



SCHOOL OF  
COMPUTER SCIENCE

Information  
Management Group

HCW— Web Evolution Technical Report 1, September 2008

## Web Evolution: Method and Materials

**Alex Q. Chen and Simon Harper**

Human Centred Web Lab  
School of Computer Science  
University of Manchester  
UK

The World Wide Web (Web) is a heterogeneous environment that is in constant evolutionary change. This includes technological changes, the management of data structures used to present the Web content, and guidelines. A lag was noticed between the time these standards and recommendations were introduced to when they were adopted by the developers. This causes a disconnection between the actual user experience, and what was expected by the technology stake-holders. In this study, we investigate the relationship that surrounds these issues, especially those involving the Web user interface. A trend was noticed that new standards and recommendations get adopted faster by the top websites than the random websites. The top websites on average get adopted one year faster than the random websites for a major (X)HTML standards, while it will take on average two years for a graphical format to get adopted. A dip in JavaScript usage was noticed for the past year (2007-2008), although a continuous increase in AJAX usage was observed, and a growth was predicted to continue for CSS. After ten years < 10% of the websites conform to the WCAG. By understanding these evolutionary trends we can inform and predict Web development into the future.

HCW

Human Centred Web

## Web Evolution

The Web Evolution project investigates the reason behind the lag between when the time standards and recommendations are introduced till when they are adopted by developers. Focused on the relationship that surrounds these issues, this project will identify the trends and provide us with graphs demonstrating how the Web has been evolving over the past ten years. By understanding these evolutionary trends we can inform and predict Web development into the future. <http://hgw.cs.manchester.ac.uk/research/>.

### Web Evolution Reports

This report is in the series of HCW Web Evolution technical reports. Other reports in this series may be found in our data repository, at <http://hgw-eprints.cs.man.ac.uk/view/subjects/web-evolution.html>. Reports from other Human Centred Web projects are also available at <http://hgw-eprints.cs.manchester.ac.uk/>.

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Synopsis . . . . .	2
<b>2</b>	<b>Purpose and Research Questions</b>	<b>3</b>
<b>3</b>	<b>Experimental Data Collection</b>	<b>4</b>
3.1	Methodology for Selecting Websites . . . . .	5
3.2	Capturing the Webpages . . . . .	6
3.3	Web Mining . . . . .	7
3.3.1	Types of HTML Standard Detection . . . . .	7
3.3.2	WCAG 1.0 Conformance Detection . . . . .	8
3.3.3	Graphical Format Usage Detection . . . . .	10
3.3.4	Client-side Scripting Usage Detection . . . . .	12
3.3.5	Cascading Style Sheets Usage Detection . . . . .	12
3.3.6	AJAX Usage Detection . . . . .	12
3.4	Overall Process . . . . .	13
3.5	Issues Relating to Captured Data . . . . .	14
3.6	Conclusion . . . . .	15
<b>4</b>	<b>Results and Discussion</b>	<b>15</b>
4.1	HTML Standards . . . . .	15
4.2	WCAG 1.0 Conformance . . . . .	21
4.3	Graphical Formats Results . . . . .	24
4.4	Client-side Scripting Results . . . . .	30
4.5	Cascading Style Sheets . . . . .	33
4.6	AJAX . . . . .	34
4.7	Further Analysis . . . . .	35
4.8	Analysis Overview . . . . .	38
<b>A</b>	<b>Data Corpus</b>	<b>41</b>
A.1	Data Corpus . . . . .	41

---

**Human Centred Web Lab**  
School of Computer Science  
University of Manchester  
Kilburn Building  
Oxford Road  
Manchester  
M13 9PL  
UK

**Corresponding author:**  
Alex Qiang Chen  
tel: +44 (161) 275 7821  
chenqa@cs.man.ac.uk

## 1 Introduction

The Web is a medium that provides an environment where files (this can be in the form of graphical formats, plain text, or audio) are interlinked, and can be accessed publicly via the Internet. It is the largest existence of hyper linked hypertext documents, and it is constantly changing and growing. From the beginning when the Web was created, HyperText Markup Language (HTML) was defined as the data structure format to be transmitted over the network, and Hypertext Transfer Protocol (HTTP) was created for traversing hypertext links [5]. Although the first Web browser for Windows operating system (Mosaic Web browser) was released in 1993, the widespread use of the Web really began only around 1995; this was around the same time when Microsoft released the Internet Explorer as part of Windows 95 [4].

At the rate that the Web is evolving, it is difficult to keep abreast of the changes to the Web content and technologies. Thus the guidelines, recommendations and standards are frequently being revised and generated by the World Wide Web Consortium (W3C) to improve accessibility to Web content, and to provide better Web experience. Often the standards, guidelines and recommendations take time to be accepted, and were slow to be adopted by Web developers, authors, and user-agents. Hence a lag was noticed between the time these standards and recommendations are introduced to when they were adopted. This causes a disconnection between the actual user experience, and what was expected by the technology stake-holders. Thus, understanding the evolution of the Web is essential as it will help us to understand the relationship between the underlying standards, recommendations, guidelines and their adoption time. This study attempts to understand these issues while focusing on the human factors surrounding the evolution of the Web user interface.

A wide variety of technologies, standards, guidelines and recommendations are available for Web content authors to use, and to conform. Early studies of Web content reported that besides HTML, graphical formats such as GIF and JPEG formats were used for transporting images over the Web even before 1997 [12]. More recently, the increase in popularity for the technologies such as client-side scripting, CSS and XHTML saw them included in some studies [9, 6]. Although a number of guidelines were generated by the Web Accessibility Initiative (WAI) to improve Web content accessibility, webmasters often do not find it beneficiary to take up these guidelines. This is because of the small number of user population it will benefit, thus it does not return huge economical benefits [22]. A recent study presented that a small percentage of the federal websites in the United States of America (US) and the government service websites across Europe conforms to these guidelines. The study also reported that these guidelines were keen to be taken up by the Japanese as well [24]. Although these reports highlighted the poor adoption rates to these guidelines, but it also showed that more are beginning to adopt them. From our analysis, we found that on average less than 10% of the Web conforms to WCAG which is still seems a little low, however these results do correspond to the results reported by Watanabe and Umegaki [24].

The information required to study the relationship between these issues can be found throughout the Web. However the biggest obstacle when collecting information from the Web is its size, hence an efficient method must be used. From an

empirical study in 1999 it was reported that the size of the publicly indexable Web was about 800 million pages [16]. Later in 2005, another study reported that the Web had grown enormously by more than 14 times; this was more than 11.5 billion webpages [14]. These studies demonstrated that tracking the changes of the Web can be difficult, and the Web is evolving and expanding at an exponential rate, however this did not deter researchers to be conducted on the evolution of the Web. Commonly two types of method were used to track the changes to the Web. The first method monitors the packets that pass through the corporate firewall in the proxy server [12]. This method can reliably capture the contents from the websites accessed by the server's users, but it is biased towards the needs and culture of the corporation. Thus it does not give a good representation of the Web. The other method (the more popular method) uses a Web crawler or a Web robot to crawl and fetch a snapshot of the targeted webpages, and the required data for further analysis [23, 13]. The coverage of this method depends on the quality of the webpages the Web crawler was deployed to capture. Hence to overcome this issue, commonly some kind of webpage selection methodology (e.g. page ranking [8] and tagging [3]) will be used to select the Web pages. A Web robot was chosen for this study due to the scope, comprehensiveness and flexibility of webpages it can be programmed to examine, and capture.

This study will identify the recommendations to questions such as ‘Do we rely on technology or guideline adoption?’, ‘Do we need technical intervention?’, ‘Will technical interventions be adopted into user-agents?’, or ‘Should we be led by users, by engineers, or by history?’ [15]. To do this, a long term slice of a number of popular and randomly selected websites for the last ten years (1999-2008) were captured. This will allow us to investigate if the random websites, in-general, follows the trends of the popular websites, and to identify the lag between them. Two larger sets of popular and randomly selected websites were also captured from the current Web to validate the analysis done for the long term slice of the websites for the last ten years. This report provides the background material in correlation to the research focus to be undertaken in this study. A detailed discussion relating to the issues of the results were followed. The analysis and discussions covered in this study will contribute as recommendations to inform and predict the Web development into the future.

## 1.1 Synopsis

This report was structured as follows:

**Section 2: Purpose and Research Questions** describes the motivation behind this research.

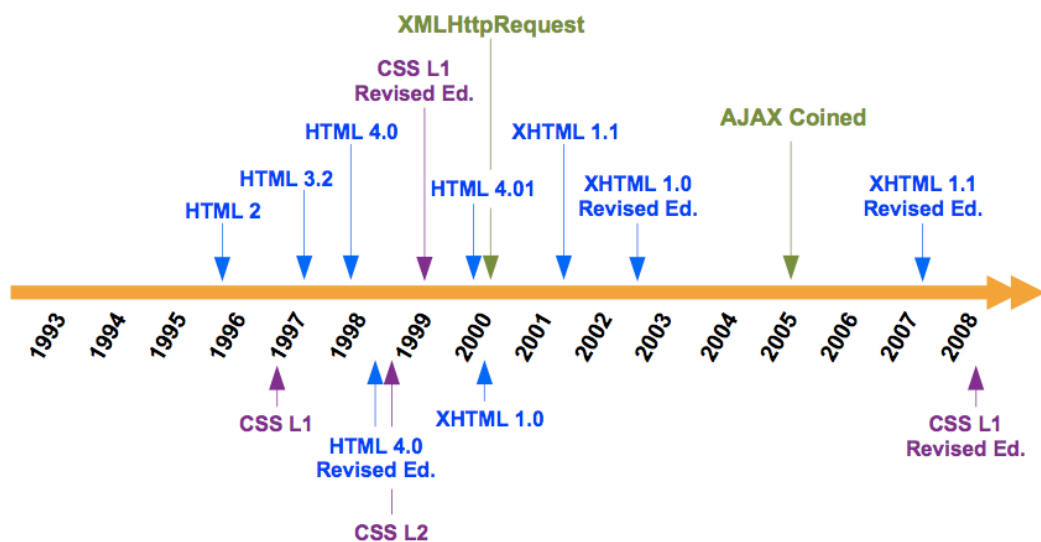
**Section 3: Experimental Data Collection** discusses the methods chosen and applied when conducting this study.

**Section 4: Results and Discussion** discusses the results collected and the more in-depth analysis on the related issues.

The series of Web Evolution technical reports consist of two reports. Besides this report, the “Web Evolution: Code and Experimental Guide” report [7] covers the instructions for the usage of the codes, its process flows, and to configure the codes for customisation.

## 2 Purpose and Research Questions

Since 1994 the W3C has been constantly revising and introducing new recommendations for the Web. Thus before discussing about the results from the analysis we had conducted, let us first look at the milestone of the Web standards. Figure 1 presents the timeline of the major HTML and CSS standards when they became a Web standard.



**Figure 1:** Web standards milestone

This study investigates into the underlying issues relating to the disconnection between the actual user experience, and what was expected by the technology stakeholders. These issues were commonly caused by the rate that the Web is evolving. Thus it makes it difficult to keep abreast of the changes done to the Web content and the technologies. Due to this, guidelines, recommendations and standards were frequently being revised and generated to improve accessibility to Web content, and to provide better Web experience. Often the standards, guidelines and recommendations take time to be accepted, and were slow to be adopted by Web developers, authors, and user-agents. This had caused a lag between the time these standards and recommendations were introduced to when they were adopted. Understanding the evolution of the Web is essential as it will help us to understand the relationship between the underlying standards, recommendations, guidelines and their adoption

time.

An attempt was made to identify the recommendations to our questions such as ‘Do we rely on technology or guideline adoption?’, ‘Do we need technical intervention?’, ‘Will technical interventions be adopted into user-agents?’, or ‘Should we be led by users, by engineers, or by history?’ [15]. This study also intends to understand if the general Web contain the similar characteristic as the top websites. Hence it will give a good projection for future work in this area relating to these two groups of websites. From our findings, further analysis were also conducted to better understand the relationships between technologies such as JavaScript and Flash have in common with the AJAX model. How did CSS affect the way Web design has evolved with the increase in graphical usage? We believe from this study, it will provide a better understanding and recommendations for future work encompass the Web user interface.

### 3 Experimental Data Collection

Understanding the evolution of the Web through analysis of historical data is thought to be beneficial for those investigating issues with the current web user interface [15]. In this study, we will use the recommendations suggested in [15] together with our methodologies. In order to investigate into these issues, data will be required to be downloaded from the Web for analysis. There are two methods commonly used to capture information from the Web: (1) monitor the packets that passes through the firewall in a proxy server, and (2) use a Web crawler or a Web robot to crawl and fetch a snapshot of the targeted webpages. Our chosen method uses Web robots to capture the data from our targeted webpages. A Web robot is a Web application that will crawl a set of selected webpages and return with the captured source code of the webpages and the necessary data for further analysis. However, this will only capture the current version of the webpage. In order to capture the historical data of a website for the past ten years, the Internet Archive<sup>1</sup> data which keeps a snapshot of Alexa’s<sup>2</sup> Web crawled history [1] was used. Alexa is one of the largest Web crawls, it ranks the websites that were submitted to them based on their traffic rankings from Alexa tool bar<sup>3</sup> [2]. In this project all the Web applications used to carry out our tasks were written using PHP: Hypertext Preprocessor (PHP). PHP version 5 was chosen due to its simplicity and capability as a server-side scripting language.

A separate application was used to conduct the Web mining processes to collect useful data from the captured webpages for specific analysis. From the analysis, trends can be identified, and together with the other results, answers to the questions that motivate this research can be determine. In order to investigate if the top websites gives a good representation of the Web in general, two sets of data will be required; one for the top websites and the other consist of random websites. Another similar pair of data set will be collected to investigate the historical data over the past ten years, and the current Web to verify the trends deduced from the historical

---

<sup>1</sup><http://www.archive.org>

<sup>2</sup><http://www.alexa.com>

<sup>3</sup><http://www.alexa.com/site/download/>

data. Therefore four sets of data were captured for this study. Two sets were used to look at a long term slice of websites for the past ten years, another two larger sets were used to give an in-depth look of the current Web, and to verify the analysis done for the two sets of historical data. In the next few sections the methodologies that were employed to select the websites, capture the source code of the webpages, and Web mining processes will be discussed.

### 3.1 Methodology for Selecting Websites

Four sets of data were captured for analysis in this study. Two sets were used to look at a long term slice of websites for the past ten years, one was for the top twenty websites, and the other was a set of five hundred randomly selected websites. Another two larger sets were used to give an in-depth look of the current Web, and to verify the analysis done for the two sets of historical data. For these two sets of data, one was for the top five hundred websites, and the other was a set of five thousand randomly selected websites.

Our top twenty websites URLs were taken from the Alexa global top 500<sup>4</sup> on 13 June 2008, and our top five hundred websites URLs were taken from Alexa global top 500 on 24 July 2008. Since our Alexa top twenty websites data looks at the long term slice of websites for the past ten years, the archives of these websites were captured from Internet Archives servers. Hence our top twenty websites were not the top twenty websites from Alexa global top 500 list, but the top twenty websites in that list with at least one archive from Internet Archive servers between year 1999 and 2000. An assumption that a existing website with archives available between year 1999 and 2000 will have archives from at least year 2000 to June 2008. This will reduce the workload of the script and the server to check for a archive in every year. Our Alexa top five hundreds websites were taken directly from the same list as this set of data looks only at the current Web. Thus our Alexa top five hundred websites data is a superset of our Alexa top twenty websites data as shown in equation 1.

$$Top20websites \subseteq Top500websites \quad (1)$$

Both sets of randomly selected websites were selected from Google Directory<sup>5</sup>. The Google Directory uses the data from the Open Directory Project<sup>6</sup>, but employs the Page Ranking algorithm [19] to rank the websites in each directory [18]. Depending on the number of targeted websites required, this amount will be divided into  $x$  number of websites and spread out as equally as possible across the fifteen directories at Google Directory. In each directory, the most populated subdirectory will be chosen, and in each subdirectories, the first  $x$  number of websites will be selected. Due to the Page Ranking algorithm, the top  $x$  websites of that subdirectory will also be ensured to be quality websites from that subdirectory. The results of this method depends on the Google ranking algorithm to rank the websites in each directory. Hence every time when this script is re-run, different websites will be selected depending on how Google ranking ranks the website.

<sup>4</sup>[http://www.alexa.com/site/ds/top\\_sites?ts\\_mode=global&lang=none](http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none)

<sup>5</sup><http://www.google.com/dirhp>

<sup>6</sup><http://www.dmoz.org>

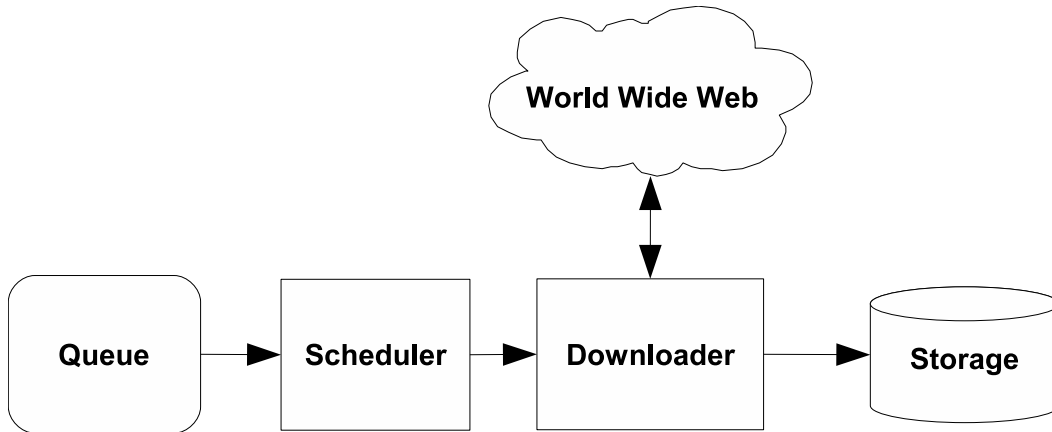
$$Random500websites \cup Random5000websites \quad (2)$$

Additional effort were taken to ensure that when selecting our sets of random websites, these websites should not be part of our top websites list. This concept is illustrated in equation 3 where the *Randomwebsites* refers to our random five hundred websites union with our random five thousand websites (refer to equation 2), and the *Topwebsites* refers to our Alexa Top twenty websites and Alexa Top five hundred websites.

$$Randomwebsites \cup Topwebsites \quad (3)$$

### 3.2 Capturing the Webpages

A Web robot was employed to carry out the capturing of information from the targeted websites in this study. The Web robot was customized to fetch only the target webpage's source code along with the required external scripting and styling source code files. This information were stored on the local machine that was used to deploy the Web robot. The suggested general modules of the Web robot is presented in figure 2. First the selected domains were stored in the queue where it will be feed into a scheduler to manage the Web robot process, and carefully not to overload the network traffic. Once the scheduler gives the go ahead, the URL will be fed to the downloader to proceed with its task. Finally the downloader will store the captured data on the machine's local hard disk.



**Figure 2:** Modules in the Web robot

When storing the downloaded data, each sets of data (i.e. Alexa Top twenty websites, random five hundred websites, Alexa Top five hundred websites and random five thousand websites) were stored in a separate folder.

After the targeted Web pages were captured, the next part will involve a process called Web mining. This process involves retrieval of the required data from the captured webpages for further analysis. In the next section this will be covered in greater depth.

### 3.3 Web Mining

Web mining is a stage that involves finding for the interesting information required from the captured webpages for further analysis. Each webpage and its required external files were either parsed using the PHP native Document Object Model (DOM) HTML parser, or searched using the Perl's regular expression syntax to retrieve the required information. Now let us examine how the different types of analysis and data were collected.

#### 3.3.1 Types of HTML Standard Detection

Two methods were employed to detect the type of HTML standards used by a webpage. (1) Whenever available, the document type (DOCTYPE) was detected using the following Perl's regular expression syntax. Commonly the DOCTYPE will be printed at the top of the webpage's source code. Listed below are some examples of the possible methods used by a webpage to declare the document's intended HTML standard. The first example given was used to declare a HTML 3.2 Final webpage, the second example shows the DOCTYPE for HTML 4.01 Transitional webpage, and the third example is the DOCTYPE for XHTML 1.0 Strict webpage.

Example 1:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
```

Example 2:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional
//EN" "http://www.w3.org/TR/1999/REC-html401-19991224/
loose.dtd">
```

Example 3:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
```

From the above examples some similarities in declaring the DOCTYPE of a webpage can be observed, hence the following syntax was devised and used to capture the DOCTYPE from a webpage source code. This syntax will not only return the version of the HTML standard, but also the document's sub-specification (Transitional, Strict...).

```
#!/DOCTYPE\sHTML\sPUBLIC\s"-\/\w+\/\DTD\s([\sa-z0-9\.]
+\/\EN"/i
```

Since this method (declaring the DOCTYPE) was not strongly enforced by most major user-agents, hence HTML documents can be written without it and still being parsed properly by them. Therefore another method was used whenever the DOCTYPE was not available. (2) This time the HTML document were parsed,

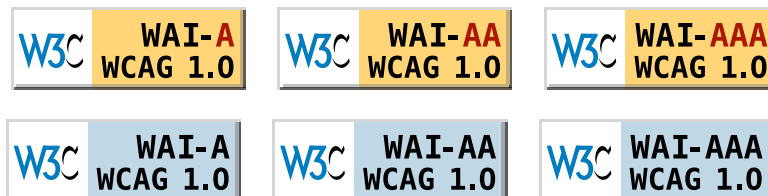
and the type of tags used were examined to determine the type of HTML standard used. However as it was not possible to know what was the actual HTML standard the Web author/developer was intending to use, and since some of our documents were historical data, we will assume that if a document uses HTML syntax that is so simple, and it was not possible to differential between the different versions of HTML standards, then we will assume it to be a HTML 2 document. So by default we will assume a webpage to be a HTML 2 document, unless a different standard was detected. The HTML elements that were used to defined for the different HTML standards in this project were listed in table 1.

HTML Standards	Elements
HTML 3 [10]	listing, plaintext, style, xmp
HTML 3.2 [21]	font, div, script
HTML 4.0, HTML 4.01	abbr, acronym, applet, basefont, bdo, button, col, colgroup, center, del, embed, fieldset, frame, frameset, iframe, ins, label, legend, noframes, noscript, object, q, s, span, tbody, tfoot, thead, u
XHTML 1.0	-
XHTML 1.1	rb, rbc, rp, rt, rtc, ruby

**Table 1:** HTML standards and elements used during our Web mining process for HTML standards detection [20, 17, 11]

### 3.3.2 WCAG 1.0 Conformance Detection

To check if a webpage is compliant to the WAI's WCAG 1.0 guidelines, we will search for a display of conformance to these guidelines on a webpage. For this process, since it was not intended to create a validation tool and due to the scope of this study, thus the above methodology was employed. A couple of techniques were used for this detection, the first technique looks for any display of WCAG 1.0 conformance logo, and the type of logo for the specific guideline level of conformance supplied by W3C. The W3C provides the logos in two colours, the original colour and a blue colour as seen in figure 3.3.



**Figure 3:** WCAG 1.0 Conformance Logos

The respective logo's level is usually displayed if a page is conform to a certain level of conformance in the WCAG 1.0 guidelines. Presented below are the sample

HTML codes provided by the W3C for the display of the above logos as a proof of conformance on a webpage.

Level A Conformance:

```
<a href="http://www.w3.org/WAI/WCAG1A-Conformance"
  title="Explanation of Level A Conformance">
  </a>
```

Level Double-A Conformance:

```
<a href="http://www.w3.org/WAI/WCAG1AA-Conformance"
  title="Explanation of Level Double-A Conformance">

  </a>
```

Level Triple-A Conformance:

```
<a href="http://www.w3.org/WAI/WCAG1AAA-Conformance"
  title="Explanation of Level Triple-A Conformance">
  </a>
```

The above codes gave an illustration of what was expected within a HTML code that display these conformance logos, however Web developers/authors may change these codes slightly to meet their design specifications. Therefore two checks were used to detect the display of the conformance logos that were generic to the two logo colours, and for more flexibility to the codes used. The following regular expressions were used to detect for the presence of WCAG 1.0 level A conformance logo.

- (1) `/href\s*=\s*[\'\"]*http\:\:\/\/www.w3.org\/WAI\/WCAG1A-Conformance[\'\"]*/i`
- (2) `/src\s*=\s*[\'\"]*http\:\:\/\/www.w3.org\/WAI\/wcag1A/i`

The next two regular expressions were used to detect for the presence of WCAG 1.0 level AA conformance logo.

- (1) `/href\s*=\s*[\'\"]*http:\/\/www.w3.org\/WAI\/WCAG1AA-Conformance[\'\"]*/i`
- (2) `/src\s*=\s*[\'\"]*http:\/\/www.w3.org\/WAI\/wcag1AA/i`

The following regular expressions were used to detect for the presence of WCAG 1.0 level AAA conformance logo.

- (1) `/href\s*=\s*[\'\"]*http:\/\/www.w3.org\/WAI\/WCAG1AAA-Conformance[\'\"]*/i`
- (2) `/src\s*=\s*[\'\"]*http:\/\/www.w3.org\/WAI\/wcag1AAA/i`

Some websites may choose not to display the conformance logo, but display their Web content accessibility conformance in plain text. Thus another technique was used to cope with this issue. Two methods were configured for this experiment. This first method was a pessimistic view that scans for the last 100 characters on each webpage for a display of conformance, and the second method was an optimistic view that scans the entire webpage for a display of conformance. Words such as ‘accessibility’ or ‘WCAG’, or just for the display of level of conformance were checked. When checking for the level of conformance, except for level A conformance, both level double A and level triple A conformance were conducted. This was because checking for level ‘A’ conformance alone, it can be easily mixed up with the letter A. The following regular expression was used to check for the display of text for the accessibility conformance.

```
/(Accessibility[-_\\s,]*|wcag(\\s?1\\.0)?[-_\\s,]?|\\s(AA|AAA)\\s)/i
```

From the two techniques presented, different forms of results will be gathered when searching for the accessibility conformance of a webpage. Thus the final results used for this analysis will be in the form of either the webpage was conformed or was it not to the accessibility guidelines. Websites that conforms to the guidelines, but do not display their conformance on their website will be assumed to be not conform.

### 3.3.3 Graphical Format Usage Detection

In this part of the report, the detection for different types of graphical format will be covered in detail. There are many ways a graphic can be formatted and prepared to be portable over the Web. Only the more popular formats, and the formats suggested by W3C will be covered in this study. To detect the usage of the different types of graphical formats, the extension of the different formats must be include in the regular expressions used. However for some graphical formats multiple file extensions may exist. So lets first look at the different types of graphical formats we will be covering in this study, and its possible file extension(s) used for portability in table 2.

Now that the foreseeable different types of file extensions for each graphical format was identified, now lets look at the few ways that a Web developer/author can include a graphic to a webpage. The generic method include either attaching

Graphical Formats	File Extensions
GIF	.gif
JPEG	.jpg, .jpeg, .jpe
PNG	.png
Flash	.swf, .flv, .f4v, .f4p, .f4a, .f4b
SVG	.svg
SMIL	.smil

**Table 2:** Graphical formats and file extensions

the graphic within an HTML code or in a style coding as a background. Hence from the webpage's source code, the following regular expressions can be used to detect a graphical format for the two generic method and the file's extension.

First for GIF graphical formats,

- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\gif)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\gif\s*)\)/i`

For JPEG and its possible file extensions,

- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\jpg)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\jpg\s*)\)/i`
- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\jpeg)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\jpeg\s*)\)/i`
- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\jpe)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\jpe\s*)\)/i`

For PNG formats,

- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\png)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\png\s*)\)/i`

For Flash and its possible file extensions,

- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\swf)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\swf\s*)\)/i`
- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\flv)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\flv\s*)\)/i`
- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\f4v)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\f4v\s*)\)/i`
- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\f4p)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\f4p\s*)\)/i`
- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\f4a)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\f4a\s*)\)/i`
- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\f4b)[\'\\""]*/i`
- (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\f4b\s*)\)/i`

For SVG,

- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\.svg)[\'\\""]*/i`  
 (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\.svg\s*)\)/i`

Finally for SMIL,

- (1) `/[\'\\""]*([-_~\:\a-z0-9\\\/\.\.]+\.\.smil)[\'\\""]*/i`  
 (2) `/\((\s*[-_~\:\a-z0-9\\\/\.\.]+\.\.smil\s*)\)/i`

If any of the above regular expressions was true, the presences of the respective graphical format would be assumed to be used on the webpage.

### 3.3.4 Client-side Scripting Usage Detection

Detecting the use of client-side scripting and styling will allow trends for the respective standards to be identified. For JavaScript, this will help to understand the usage of it and how it has effected the growth or decline in popularity of other standards and recommendations (E.g. AJAX...). To detect the usage of client-side scripting, the HTML codes will be parsed and the “<script>” tag will be examined. Under the “script” elements, both the “type” and “language” attributes will be searched using the regular expression “/javascript/i” for the existence of JavaScript. If this exist, or none of these attributes were defined, then an assumption will be made that the client-side scripting language was JavaScript. However if at least one of these attributes was defined and JavaScript was not found, then VBScript will be assumed.

### 3.3.5 Cascading Style Sheets Usage Detection

Three methods were employed to conduct the detection of CSS. (1) Detect within the HTML code if the element “style” is used. (2) All the elements defined in a HTML document will be checked if the attribute “style” was used. This method was employed to detect if CSS was applied to the individual HTML tags for specific display control. (3) The next method checks for the attachment of external CSS files by using the following regular expression.

```
/<link\s[-_~\:\a-z0-9\s]*rel\s?=\s?[\'\\""]?stylesheet[\'\\""]
  ?[-_~\:\a-z0-9\s]*type\s?=\s?[\'\\""]?text/css[\'\\""]?/i
```

As long as one of the above methods was detected, an assumption will be made that the presence of CSS usage exist.

### 3.3.6 AJAX Usage Detection

AJAX growing popularity may be benefited from the fruits of other Web technologies such as JavaScript. Thus analysing the usage of AJAX will help us to understand its usage trend and how it was affecting the other related Web technologies. On the client-side, AJAX uses JavaScript and the asynchronous technology to communicate with the server. Although there are numerous ways in which one can determine the presence of AJAX, our method employed consist of the following two techniques.

(1) The detection of “XMLHttpRequest” in JavaScript, and (2) the usage of HTML element “iframe”. The detection of HTML element “iframe” was used because this element uses the asynchronous technology, and can be dealt with by JavaScript. Once the external JavaScript file’s codes were concatenated with the JavaScript codes embedded within the HTML code, the following regular expressions can be applied to search for the “XMLHttpRequest” within them.

```
/XMLHttpRequest\(/i
```

To search for the “iframe” element, the following regular expression was used; although parsing the HTML code will do the job as well.

```
/<iframe\s/i
```

As long as one of the above method was detected on a webpage, an assumption would be made that the presence of AJAX exist.

### 3.4 Overall Process

The processes and methods required to collect the necessary data can be quite a lot due to the volume of websites analysed in this study. This overview wraps up the overall processes that is required to go through when collecting each set of data. Four general stages will be required to complete the entire process for each set of data. As shown in figure 4, in the first stage the Web robot will be sent out to extract or select the targeted URLs to be captured. Then in the second stage the Web robot will be deployed to capture the source code of the targeted websites along with the necessary external files. In the third stage, this consist of two parts: an automated, and a manual verification of data process to ensure they were captured correctly. If an error was detected, stage two and three will be required to be repeated before one can proceed to the final stage. The purpose of last stage is to collect the required information from the captured source code so that further analysis can be done. None of the stages in figure 4 should be bypassed to ensure repeatability. To ensure integrity of the data captured, stage three should be done thoroughly. The capture data collected when this study was conducted were organised in a different directories according to their data sets. This was covered more in Appendix A.

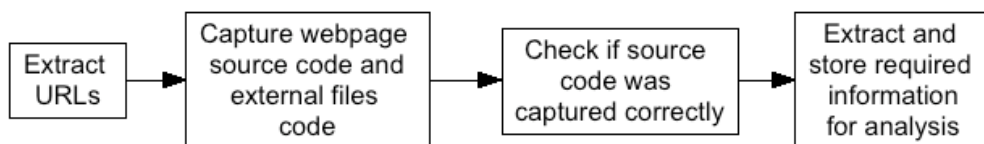


Figure 4: Overall Process Flow

### 3.5 Issues Relating to Captured Data

The data captured for this study were not all perfect. Some of the historical data were not available from Internet Archive's servers, or the websites selected that were no longer existent. Due to this, our historical data sets require some form of normalisation after the data was captured.

Table 3 list the total number of websites retrievable for each sets of data during our capturing process. The missing data in our Alexa top five hundred websites data was due to four no longer existing URLs provided by Alexa global top five hundred<sup>7</sup> on the 24 July 2008.

Year	Alexa top 20 websites	Alexa top 500 websites	Random 500 websites	Random 5000 websites
Jan 1999	17	-	320	-
Jul 1999	15	-	76	-
Jan 2000	14	-	176	-
Jul 2000	19	-	452	-
Jan 2001	15	-	444	-
Jul 2001	18	-	380	-
Jan 2002	18	-	380	-
Jul 2002	17	-	452	-
Jan 2003	20	-	479	-
Jul 2003	20	-	459	-
Jan 2004	18	-	409	-
Jul 2004	20	-	483	-
Jan 2005	20	-	487	-
Jul 2005	20	-	463	-
Jan 2006	20	-	478	-
Jul 2006	20	-	473	-
Jan 2007	20	-	467	-
Jul 2007	16	-	440	-
Jun 2008	20	496	500	5000

**Table 3:** Total number of websites obtainable during capturing process

When retrieving archives from Internet Archives, it was noticed that between July 1999 and February 2000, very little archives were available. As seen in table 3 very few data were captured between these times for both our data for Alexa top twenty websites and random five hundred websites.

To deal with the inconsistency of the data capture, some form of normalisation was required to be applied to these data before further analysis were done. In this study, percentage was applied to all the data before conducting any further analysis. Thus all the results from the analysis presented were in the form of percentages. Another gap in the captured data can be noticed. These are archives for the interval

<sup>7</sup>[http://www.alexa.com/site/ds/top\\_sites?ts\\_mode=global&lang=none](http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none)

of January 2008 that were not available from Internet Archive during the time the data were downloaded, thus a gap between July 2007 and June 2008 can be noticed.

### 3.6 Conclusion

Different methods were applied in this study to understand the evolution of the Web. Four major stages were used to successfully capture the selected websites, and to extract the required data for analysis. From the discussion above, due to the existence of some missing data from the Internet Archives, the historical data sets had to be normalised due to the inconsistent volume of data captured between the intervals. Percentage was later applied to all the data captured for normalisation before analysis were conducted. Besides the dip in historical data available from the Internet Archives between July 1999 and January 2000, most of the data retrievable were of acceptable volume. Finally regular expressions and parsing the HTML codes were techniques used together with our methodologies to extract the necessary data for our analysis required by this study. Refer to [7] for more in-depth explanation of the flow of the capturing process, and the purpose of the codes.

## 4 Results and Discussion

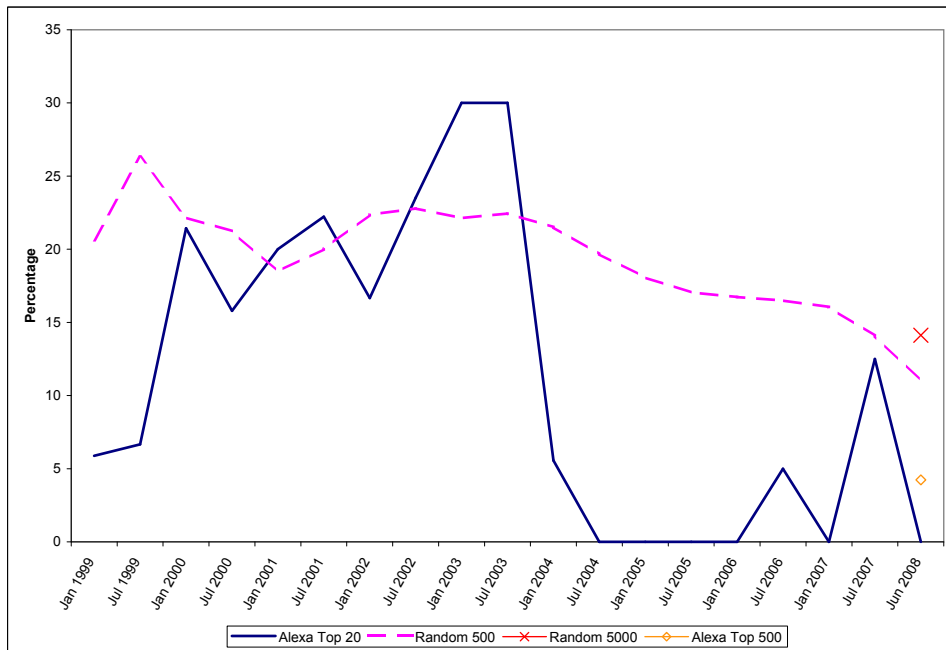
In this section, possible analysis, trends and conclusions from our results were presented. The presentation of our discussions will be structured into the four categories discussed; W3C standards, graphical formats, client-side scripting, and guidelines, followed by the further analysis done to enhance our understanding for some trends.

### 4.1 HTML Standards

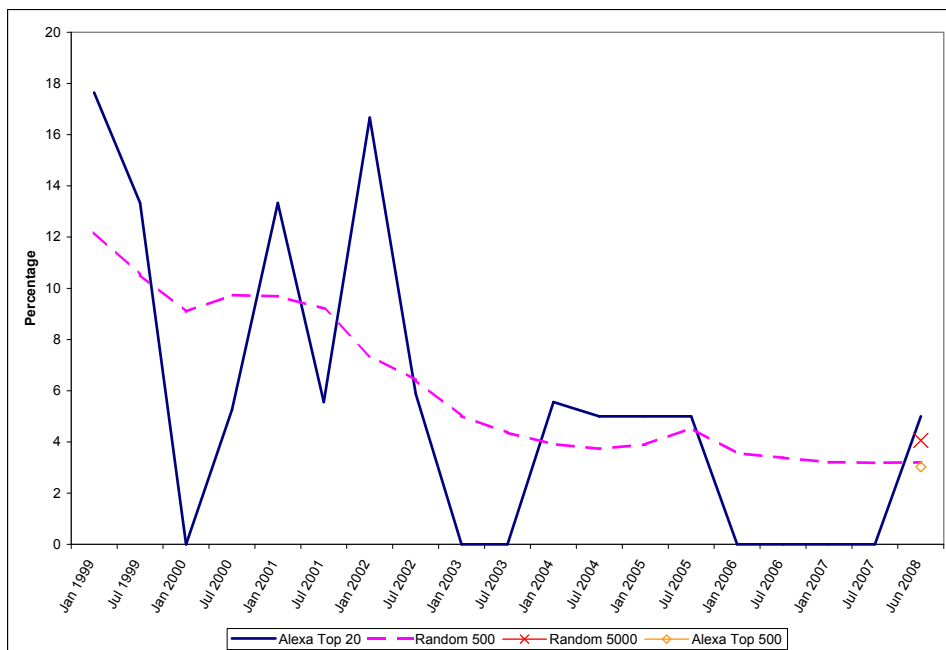
A gradual decline in HTML 2 usage for the last ten years was noticed from the graph in figure 5. This was seen for both the random five hundred websites and the Alexa top twenty websites sets of data. Further analysis was done to determine if these two sets of data have any correlation applying Pearson correlation. A significant relationship between Alexa top twenty websites and random five hundred websites was noticed,  $r = .52$ ,  $p$  (two-tailed)  $< .05$ . Both the random five thousand websites and the Alexa top five hundred websites results were close for the current Web analysis. Thus a similar usage for this standard was forecasted.

Similar to HTML 2 standards, the HTML 3 standards usage also exhibited a decline for the last ten years. Figure 6 showed that a correlation between both the the random five hundred websites results and Alexa top twenty websites results. When applying Pearson correlation, a significant relationship was noticed between them,  $r = .68$ ,  $p$  (two-tailed)  $< .01$ . From this experiment, both the random five thousand websites and Alexa top five hundred websites demonstrated very close percentages for the current Web. This proves that HTML 3 is slowly losing its popularity with Web developers/authors.

The HTML 4 standards has been heavily used by the Web as shown in figure 7, but a declining trend is predicted. A significant relationship between the random five hundred websites results and the Alexa top twenty websites results was noticed

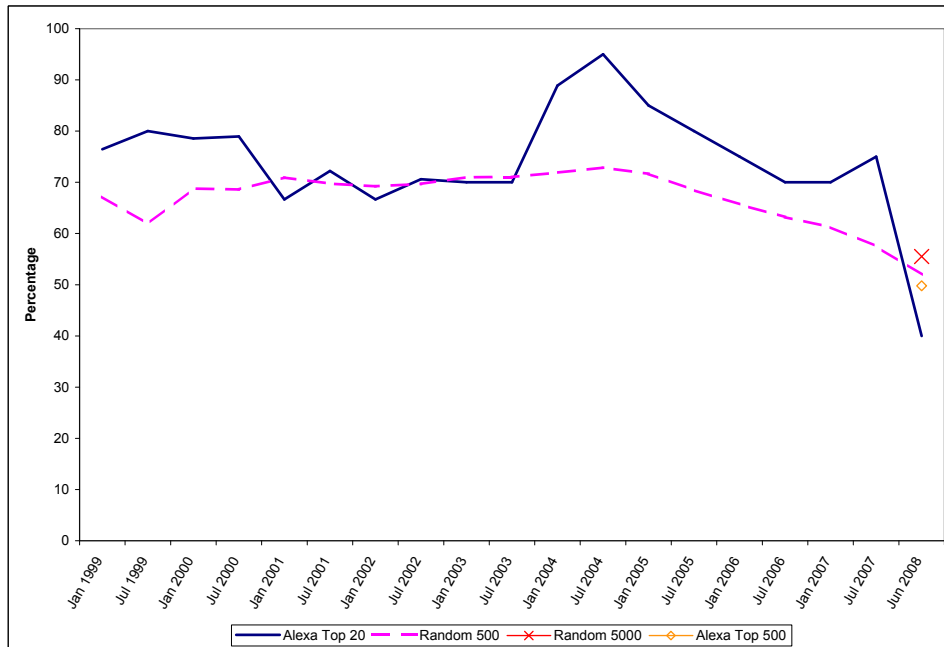


**Figure 5:** HTML 2 usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites



**Figure 6:** HTML 3 usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

when applying Pearson correlation,  $r = .62$ ,  $p$  (two-tailed)  $< .01$ . Our prediction for this trend was justified using both the Alexa top five hundred websites results and the random five thousand websites results presented.

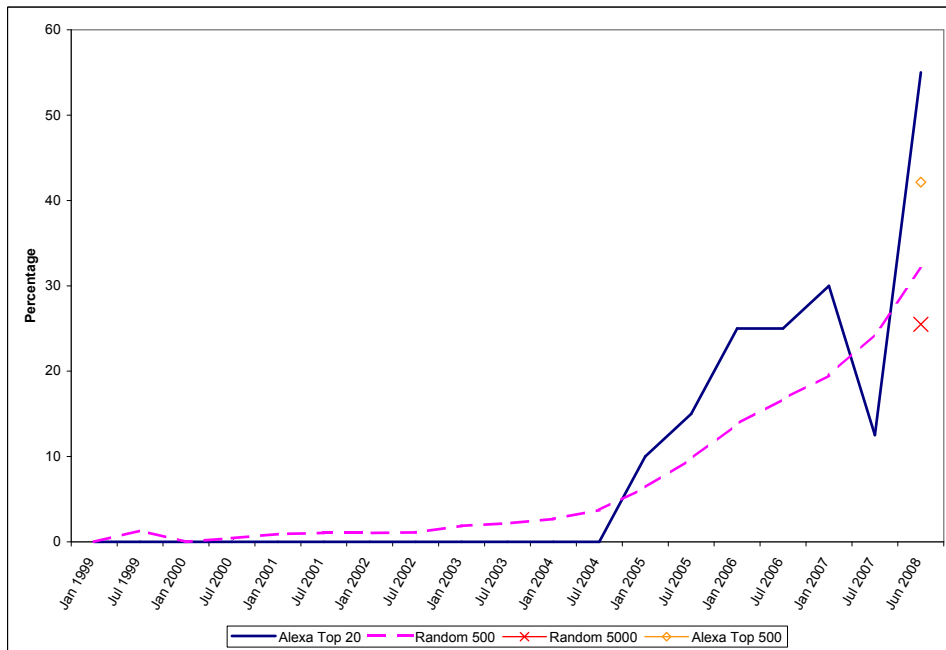


**Figure 7:** HTML 4 usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

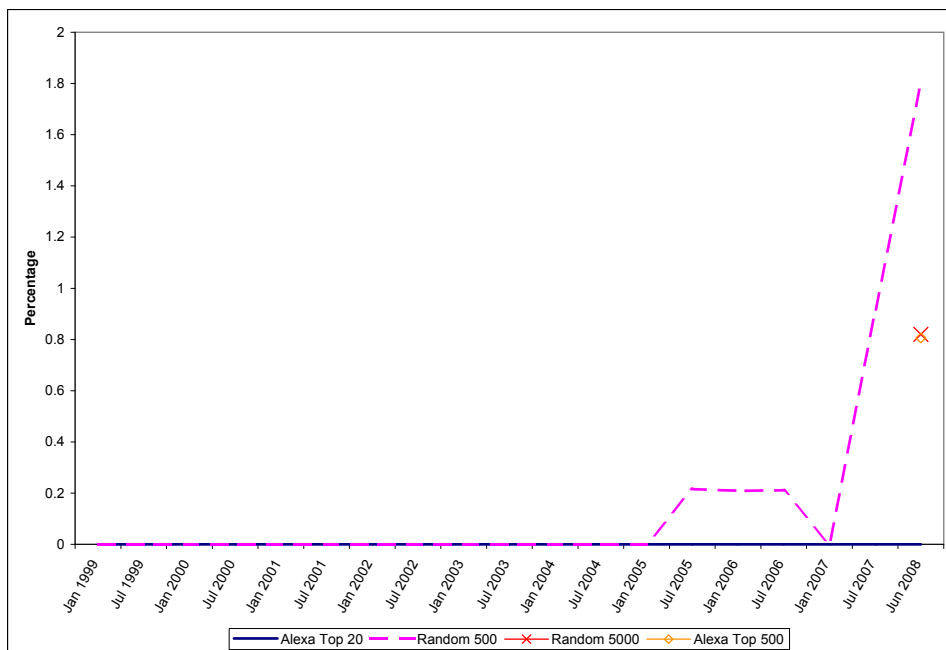
Since the release of the revised edition for the XHTML 1.0 standard in August 2002, an increase in usage for this standard was noticed around 2004 as shown in figure 8. There was also a significant relationship between the random five hundred websites and the Alexa top twenty websites results when Pearson correlation was applied,  $r = .92$ ,  $p$  (two-tailed)  $< .01$ . From the graphs, based on the verification of the Alexa top five hundred websites and the random five thousand websites results, a growing trend was predicted to continue.

The recent release of the revised edition of the XHTML 1.1 standards in February 2007 as seen in figure 1 explains the reasons behind these poor adoption rates for this standard during the time when this study was conducted. From our data collected as presented in figure 9, and from our adoption trends of the previous HTML standards, a growth in usage for this standard was predicted. This prediction cannot be validated, and future work for this standard is required, and to prove if our prediction was correct.

From the above analysis, besides the XHTML 1.1 standards, all the rest of the major W3C standards demonstrated a significant correlation between the Alexa top websites and the random websites results,  $p$  (two-tailed)  $< .05$ . Therefore we can conclude that from these results the Alexa top websites do give a good idea of how the Web is evolving for the W3C standards in general.

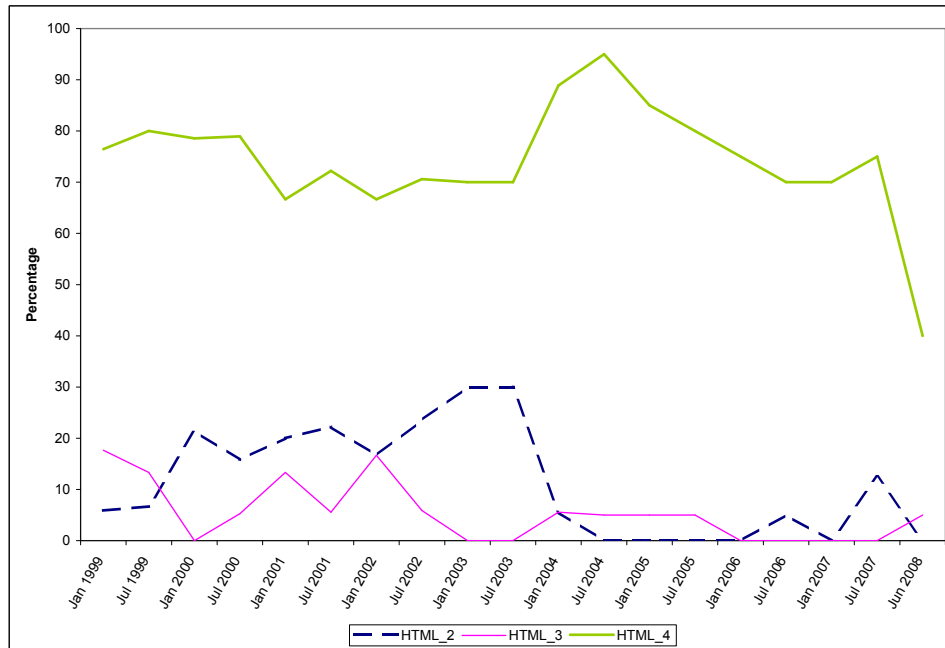


**Figure 8:** XHTML 1.0 usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites



**Figure 9:** XHTML 1.1 usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

Further analysis were done to understand the relationships between the W3C standards for the random five hundred websites, and the Alexa top twenty websites. Figure 10 shows the usage of HTML 2, 3 and 4 for Alexa top twenty websites over the past ten years. Both HTML 2 and 3 exhibits a gradual decline in usage, while HTML 4 was increasing before the year 2006, and from year 2007 a rapid roll off was noticed.



**Figure 10:** HTML 2, 3, 4 usage percentage for Alexa top 20

The graph in figure 11 for the random five hundred websites for the last ten years also showed trends similar the Alexa top twenty websites results. Again from the discussions above, a decline in usage for these standards were expected. It was also observed from figure 10 and figure 11 that the influencing factor for their decline were not from either the HTML 2, 3 or 4, but by some other standards or recommendations.

Next lets examine how will HTML 4 perform when plotted against XHTML 1.0 and 1.1. From figure 12 it was observed that around the same time when HTML 4 began to roll off, a significant increase in the usage of XHTML 1.0 was also notice. The pattern of the declining HTML standard; HTML 4, exhibits a mirroring image of the XHTML 1.0's graph. Thus for the top websites, it was observed that a major shift in usage (> 50%) from HTML 4 to XHTML 1.0 had already occurred. This converting trend is expected to continue as it can be verified by the discussion presented earlier for the individual HTML standards trends.

Finally in the last analysis, HTML 4 was again plotted against XHTML 1.0 and 1.1 for the random five hundred websites results. This was to check if the same trends exist for the random websites results. From figure 13 showed that this

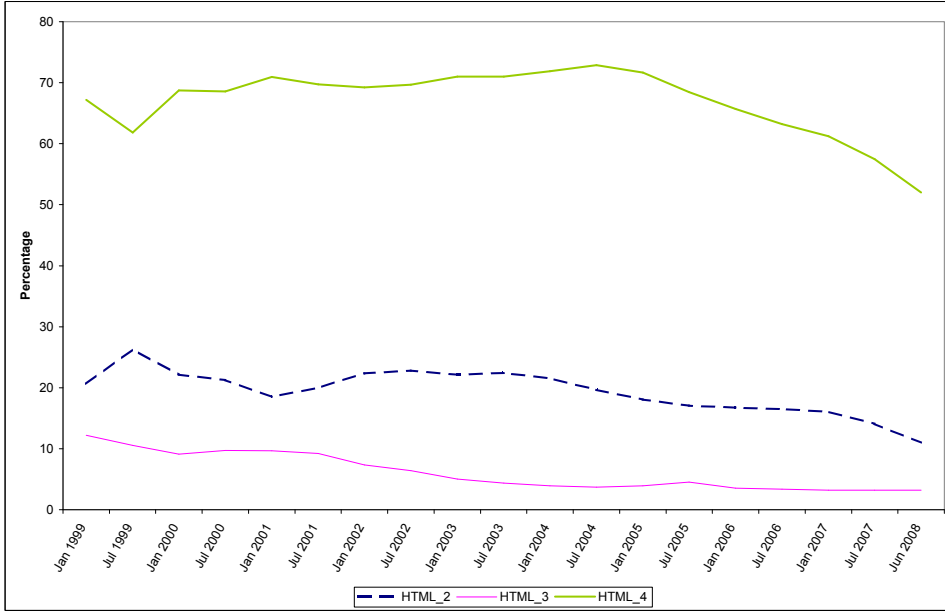


Figure 11: HTML 2, 3, 4 usage percentage for random 500

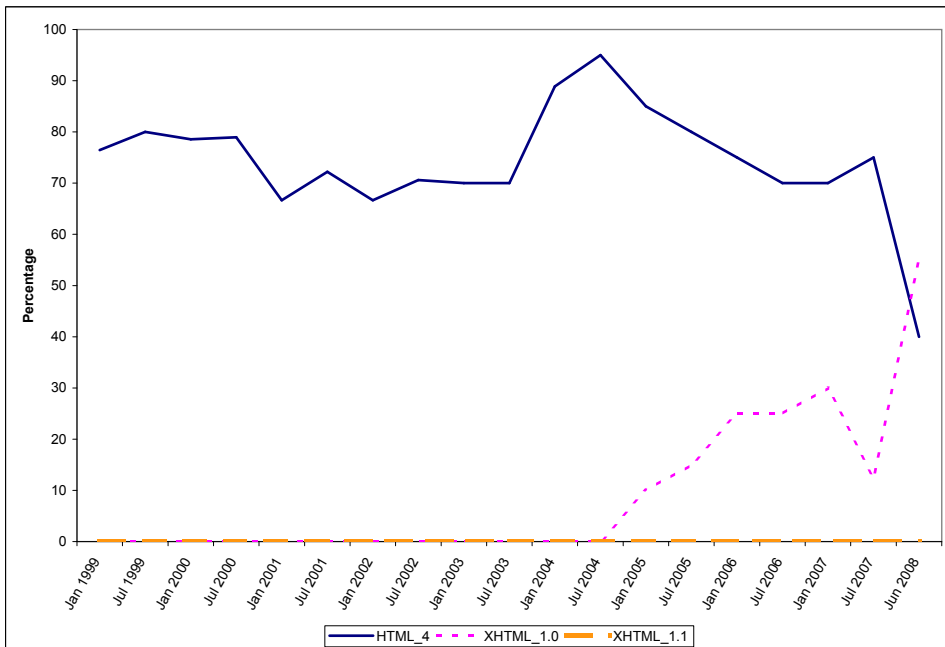
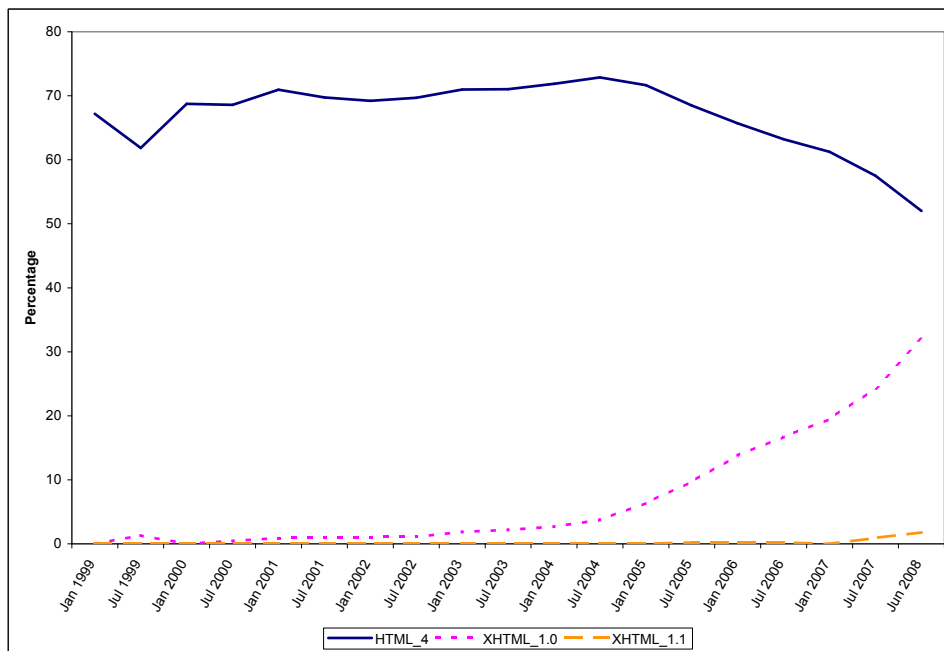


Figure 12: HTML 4, XHTML 1.0 and 1.1 usage percentage for Alexa top 20

set of results also exhibited similar trends, as one would expected this from their individual standards analysis discussed earlier. Although the random five hundred websites have not been adopted as much as the Alexa top twenty websites to the XHTML 1.0 standards, but they were showing similar trends to the Alexa top twenty websites W3C standards.



**Figure 13:** HTML 4, XHTML 1.0 and 1.1 usage percentage for random 500

To summarise the discussions presented in this section, the major W3C standards on average seems to be led by the Alexa top websites. From the discussions presented earlier, the Alexa top websites in-general does give a good representation of how the Web is evolving for the major W3C standards. On average the Alexa top websites adopts to a new standard one year faster than the random websites, and a growth in XHTML 1.0, and 1.1 usage was predicted. A trend was also notice that the usage of HTML 4 is decreasing, and websites that were previously using it are replacing their webpages with XHTML 1.0.

## 4.2 WCAG 1.0 Conformance

As discussed previously, the WCAG will be the only guideline analysed in this study for Web content accessibility conformance. Also discussed earlier, we will look for only the display of conformance to these guidelines on a website to detect if it is conformed to it. This can be in the form of displaying a logo or in plain text. The level of conformance was also search during this experiment, such as A, double A, or triple A.

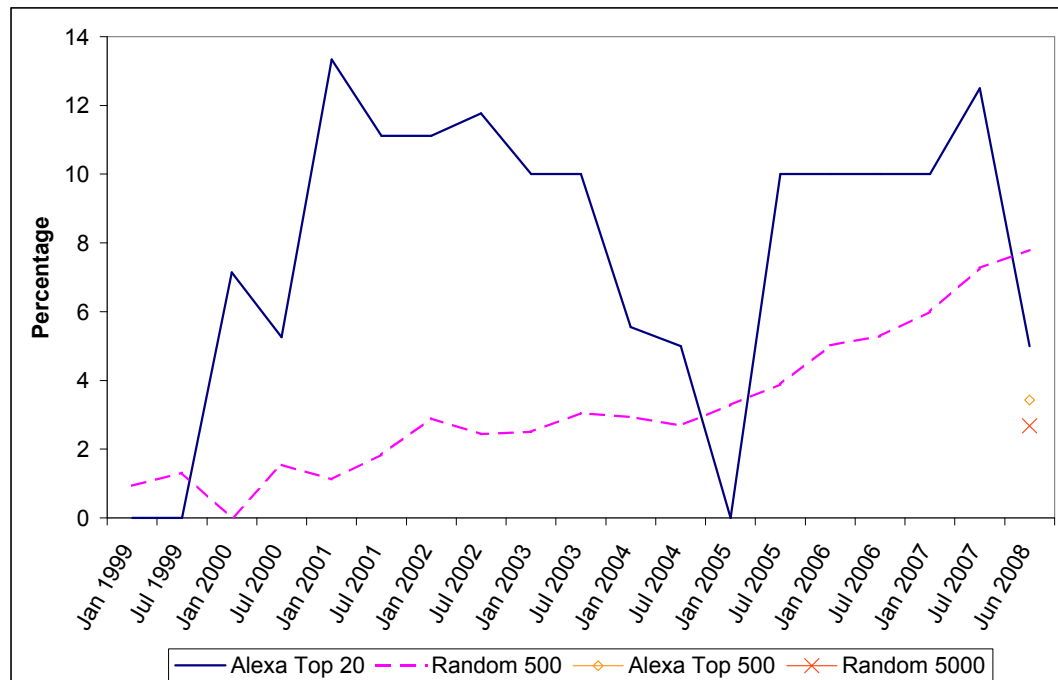
Table 4 showed a poor conformance results collected using our method (see

Year	Alexa Top 20		Alexa Top 500		Random 500		Random 5000	
	Pessimistic	Optimistic	Pessimistic	Optimistic	Pessimistic	Optimistic	Pessimistic	Optimistic
Jan 1999	0	0	-	-	0	0.94	-	-
Jul 1999	0	0	-	-	0	1.32	-	-
Jan 2000	0	7.14	-	-	0	0	-	-
Jul 2000	0	5.26	-	-	0	1.55	-	-
Jan 2001	0	13.33	-	-	0	1.13	-	-
Jul 2001	0	11.11	-	-	0	1.84	-	-
Jan 2002	0	11.11	-	-	0	2.89	-	-
Jul 2002	0	11.76	-	-	0	2.43	-	-
Jan 2003	0	10	-	-	0.21	2.51	-	-
Jul 2003	0	10	-	-	0.22	3.05	-	-
Jan 2004	0	5.56	-	-	0	2.93	-	-
Jul 2004	0	5	-	-	0	2.69	-	-
Jan 2005	0	0	-	-	0	3.29	-	-
Jul 2005	0	10	-	-	0	3.89	-	-
Jan 2006	0	10	-	-	0	5.02	-	-
Jul 2006	0	10	-	-	0	5.29	-	-
Jan 2007	0	10	-	-	0	6	-	-
Jul 2007	0	12.5	-	-	0	7.27	-	-
Jun 2008	0	5	0	3.43	0	7.8	0.06	2.68

**Table 4:** WCAG 1.0 conformance results in percentage for both pessimistic view and optimistic view. Notice that for both Alexa Top 500 and Random 5000 data, only June 2008 was presented, this was because these sets of data only looks at the current Web as discussed earlier.

section 3.3.2) for the pessimistic view. From the results presented by Watanabe and Umegaki [24], we suspect that more of these websites may be conform to the WCAG 1.0 guidelines, but either not all of them display their conformance on their websites or they may be in the form of plain that were displayed before the last one hundred characters. Thus our optimistic view experiment showed some more promising results that were closer to those mentioned by Watanabe and Umegaki. Pearson correlation was applied to check if there was any correlation between the Alexa top twenty websites and the random five hundred websites results, but no significant relationship was found as one would expect from figure 14. With these results, as seen in figure 14 demonstrated that the Alexa top twenty websites were quicker to be adopt by these guidelines then our random five hundred websites. It also shows that the Alexa top twenty websites led the trend for conformance, while our random five hundred websites were gradually catching up. Hence it can be concluded that more websites prefer to display their conformance via plain text,

and this can be found in any part of a webpage. After validating the historical data results with the current Web results, no increase in conformance was forecasted.



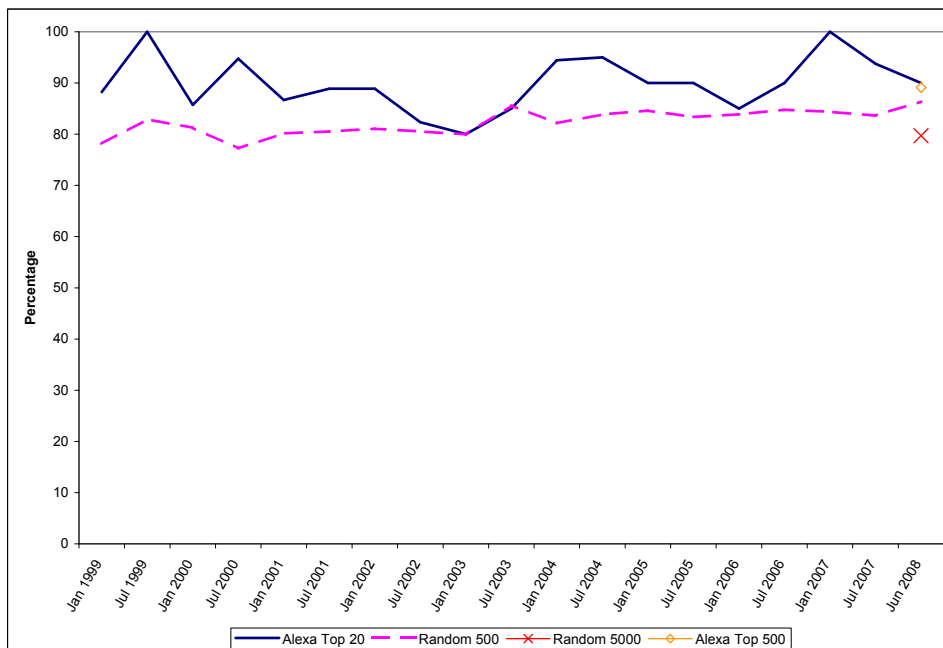
**Figure 14:** WCAG 1.0 conformance percentage (Optimistic view) for Alexa top 20 and 500 websites

These results demonstrated that little increase in the conformance to the WCAG 1.0 guidelines had been achieved for the last ten years, since May 1999 when it became a W3C recommendation. The adoption rates for WCAG 1.0 never seem to improve when comparing with the other W3C recommendations discussed earlier. A lot more research is required to understand the reasons for these trends, and more aiding tools are suggested to make these guidelines seem easier to be taken up or conformed by more websites. One of the reasons for the low conformance rate is due to the small user population that it will benefit, thus the economical benefits return is not huge. Another reason for this low conformance rate could be due to the many different types of Web content accessibility guidelines available, thus it may seem confusing, difficult, and not beneficiary for Web developers/authors to conform to them when their economical return are low. However in a recent report by Yesilada et. al., suggests that both people with or without disabilities experience similar limitations, and barriers when interacting with websites on mobile devices [25]. This claim may put forward a better case for Web developers/authors to conform to these guidelines as it will benefit a larger user population.

### 4.3 Graphical Formats Results

There are many different types of graphical formats available to be used over the Web, however as mentioned earlier, only JPEG, GIF, SVG, PNG, SMIL and Flash formats will be covered. The results collected for the individual graphical format will be discussed first, followed by the further analysis for these formats. To begin lets look the results from our analysis conducted on GIF.

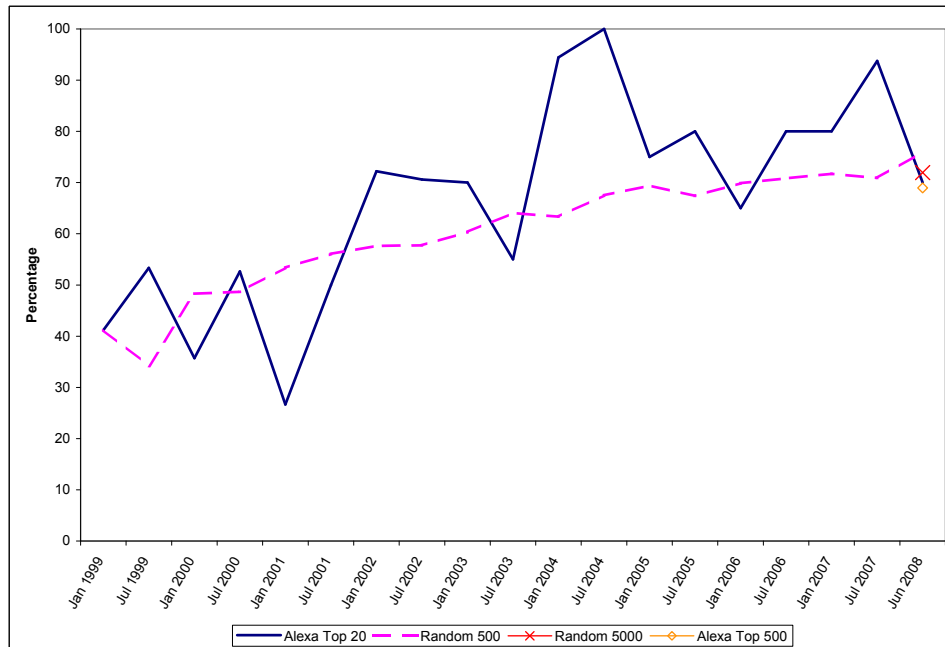
Figure 15 shows the results for the usage of GIF for our four sets of websites. Although both the random five hundred websites and the Alexa top twenty websites results displayed a gradual take up trend, but no significant correlation was found between the two sets of results when Pearson correlation was applied. Using the Alexa top five hundred websites and the random five thousand websites results for verification, the usage of GIF was expected to remain unchanged for the near future.



**Figure 15:** GIFs usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

In figure 16 shows a growth in usage trend for JPEG in the past ten years. This trend was predicted to continue as verified by the Alexa top five websites and the random five thousand websites results. Although JPEG has been implemented by Web browsers since 1996, but it took three years for more than fifty percent of the Alexa top twenty websites to have an adopt it. The lag between the Alexa top twenty websites results and the random five hundred websites results were quite small, it took only around one year later for more than fifty percent of the random five hundred websites use the JPEG graphical format. To analyse if a correlation exist between the top websites and the random websites, Pearson correlation was applied. There was a significant relationship between the Alexa top twenty websites

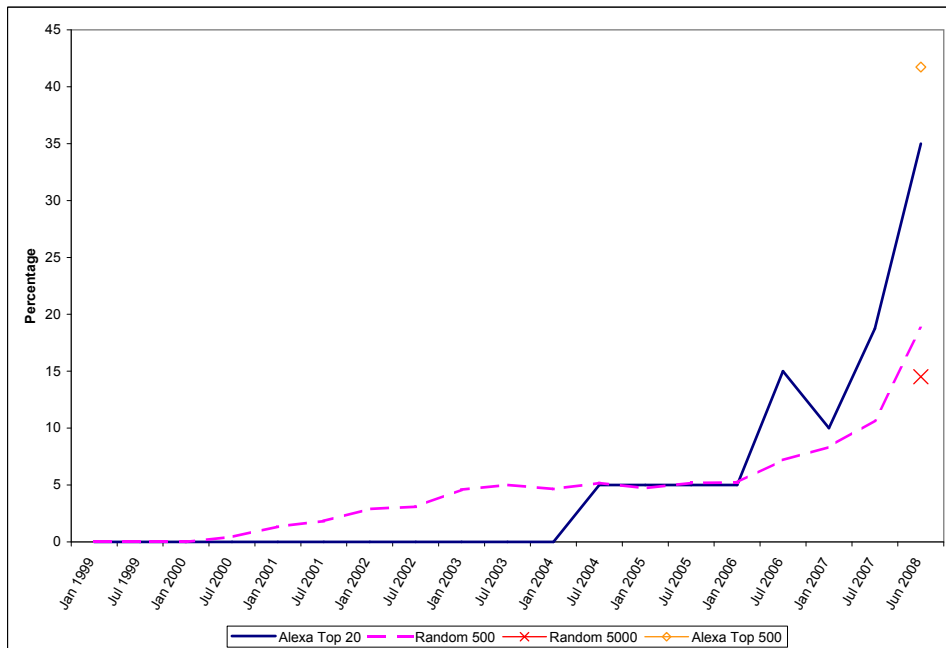
and the random five hundred websites results,  $r = .67$ ,  $p$  (two-tailed)  $< .01$ .



**Figure 16:** JPEGs usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

PNG became a W3C recommendation since October 1996, and the second edition was released in November 2003. As shown in figure 17, the random five hundred websites initial take up this graphical format much earlier then the Alexa top twenty website. However after the release of the second edition, the Alexa top twenty websites quickly pick up, and led the adoption trend for this graphical format. When Pearson correlation was applied to the two sets of data, a significant relationship between the Alexa top twenty websites and the random five hundred websites results was observed,  $r = .93$ ,  $p$  (two-tailed)  $< .01$ . Hence from this experiment, it can be seen that both the Alexa top twenty websites and the random five hundred websites learned from each others usage trends. From the results of the Alexa top five hundred websites and the random five thousand websites, it shows a growing trend for the usage of PNG. Based on our current Web analysis for the Alex top five hundred websites and the random five thousand websites, this trend is predicted to continue. The increase in usage for this type of graphical format may not be the results of other technologies, but by the capability of this type of format itself. Hence after the release of its second edition a significant increase in usage was observed.

The SVG format became a W3C recommendation from in 14 January 2003, but from the graphs shown in figure 18, a poor adoption rates were observed. Due to the poor results analysed, it was not justifiable to conduct a correlation test with these results. However looking that the Alexa top five hundred websites and the random five thousand websites results, together with the results for the random

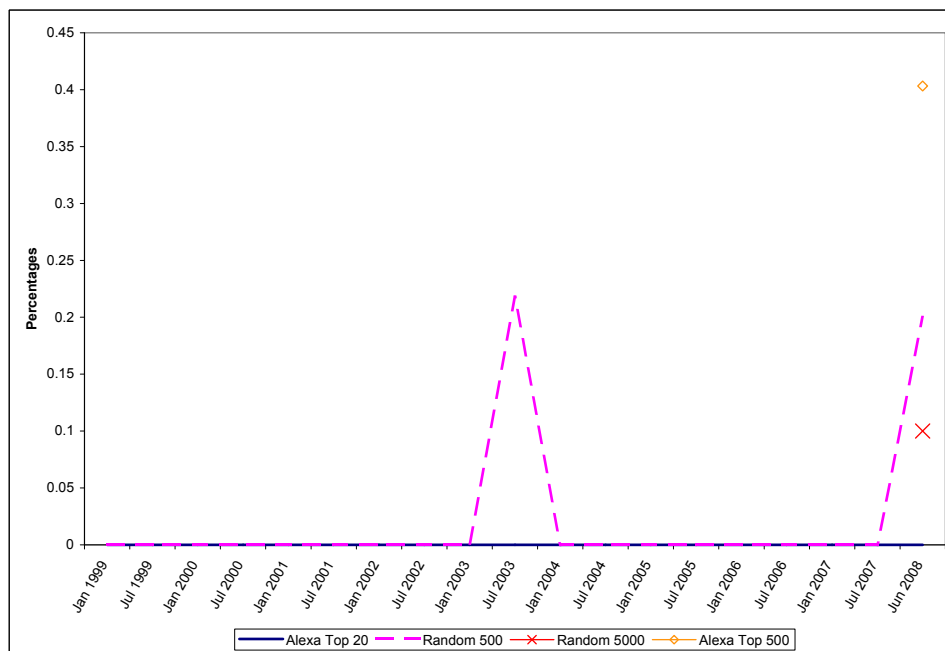


**Figure 17:** PNGs usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

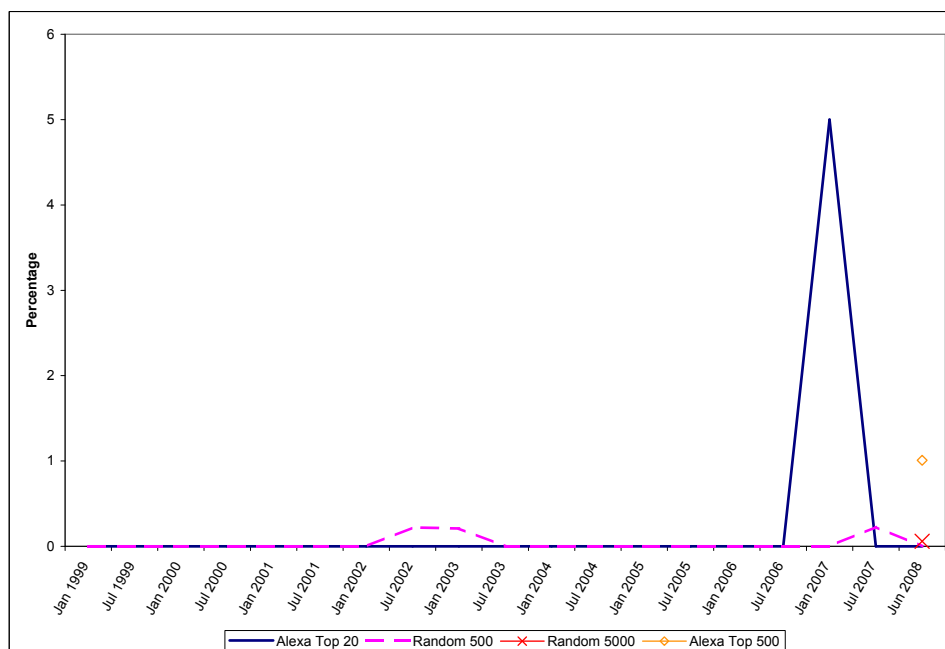
five hundred websites in June 2008, some websites has started to use this type of graphical format. Further research is required to identify if this shows an increase in this type of graphical format or is it just some random websites trying out this type of graphical format. As observed from the other graphical formats trends, a significant increase will be expected after its revised or second edition is released.

In order to synchronise multimedia over the Web, the W3C had introduced the SMIL format as a recommendation for synchronised multimedia. From figure 19 one can notice from the graphs that the usage of this type of format is very poor, and a few attempts by websites to take up this format can be observed. Using the data from the Alexa top five hundred and top twenty websites, together with the random five hundred and five thousand websites, no increase in usage was expected. Again due to the poor usage, no correlation test was conducted as it was not justifiable.

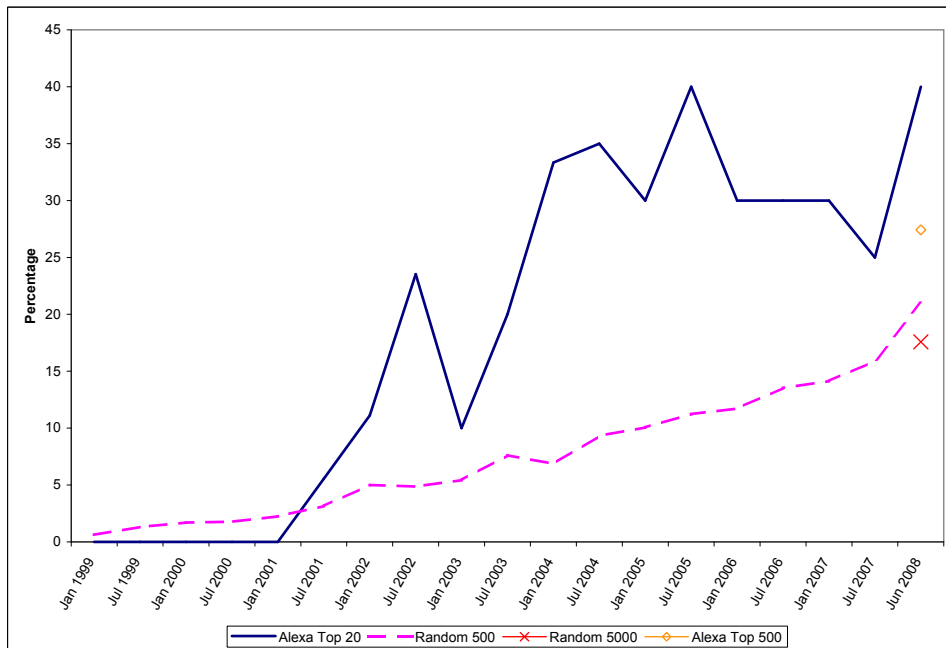
The last type of graphical format covered in this study is Flash. From figure 20, the usage of this for this type of graphical format for the last ten years were presented. A steady growth in the usage was noticed from Alexa top twenty websites and the random five hundred websites results. Based on the Alexa top five hundred websites and the random five thousand websites results to verify these claims, a continuous growing trend was forecasted. Looking at the graphs a correlation between the Alexa top twenty websites and the random five hundred websites results was suggested. Hence Pearson correlation was applied, and a significant relationship between the Alexa top twenty websites and the random five hundred websites results was noticed,  $r = .84$ ,  $p$  (two-tailed)  $< .01$ .



**Figure 18:** SVGs usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites



**Figure 19:** SMILs usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites



**Figure 20:** Flash usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

From the above results and discussions, based on the correlation results, besides GIF, all the other graphical format had a significant relationship was noticed between the Alexa top twenty websites and the random five hundred websites results. On the average it will take about two years for a new graphical format to get adopted. These results demonstrated that when analysing a graphical format usage trend, the Alexa top websites do gives a good indication of how the random Web is evolving.

Further analysis was also done to see how the different graphical formats fair against each other. Figure 21 plots the results collected of the different graphical formats results from the Alexa top twenty websites. No relationship was noticed between the different graphical formats, but when observing the Alexa top twenty websites results, it was noticed that it is more likely for a graphical format to be used side by side with another graphical format than to be replaced.

A similar analysis was also done for the random five hundred websites data set for the different types of graphical formats. As expected, a similar trend was noticed when the results were collected for the Alexa top twenty websites. However again no relationship was noticed between the different graphical formats, and it was more likely that a new graphical format will be used side by side with existing graphical format than to be replaced. However this set of results were more consistent since a larger set of websites were examined.

To conclude this section for the results and discussions for the different types of graphical formats, it can be observed that the Alexa top websites do give a good representation of how the Web in-general was evolving for this type of analysis. On

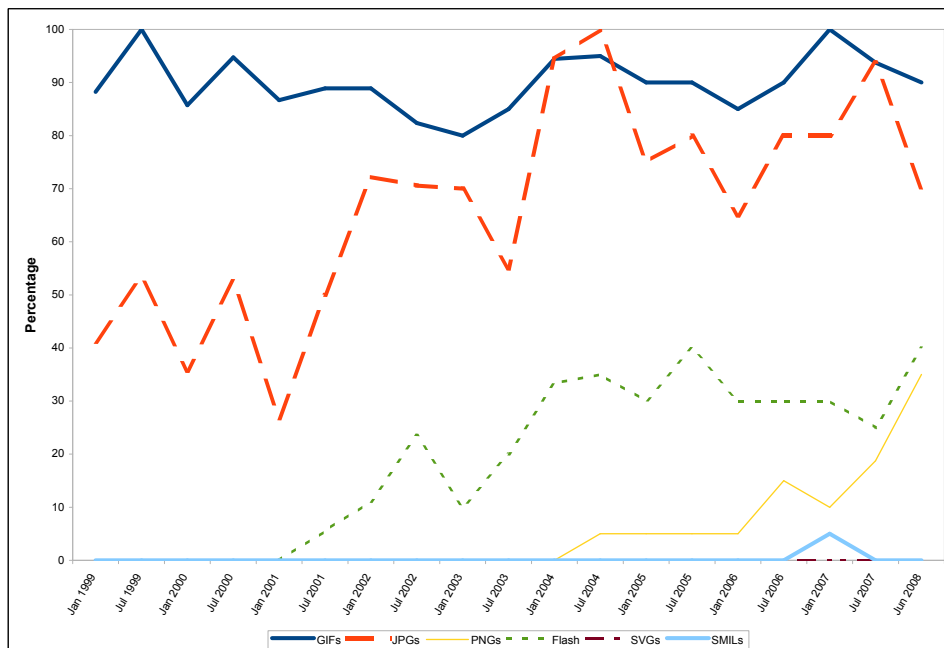


Figure 21: Graphical usage percentage for Alexa top 20 websites

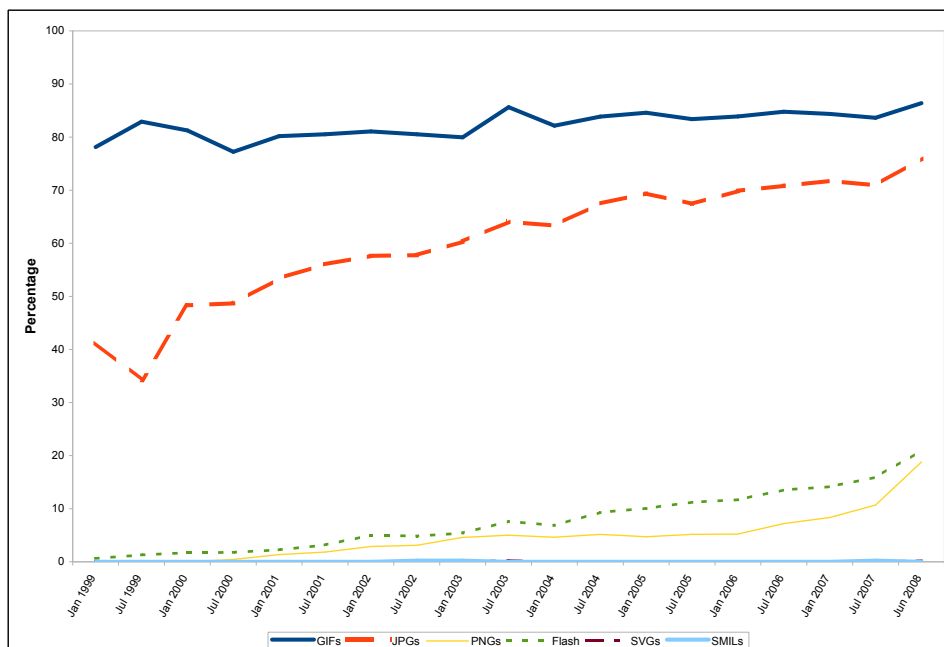
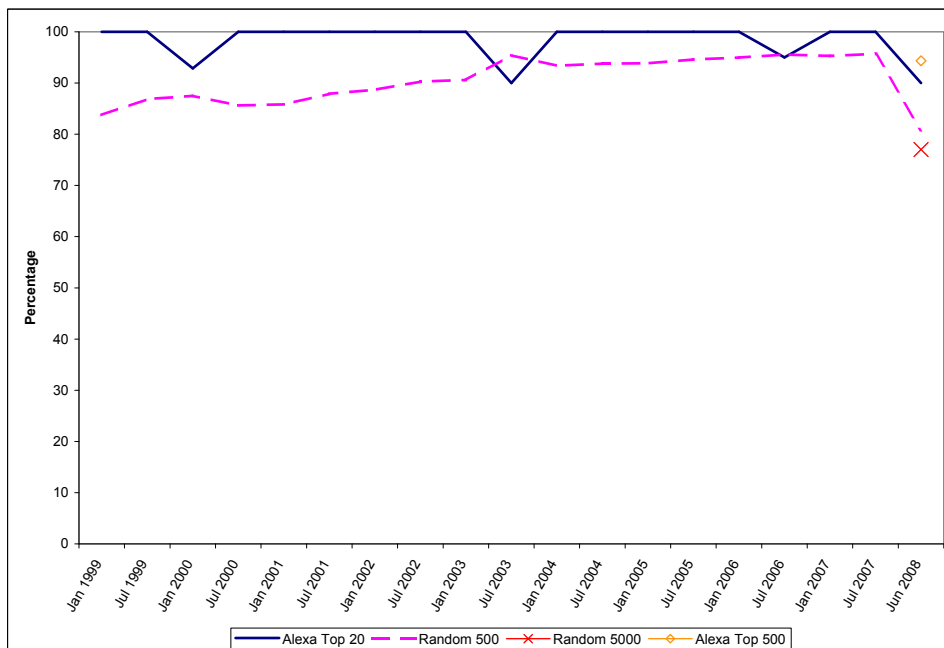


Figure 22: Graphical usage percentage for random 500 websites

the average it will take about two years for a new graphical format to get adopted by the Web from the time it was released. It is also more likely for a graphical format to get adopted, and to be used side by side with the older formats, then to be used as a replacement. Finally the Alexa top websites and the random websites do learn from each other trends when it comes to taking up new graphical formats as seen previously.

#### 4.4 Client-side Scripting Results

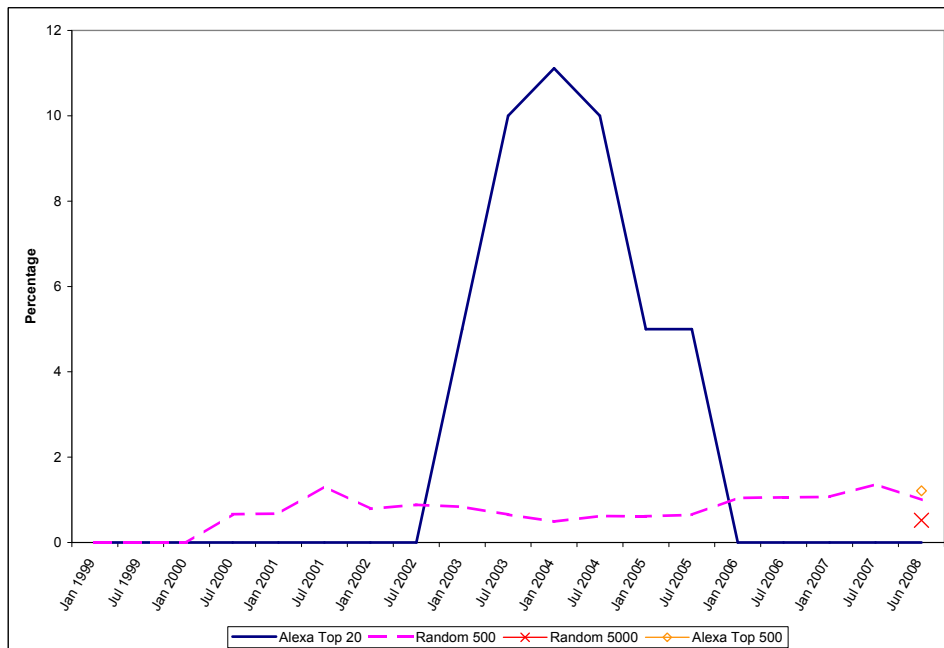
Client-side scripting plays a vital row in Web development. As discussed earlier it provides the means for Web developers/authors to control the appearance of the website, and to reduce servers work load that will help to make better use of the network traffic. Two types of client-side scripting were covered in this study; JavaScript and VBScript. Beginning with JavaScript lets looked at our analysis results in figure 23.



**Figure 23:** JavaScripts usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

It was observed that most of the Alexa top twenty websites uses some JavaScript in their website design, but a decline in usage was also noticed since July 2007. Our random five hundred websites results also demonstrated similarly trend. By using our Alex top five hundred websites and the random five thousand websites results to validate these claims, a continuous decline in JavaScript usage was predicted. Further analysis is required to understand more about what causes this decline. Some of the possible analysis will be discussed later in section 4.7.

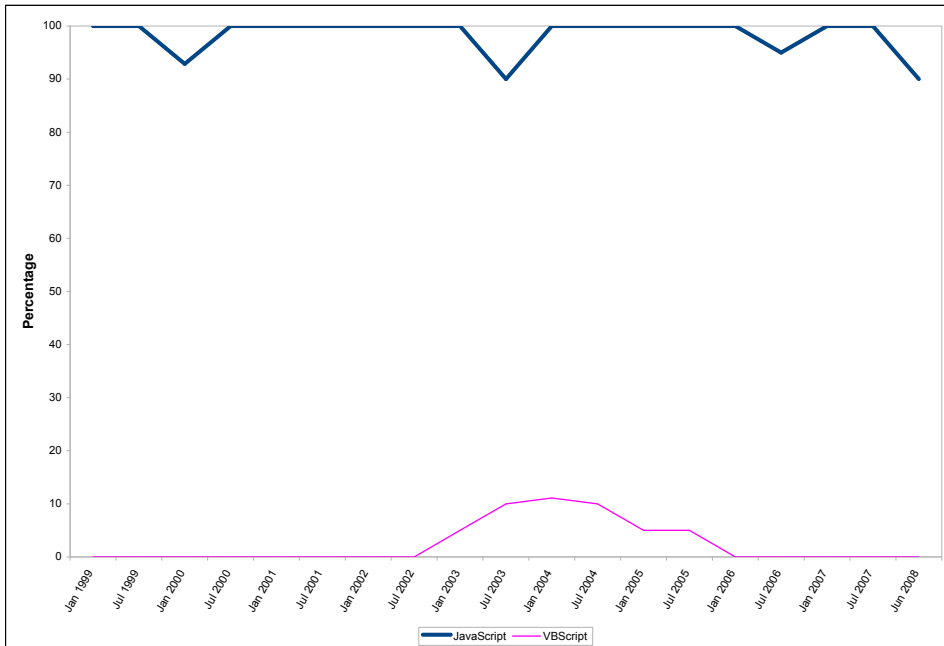
VBScript is the other client-side scripting language covered in this study. Commonly VBScripts will only be executed when run in Microsoft Internet Explorer. Due to its poor adoption rate by other user-agents, poor usage by Web developers/authors is expected for this type of client-side scripting.



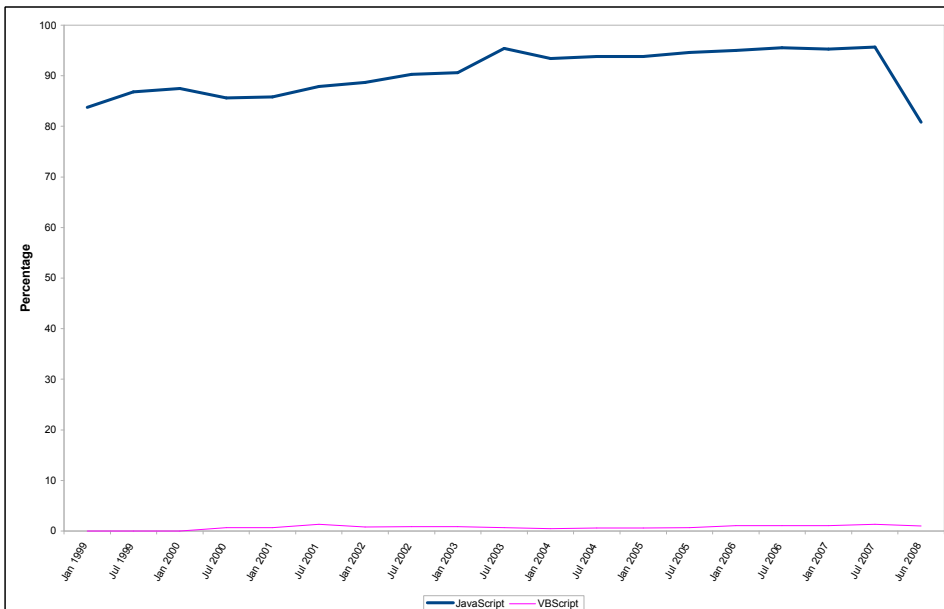
**Figure 24:** VBScripts usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

The graphs shown in figure 24 reflected our view based on both the Alexa top twenty websites and the random five hundred websites results. This type of client-side scripting language never seem to get adopted by the Alexa top twenty websites, even though a pick up in usage was notice between July 2002 and January 2006. Initially a gain in its popularity was noticed in year 2000 from the random five hundred websites results, but its usage percentage remained almost the same after that. Based on the results from the Alexa top five hundred websites and random five thousand websites, a similar percentage of usage for VBScript was expected for the near future, but no significant increase for foreseeable.

To conclude this section, figure 25 and figure 26 demonstrated a huge difference between the usage of JavaScript and VBScript for both the Alexa top twenty websites and the random five hundred websites results. When comparing VBScript with JavaScript, VBScript seems to never get adopted by Web developers/authors. However it was notice that the usage of JavaScript for both sets of data had begin to roll off since July 2007, hence further analysis such as plotting JavaScript with AJAX, is required to understand more about the reason for this trend.



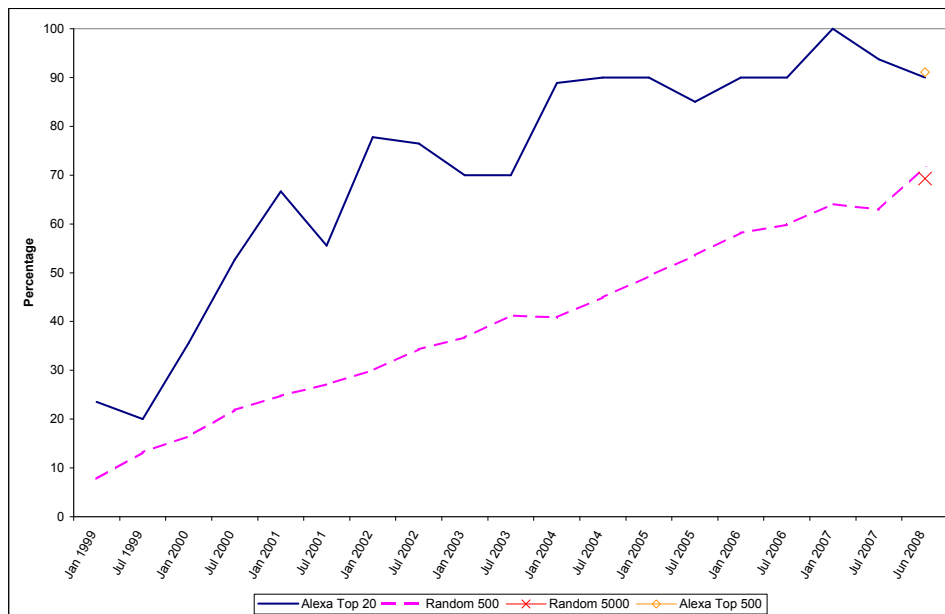
**Figure 25:** Scripting usage percentage for Alexa top 20 websites



**Figure 26:** Scripting usage percentage for random 500 websites

## 4.5 Cascading Style Sheets

The information collected for the styling of Web content analysis was done for CSS in general. This was because we were looking at the general CSS usage and not the specific version of CSS usage. Figure 27 shows the results in percentage for the usage of CSS for all the four sets of data collected. A steady growth was noticed and we predict that this growth will continue for the next year. When analysing the trend for the usage of CSS to be more than 50%, one can notice from Figure 1 and figure 27 that it took the Alexa top twenty websites about four years to achieve it, and the random 500 websites about nine years for it to cross the more than 50% mark. Although both showed similar trends, a four years lag was noticed between the top websites and the random websites. Hence for this type recommendations to get adopted by more than 50% of the Web, an adoption time of about four to nine years is required, and less than four years for Web technologies that surrounds it to get developed. When using Pearson correlation, a significant relationship was noticed between the Alexa top twenty websites and the random five hundred websites,  $r = .89$ ,  $p$  (two-tailed)  $< .01$ .

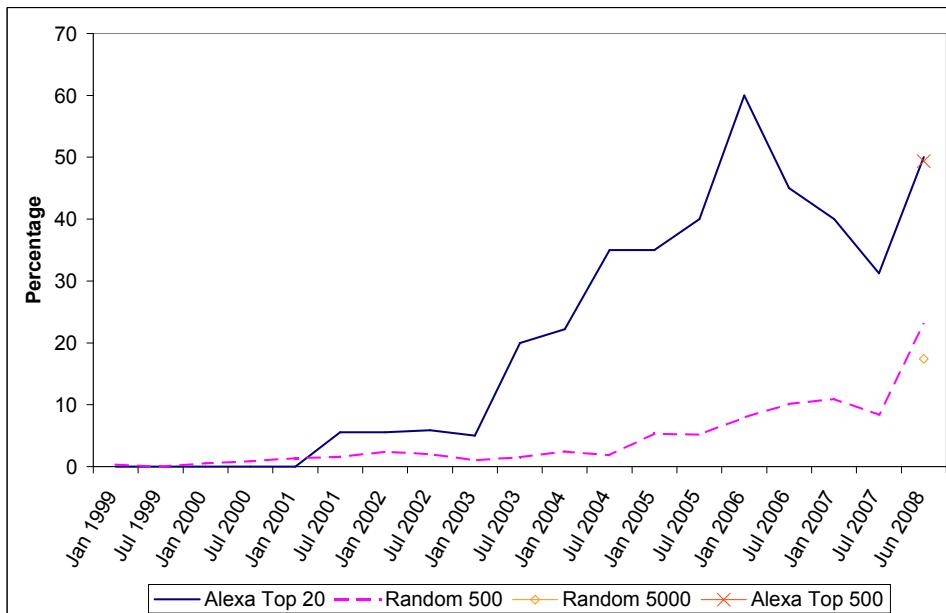


**Figure 27:** CSS usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites

Styling of plain text can control the way how a webpage is presented in a user-agent, however the data structure of the plain text is equivalently important to realised its full capability. The individual standards will be discussed first before covering the further analysis surrounding these standards.

## 4.6 AJAX

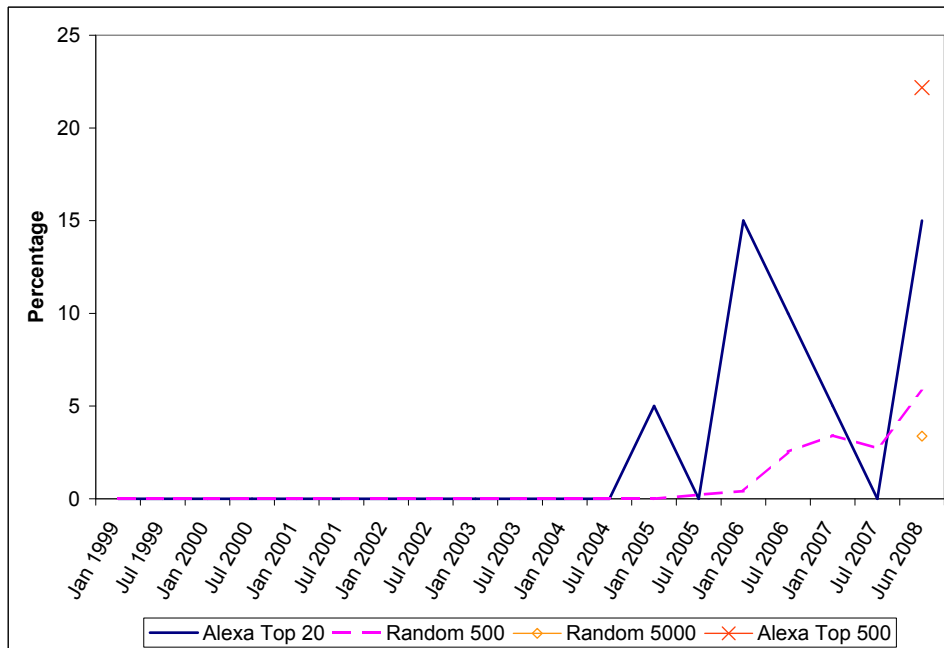
AJAX is a model created to take advantage of the popularity and capability of JavaScript, the asynchronous technology, and XML. Using the methodology discussed earlier in section 3.3.6, data were extracted to identify the usage of AJAX. In figure 28 it shows that a growing usage trend for the AJAX model using the combined result of the iFrame element and the XMLHttpRequest object. The Alexa top twenty websites led the way in the usage of AJAX, while the random five hundred websites grew in popularity gradually. Pearson correlation was applied to these results for a correlation test, and a significant relationship between them was noticed,  $r = .75$ ,  $p$  (two-tailed)  $< .01$ .



**Figure 28:** AJAX detection based on the combination of iFrames and XMLHttpRequest usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites.

The analysis break down for the two methods used to detect the usage of AJAX were presented in figure 29 and 30. The first analysis was done by searching for the usage of the XMLHttpRequest object within the JavaScript source code, and the second analysis was done by searching the use of iFrame elements within the HTML code.

The usage results for AJAX detection using the XMLHttpRequest object within JavaScript was presented in figure 29. It shows that the Alexa top twenty websites led the trend while the random five hundred websites exhibited a similar trend. Pearson correlation was used to check if the both sets of results have any correlation. There was a significant relationship between the both sets of results,  $r = .64$ ,  $p$  (two-tailed)  $< .01$ . These trends were verified using the Alexa top five hundred websites



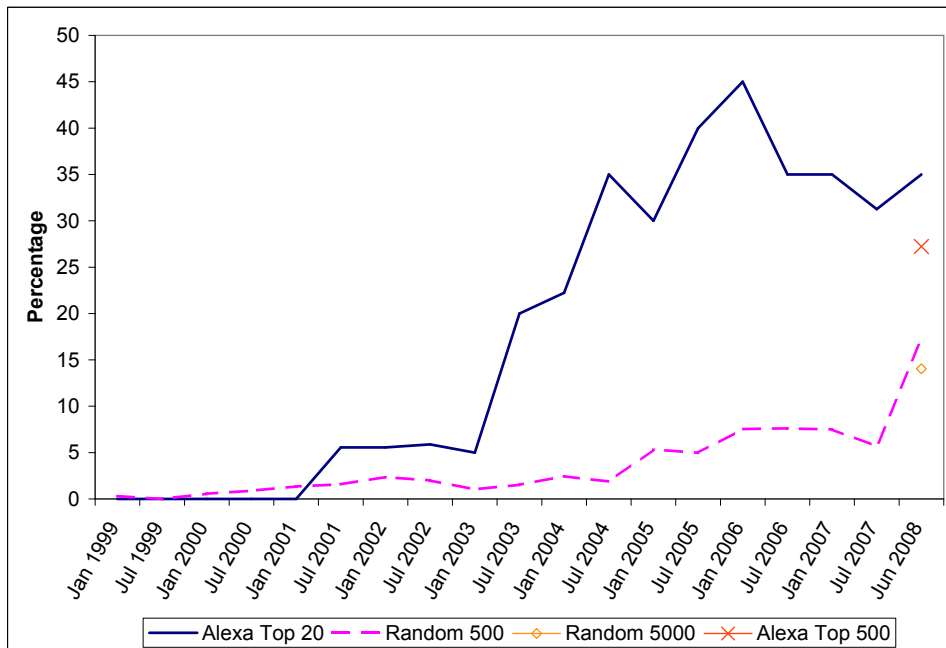
**Figure 29:** AJAX detection based on the XMLHttpRequest usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites.

and the random five thousand websites results. Using these results a forecasted increase in the XMLHttpRequest object usage was concluded.

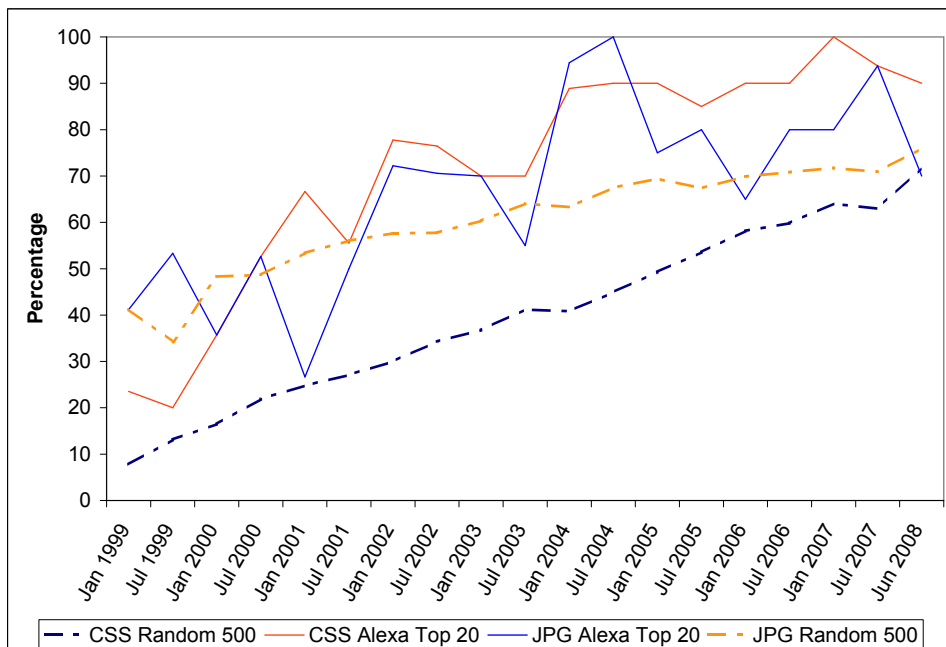
The next analysis for AJAX detection was searching for the use of iFrame element(s) within the HTML code. These results showed that initially the random five hundred websites led the trend, but the Alexa top twenty websites were quick to pick up, and eventually surpassing the random five hundred websites to lead the usage trend. Applying Pearson correlation to the both sets of results gave a significant relationship between them,  $r = .70$ ,  $p$  (two-tailed)  $< .01$ . Based on the Alexa top five hundred websites and the random five thousand websites results, a gradual increase in usage of iFrame element(s) can be expected.

## 4.7 Further Analysis

Now that we had discuss all the results from the individual standards and recommendations analysis, further analysis were also conducted to understand more about the reasons behind some of these standards and recommendations usage trends. A few analysis were suggested to be done so that better understanding between the relationships and reason behind these standards and recommendations trends. The first analysis was done between CSS and JPEG because through visual observation the CSS (figure 27) usage seems to possess similar growth trend patterns with JPEG (figure 16). This analysis will explain the growth in usage for some graphical formats and what effected it.

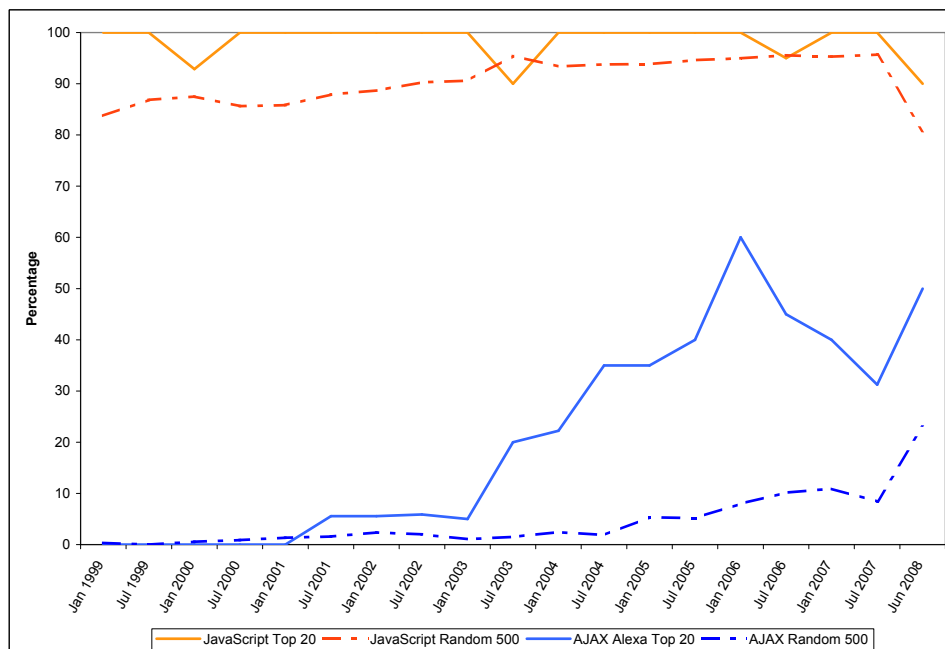


**Figure 30:** AJAX based on the iFrames usage percentage for Alexa top 20 and 500 websites, and a set of 500 and 5000 random websites.



**Figure 31:** CSS VS. JPEG percentage for Alexa top 20 and 500 websites

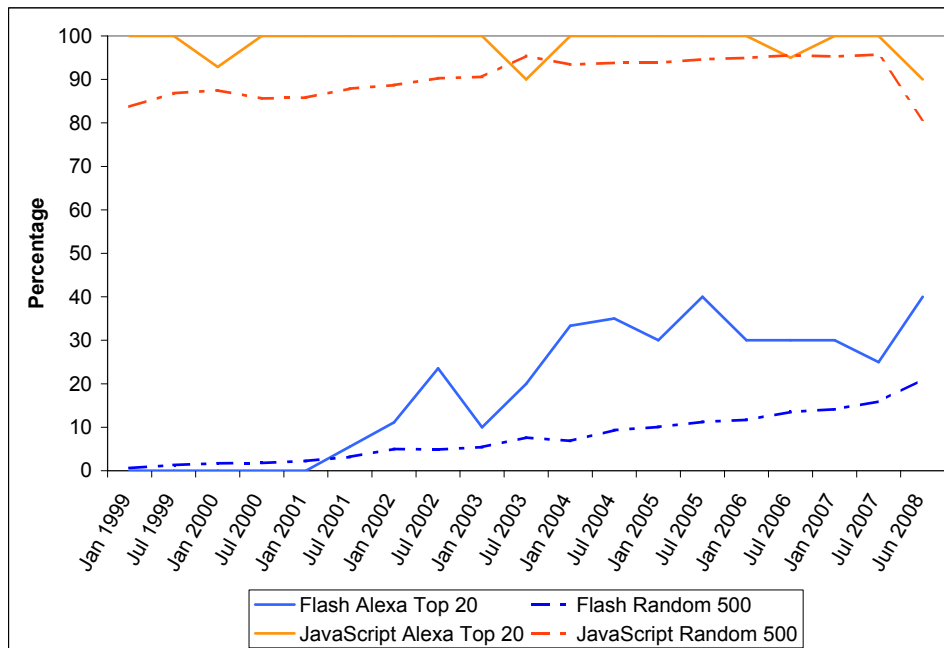
Both CSS and JPEG demonstrated similar trends in figure 31. Pearson correlation was used to determine if a correlation exist between these recommendations for the Alexa top twenty websites, and the random five hundred websites results. A significant relationship was noticed for the Alexa top twenty websites between the CSS results and the JPEG results,  $r = .75$ ,  $p$  (two-tailed)  $< .01$ . There was also a significant relationship for the random five hundred websites between the CSS results and the JPEG results,  $r = .95$ ,  $p$  (two-tailed)  $< .01$ . Hence there was clearly a significant correlation between CSS and JPEG results, and they possess similar usage trends. CSS allows the Web developers/authors the flexibility, and more control over the webpage's presentation, thus this led to an increase in the usage for different types of graphical formats such as JPEG. This trends suggest that the take up of JPEG may have benefited from the increase of the usage of CSS.



**Figure 32:** AJAX VS. JavaScript percentage for Alexa top 20 and 500 websites

Two other analysis were conducted to understand reasons for the decline of JavaScript usage trend. From these analysis, we hope to understand why introducing a model that uses existing standards, and recommendations may not necessary improves the technology's popularity. The first one was between AJAX and JavaScript as shown in figure 32. The results from this analysis gave an interesting view of JavaScript and AJAX usage trends. It can be noticed that even when the usage of AJAX was increasing, a decline in JavaScript usage was still observed. Therefore another analysis was carried out between Flash and JavaScript as seen in figure 33. Again it was noticed that around July 2007, an increase in Flash usage was noticed around the same time when JavaScript began to roll off. This analysis supplies a reason for the trend of the increase in AJAX popularity, and the con-

trary results for the usage of JavaScript trend. This showed that the AJAX model may be gaining popularity, but it may not be enjoying the full increase, but it was sharing it with another technologies such as Flash that is capable of providing the asynchronous model.



**Figure 33:** Flash VS. JavaScript percentage for Alexa top 20 and 500 websites

To conclude the section on further analysis, the increase in usage for graphical formats such as JPEG may be the benefiting from the fruits of the CSS usage increase. This is because the CSS allows the Web developers/authors more flexibility and control over the webpage presentation, thus this allows better use of graphics. The other analysis conducted surrounding AJAX conclude that it is not enjoying the full popularity of the asynchronous technology revolution, but it is sharing it with other Web technologies such as Flash. Since the roll off of JavaScript has just began in 2007 further research will be required to determine if this is a true decline or was it just a dip in usage.

#### 4.8 Analysis Overview

The discussions on the analysis presented above highlights and predicts possible trends for individual standards, and recommendations when done separately. However when plotted against each other, further understanding for the reasons behind these trends were better explained. As discussed earlier, the Alexa top websites does give a good representation of the random Web when analysing the W3C standards and graphical formats. From the analysis results done on the Web content accessibility conformance, no increase in conformance for the WCAG 1.0 guidelines was

forecasted. Further analysis were conducted to understand more about the reasons behind some of these trends such as AJAX, CSS, Flash and JPEG. It was noticed that more websites were taking up the asynchronous model, but this trend was shared between AJAX and Flash. CSS has given the Web the flexibility and the control over the presentation of Web contents. Due to this, a usage growth for this standard was predicted, and graphical formats such as JPEG usage also benefiting from it.

## References

- [1] A snapshot of cyberspace. *Library of Congress Bulletin*, 57(11), November 1998.
- [2] Alexa Internet, Inc. Alexa company info. <http://www.alexa.com/site/company>, July 2008.
- [3] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The connectivity sonar: detecting site functionality by structural patterns. In *HYPertext '03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pages 38–47, New York, NY, USA, 2003. ACM.
- [4] BBC News. Fifteen years of the web. <http://news.bbc.co.uk/go/pr/fr/-/2/hi/technology/5243862.stm>, August 2006.
- [5] T. Berners-Lee. WWW: past, present, and future. *Computer*, 29(10):69–77, 1996.
- [6] K. Chandra, S. S. Chandra, and S. S. Chandra. A comparison of VBScript, Javascript, and Jscript. *J. Comput. Small Coll.*, 19(1):323–335, October 2003.
- [7] A. Q. Chen. Web evolution: Code and experimental guide. Technical report, The University of Manchester, September 2008.
- [8] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pages 200–209, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- [9] J. Clark. *Building Accessible Websites*. New Riders Publishing, United States of America, 2003.
- [10] Connolly. HTML 3.0 draft. <http://www.w3.org/MarkUp/html3/>, December 1995.
- [11] D. Connolly. <http://www.w3.org/MarkUp/html-spec/>, September 1999.
- [12] F. Douglass, A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of change and other metrics: a live study of the world wide web. In *USITS'97: Proceedings of the USENIX Symposium on Internet Technologies and Systems on USENIX Symposium on Internet Technologies and Systems*, pages 14–14, Berkeley, CA, USA, 1997. USENIX Association.

- 
- [13] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 669–678, New York, NY, USA, 2003. ACM Press.
- [14] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903, New York, NY, USA, 2005. ACM Press.
- [15] S. Harper. Web evolution and its importance for supporting research arguments in web accessibility. In *Web Science Workshop (WSW2008)*. International World Wide Web Conference, 2008.
- [16] S. Lawrence and L. C. Giles. Accessibility of information on the web. *Nature*, 400(6740):107–107, July 1999.
- [17] J. Meiert. HTML elements index. <http://meiert.com/en/indices/html-elements>, June 2008.
- [18] A. Ntoulas, J. Cho, and C. Olston. What’s new on the web?: the evolution of the web from a search engine perspective. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 1–12, New York, NY, USA, 2004. ACM Press.
- [19] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, October 2001.
- [20] S. Pemberton and et al. XHTML 1.0 (second edition). <http://www.w3.org/TR/xhtml1/>, August 2006.
- [21] D. Raggett. HTML 3.2 reference specification. <http://www.w3.org/TR/REC-html32>, Jan 1997.
- [22] J. T. Richards and V. L. Hanson. Web accessibility: a broader view. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 72–79, New York, NY, USA, 2004. ACM.
- [23] M. Toyoda and M. Kitsuregawa. Extracting evolution of web communities from a series of web archives. In *HYPertext '03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pages 28–37, New York, NY, USA, 2003. ACM Press.
- [24] T. Watanabe and M. Umegaki. Capability survey of japanese user agents and its impact on web accessibility. In *W4A: Proceedings of the 2006 international cross-disciplinary workshop on Web accessibility (W4A)*, pages 38–48, New York, NY, USA, 2006. ACM.
- [25] Y. Yesilada, A. Chuter, and S. L. Henry. Shared web experiences: Barriers common to mobile device users and people with disabilities (table format). <http://www.w3.org/WAI/mobile/experiences>, July 2008.

## A Data Corpus

### A.1 Data Corpus

The storage of the downloaded data were organised in an fashion according to its data set in separate folders under the “db” directory. Table 5 is the list of the folders and its respective data set.

Folder Names	Data Sets
alexaTop20_webpages	Alexa top 20 websites
alexaTop500_webpages	Alexa top 500 websites
Random500_webpages	Random 500 websites
Random5000_webpages	Random 5000 websites

**Table 5:** Data corpus folders

These data includes the website HTML content and its related external files such as JavaScript and CSS. Hence a few types of file extensions can be found in these folders. Table 6 presents a list of the possible type of files and its extensions.

Type of files	File extension
HTML content	.aaf
External JavaScript	.js
External CSS	.css

**Table 6:** Possible data file extensions