension mechanisms using XML Namespaces, either directly (in the 2.\* branch) or through RDF (in the 1.\* branch). Most syndication software supports both branches. "The Myth of RSS Compatibility", an article written in 2004 by RSS critic and Atom advocate Mark Pilgrim, discusses RSS version compatibility issues in more detail.

The extension mechanisms make it possible for each branch to copy innovations in the other. For example, the RSS 2.\* branch was the first to support enclosures, making it the current leading choice for podcasting, and as of 2005 is the format supported for that use by iTunes and other podcasting software; however, an enclosure extension is now available for the RSS 1.\* branch, `mod_enclosure`. Likewise, the RSS 2.\* core specification does not support providing full-text in addition to a synopsis, but the RSS 1.\* markup can be (and often is) used as an extension. There are also several common outside extension packages available, e.g. one from Microsoft for use in Internet Explorer 7.

The most serious compatibility problem is with HTML markup. Userland's RSS reader—generally considered as the reference implementation—did not originally filter out HTML markup from feeds. As a result, publishers began placing HTML markup into the titles and descriptions of items in their RSS feeds. This behavior has become expected of readers, to the point of becoming a de facto standard, though there is still some inconsistency in how software handles this markup, particularly in titles. The RSS 2.0 specification was later updated to include examples of entity-encoded HTML; however, all prior plain text usages remain valid.

As of January 2007, tracking data from [www.syndic8.com](http://www.syndic8.com) indicates that the three main versions of RSS in current use are

0.91, 1.0, and 2.0, constituting 13%, 17%, and 67% of worldwide RSS usage, respectively.<sup>[25]</sup> These figures, however, do not include usage of the rival web feed format Atom. As of August 2008, the syndic8.com website is indexing 546,069 total feeds, of which 86,496 (16%) were some dialect of Atom and 438,102 were some dialect of RSS.

### RSS MODULES

The primary objective of all RSS modules is to extend the basic XML schema established for more robust syndication of content. This inherently allows for more diverse, yet standardized, transactions without modifying the core RSS specification.

To accomplish this extension, a tightly controlled vocabulary (in the RSS world, "module"; in the XML world, "schema") is declared through an XML namespace to give names to concepts and relationships between those concepts.

### CURRENT USAGE OF RSS

Several major sites such as Facebook and Twitter previously offered RSS feeds but have reduced or removed support. Additionally, widely used readers such as Shiira, FeedDemon, and Google Reader have been discontinued having cited declining popularity in RSS. RSS support was removed in OS X Mountain Lion's versions of Mail and Safari, although the features were partially restored in Safari 8. Mozilla removed RSS support from Mozilla Firefox version 64.0, joining Google Chrome and Microsoft Edge which do not include RSS support, thus leaves Internet Explorer the last major browser to include RSS support by default.

## IV. MAIN TYPES OF RSS FEEDS

There are five main types of RSS feeds you want to save and access:

### 1. Feeds from other businesses in your industry

If you're in business, you will want to keep up with what your competitors are doing. By adding your direct competitors to your RSS feed, you will be able to know when they are launching a new site, releasing a new product or using a new marketing tactic. Having this early perspective on their activities can help you determine what you need to do to support or counteract their actions. You also can get good ideas for approaches you might want to implement in your business as well.

### 2. Feeds from Industry Leaders

In every industry, there are movers and shakers. These people are the important game changers, the ones everyone listens to when they speak. By adding these industry leaders and their sites to your RSS feeds, you will always know what's happening in your industry.

You'll be on the front line when new trends happen and important changes are made. You will know when industry-sponsored conferences and training material become available. You'll also be one of the first to know of learning, marketing, speaking opportunities.

### 3. Feeds about Trends

You can get really good ideas by watching trends. You can choose to watch trends in any arena — marketing, retail, entertainment, lifestyle — whatever interests you. The great thing about watching trends is they often give rise to business innovation, even if your company is very different and in a different industry entirely.

### 4. Feeds about Marketing

If you have a business, it is wise to subscribe to at least a few marketing feeds. You can get relevant information on new sites to participate in, new promotional strategies, and new marketing avenues. None of us, individually, can keep track of all the new opportunities, but we can harness the effort of others by keeping up with them in RSS.

### 5. Inspirational and Creative Feeds

I subscribe to about 25 creative and inspirational feeds, consisting of quotes, photographic images, furniture design, artwork (especially that done in glass and clay), and high fashion. Not only do I learn interesting things about each of these industries, these feeds are like candy for the eye and the brain. If I ever need a quick break or a way to refresh my thinking, visiting these inspirational and creative feeds helps me shift my mindset and get inspired.

## V. PROS AND CONS OF RSS AGGREGATOR

### Pros Of Using RSS

The benefits for a webmaster who opts to implement RSS feeds on their website are numerous:

#### 1. Saves Time

RSS feeds save time. RSS subscribers can quickly scan RSS feeds, without having to visit each and every website. Subscribers can then click on any items they are interested in, to get additional information.

#### 2. Timely

RSS feeds are timely. RSS feeds will automatically update themselves any time new information is posted, so the information your subscribers receive via their RSS reader or news aggregator is timely.

**3. Spam Free**

RSS is free of spam. Subscribers don't have to worry about wading through huge amounts of spam in an attempt to get to the information they are actually interested in.

**4. Opt-In**

The RSS subscriber chooses what they want to see, and what information they wish to receive. Knowing they have full control, and that they do not have to provide any personal information to subscribe, they will be more likely to opt-in.

**5. Unsubscribing Is Easy**

It is also easy to unsubscribe from an RSS feed. If they do not like information contained in an RSS feed, they can simply remove the RSS feed from their RSS reader or news aggregator in order to unsubscribe.

**6. Alternate Communication Channel**

RSS provides you with an alternate communication channel for your business. And the more channels you provide, the more opportunities you have to connect with your customers and potential customers.

**7. Expands Audience Through Syndication**

The very nature of RSS is that it is designed specifically for syndication (i.e. publication by others). And wide-spread syndication can expand a company's reach and strengthen the company brand.

**8. Can Increase Backlinks**

When an RSS feed is syndicated, it can increase the number of links back to the original website. And additional incoming links will often help a website rank better in organic search rankings.

**9. Increases Productivity**

RSS increases productivity, allowing people to quickly scan new posts and headlines, and only clicking through and spending time on the items of interest.

**10. Competitive**

Whether you decide to implement RSS feeds or not, your competitors likely will. So one way to remain competitive is to implement RSS feeds and other web 2.0 technology, and not allow your competition to get ahead of you.

**Cons Against Using RSS****1. Not Widely Adopted Yet**

Outside of technical circles, RSS has not yet been widely adopted. While it is becoming more and more popular, it is still far from being a mainstream technology.

**2. Content Can Easily Be Copied**

Content contained in an RSS feed can easily be copied and replicated, regardless of whether you want it to be or not. Few aggregators respect the copyrights of content contained in an RSS feed.

**3. Tracking Subscribers Is Difficult**

It is very difficult to accurately track the number of subscribers who read an RSS feed or the items contained in an RSS feed. This is due in part to the fact that at its heart, RSS is all about achieving the widest syndication possible.

**4. Source Origination Difficult**

It is sometimes difficult to discern the origin of an RSS feed item. When an item is syndicated, the source is not always indicated. The metrics available are not always reflective of the traffic received.

Weigh the pros and cons of implementing an RSS feed as a communication channel, and determine whether the benefits outweigh the risks in your own situation.

**VI. NEED FOR RSS AGGREGATOR**

An aggregator is responsible for the convenience of RSS feeds. The RSS aggregator checks websites for new content automatically. It immediately pulls that content over to your feed so you don't have to go and check each website individually to find new content.

The aggregator even keeps track of what you have and have not read by listing the number of articles or pieces of content for each website you are following that has not been seen. This helps you to more quickly see what's going on within the websites that interest you

**VII. EXISTING SYSTEM**

A general RSS Aggregator displays the content from the RSS feed. Updates from various news feeds can be accessed. The news aggregator will automatically check the RSS feed for new content, allowing the content to be automatically passed from website to website or from website to user. This aggregator displays the news feed in a direct form and this may not be very captivating and informative. A link is provided to be directed to that particular website in order to know more about the updated news content.

**VIII. PROPOSED SYSTEM**

We introduce bookmarks for the users in order to save a particular website or even to save their favorite content from a specific website.

By saving news feeds, we provide the users to check out their favorite content easily. Dictionary

Text-Summarization is provided in such a way that only a summarized version of the online content is provided, such that the reader does not have to go through the entire information. Instead the main necessary data is given out.

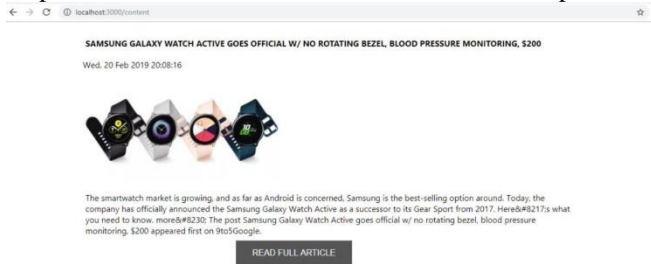
**Scope of Proposed System**

The scope of our project is to provide an easy and efficient way to understand and save the content for our users in an easily accessible manner. Instead of clicking on links to be directed to the main website and read through the entire news article, we provide an interesting approach to check out their online data. With images and titles displayed along with the time a date provided, the user can easily identify when his saved data has been updated. According to the fast moving needs of this generation, news and other required information must also be fast paced and easy to go through in quick time.

**IX. RESULTS**

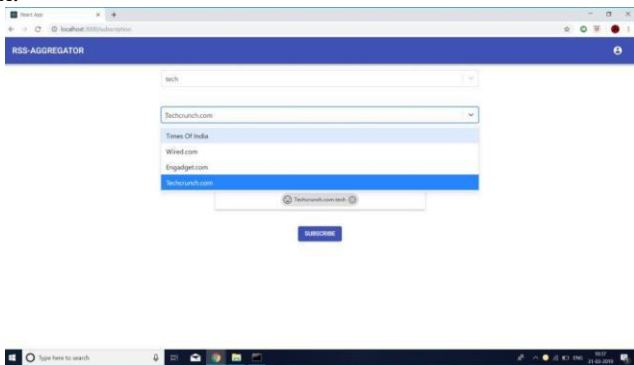
**Article page:**

This page displays the content title, date and time of when it was last updated. A brief introduction of the article is also provided



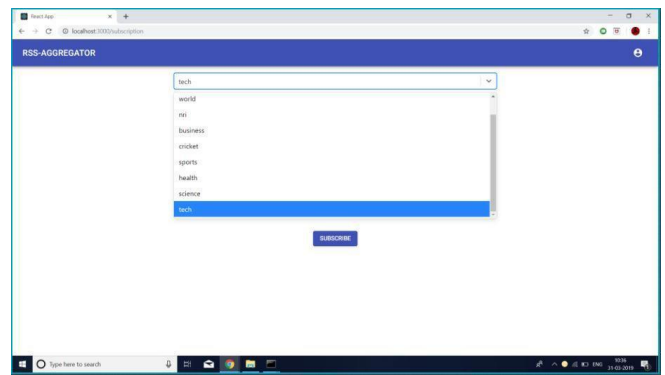
**Selecting websites:**

The various news websites can be selected from the dropdown box.



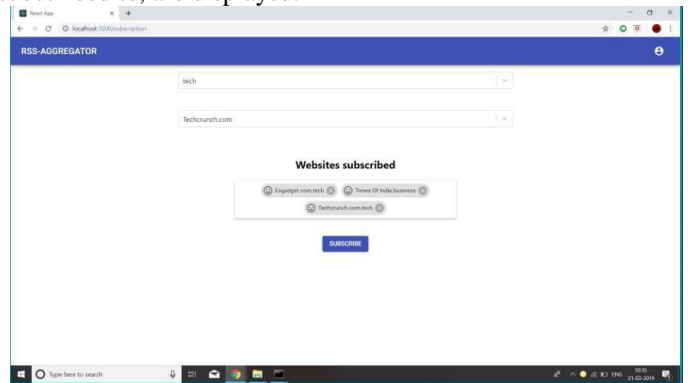
**Selecting genres:**

The user can choose from a variety of genres based on which the news can be filtered out.



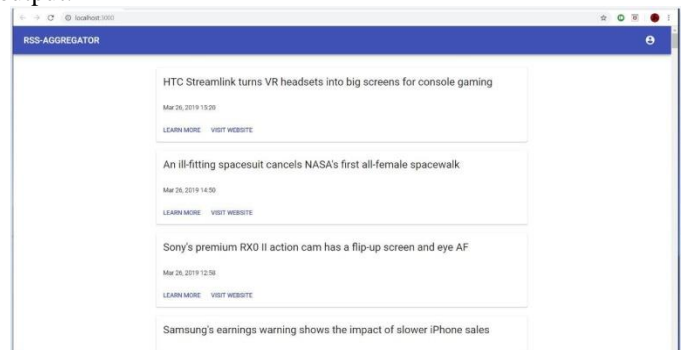
**Websites Subscribed displayed:**

Based on the user's selections, the websites that have been subscribed to, are displayed.



**Output:**

The articles filtered based on the user's requirements are given as output.



**X. CONCLUSION**

Through our project we have tried to provide a platform through which a user can easily save his favorite online content and can be notified whenever updates occur. As we all know that there are many projects that have attempted this topic, we wish ours to be the best one. The purpose of this project is to provide an application that keeps track of what you have and have not read

by listing the number of articles or pieces of content for each website you are following that has not been seen.

The conclusion of the project is to develop an algorithm that will be used to develop an accurate analysis of the user's preferences. RSS Aggregator can be developed using various types of service and in the future we would like to expand to those as well. A database will be developed, which will store all types of necessary features particularly for this purpose. A usable system will be designed, developed and deployed to the public so that they can use this to obtain the results desired.

#### REFERENCES

- [1] O'Reilly. Developing a React Edge , 2013.
- [2] Shelley Powers. Learning Node: Moving to the Server-Side, 2012
- [3] Eric Elliot. Programming JavaScript Applications Robust Web Architecture with Node, HTML5, and Modern JS Libraries, 2013
- [4] Eric Bush. Full-Stack JavaScript Development: Develop, Test and Deploy with MongoDB, Express, Angular and Node on AWS, 2016
- [5] Rubayeet Islam. MongoDB Web Development, 2011
- [6] Jasim Muhammed. Building Cross-Platform Desktop Applications with Electron, 2017
- [7] Mardan,Azat.Pro Express.js.Master Express.js: The Node.js Framework For Your Web Development, 2014
- [8] Valentin Bojinov. RESTful Web API Design with Node.JS , 2016
- [9] Basarat Syed. Beginning Node.js , 2014
- [10] Mario Casciaro. Node.js Design Patterns - Second Edition: Master best practices to build modular and scalable server-side web applications , 2016
- [11] Sandro Pasquali, Kevin Faaborg. Mastering Node.js - Second Edition, 2017
- [12] Ambily K.K. Beginning Web Application Development with Node, 2015
- [13] Pedro Teixeira. Professional Node.js: Building Javascript Based Scalable Software , 2012
- [14] Azat Mardan. Practical Node.js: Building Real-World Scalable Web Apps, 2014
- [15] Guillermo Rauch. Smashing Node.js: JavaScript Everywhere 2nd Edition, 2012
- [16] Ethan Brown. Web Development with Node and Express: Leveraging the JavaScript Stack 1st Edition, 2014
- [17] Jim Wilson. Node.js the Right Way: Practical, Server-Side JavaScript That Scales 1st Edition, 2013
- [18] Andrew Hunt. The Pragmatic Programmer: From Journeyman to Master 1st Edition, 1999