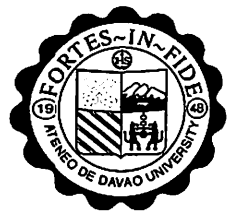

ATENEO DE DAVAO UNIVERSITY
Computer Studies Division

Senior Project

REALLY SIMPLE SYNDICATION (RSS) DIGITAL TECHNOLOGY NEWS LIBRARY INDEXING AND SEARCHING

Group Name:	BGG
Proponents:	AQUINO, HAZEL JANE G. GULFO, SAHARA MAY S. NGO, AILEEN MAY O.
Course:	Bachelor of Science in Information Technology
School Year:	SY 2005 – 2006

REALLY SIMPLE SYNDICATION (RSS)

DIGITAL TECHNOLOGY NEWS LIBRARY

INDEXING AND SEARCHING

An Independent Research

Presented to

The Faculty of the Computer Studies Division

Ateneo de Davao University

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science Major in Information Technology

By

Hazel Jane G. Aquino

Sahara May S. Gulfo

Aileen May O. Ngo

SCHOOL OF ARTS AND SCIENCES

ATENEO DE DAVAO UNIVERSITY

MARCH 2006

TABLE OF CONTENTS

ABSTRACT

In this day and age, people who would like to do some research most often go to the World Wide Web to look for their topics. This is due to the fact they can access numerous information on the web and when they look for a certain topic all they have to do is type their search term, and they can already choose from the results that a search engine would return. The problem with this is that not all the searches being returned are updated. An RSS digital news library, that provides articles on the latest developments about computer related topics of the curriculum for the courses under the computer studies division, is just be the key to accessing up to date information regarding a certain topic. In this paper is a discussion of the methodology that is being employed to incorporate the RSS technology in a digital library.

Keywords:

Digital library, Really Simple Syndication or Rich Site Summary (RSS)

CHAPTER I

INTRODUCTION

1.1 Background of the Study

When searching for topics on the internet or a library, it is often not easy to find the things that a person is really looking for. In a library, for instance, books are often not updated. As a result, a person does not always get the information that he is looking for. In the case of regular search engines, they look for topics based on keywords. If a search term happens to match a single word in an article, the search engine would include that article in the results.

RSS is a new technology that is slowly gaining much attention. The group would like to take advantage of this useful technology to build an RSS digital library that would cater to the different topics or subjects that CS students would take. One noteworthy feature of this library is that all articles are online, thereby eliminating the need to store hundreds, maybe thousands, of documents on a server or computer at the user's end. With RSS, one can make sure that all information a person gets from the digital library is up to date.

1.2 Technology Application Context

The application that the group developed is a digital library that incorporates RSS feeds into its content. All sources for the digital library are online, so the there is no need to place all documents or resources in a server or computer at the user's end. Information about the articles is stored on a

database. When a user visits the site, he can view the feeds currently stored in the database. If the user is looking for a particular feed, he will go to the search engine, types a search term and the articles that match the user's query will appear. The results are RSS feeds that contain the topic/s the user is looking for. The user needs to view the feed from the feed reader in order to properly see its contents.

1.3 Objective of the Study

The project aimed to utilize the RSS technology in order to create a digital news library that would provide news articles on the latest developments or updates in a particular technology. Specifically, this study has:

- Made computer related subjects, which are part of the curriculum for the courses under the computer studies division, available on the Internet through the integration of RSS technology in the digital news library.
- Incorporated the Google Web API in the application to gather RSS feeds in the World Wide Web through the search functionality
- Automatically indexed RSS feeds in the database eliminating the need for a human editor to extract relevant metadata from the document
- Allowed users to search through the database for the topic that they are looking for and to refine their search through the use of search filters

- Archived RSS feeds according to the number of days, months or years as specified by the administrator

1.4 Significance of the Study

This study is significant for all people particularly students under the computer studies division. Through an RSS digital library, they can have access to updated information on developments pertaining to computer technology. Some resources about computer technology in the library are already obsolete. This digital library will provide an environment wherein users can access information easily. Everything is updated because of the RSS technology that is implemented in the digital library.

When researching about certain topics that they will be discussing in class, the CS students would most likely find the digital library helpful since the topics covered in this library would follow the curriculum for the Computer Science, Information Management and Information Technology courses. The application also has a search capability thereby making it easier for users to find the information they are looking for.

1.5 Project Output and Scope

After doing the study on RSS feeds and digital libraries, the group was able to develop a web-based digital library that contained news articles about topics which are covered in the curriculum for courses under the computer studies division and can index and search RSS feeds. This library also has an online RSS feed reader so that a person who visits the site does not have to

download a software in order to interpret the feeds. A database at the back end of the application serves as the storage for all indexed feeds.

Topics covered in the digital library would be limited to those which are part of the curriculum for computer studies courses. The site contains only RSS feeds which were previously indexed by the program and feeds which the user/s added to the database. These feeds are mostly news articles about new and upcoming technologies, developments in computer related subject matter.

This application does not deal with whether or not the content of a feed is congruent with the description that was provided. The program does not dwell on this problem since it focuses on the searching and automatic indexing of RSS feeds through the use of web crawlers and schedulers.

The reader does not check whether or not the URL of the feed can be found or not since this depends on a variety of factors that cannot be accurately determined. It could be that the server on which the feed is hosted is down or it could also be that the feed has not been maintained.

Lastly, the application can only be properly displayed using the Mozilla Firefox browser. This is due to some design components used in the user interface that the Internet Explorer does not support.

1.6 Definition of Terms

1. RSS (Really Simple Syndication or Rich Site Summary) – RSS is an XML format for sharing content among different Web sites such as news items. RSS is usually used to allow a website or computer program to efficiently

gather information from a site to display. It allows a web developer to share the content on his/her site. RSS repackages the web content as a list of data items, to which you can subscribe from a directory of RSS publishers.

2. Digital library - A digital library is a collection of documents in organized electronic form, available on the Internet or on CD-ROM disks. Depending on the specific library, a user may be able to access magazine articles, books, papers, images, sound files, and videos.
3. Extensible Markup Language (XML) – XML is a standard for creating markup languages which describe the structure of data. It is not a fixed set of elements like HTML, but rather, it is like SGML (Standard Generalized Markup Language) in that it is a metalanguage, or a language for describing languages. XML enables authors to define their own tags.
4. Indexing – Indexing is often used to refer to the automatic selection and compilation of 'meaningful' words from a website into a list that can be used by a search system to retrieve pages.
5. Searching – Searching is the process of purposefully trying to look for some object or information, sometimes with the help of a search system or search engine, sometimes using an information retrieval system, sometimes by submitting a formal query, often following some search strategy or plan.
6. RSS Reader – An RSS reader reads RSS feeds at a certain interval and copies RSS feed messages to its database or some kind of file structure.

7. Web crawler – A web crawler is also known as a web spider or robot. It is a program which browses the World Wide Web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.
8. Scheduler - A scheduler provides scheduling policies, which are rules, procedures, or criteria used in making process scheduling decisions.
9. Simple Object Access Protocol (SOAP) - A standard for exchanging XML-based messages over a computer network, normally using HTTP.
10. Web Services Description Language (WSDL) - An XML format published for describing Web services. A WSDL definition describes how to access a web service and what operations it will perform.