Table of Contents

1. Introduction

1.1 Summary

The Funnel and Filter site project will allow a client to register an interest in a particular football club. The site will then search suitable news sites for stories about the nominated football club and will return any finds in the form of an html page of annotated links. It should be possible for the client to identify and thus "filter out" unworthy sites. It should be possible for the user to nominate a site for inclusion in the news sites searched. The site http://www.goonernews.com/ (1) is an example of some of what is required but only supplies Arsenal news stories and doesn't allow unworthy news sites to be filtered out.

Be it in the work place or at home football websites are one of the most predominately accessed website on the web. Football or soccer in some countries is the largest sport in the world. It has the highest level of participants of any sport. Football has a strange effect on people. During a World Cup year it is common for whole countries to cease working to support their national team. Furthermore, level of fan dedication at club level sometimes can beggar belief. Manchester United is estimated to have over 75 million fans all over the world. This in turn means literally thousands of websites, news feeds, newspapers and fan websites will have new stories about Manchester United every day.

This means there is a wealth of information to be gathered about football clubs from all over the internet.

Various people access these sites throughout the day, from the part time fan who might just want to browse the headlines for any stories, to the fanatics who read every story related to their club, constantly checking for news stories and posting on the forums and messages boards. Some even go as far as setting up their own messages boards.

The scope for a commercially viable "filter and funnel" website related to football is obvious. Football club fans rarely want to read about other football clubs or players belonging to other clubs. This ensures a user base theoretically as large as the amount of

fans related to that football club who use the internet. This user base will also have a strong brand loyalty seeing as it is the basis of being a football fan is loyalty.

Rather than dedicate the website to one particular team the premise of the proposed functionality of the website is to gather information related to the users interest. To maximize the user base of the proposed website a process of extraction and encapsulation of presented data is a core component to its functionality.

The premise is simple. The user selects which football club is of interest to them. From this user input the website gathers information related to the user's interest. This will guarantee a very large user base which is the ultimate goal of any website.

Seeing as this website is a proof of concept the scope of teams that the end user will be able to search is limited to English Barclays Premier League. To aid the extraction process of relevant information a list of websites associated with each of the Premier League teams will be held in a database. This will improve search and retrieval times.

The primary source of information will be gathered from RSS feeds. RSS feeds are used by news syndicates to disseminate current information in a standardize fashion via the use of XML and related technologies. This provides the developer a standardized format to extract news from the website using DOM API specific technologies. All major programming languages provide API specific libraries for parsing the common interoperability standard of XML (RSS).

A secondary source of information will be obtained from websites that use JSON as their feed type. The concept is similar to RSS feeds in that the message is sent in a data interchange format. Rather than use XML as its interchange method it uses a concept similar to spread sheets to transfer its information. Websites such as Google use this format.

Relevant data will then be obtained and stored in the database after extraction. This information can then be displayed on the filtering webpage.

1.2 Motivating Factors

There are a few motivating factors for my choosing of this project.

I. Firstly it will broaden my skill set significantly to enable me to better challenge for a job in the work place. The role the internet and websites play in the current computing world is increasingly dominant. As the industry marches indomitably towards the era of cloud computing and centralization of information it is beneficial to direct my research into the related technologies.

II. I have always had an interest in data gathering techniques of the major social networking sites. These sites have many interesting and diverse techniques of information gathering. I hope to integrate some of these methodologies for data retrieval into my website. Having studied two modules related to databases and relational database queries i feel an affinity to database and database design.

III. Lastly I am a fan of football myself and see the worth of developing such a project. As any football fan will tell you the wealth of information pertaining to a club available on the internet is vast. Having a single website which gathers this information could be a very viable and functional product

1.3 Research Objectives

Before I started doing my research I had to figure out what it was exactly I was going researching. So I decided to set myself a couple of goals

I. An Investigation into user requirements
II. Research into database management
III. Was this actually possible
IV. Where and how would I get the news articles from
V. What development language would I use
VI. What server would I use
VII. Prototype development
VIII. Testing

1.3 Research Methodologies

When I was doing my research I didn't really stick to any clear cut plan or method. First thing I did was talk to two web developer friends of mine to see what they thought of the project and could it be done. I also asked them for any advice, hints or tips they could give me. They set me in the direction of first of all finding a development language to choose and then find a suitable back end database manager. I stayed in contact with them throughout the duration of the project.

From there I decided to turn to the internet for more information on trying to solve my goals. The internet contains a vast array of information on every topic and found it easy to find information.

I also decided to hold informal interviews with a couple of my class mates and friends. I held these interviews so that I could find out more information about what other people expected from a football orientated news website rather then just what I thought was best. Just because I thought something was good doesn't mean other people would like it. From holding these interviews I also got a vibe that a lot of people would have an interest in my fyp and would be willing to used it once it was up and running.

I also decided to email other football news sites to see what kind of methods they used when searching for their news stories and also in the various techniques they used.

1.4 Overview of report

In the first section of my report I want to give you a description of what my project is about and how I went about researching it.

The second chapter outline what I found out about various architectures, languages, technologies and other invaluable information I learned during my research of the project.

The third section tells how I went about finding out exactly what I and the users wanted the system to do as well as explaining how I went about and did it.

In the fourth section I give an evaluation of the application as well as describe the different methods I used for testing it.

The fifth section contains my conclusion of the project as a whole the things iv learned and the things id change if undertaking such a project again.

2. Research

2.1 General Research

When I decided on this project I had various ideas of how to proceed with it. My initial thought were how and where do I abstract the stories from the website? Do I use a web scraping techniques to scrape the stories off the websites and consolidate them into my database? Do I manually trawl football websites searching for news stories and link to these? Do I use RSS feeds to gather information for use in the website's database? I decided that it was best to finalize this decision first before I could progress any further.

In retrospect the second option was disadvantageous as first of all I had no interest in trawling through the web looking for news stories on various different clubs. As previously stated in the introduction I wanted the website to be highly automated with practically no human interaction. This view was also shared by my supervisor. The data mining technique came down to RSS feeds/JSON or data scraping. Next question posed was what exactly is data scraping and how do RSS feeds/JSON work?

2.2 Interviews

From the interviews with my friends and colleges I found out what they expected from the site. They expected it to be simple and easy to use while also providing the information that they were looking for with ease of access.

2.3 Data scraping

Data scraping is the technique of using a piece of software to extract information from websites. A data scraping "bot" searches the net for sites that match the search parameters. When it finds a web page, it scrapes the data required and stores it in the database of the website. The quality of the information retrieved depends on the sophistication of the bot's data retrieval algorithm. This was heading towards the right direction for what i required. However, one thing was still discouraging about this technique. The issue was, could I implement the data scraping bot to harvest data from the differently formatted web pages. Another consequence relating to the use of bots was,

depending on the terms and conditions of some websites, these bots can be deemed illegal and in turn blocked. These implantation issues meant that this idea was put on hold until further research of other data retrieval techniques. (1)

2.4 RSS Feeds

Initial research on RSS feed was a common student option – Wikipedia. It stated: "RSS is a family of web feed formats used to publish frequently update works - such as blog entries, news headlines, audio and video in a standardized format" This resolved the two main concerns that I had from data scraping. Firstly there was the element of “standardization” to all the RSS feeds. This meant all the information delivery would be structured in a similar manner and secondly it was completely legal. RSS feeds are sent in an XML format and are easily manipulated.

Because of this standardised format it will be significantly easier for my script to search through and extract the relevant information. Using RSS feeds also means that I can store the information in a database and query it when a user accesses my website. This is more economical than conducting a live web search or data scraping activity when the website is accessed. Using RSS feeds also means that I check for recent or updated news stories. It is then up to the script whether or not the story is added to the database. (2)

2.5 Emails

While researching these topics it occurred to me that when implementing the search i needed to define broader parameters. It wouldn't just be enough to search for the clubs names. i would have to include various nicknames, team players, stadium and other associated terms with individual clubs. This was to ensure that i completed a comprehensive search. I got the idea to email sites similar to mine and ask them for information on how their sites functioned, what they did to get their stories and see if they would mind giving me any ideas to help me with mine.

Two sites that quite similar to mine in their contents are “Goonernews” (3) and “QPRReport” (4). Goonernews is a news site based solely on Arsenal. QPRReport is a blog

site that contains the latest stories for QPR. Goonernews never returned an email to me. However Mike form QPRReport kindly replied to my cries for assistance and direction. I understood from his email it was just him putting in a lot of work himself and doing manual searching for a lot of the stories. He did say that he uses a lot of the same websites to search for stories. I also emailed one of the main football news sites that I use, Football365 (5). I was going to email another site that I use, Teamtalk (6). However on obtaining contact information for both these sites i realised that they are both owned by BskyB. I emailed BskyB directly and similar to goonernews, never received a reply.

I did not let this lack of replies dishearten me. However one other thing I now had to consider was how many clubs I had to limit my website to. As my research had shown there was no “God like” all-encompassing website, if you will, that contained all the news stories for every club. Instead I would have to use various sites to ensure that I get a wide variety of topics covered for each club.

2.6 Websites

Obviously each club has their official site. I would need to use these as well as the official Premiership site. There is also the general football websites like football365 and teamtalk as well as the newspaper websites like the Sun, the Times, and the Independent etc. I have also decided to include the Fan websites such as Goonernews and QPRreport in case they have stories that I don't get.

Using numerous websites to find stories has both it pros and cons. Although it means that I will get a more comprehensive search of the web for stories it also brings with it the trouble of duplicate stories.

I am not entirely certain yet on how I am going to deal with duplicate stories. I could use a “first come first serve” basis. This is where the first source of a particular story that I get will be the only copy. Another option is to show all the stories from the different sources and allow the user to pick which story they would like to read themselves.

From my research on the Internet I have not been able to find another website that has the functionality of mine. Most of the football story sites and news sites allow you to pick a topic that you would like to read about. These sites however, only give you their point of view. The fan websites only give you news for the clubs that they are dedicated to whereas my site will allow you to choose your club. It will also give you a list of websites from which you can pick from to read stories while not showing the user stories from what they deem to be unworthy sites.

2.7 Implementation Technologies

2.7.1 PHP

Php is a server side open source scripting language designed specifically for creating dynamic web pages. Server side means that all of the work is done on the web server rather than on the client machine,. Open source means that it free to download and use. Scripting languages are used to perform the same tasks repeatedly. Php is rich in features that are there to make web design and programming easier. At the last count php was in use over 20 million domains(ie over 20million websites used php as their programming language.) This figure is growing which shows that php is doing something right.

Server side scripting allows for the creation of dynamic websites. Server side scripting has the potential to allow for a high level of customization in response to users requests. When a script is executed it is processed on the server and only the result is returned to the web browser. The script itself does not get included in the result so it is transparent to the user. This is an excellent hacking prevention as any would be hacker cannot see the code that was executed which means sensitive information like database login or structure details never leave the server and are hidden from the user at all times.

Being open source means that it is free to download and use. This means that if you were a company looking to create a website php would be an option because free stuff is very cost effective. This cost effectiveness is also applicable to students looking to create websites as students and poor finances go hand in hand. The open source community is very different from that of the corporate community such as Microsoft and adobe.

Because it is open source there is no big corporate website you can log onto to find solutions for any problems you might have. Instead there are many forums out there that contain a host of solutions to problems that other users have had and how they resolved those problems. From using these forums I felt that any problems I had it was easy to find a solution to as you got a step by a step guide through the solution from somebody who had the same problem rather than from a developer trying to solve problems for other people.

Php works well on several web servers but it works best with the Apache web server. Like php, Apache is also free, open source and popular as well as being considered very stable and one of the best servers out there.

When using php it is quite common to use mysql as the database application for your website. Together these two are commonly referred to as the dynamic duo. When using MySql with php, php provides the application part while MySql provides the database management. Php has built in processes to deal with MySql so all you have to do is give php your login details for the database and it worries about creating the connection.

. The PHP Data Objects (PDO) extension will provide a consistent interface for accessing databases. Each database driver that implements the PDO interface can expose database specific features as regular extension functions. (7)

2.7.2 MySql

MySQL is a relational database management system(RDBMS) providing multi-user access to a number of databases. All major programming languages with language specific APIs include Libraries for accessing MySQL databases. In addition, an ODBC interface called MyODBC allows additional programming languages that support the ODBC interface to communicate with a MySQL database. The use of MySql is preferred over propierity DBMS as it is a open source product without the need for licensing. MySql database is stored outside the scope of the website implementation thereby increasing security, unlike other free technologies such as Microsoft Access or SQLLite which store the databases in a local file. (8) mysql

2.7.3 Apache

The Apache web server is a computer program that delivers (serves) content using the Hypertext Transfer Protocol. Apache supports a variety of features, many implemented as compiled modules which extend the core functionality. These can range from server-side programming language support to authentication schemes. Some common language interfaces support Perl, Python, TCU, and PHP. (9) As of January 2010 it holds 53.84 percent of the market share for web servers. - http://news.netcraft.com/archives/2010/01/

2.7.4 Curl

A free and easy to use client side URL transfer library, supporting many protocols such as FTP, FTPS, HTTP, HTTPS, SCP, SFTP, TFTP, TELNET, DICT, FILE, LDAP and LDAPS. libcurl supports HTTPS certificates, HTTP POST, HTTP PUT, FTP uploading, kerberos, HTTP form based upload, proxies, cookies, user + password authentication, file transfer resume, http proxy tunneling and more. PHP provides built in support for Curl allowing for easy HTTP interaction between HOST application and the news syndicates. Curl implementation has substantial performance increase over standard PHP API functions making it an ideal choice for application / news syndicate interaction. (10)

2.7.5 RSS

RSS (most commonly expanded as "Really Simple Syndication") is a family of web feed formats used to publish frequently updated works. A standardized XML file format allows the information to be published once and viewed by many different programs (2)

2.7.6 JSON

JSON, short for JavaScript Object Notation, is a lightweight computer data interchange format. It is a text-based, human-readable format for representing simple data structures and associative arrays (called objects).

The JSON format is often used for serialization and transmitting structured data over a network connection. Its main application is in Ajax web application programming, where it

serves as an alternative to the XML format. AS some news syndicates use JSON as an alternative to RSS such as Google news it is a required technology for the application. (1)

2.7.7 JavaScript

JavaScript is an object oriented scripting language used to enable programmatic access to objects within both the client application and other applications. It is primarily used in the form of client-side JavaScript, implemented as an integrated component of the web browser, allowing the development of enhanced user interfaces and dynamic websites. JavaScript is a dialect of the ECMAScript standard and is characterized as a dynamic,weakly typed, prototype-based language with first-class functions. JavaScript was influenced by many languages and was designed to look like Java, but to be easier for non-programmers to work with. Javascript supports the interaction with the DOM API and related technologies that support this functionality such as XML, HTML and JSON. (11)

2.7.8 Document Object Model

The Document Object Model (DOM) is a cross-platform and language-independent convention for representing and interacting with objects in HTML, XHTML and XML documents. Aspects of the DOM (such as its "Elements") may be addressed and manipulated within the syntax of the programming language in use. (12)

2.7.9 Ajax

Asynchronous JavaScript and XMLis a group of interrelated web development techniques used on the client-side to create interactive web applications. With Ajax, web applications can retrieve data from the server asynchronously in the background without interfering with the display and behaviour of the existing page. The use of Ajax techniques has led to an increase in interactive or dynamic interfaces on web pages. Data is usually retrieved using the XMLHttpRequest object. (13)

2.7.10 Cascading Style Sheets

Cascading Style Sheets (CSS) is a style sheet language used to describe the presentation semantics (that is, the look and formatting) of a document written in a markup language.

Its most common application is to style web pages written in HTML and XHTML, but the language can be applied to any kind of XML document, including SVG and XUL.

CSS is designed primarily to enable the separation of document content (written in HTML or a similar markup language) from document presentation, including elements such as the layout, colors, and fonts. This separation can improve content accessibility, provide more flexibility and control in the specification of presentation characteristics, enable multiple pages to share formatting, and reduce complexity and repetition in the structural content (such as by allowing for table less web design). (14)

2.7.11 HTML

HTML defines the structure of all web pages. The Hyper Text Markup Language defines the structure and layout of web pages. A predefined document type definition defines the semantics of the elements and their attributes. It defines the structure of a compound document including such elements images, audio, java applets and general text in form of paragraphs and headings.

I decided to go with these technologies as they are part of the fastest growing open source enterprise software stacks the LAMP stack. (Figure 9)

3. Design and Implementation

3.1 Lifecycle and Process

When developing software there are various different types of approaches that can be taken, these approaches can be referred to as “Software Development Process Models”. Each different software process follows its own particular life cycle. The process that I chose during the development of the website was the “Waterfall Model”. in the waterfall process there are five phases.

I. Requirement Analysis
II. System Design
III. Implementation
IV. Verification
V. Maintenance

Each of the 5 phases cascade into the one below and the next phase is only begun when the previous phase has been completes and fully signed off on. After moving onto the next phase in the model it is not possible to go back ie. Once the requirements analysis has been signed off on they are then considered to be set in stone.

As you go through the waterfall process it is given that the more time spent on the earlier phases like requirements and design is good practice. It has been shown that the earlier a bug is realized with the development process the more inexpensive it is to fix. It is easier to fix a bug realized at design stage then it is to fix at the implementation stage when other components have been written to interact with the component that carries the error as these components may then have to be refactored once the error is fixed.

Compared to that of other processes such as the Rational Unified Process (RUP) the waterfall is considered to be quite simplistic and structured in its approach to software processing. In its linear format the waterfall has easily understandable and explainable stages.

3.2 Requirements Analysis

Before any software development can be started it is important to have a list of well defined and thought out requirements. You cannot develop if you do not know what it is exactly you want to develop. I decided on a couple of my basic requirements from reading the project description. While I was interviewing my fellow students and colleges I also developed use cases. This allowed me to come up with more requirements that they would like to have been integrated into the system. From talking to web developers I was also able to better define some of my requirements. My main requirement was to obtain the latest relevant information for each club while also trying to ensure that there was no duplication in the news articles which would appear on the site.

3.2.1 Actors

An actor is "an external entity of any form that interacts with the system. Actors may be physical devices, humans or information systems" (Bennett, McRobb, & Farmer, 2006)

Below I have comprised a table to list and describe the actors that are involved in the involved in the application

Actor	Description
General User	The general user is the typical football fan who I hope will access my website. They will be able to select their favorite club to get the latest up to date news stories
Admin	For nearly every website almost some input from an administrative person is required at some stage. This person will be me.
News Sites	These are the other websites from which I will obtain the rss feeds for the news articles that I will upload to my website for the general user to read.

3.2.2 Use Cases

To help me to develop my requirements I had to come up with a set of use cases that I thought a typical user would go through when using my website. A use case is a description of a set of sequences of actions, including variants, that a system performs to yield an observable result" (Booch, Rumbaugh, & Jacobson, 1999).

A use case is a description of a task that an actor should be able to perform from start to finish, while also showing the response of the system while that task is undertaken. A full list of use cases can be seen in the appendix.

3.2.3 Non Functional Requirements

A non functional requirement is a requirement that specifies the system qualities that must exist. "The nonfunctional requirements should provide specific measurements that the software must meet. The maximum number of seconds it must take to perform a task, the maximum size of a database on disk, the number of hours per day a system must be available, and the number of concurrent users supported are examples of requirements that the software must implement but do not change its behavior" (Applies Software Project Management – O'Reilly) While researching these type of requirements I came up with a list of the following that I believe are relevant to my project.

3.2.3.1 Efficiency and Performance

When developing any system efficiency is one of the key requirements that must always be considered. In relation to my website there is no point in the user selecting their favorite club and having to wait a couple of seconds for the page to load. Php and MySql were both designed with speed in mind so providing I get my procedure right in the code and have my database organized correctly I would hope that the results should be returned almost instantaneously. Different people have different bandwidth allowances and also depending on the time of day the line could be busy so I have to keep the information returned to the user as efficient as possible

3.2.3.2 Availability

Availability refers to the uptime of a system ie how long it is available for during a given time period. Although I cannot directly influence this I had to take it into account when selecting my web server. There is no point of selecting a server that is known for constantly being down or crashing. Also when writing my code for the website I have to try make sure the code itself does not cause the site to crash.

3.2.3.3 Human computer interaction

Although not an obvious one it is very important when creating something that well be made available to a large target market that the users can feel confident when interacting with it. There is no point in doing all of the hard work behind the scenes only to return a result to the user that does not look nice, usable and readable. Especially when it comes to websites design is very important that the finish product looks the part. As I am doing a football website I tried to make the whole website as football themed as possible while also keeping it simple. I also have to make sure that website is compatible across multiple browsers. Not everybody is going to be using the same browser.

3.2.3.4 Security

With any website security should always be a key issue. Although the web site does not have a user login that is one security issue I do not have to contend with. However I do have to make sure that any user input is encapsulated within a "`mysql_real_escape_string`" function or some other php security derivative. This is prevent any unauthorized access to my database by any would be hacker.

Selecting php as stated in the implementation technologies means that all of the process information is kept on the server and only the result is returned to the user so they cannot see any of the code. This in itself is a security feature.

3.3 System Architecture

My system is going to be bas on the three tier architecture. This is called client server architecture. It gets the name three tiered because it is based on three different levels. The presentation tier, the application tier and the data tier

3.3.1 Presentation tier

The presentation tier is the part that the user sees. In the case of my project this is going to be the webpage themselves. The user only interacts directly with this tier,

3.3.2 Application tier

Also known as the logic tier or the data access tier. This tier contains the php code that will be used to execute a request from the user submitted through the presentation tier. It controls the applications functionality and does all the work from turning the user request into a request that can be read by the database in the tier below while also transforming the results from the request given by the data tier into presentable data for the presentation tier.

3.3.3 Data tier

This tier contains that database for the application. This is where all of the information is stored. Every news story within each rss feed will be stored here to allow easy retrieval when required.

In the three tiered architecture each tier only interacts with the tier(s) that is direct above and/or below it. Using this type of architecture also allows for easy modifiability and good flexibility.

3.4 Analysis and Design

After extensive research I decided that the rss road was the most efficient road to go down. In rss feeds each topic must be put within its own “item” tag. And within each item tag there must also be a title tag, description tag and a link tag. Every other tag that you may see in rss feeds is optional. However these tags gave me enough information that I needed. I decided that the most efficient way to find my stories was to split the feed up

into its individual items(Figure 5) and from these items extract the information I needed. From each feed I extract the title description and link to the article. After this I then create a loop which cycles through an array populated with the club id of each club(Figure 6). I also use this club id as the key when querying the database. All information related to any club can be got from their key.

Using this id I query the database to find all of the relevant information for an individual club(Figure 7). I then compare this information to that of the information given in the description of the story from the feed. If the matches found meet certain criteria i.e. the club name and a player from that club are found then that story is checked is see if it is already in the database. I use the link related to that story and check if it exists in the database already. If it does not it is added to the database with a reference also put in the club_news table to reference that story to that club. The loop then moves onto the next club and checks the same description against the relevant information on that club. Again if it finds a match that matches the criteria a reference is added in the club_news table for that club to that same news story. Once a match is found that meets the criteria it moves onto the next club until the all of the clubs have been cycled through. Once all of the clubs are finished it moves onto the next item in the feed and does the same again. Once all of the items in a feed have been processed I then moves onto the next feed and starts again on that feed.

That is how I find and insert the stories into the database. All that's left is for the user to come along and select their club. A sql query is sent to the database selecting the news stories relating to that club using the club id as a key and the results are printed to screen using php, html and a css style sheet.

One of the other scripts that I have written that runs in the background is a script that automatically goes and fetches the official list of premier league clubs and populate them to the clubs table within the database. The old list of clubs is deleted so as to avoid any ambiguities Once this is done another script is then run using those clubs it goes to the premier league site again and populates the database this time with a list of players

(Figure 8) for each individual club in the club_players table. This script is particularly useful as it means that I do not have to manually enter their names. As well as being efficient it also removes the chance of human error when performing this task.

4. Evaluation and Testing

"Software Testing is the process of executing a program or a piece of software with the intention of finding errors"

Myers, G. J. (1979). The art of software testing. New York: Wiley.

The site was put live a few days previous to demo day when I got a couple of my friends mainly the ones I had interviewed to go to the site and tell me what they thought of it. As there is very little user input into the site(all they can do is click) there was not mush testing from the user required. The only thing they could really inform, me on was whether or not the stories presented to them matched their chosen clubs. This is how I checked the accuracy of my criteria for linking stories to clubs.

From this test I also got some feedback on the look and feel of the site. Everybody who tested it was happy with it with the general consensus being that the football theme was a good idea but it also didn't get in the way of the main point of the site. I was quite happy with this feedback.

Before I put the website live I looked at various different style in which I could present the news. The first option (Figure 1) that I considered as my home page however I found various faults with this page. One thing I wanted to make sure I did was put the crest of each of the clubs on the homepage. As any football fan knows it is easier to pick out their teams crest then the team name, Given the green background the crests would not have been very visible and nice on the eye. I also felt that this page carried too much detail for what I was looking for I also found this problem with my second prototype(Figure 2). My third and final prototype was the one I decided to use. Although very simplistic it does what I want it to do. It is easy on the eye while also clearly showing the club crests cleanly in both the homepage(Figure 4) and the results pages(Figure 3). I am very happy with this selection for my website.

As I know had the website up and running I was able to test it. One of the main problems I had was the sorting of the stories. I was able to edit my regular expressions for the matching criteria but unfortunately a couple of missed placed stories still got through.

Another issue that I saw coming with this during my requirements was what to do with the same stories from different web sites, ie Sir Alex Ferguson quits in the morning, that story will be all over every soccer website. After careful consideration I decided it wasn't up to me which site got to put their story in their first it was up to the user which one they wanted to read. In my original spec I had planned to allow for the user to pick and choose from the various sources that were available to tem however due to time constraints this was not applicable.

4.1 V-Model testing

For my testing during development I used the V-Model. In the V-Modelthere are 4 different stages.

4.1.1 Unit Testing

This is the testing of individual segments of code before they are inserted into the overall project. This testing is extremely useful as usually any errors found from this type of testing are easily managed as the developer is only looking at a segment of the code rather than having to trawl through the whole application.

4.1.2 Integration Testing

Integration testing takes place when segments of the code get tested together, this is used to test the coupling and cohesion between the classes. This testing is usually used to expose faults within the interface of an application.

4.1.3 System Testing

After integration testing has been completed the next step is system testing. This is the testing of the system against the system specifications. The system specification is realized from the use cases, what the system was supposed to do compared to what the system

actually does. Lack of detail in the requirements can sometimes lead to ambiguities between the system designer and the system developer.

4.1.4 User Acceptance Testing

This type of testing is used by the consumer to decide whether or not to accept the system. When testing this, the client looks at many different aspects. Does it meet performance requirements? Does it meet functional requirements, It is easy to use, Can it be integrated safely into existing system. These are only some of the things a client must consider. There is always the financial consideration that has to be taken into account however like in my project using open source software can help keep the cost down.

5. Evaluation and Conclusion

Overall I am happy enough with the project. At the beginning my knowledge of MySql was little and php even less however after completing this project I now feel quite confident in dealing with both. When I stated this project I like many of my fellow students was worried that come demo day I would have nothing to show however this was not the case and once I got my head around the basics of programming in php and interacting with the MySql database the project just seemed to get easier and that worry lifted.

Unfortunately due to time constraints I was not able to implement all of my aims for this project most notably the client being able to filter out unworthy news feeds; however I do feel that if given a bit more time I would be able to implement this. It is something I like to think that after the exams I would look at finishing.

Would I do anything different if doing it again?

Yes there is one thing that I would change and that is the way I developed it. When programming the application I used the method I was thought in first year Imperative procedural programming. Although this is my preferential method of programming for a project like the one I just undertook it is not the most efficient method, unfortunately during all of my research I forgot to compare it to the object orientated(OO) paradigm. If I had I would have noticed that this was the better programming method in which to do this project.

Had I done that research before I started coding would I have chosen OO?

That I do not know as I said in the previous paragraph I was more comfortable with procedural personally and taking into account that this was for my final year project and I already lack knowledge in both php and MySql I am happy that I chose procedural as at least I was confident in some part of the project from the beginning. However if I was to start the project again tomorrow and it was just for fun I think I would try to use the OO paradigm but this time I also feel more confident in my php and MySql ability.

All in all I am happy with the project ht main goal of the website was to return a list of news stories to the user related to their club choice. I also learned two valuable new skills in php and MySql which is the basis for any good dynamic website. After undertaking this project I feel that a career in web development is a possible option for me in the future as well as being one that I would be interested in exploring.

Bibliography

1. Web scraping. *wikipedia.* [Online] http://en.wikipedia.org/wiki/Web_scraping.

2. RSS. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Rss.

3. *Goonernews.* [Online] http://www.goonernews.com/.

4. *QPRReport.* [Online] http://qprreport.blogspot.com/.

5. *Football365.* [Online] http://www.football365.com/.

6. *Teamtalk.* [Online] http://www.teamtalk.com/.

7. *Php.net.* [Online] http://www.php.net/manual/en/preface.php.

8. *MySQL.* [Online] http://www.mysql.com/.

9. Apache. *wikipedia.* [Online] http://en.wikipedia.org/wiki/Apache_HTTP_Server.

10. Curl. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Curl_%28programming_language%29.

11. Json. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Json.

12. DOM. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Document_object_model.

13. Ajax. *Wkikpedia.* [Online] http://en.wikipedia.org/wiki/Ajax_%28programming%29.

14. Cascading Style sheets. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Css.

Appendix

Figure 1

Figure 2

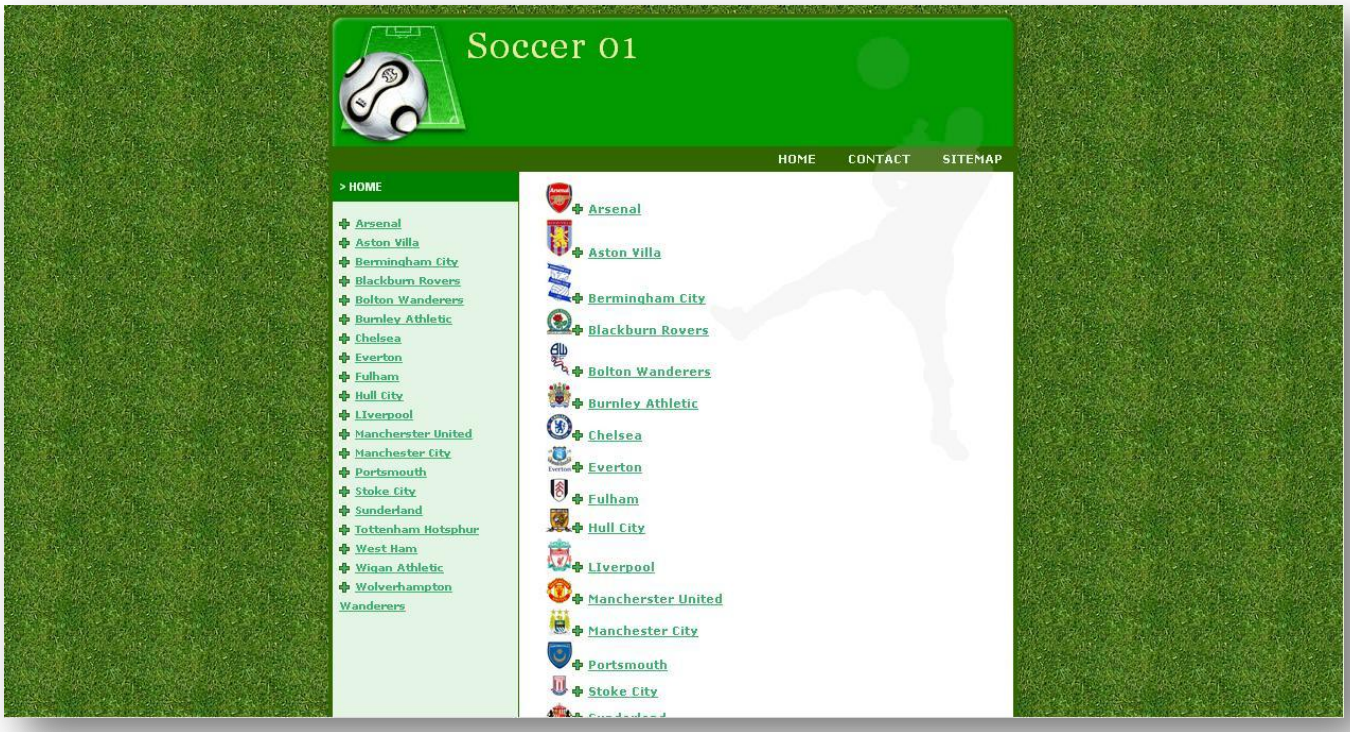

Figure 4

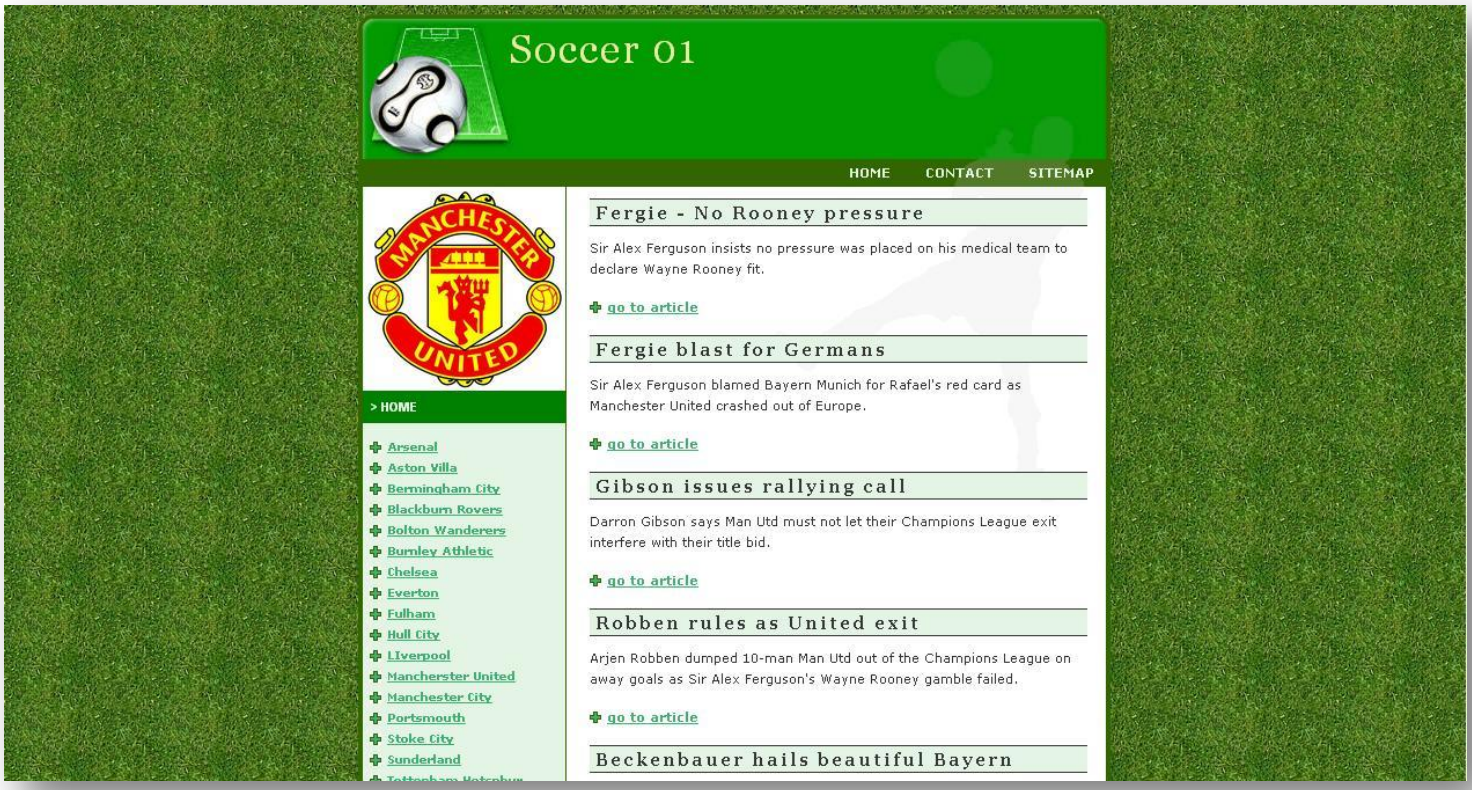

Figure 3

In this section I am going to submit a couple of code fragments and explain what they are doing and why I chose that particular method for doing it that way.

FeedReader.php

This script is used to take in the rss feeds from the various news sites, extract the relevant information and store it in the database with its relevant details

```
$curl_handle=curl_init();
curl_setopt($curl_handle,CURLOPT_URL,$feed);
curl_setopt($curl_handle,CURLOPT_CONNECTTIMEOUT,2);
curl_setopt($curl_handle,CURLOPT_RETURNTRANSFER,1);
$buffer = curl_exec($curl_handle);
curl_close($curl_handle);
$xml = new SimpleXMLElement($buffer);
$result = $xml->xpath('//item');
```

In php the "$" symbol is used to declare a variable. So in the first line shown a variable called "`curl_handle`" this is assigned to the "`curl_init` "method donated by the open close brackets. The "curl_init" method tells php to load the curl module. Once loaded on the next line I use the "`curl_setopt`" method and pass the reference to "`curl_handle` "and I set the constant "`CURLOPT_URL`" to the website from which a particular feed comes from. On the next line again using the "`curl_setopt`" method I again pass in the reference to "`curl_handle`" the second parameter this time is the "`CURLOPT_CONNECTTIMEOUT`" this tells curl that after 2 seconds that if it couldn't fetch the page drop the connection. This feature is there in case there is a problem with the feed curl does not keep looking for the same feed that might not be there anymore. The next line uses the same method but with the constant "`CURLOPT_RETURNTRANSFER`" and I set it to one. This parameter returns a true or false value. By setting it to one I tell it that I want it to return something whereas setting it to zero curl would not return anything from the page it would just make the request. On the following line I create a variable called "buffer". This variable is used to store the result from the curl query i.e. the rss feed. And then I execute the query. The next line just tells curl that it can close the connection to the webpage. In the next line a variable called "xml" is created. This variable takes the result of the buffer variable being passed through "`SimpleXMLElement`" method which takes the

result from the curl query and converts it to xml. I then use the "xpath" expression to select each "item" from the feed. The double forward slash preceding item notates that you want to select more than one

Figure 5

FeedReader.php

```
$clubnums = array(1,2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
17, 18, 19, 20);
```

Here a variable of array type called "clubnums" is declared and populated with each individual club id. The club number in this array is a reference to that of the club id in the database.

Figure 6

FeedReader.php

```
$info = mysql_query("select Clubs.Club_ID, Club_Name, Manager_Name,
first, last, Stadium_Name from Clubs, player, Club_Manager,
Club_Stadium where Clubs.Club_ID = player.Club_ID and Clubs.club_id =
Club_Manager.Club and Clubs.Club_ID = Club_Stadium.Club and
Clubs.Club_id =$clubid",$connection);

while ($row = mysql_fetch_assoc($info))
```

This fragment contains a MySql query it querys the database for a "Clubs.Club_ID, Club_Name, Manager_Name, first, last, Stadium_Name" from various tables using the variable "clubid" as the key. This query is the query used to find all of the relevant data for a particular club.

Figure 7

Loaddatabase.php

This script is used to populate my database with a list of players for each club

```
$link =
"http://www.premierleague.com/dynamicxml/stats/2009/PlayerIndexSquadList/Squa
dList_1_2009_";
$teamNumbers =
array(391,392,579,578,580,395,398,593,598,402,404,403,605,409,410,638,584,408
,618,633);
```

So a variable called “link” is declared and given the value of a webpage. This webpage when concatenated with a number from the”teamNumbers” array displays a list of the players officially registered to each team.

Figure 8

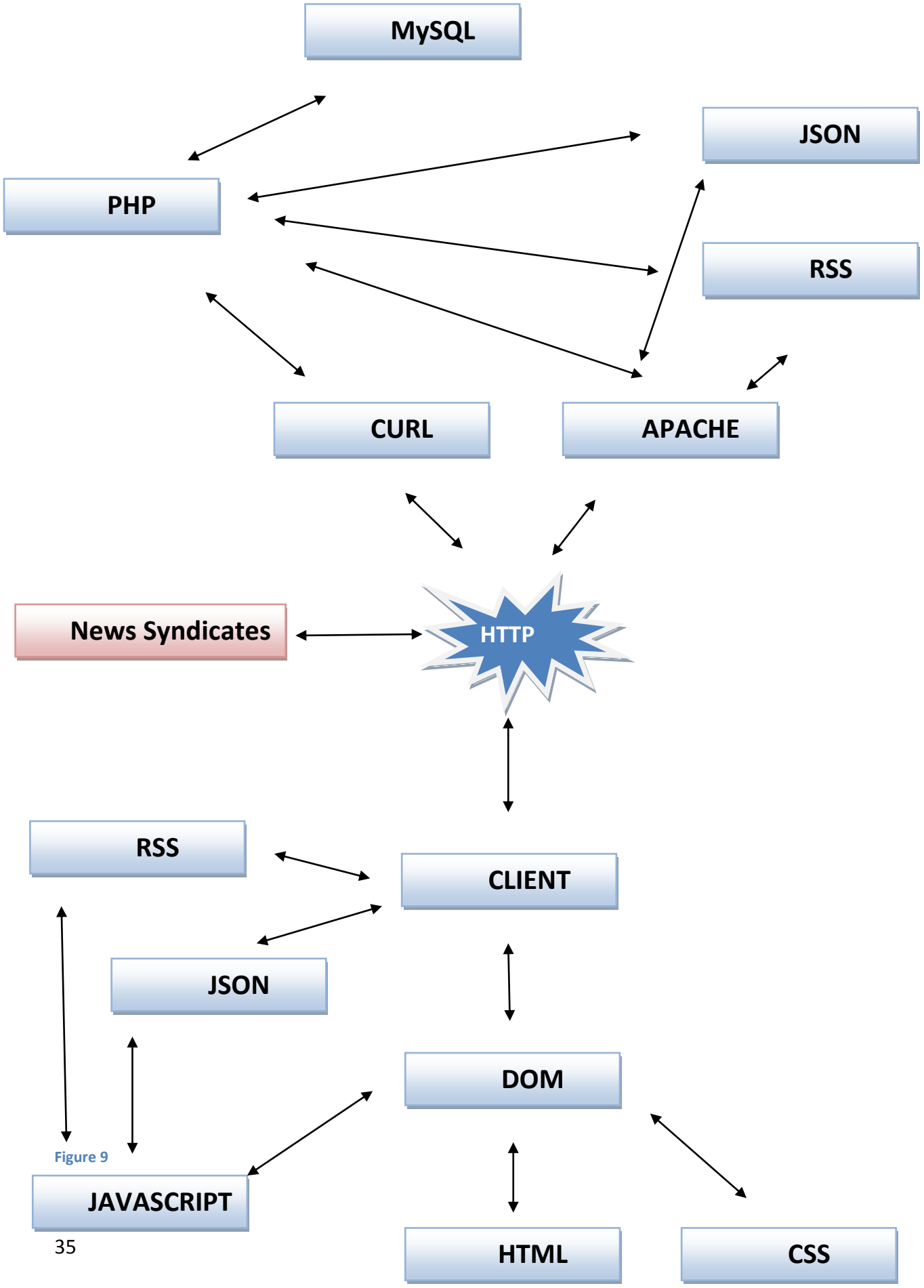
MySQL
JSON
PHP
RSS
CURL
APACHE
HTTP
News Syndicates
RSS
CLIENT
JSON
DOM
JAVASCRIPT
HTML
CSS

Figure 9