Web 1.0 to Web 2.0: an observational study and empirical evidence for the historical r(evolution) of the social web

This item was submitted to Loughborough University's Institutional Repository by the/an author.

Citation: SYKORA, M.D., 2017. Web 1.0 to Web 2.0: an observational study and empirical evidence for the historical r(evolution) of the social web. International Journal of Web Engineering and Technology, 12 (1), pp. 70-94.

Additional Information:

- This paper was accepted for publication in the journal International Journal of Web Engineering and Technology and the definitive published version is available at http://dx.doi.org/10.1504/IJWET.2017.084024.

Metadata Record: [https://dspace.lboro.ac.uk/2134/25153](https://dspace.lboro.ac.uk/2134/25153)

Version: Accepted for publication

Publisher: © Inderscience

Rights: This work is made available according to the conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. Full details of this licence are available at: [https://creativecommons.org/licenses/by-nc-nd/4.0/](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Please cite the published version.
WEB 1.0 TO WEB 2.0: AN OBSERVATIONAL STUDY AND EMPIRICAL EVIDENCE FOR THE HISTORICAL R(EVOLUTION) OF THE SOCIAL WEB

Author Name*
Institutional Address
*corresponding address / email

ABSTRACT
Applications such as Twitter, Facebook or Youtube have taken the internet community by storm and have literally initiated a revolution in online communication. These social media applications are also often referred to by the general, somewhat vague but yet eloquent term – Web 2.0. The term implies a perceived second generation of World Wide Web, i.e. Web 1.0 → Web 2.0. Nevertheless the reference to a "second" version of the Web is misleading, since there wasn't any specific technical update of the World Wide Web. Hence, this paper empirically explores social media in a historical context as it has evolved and driven a revolution in online communication over the last decade. A unique historical dataset, made available by the Wayback Machine Internet Archive project is employed in order to provide an accurate historical record of the primary web design and related elements that drove the evolution towards a more social and interactive web. This study presents a unique contribution to related academic literature.

KEYWORDS
Web 2.0; World Wide Web; Social Media; History of the Web; Web Engineering; Evolution of the Web; Information Analysis; Information Retrieval

1. INTRODUCTION

Despite some ongoing criticism of Web 2.0 (Zimmer 2008), the term and concept behind it provides a valuable insight into the way in which Web has evolved over time (Millard and Ross 2006). Alexander (2006) suggested that, “the term Web 2.0 assumes a certain interpretation of Web history, including enough progress in certain direction to trigger a succession [i.e. Web 1.0 → Web 2.0]”. The most recent generation of websites have been considered by some to be fundamentally different from the ones found on the early Web (see section 2.1), these have been grouped together under the term Web 2.0, and now widely recognised as social media. The name is arguably misleading, since it implies a designed version and a discrete evolution, although presumably the emergence of Web 2.0 applications did not occur as a sharp break with the old but, rather, the gradual emergence of a new type of practice. However, the concepts behind it provide a valuable insight to the simple observation that the Web and how it’s used has evolved over time. Within this paper an original retrospective and historical observational study of seven major websites is undertaken in order to investigate, what it is that has changed on actual websites over time. Given the current, interactivity and pervasive social nature of the web (Gustafson 2015), this study specifically avoids an analysis of the most recent years. Instead the focus of the historical analysis is on the last decade, which is the period during which most of the changes and uptake of social and other Web 2.0 features took place (O’Reilly 2005). The study is looking back as far as the data-source permits, reaching all the way back to the late 90s. Social media has arguably changed the web and its communities forever (Shirky 2010), and although there is now talk of Web 3.0 and even Web 4.0 (Aghaei et al. 2012), currently the impact of Web 2.0 is probably felt the most and importantly the changes during the last decade have come to define the nature of the web today. An understanding and historical record of the primary web design techniques that drove the evolution towards a more social and interactive web is a unique contribution to current academic literature and this is the first kind of its study, as far as the author is aware.

The rest of this paper is organized as follows: in section 2, background literature and relevant historical
overview is provided, section 3 discusses the framework and methodology employed for the empirical study. Section 4 presents results, with further discussion and limitations in section 5, and finally section 6 concludes this paper.

2. BACKGROUND

Web 2.0 stands for the paradigm that a considerable portion of the Web has become read and write (Vosen and Hagemann 2007), where instead of a single agent (i.e. administrator) updating a website, everybody can now interactively update, upload and collaborate on content of webpages (Tapscott and Williams 2008). The term Web 2.0 was introduced by the president of O'Reilly Media (Tim O'Reilly) in his influential blog post in 2005 (O'Reilly 2005), following a spontaneous conference debate at which O'Reilly and other industry and academic domain experts discussed the changes in development and usage of World Wide Web taking ubiquitous hold at the time. Their conclusion was that the World Wide Web is becoming a collaborative platform, and Web 2.0 could be understood to represent this platform in its own right. In other words, collaboration platforms needed little to no human intervention as they became mostly automated thanks to wide adoption of database backed programming techniques, asynchronous client-server communication (AJAX), or HTML / CSS standardization among other things. The original definition was given by example1 which makes it intrinsically subjective, and despite all his [O'Reilly's] efforts it became frequently misunderstood. Still O'Reilly managed to convey main elements of the “new” conventions and principles expected in web-sites that would be characteristic of a symbolic second generation of the Web. Essentially his Web 2.0 is based on a set of seven principles which the reader can also study further in O'Reilly (2005), and Anderson (2007). The primary principle that resonates throughout the definition is that previously most systems were characterised by a small number of pre-approved participants and were not fully interactive, while Web 2.0 by contrast is characterised by online communities, open and very easy sharing, interactivity and collaboration (O'Reilly 2005, Vosen and Hagemann 2007, Shirky 2009). While technical aspects such as shorter release cycles, using innovative approaches of perpetual beta testing, and lightweight programming models, based on RSS, Rest APIs and other loosely coupled systems were considered important side developments.

2.1 Early Days of the WWW

Tim Berners-Lee, the innovator of the World Wide Web2, in an IBM developer works interview (Laningham 2006) when asked what opinion he had of Web 2.03, argued that there was little need for the term at all, since [as he understood] it would imply that collaboration on a wide and interactive scale was not the original goal set out with the World Wide Web, which in fact was his intention from the beginning anyway4. At the same time in Berners-Lee and Fischetti (1999), he acknowledges that all his efforts in this direction went astray and also explains the reasons for this; “...Part of the reason, I guessed was that collaboration required much more of a social change in how people worked [...] As a medium it grew very

---

1 e.g. mp3.com → Napster / P2P, Britannica Online → Wikipedia, Personal Websites → Blogging, Ofoto → Flickr
2 Berners-Lee submitted the first proposal for the WWW during March 1989 while working as a fellow at CERN.
3 Question asked by Laningham: “You know, with Web 2.0, a common explanation out there is Web 1.0 was about connecting computers and making information available; and Web 2.0 is about connecting people and facilitating new kinds of collaboration. Is that how you see Web 2.0?”
4 Berner-Lee's answer: “Totally not! Web 1.0 was all about connecting people. It was an interactive space, and I think Web 2.0 is of course a piece of jargon, nobody even knows what it means. If Web 2.0 for you is Blogs and Wikis, then that is people to people. But that was what the Web was supposed to be all along. And in fact, you know, this Web 2.0, means using the standards which have been produced by all these people working on Web 1.0 [...] So Web 2.0 for some people it means moving some of the thinking client side so making it more immediate, but the idea of the Web as interaction between people is really what the Web is. That was what it was designed to be as a collaborative space where people can interact. Now, I really like the idea of people building things in hypertext, [...] and I think that Blogs and Wikis are two things which are fun, I think they’ve taken off partly because they do a lot of the management of the navigation for you and allow you to add content yourself.”
global and became more a publication medium but less of a collaboration medium.". It took many years for the Web to evolve into what it is today, and even though human nature is a social and collaborative one by default (Shirky 2010), the collaborative usage of World Wide Web did not really take hold until the early 2000s. This is not to say that earlier efforts didn't exist.

Here is an excerpt (Berners-Lee and Fischetti 1999) describing Berners-Lee's efforts for a read and write Web in the very early 90s: “Although browsers were starting to spread, no one working on them tried to include writing and editing functions. There seemed to be a perception that creating a browser had a strong potential for payback, since it would make information from around the world available to anyone who used it. Putting as much effort into the collaborative side of the Web didn’t seem to promise that millionfold multiplier.” (on p. 61), he further concludes; “Without a hypertext editor people would not have the tools to really use the Web as an intimate collaborative medium.” It is noteworthy that as early as 1995 there were efforts by W3C to understand the Web in terms of its collaborative potential – a workshop on WWW and Collaboration (W3C 1995) was held. Unfortunately many early efforts by W3C received little attention from the wider academic community. This was, in part, due to the nature of W3Cs’ work – a commitment to develop WWW technology standards – as adoption of these standards was exceptionally slow, especially during the 90s (Zeldman 2007). But also, in part, due to the initial experiments with WWW collaborative tools not gaining enough traction, for example Gramlich (1995) listed major technical barriers and quite simply the lack of user participation as failures to his project. It is clear that ideas behind collaborative use of WWW, i.e. what Web 2.0 stands for, were not recent but have been around for a while. Tim Berners-Lee's vision of the World Wide Web was for a tool which created, gathered and allowed to share knowledge through human interaction and collaboration. It seems that Berners-Lee himself did not quite understand how this would exactly become possible, therefore it might be useful to look at the Web as Charles Leadbeater (Leadbeater 2009) suggests – “Web 2.0 is simply a stage of development in which the Web is progressing towards this goal”.

2.2 The Uptake of Web 2.0

In many respects the burst of the 2000’s dot-com bubble had long lasting effects on the internet, ranging from economical to trust issues. Even to this date the NASDAQ index has not yet recovered to its average values from before the dot-com bust. The crisis had at least three profound effects on the World Wide Web; 1- Trust in online businesses was lost, 2- Investments were pulled out from technology companies hence innovation slowed down, 3- The companies that survived the bubble were considered to be doing something better than the rest. As O’Reilly notes in his influential blog post (O’Reilly 2005) – “the need for Web 2.0 was spawned by an apparent realisation that the dot-com collapse marked some kind of turning point for the Web”. Indeed, to some extent just jargon, but nevertheless Web 2.0 was often used to convey this new situation in which the World Wide Web will not “fail us again”, like it had in March 2000. Slight technological advances, but mainly improvements in trust, and emphasis on economically sustainable business models were behind a worldwide internet recovery. In the list below, a number of causes for the uptake in collaborative Web usage are suggested – in terms of what it is that has changed since the events of the dot com crash. The list below drives the construction of the framework for this study, presented in more detail in section 3.2.

---

5 Some early, example case-studies of first collaborative usage are reported in Tapscott and Williams (2008) and Boyd and Ellison (2007) provide a comprehensive review of the evolution of social networking sites.
6 We would like to note; Berners-Lee describes a relatively simple hypertext editor and not page edit functionalities directly plugged to storage systems of web-applications which would not become more common until late 90s.
7 For further very early work on collaboration we refer the interested reader to the relevant working group archive, accessible at http://www.w3.org/Collaboration/
8 A highly speculative period covering approximately 1995-2000 during which Internet related companies became significantly overvalued. The turning point (bubble burst) is taken to be NASDAQ’s price peak from March 10th 2000.
9 A good measure for the performance of technology companies is the NASDAQ composite stockmarket-index, due to its bias towards technology stocks.
Trust in the Web grew with better legal regulations and numerous security improvements of online cash transactions. Also the emergence and growing popularity of online shopping services such as PayPal, Amazon or Ebay and retail-store chains (Tesco, John-Lewis, etc.) had the effect of large number of online users becoming more confident in online transactions and in some way making these more commonly accepted within wider society (Hoffman et al. 1999, Chen and Barnes 2007). Finally trust also emerged due to standardisation of the underlying Web mark-up technology (HTML, CSS, etc...), which made the Web experience more consistent among browsers (Zeldman 2007).

Viability of online business models became a tangible reality (Li and Bernoff 2008). Online advertising became a business model capable of supporting practically any websites with enough visitor numbers. Assuming a common click-through rate of around 2%, most higher to average traffic websites can usually generate considerable profits purely from online advertising 10. Other Web based income generating revenue models are discussed in Lindmark (2009). Further to this, many niche and smaller businesses which built an online presence were now able to implement their proved and tested business models with online cash payment facilities, largely thanks to the increased trust into the Web. Innovative new business models also developed, such as the group buying model (Wang 2009), where a number of consumers act together in an online community to negotiate better deals and discounts for products and/or services. Finally let us not forget the overall internet economics – that is web-design, online advertising, infrastructure, database and data modeling and more recently cloud computing, and cloud data storage have made the internet industry substantial in terms of economic scale and turnover (Lindmark 2009, pp. 28-29). Also, see (Lindmark 2009, pp. 47-56) for a discussion on Web 2.0 employment and venture capital investments, both of which are substantial.

Standardisation of Web mark-up made it also possible to employ more advanced and powerful client-side scripting techniques. One of these is the well known AJAX (Asynchronous JavaScript and XML) technology which made it possible for web-pages to only send partial-page server requests to a server 11. In other words, the entire webpage in a browser did not need updating, as a small part of the page could be updated with new data from the server asynchronously based on, for example, some user activity in the rest of the page. This programming technique had a major effect on interactivity within Web applications, and perceived latency for web-pages was significantly reduced. This also meant that input / output complexity – multiple levels of input became possible.

Development became easier – productivity tools and easy to use server side scripting languages as well as client side scripting libraries appeared 12. These are often open-source and therefore initial Web development start up costs can be small. Further to this, the default settings in these frameworks allow to use many elements of Web 2.0, “out of the box”.

Architecture / Integration advances made the Web a true application platform, since these advances make it possible to loosely integrate a number of entirely separate web-pages together. At the core of this are API (Application Programming Interface) web-services technology and RSS (Really Simple Syndication) which made information exchange between websites possible, and hence these applications gained a global (above a single web-application) scope (see Schroth and Janner 2007, Wilde 2010). Thanks to these integration technologies but also to the standardisation efforts, World Wide Web content became compatible over a heterogeneous set of mobile platforms. In more technical terms, browsers act as abstractions for platform, this is extremely significant since it stipulates a shift to completely online applications – in fact this is

10 In a comScore Inc. research paper it has been shown that display based online advertising is actually very effective (Fulgoni et al. 2010). Also, according to (Slot and Frissen 2007) as of 2007, 67% of the top 137 web 2.0 websites were using advertisement as their business revenue model.

11 Frames have also been used to this effect; however, their implementation tends to be cumbersome and inconsistent. Flash (the ubiquitous plug-in from Macromedia, now Adobe), and Microsoft's Silverlight are also alternatives to AJAX. More recently HTML 5 has introduced several features which facilitate AJAX like behaviours. Google Gears technology has also provided some of this functionality.

12 Ruby (Ruby On Rails), Python (CherryPython, Django), PHP (Drupal, Wordpress), JavaScript (jQuery, Scriptalicious). In brackets are examples of open source frameworks and libraries, widely used in web 2.0 application development.
already happening with laptops, often referred to as net-books being produced partly to satisfy this new demand.

Due to the latency, interactivity, trust, standardisation and architectural developments as highlighted in the points above, collaborative and social use of the World Wide Web has emerged to be worthwhile. This is important since in aggregate it makes social media possible. This is not to say that collaborative-capable elements materialised all at once, instead they transpired progressively. The next section will describe the methodology involved in measuring these developments over time.

3. TRACKING THE EVOLUTION OF WEB 2.0

3.1 Data Source

In order to conduct the historical evaluation, a reliable, independent, and trustworthy historical dataset was required, and the Wayback Machine (accessible at http://web.archive.org), an online internet archive, was identified for the study’s purpose. The Wayback Machine (from now on referred to as WM) is a time-capsule library of cached websites with over 3 petabytes of stored web-page content (as of 2009), with about 20 terabytes (10^{12} \text{ bytes}) of new digital content being added each month to the archive (Murphy et. al. 2008). Web-pages are usually cashed repeatedly at various multiple month intervals. The library is a non-commercial project supported by Alexa Internet Inc and its significant uses included, as evidence in court cases (Gelman 2004, Howell 2006), and also a significant portion of the archive was donated to the US congressional library as early as 1998 (2 years into the project's existence). Since the system started to aggregate web-page snapshots in 1996, many popular websites are represented over their entire lifespan. Hence the WM is a unique dataset, a series of time-capsules of WWW at various points in its evolution, and suggests itself as a suitable tool for a historical web-page analysis. Indeed earlier studies using the WM dataset also exist. However, these have in particular investigated the evolution of design techniques and Web accessibility using historical snapshots from the WM (Fukuda et al. 2005, Hackett and Parmanto 2005, respectively). These studies used various complexity and accessibility scores derived from historical page features over statistically large samples to successfully infer interesting patterns. The reliability of data on WM has been scrutinised and explicitly validated in a separate study conducted by Murphy et al. (2008), and although not strictly generalizable, was found to be of relatively good quality. Nevertheless Murphy et al. (ibid.) highlight that significant future work on establishing WM validity is still necessary, especially if it is to see wider adoption by the research community. It is hence acknowledged, that its validity is only indicative and is a particular limitation of this study; however, given previous successful studies using the WM (e.g. Fukuda et al. 2005, Hackett and Parmanto 2005, Murphy et al. 2008), it was deemed the WM archive would be suitable for an exploratory empirical analysis looking at Web 2.0 related developments over time.

The websites selected for analysis were chosen on the basis of (i) being considered to be representative of Web 2.0 systems in several ways and (ii) having played a significant historical role. In particular, Youtube, Amazon, Flickr, Twitter, Craigslist, Digg and Yahoo were chosen. Although this selection may seem quite heterogeneous and indeed these websites all followed different pathways and design patterns throughout their

---

13 Social media and Web 2.0 are considered synonymous throughout this paper. Social media tends to be used by media scientists more frequently than the term Web 2.0, which tends to be more common with other vocational fields. Although Kaplan and Haenlein (2010) tried to define a distinction between the terms, they did not succeed to make a strong point. When the term media is constrained to Web based media, then as far as the author is aware there is nothing to suggest that social media and Web 2.0 should not be treated as synonymous terms.

14 Some web-pages are archived more frequently than others however there is a lag of 6-12 months between the cashing and the time the cached content is made publicly available on the Wayback Machine.

15 The site itself argues that they “seek to collect and preserve the digital artefacts of our culture for the benefit of future researchers and generations”.

16 In numerous countries web accessibility standards became a regulatory requirement by law and have therefore received some attention (Zeldman 2007).

17 The study analysed the accuracy of WM on a sample of travel industry websites. See Murphy et al. (2008) for a discussion of numerous issues and the full analysis of Wayback Machine reliability.
lifetime, they are all high traffic websites, significantly leveraging UGC (User Generated Content) in some ways and broadly perceived to be characteristic of Web 2.0. This is especially true for Youtube, Amazon, Flickr, Twitter, Craigslist and Digg. The choice of Amazon and Yahoo was primarily based on their longevity and subsequent ability to observe their evolution over a longer time-period.

### 3.2 Observational Framework

The overall aim of this study is to empirically shed light onto the evolution of the Web. Given the discussion in section 2.2, the list below identifies main factors of interest to Web 2.0 and ways in which these will be measured within the study to observe their emergence.

1. *Web development became easier* – this can be observed with the appearance of JavaScript Libraries, and improved XHTML use (generally much better cross-browser standardisation).
2. *Architecture / Integration* – can be gauged by the appearance of RSS, Mashups and APIs. These are understandably rough approximations, yet to some extent indicative.
3. *Standardisation* – outdated HTML, the phase-out of old mark-up / tag attributes, and the arrival of new tags and a strong emphasis on design – content separation.
4. *Viability of online business models* – cannot be investigated by this study; however, references to literature dealing with this question were provided in section 2.2.
5. *Trust* – has not been measured in this study in any direct manner. However, on a more indirect basis, significant increase in sharing of personal data on the Web can be assumed to be an effect attributable to Web users having more trust in sharing content online. Therefore Web 2.0 activities, such as sharing, commenting, or submitting content may be somewhat indicative.

Web design and presentation habits have changed markedly over time (Zeldman 2007, Comrode and Krishnamurthy 2008). Some HTML syntax has become obsolete while new development patterns such as AJAX or CSS based design, *i.e.* design from content separation have emerged (Murphy and Persson 2008, Budd et al. 2009). In this study some 40 features per page, for every website were extracted. These features were split into four streams of features in order to help organise the various aspects related to evolution of the Web. The main emphasis was on the temporal adoption of Web 2.0 elements, activities and design patterns. Most of the features presented in table 1 can be directly related to the adoption of Web 2.0 (4th stream of features, right-most column in the table). The 3rd stream of features (2nd column from the right, table 1) is represented by JavaScript, AJAX, and CSS Stylesheet adoption, as these were identified to be conductive to Web 2.0 adoption, in acting as a catalyst to the process of change. Outdated tags (in the 2nd stream, table 1) should illustrate the gradual disappearance of problematic design elements that would have acted as barriers to wider web-application standardisation and adoption.

<table>
<thead>
<tr>
<th>General (1st Stream of features)</th>
<th>Outdated HTML (2nd Stream of features)</th>
<th>Script and Styling (3rd Stream of features)</th>
<th>Web 2.0 activities &amp; elements (4th Stream of features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1- table (tr, td)</td>
<td>1- All JavaScript tags</td>
<td>1- RSS</td>
<td>Web 2.0 Elements</td>
</tr>
<tr>
<td>2- font formatting tags (b, i, u, big, small, font)</td>
<td>2- JavaScript source tags (external JS code)</td>
<td>2- API</td>
<td>1- RSS</td>
</tr>
<tr>
<td>3- center tag</td>
<td>3- Script libraries used</td>
<td>3- Mashups</td>
<td>2- API</td>
</tr>
<tr>
<td>4- menu tag</td>
<td>4- AJAX use</td>
<td>4- Podcast</td>
<td>3- Mashups</td>
</tr>
<tr>
<td>5- layer tag</td>
<td>5- CSS Styling (in-page)</td>
<td>5- Blog</td>
<td>4- Podcast</td>
</tr>
<tr>
<td>6- blink tag</td>
<td>6- CSS Styling (source)</td>
<td>6- Tag</td>
<td>5- Blog</td>
</tr>
<tr>
<td>7- marquee tag</td>
<td>7- Wiki</td>
<td>7- Like/Favourite/Unlike/Report This</td>
<td>6- Wiki</td>
</tr>
<tr>
<td>8- attributes (align, bgcolor, background, text, link, vlink,alink)</td>
<td>8- Permalink</td>
<td>7- Edit</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 – Wayback Machine study; overview of features extracted from every page analysed
External JavaScript files referenced from the page were also checked for use of client-side libraries – 14 different popular libraries were checked for, namely: JQuery, Prototype, Joose, Dojo, GWT, Processing, Scriptaculous, Midori, Pyjamas, Rico, YUI, Qooxdoo, Mootools and Mochikit. AJAX use was checked by searching for msxml2.xmlhttp, microsoft.xmlhttp, xmlhttprequest JavaScript objects. Also all related “src” attribute *.js files for any AJAX were searched. HTML frame based implementations were not checked for; however, these are not common.

3.3 Methodology

Data from the WM was downloaded in multiple phases, and collated into 7,482 (after initial cleaning) instances, where each instance describes a historical web-page using over 40 features (see table 1). The overall time-window covered by the dataset ranges from 17/10/1996 to 16/04/2010. This date range is partly conditioned by the availability of data through the WM (to be discussed further below), and also takes into account the term Web 2.0 was coined and emerged throughout 2004 and 2005 (Comrode and Krishnamurthy 2008); hence it was felt that the time-period centred around these years with a cut-off in 2010 would be a sensible range for observing most of the discernible changes in the transition to a Web 2.0.

The frequency of entries in the archive tends to be irregular, due to at least two reasons related to how the WM crawls the web. Historically, some web-pages were also down for maintenance or technical issues, such as when Amazon experienced disruption to its service during October 2000. Other pages were corrupted or temporally inaccessible from the actual WM servers. Altogether over 350 Megabytes of web-page data was extracted, parsed and analysed for the required features. The overall download methodology is summarised in figure 1.

![Diagram of Wayback Machine archive extraction methodology](image)

The methodology is essentially split into four phases. Phase 1 – A custom script fetches all historical records available from the WM archive (only cases indicated to contain substantial changes by the archive are considered). Phase 2 – For each archive entry in the list, the HTML code of the page is requested, in addition with all the .js files linked to it. Phase 3 – A DOM (Document Object Model) tree of the HTML code is generated, which allows a custom script to then extract the required tags, and allows the automated content analysis to look for keywords. Phase 4 – Saves all the extracted features of each HTML page and related .js scripts as a record into a .csv file, which can be imported into SPSS or Excel for further analysis.

The final dataset was broken down by individual websites into seven separate groups; Craigslist (N=304),

---

18 See Wayback FAQ - [http://www.archive.org/about/faqs.php](http://www.archive.org/about/faqs.php)
19 .js is the file extension usually associated with JavaScript files, and is the de-facto standard client-side language on the Web (Crockford 2008).
20 A DOM tree is a standard representation of an (X)HTML page, which allows to programmatically manipulate or access its elements (XPATH supported), see [http://www.w3.org/TR/DOM-Level-3-Core/introduction.html](http://www.w3.org/TR/DOM-Level-3-Core/introduction.html)
Twitter (N=312), Youtube (N=922), Flickr (N=1,141), Digg (N=1,141), Amazon (N=1,420), and Yahoo (N=2,247). The box-plot in figure 2 illustrates the distributions of downloaded pages over a range of dates. Yahoo is represented in the dataset for the longest time-period, over 13 years (17/10/1996 – 24/03/2010), followed by Amazon, Craigslist and the other websites.

Figure 2 – Box-plots of downloaded page frequencies for available dates (all seven websites)

3.4 Analysis

Although this study was largely automated through the use of custom scripts (section 3.3), some qualitative analysis, i.e. careful manual inspection, was also undertaken to help interpret the various observations throughout the archive. A qualitative approach was useful in order to help validate results, and to observe things that weren’t detectable automatically. Let us consider the main page of Youtube from the 1st January 2006.\(^{21}\) It contained a tag-cloud,\(^{22}\) a last five users online widget, and video listings used practically no client-side scripting yet. The Youtube Blog was just launched recently – in other words, it seemed that Youtube was just beginning to use various Web 2.0 elements more intensively, but with some reservations. In the next section it will be shown how this assessment correlates with our quantitative analysis. The second issue which is closely related to what was said above is concerned with the interpretability of such results. Any inference from the dataset was approached with caution as other contributing factors might have been unmonitored or omitted; however, in certain cases inference was relatively self-evident. Consider for example that from within the earliest entries for the title of Youtube (html based header / title tag\(^{23}\)), the slogan in the title changed from “YouTube – Your Digital Video Repository” into “YouTube – Broadcast Yourself”, this represents the symbolic shift in the site’s purpose to a more direct “Broadcast Yourself” community. This was confirmed and also noted by Burgess and Green (2009).


\(^{22}\) Tag cloud or word cloud is a visual depiction of textual content within a graphic, where words are scaled based on their occurrence frequencies, see http://en.wikipedia.org/wiki/Tag_cloud.

\(^{23}\) This is an element that makes a title appear as the title in the browser bar (or an active browser tab). As it became apparent from the study – over time this title does not tend to change much, the only time it has changed was when a website wanted to communicate a different perception about itself.
4. RESULTS

This section provides an overview of the collected historical websites dataset, followed with an analysis of substantive website-design changes, as observed throughout the available sample time range, and for each; Youtube, Digg, Amazon, Yahoo, Twitter, Flicker and Craigslist, discussed in that order. Table 2 provides a summary of the dataset for the seven websites.

<table>
<thead>
<tr>
<th>Basic Details</th>
<th>Available Dates</th>
<th>tag</th>
<th>tag</th>
<th>tag</th>
<th>tag</th>
<th>tag *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Craigslist</td>
<td>Min 11/11/1998</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Max 20/06/2006</td>
<td>25</td>
<td>45</td>
<td>1</td>
<td>259</td>
<td>2</td>
</tr>
<tr>
<td>Twitter</td>
<td>Min 30/09/2006</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Max 31/07/2008</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Youtube</td>
<td>Min 28/04/2005</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Max 01/07/2011</td>
<td>103</td>
<td>101</td>
<td>5</td>
<td>259</td>
<td>2</td>
</tr>
<tr>
<td>Flickr</td>
<td>Min 26/02/2004</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Max 01/07/2011</td>
<td>5</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Digg</td>
<td>Min 09/12/2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Max 27/10/2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon</td>
<td>Min 12/12/1998</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Max 27/10/2009</td>
<td>65</td>
<td>71</td>
<td>39</td>
<td>86</td>
<td>7</td>
</tr>
<tr>
<td>Yahoo</td>
<td>Min 17/10/1996</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Max 24/03/2010</td>
<td>39</td>
<td>106</td>
<td>14</td>
<td>151</td>
<td>6</td>
</tr>
</tbody>
</table>

* Old tags are center, menu, layer, blink and marquee. Their usage is very strongly discouraged (e.g. Zeldman 2007).

Note: u and big tags were only used by yahoo (min = 0, max = 6; min = 0, max = 4 respectively). The small tag was only used by yahoo (min = 0, max = 47) and twitter (min = 0, max = 3).

From table 2 it is evident that Yahoo, Craigslist, Amazon, and Flickr rely on depreciated html tags. This was somewhat unexpected and indeed illustrates that in some areas even leading websites haven’t modernised to more up-to-date W3C specifications, in recent times. Very rarely their usage can be justified, such as in certain client-scripting scenarios or when dealing with compatibility issues, although this is now rare (Zeldman 2007).

Table 3 shows that client side scripts are prevalent on all websites; however, external scripts are used more sparingly. Craigslist surprisingly never made use of external scripts in the sampled period, even though page level scripts were used; external script files are desired and preferred over page level scripts, since the former speeds up performance thanks to client file caching. In terms of advanced scripting, AJAX or well known client Javascript libraries were identified on all websites, with the exception of Yahoo, Amazon and Craigslist. CSS styles are prevalent on all websites, but yet again external Stylesheets are never used by Youtube.

The 4th feature stream contains features related to Web 2.0 activities (such as like, share, rate, comment…) and Web 2.0 elements (such as Blog, RSS, Podcast…). It can be appreciated from table 4 that each website contained Web 2.0 elements and activities at some point during the sampled period. Most

24 An inherent problem with this feature in the experimental set-up is that features were extracted by matching keywords in the pages content text and alt attributes (i.e. alternative texts often used on graphics). This is not the safest method of detecting these features, since some keywords can and will appear by chance.

Table 3 – Summary of websites – features from stream 3 (see section 3.2)

<table>
<thead>
<tr>
<th>Websites</th>
<th>Basic Details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>script tag</td>
</tr>
<tr>
<td>Craigslist</td>
<td>Min</td>
</tr>
<tr>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>Twitter</td>
<td>Min</td>
</tr>
<tr>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>Youtube</td>
<td>Min</td>
</tr>
<tr>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>Flickr</td>
<td>Min</td>
</tr>
<tr>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>Digg</td>
<td>Min</td>
</tr>
<tr>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>Amazon</td>
<td>Min</td>
</tr>
<tr>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>Yahoo</td>
<td>Min</td>
</tr>
<tr>
<td></td>
<td>Max</td>
</tr>
</tbody>
</table>

Table 4 – Basic summary of Web 2.0 elements and Web 2.0 activities from stream 4 (see section 3.2)

<table>
<thead>
<tr>
<th>Websites</th>
<th>Web 2.0 Elements</th>
<th>Web 2.0 Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Craigslist</td>
<td>Min</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>2</td>
</tr>
<tr>
<td>Twitter</td>
<td>Min</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>3</td>
</tr>
<tr>
<td>Youtube</td>
<td>Min</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>3</td>
</tr>
<tr>
<td>Flickr</td>
<td>Min</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>2</td>
</tr>
<tr>
<td>Digg</td>
<td>Min</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>1</td>
</tr>
<tr>
<td>Amazon</td>
<td>Min</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>4</td>
</tr>
<tr>
<td>Yahoo</td>
<td>Min</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>0</td>
</tr>
</tbody>
</table>

4.1 Youtube

The historical analysis of Youtube is based on archive pages for a period of several years (starting April 2005). The number of tables used for page layout throughout the years has been constant until the 10th April 2008, when the use of tables was scaled down from 15 to only two tables. Initially there was a table for each video (picture, title, description…), and this was reduced down to only two tables, one used for login form input, and the other for the footer menu. By 19th June 2008 table use was reduced down to only one table, and as of 1st July 2011 no tables were used. The site's layout is now entirely based on CSS. In fact throughout the analysed period Youtube landing page has used external Stylesheets. As for old and depreciated tags, from April 2006 Youtube began to use b tags instead of inline CSS, to set characters bold. Since September 2006 their use has been decreasing from 101 (at its peak), yet even today Youtube still uses 12 b tags, specifically in their menu items. This is a rather bad practice but it was not possible to discover the reasons behind such design. More predictably, Javascript became more intensely used in recent times, although it was used
Throughout the entire sample period, with AJAX found to be used on the main Youtube landing page during Oct. 2006 – Feb. 2008.

Several Web 2.0 elements were identified; Youtube had an API, RSS-feed, and a Blog, and it was possible to identify the exact dates when these were introduced. The RSS feed was introduced on 17th June 2005, an API as early as 4th August 2005, and the Blog followed on the 15th December 2005. Web 2.0 activities were detected on at least 50% of the 923 web-page entries, and included the activities of sharing, commenting, submitting, rating, tagging and liking. One of Youtube’s earliest features was commenting on videos, introduced as early as 17th June 2005. A further manual investigation of the earliest cases containing keywords relating to these activities revealed them to be part of the Youtube website, rather than accidental content within arbitrary Video descriptions.

The first video uploaded to Youtube was on 23rd April 2005, and since data from WM was available from 28th April 2005, the entire first and most interesting years of Youtube’s historical existence were analysed. It was possible to observe when exactly certain features such as commenting, rating or favourite lists were added to Youtube. Youtube had a heavy Web 2.0 footprint early on, in terms of the number of activities such as uploading / posting videos, rating videos, liking videos into favourite lists and commenting. The entry points for UGC (User Generated Content) are indeed more varied than Twitter, for example. Youtube was also early to adopt RSS feeds and to provide an API for developers. Expected patterns were observed for features relating to CSS, Javascript (AJAX), and deprecated tags. CSS and Javascript were used from the outset and AJAX was added during Oct. 2006. Table based layouts gave way to a more standardised CSS approach.

4.2 Digg

From the outset Digg has been developed as a social news website. A core activity of it being the idea of average users vetting news-stories, by either voting them up or down, or in other words digging or burying stories, respectively. The interesting thing about Digg is that as a site it has adopted a very clean and standards based design, i.e. no tables and deprecated tags were found throughout the available time period. The site was continuously improved with four major re-designs of its user interface, during July 2005, June 2006, November 2006, and June 2007. AJAX was introduced to the site in January 2005. As opposed to Flickr and Twitter, most of Digg application functionality can be accessed on the main landing page, i.e. reading stories, voting for stories, etc. In January 2005 commenting on links was introduced, the “blog this” feature (allowing to directly post articles on Blogs) was added in April 2005, in August 2005 Digg Podcasts, and on April 2007 an extensive API platform were introduced to Digg.

It was found that Digg shows a great rate of adoption of client-side scripts, and at the same time CSS is relied upon for the entire surveyed time-period, with virtually no use of outdated or non-standards HTML. Since Digg relies upon user-participation Web 2.0 based elements that facilitate and help trigger social interaction were introduced relatively early to Digg.

4.3 Amazon

Amazon differs from Yahoo (discussed next) in that it is widely regarded as an innovator in Web being used as an interactive media (O'Reilly 2006), and was in fact an early adopter of some UGC facilitating features, such as rating and commenting within its product pages (Spector 2002). This is the reason Amazon was included in this study.

The website layout on 13th October 1999 (earliest available entry from the WM) is predictably based on a simple set of tables with some bgcolor attributes. During this time it also seems that image-maps were rather popular (Yahoo was also using one in the 90s), and Amazon employed it to give the impression of rounded corners in its main menu, as achieving the same effect back than was complicated using other means. As early as 2002, Amazon was already using basic elements of CSS and some Javascript in order to

---

25 A crude technique used to split a picture into different clickable areas – i.e. `<map>` tags.
detect a wide set of browsers and execute browser specific code where necessary. The Amazon page used an AJAX based element on the main page from mid-2011 – however AJAX hasn’t been detected in earlier pages, and must have been introduced recently. In 2001 product wishlists which are shareable between customers were introduced. In August 2006 Amazon podcasts became available, and an API for developers to build on top of the Amazon platform was provided by Amazon at the time. As of 2007 a corporate Amazon Blog went online. Around March 2008 a tag cloud on the main page was introduced for a short time period, and was subsequently removed. A similar tag cloud, for navigation, was also observed on Youtube.

Amazon has shown a pattern similar to other, already discussed websites. CSS or Javascript usage emerges and increases over the years, as the use of older tags becomes less relied upon. As for the collection of UGC, Amazon it seems has indeed embraced social and collaborative elements of the Web, in a Web 2.0 exemplar way. Many keywords relating to sharing / openness and Web 2.0 specific elements such as Blogs, Podcasts and Tag Clouds were detected. Quite early on Amazon was experimenting with how to engage its wide user-base into online social networks of shoppers that might help make online shopping a more social activity. During 2005 the “Friends & Favorites”26 social network was introduced within Amazon and users were able to build profile pages and connect in a number of social ways (see figure 3).

4.4 Yahoo

The range of dates represented by the historical sample for Yahoo is the widest, and covers a period of over 10 years. The data facilitates an analysis of the late 90s and the pre Web 2.0 era. Yahoo is generally perceived

---

to be an administered web-portal\textsuperscript{27}, rather than a social media or Web 2.0 website, and hence it was expected that a Web 2.0 footprint as measured over time would be relatively small, possibly but not necessarily increasing somewhat in more recent years, as well as some decrease in depreciated HTML elements was expected as Yahoo strived for HTML standardisation. Unfortunately after an extensive manual inspection, 591 page records from the period between August 2006 and February 2010 had to be excluded from analysis, due to an issue with WM. Consequently there is no page representation for years 2008 and 2009. During the early years of Yahoo, in the months of October, November and December 1996 Yahoo wasn’t using any table tags and page design was extremely basic\textsuperscript{28}. These were still very much the early days of the Web (IE 3 and Netscape Navigator 3 both came out in August 1996), with simple browsers not facilitating much in the way of more complex designs\textsuperscript{29}. The first major redesign discovered by our analysis took place in January 1997 with Yahoo introducing tables to lay out the still very basic HTML content into a more compact visual display. Two years later (February 1999) Yahoo introduced the use of \textit{bgcolor} attributes and table based layout now received a more colourful design. During March 2001, basic in-page cookie checking scripts were introduced. It wasn’t until July 2002, and November 2004 when Yahoo’s visual design was polished considerably; however, the design was still heavily based on tables and with entirely no use of CSS. Only towards the end of 2005 CSS was introduced. As of February 2010 Yahoo layout was finally based on CSS, with old tags now completely removed, and JavaScript was used to some degree\textsuperscript{30}. Web 2.0 elements, other than arbitrary mentions in the content of articles, haven’t been detected throughout the content analysis. Only, on a more recent page (1st July 2011) a corporate Blog is finally available, and there also doesn’t seem to be much in terms of keywords pointing to possible Web 2.0 activities on the main page.

It is well known that Yahoo provided extensive capabilities to contribute UGC within its child projects such as Flickr, Delicious, Yahoo! Answers, etc. and supports a strong API. However, the main landing page itself has kept most of its classical portal image, with practically no Web 2.0 style content sharing. Yahoo, provides an API to much of its data, and has opened up considerably as a platform, yet little of this is observable on Yahoo’s main landing page, with no direct encouragement for contributing UGC. Yahoo was also slow to adopt a CSS based design and table based layouts seemed to have been preferred by Yahoo designers. During the year 1996, design was extremely simple as there were numerous compatibility and rendering issues associated with many HTML elements, and CSS was mostly an unsupported standard\textsuperscript{31}.

4.5 Twitter

A look at the 21 months of data for twitter, from September 2006 to July 2008 reveals a number of redesigns to the main Twitter page. As expected Twitter has phased out tables and never looked back. Between 28\textsuperscript{th} and 31\textsuperscript{st} October 2007 Twitter removed tables completely\textsuperscript{32} from its page design. Before then tables were used to lay-out the content on the main page, where table layouts themselves were slightly re-designed several times over the sample period\textsuperscript{33}. The removal of tables during 31\textsuperscript{st} October 2007 was accompanied with other changes in which old and depreciated HTML tags were removed from the pages’ design and as for visual changes, Twitter updates were now fully hidden behind a login wall. There were also some changes to the number of referenced CSS files, as well as CSS related tags, but some form of CSS was used throughout the entire sampled period, also during the earlier periods of table based layout. AJAX was detected to be

\begin{footnotesize}
\begin{itemize}
\item\textsuperscript{27} Yahoo is best known for its web portal, from which one can access their search engine, Yahoo directory, Yahoo Mail, Yahoo News, Yahoo Maps, Yahoo Videos, etc... Yahoo also operates Flickr (a photo sharing application), however this website is analysed separately. The focus of interest is the design of Yahoo’s main page.
\item\textsuperscript{28} The HTML code of the page contained only some \texttt{<img>, <a>, <p>, <hr>, <b>, <ul>, <li>,<center> and <font>} tags. A page example from this period is available for inspection under this url: http://web.archive.org/web/19961022175643/http://www10.yahoo.com/
\item\textsuperscript{29} One of the main reasons why Yahoo wouldn’t be using tables in the early days was that I-pages took longer to load with the overhead code and 2- some browser versions weren’t able to render more involved table layouts.
\item\textsuperscript{30} Yahoo uses their YUI custom open-source Javascript library http://developer.yahoo.com/yui/
\item\textsuperscript{31} See http://www.w3.org/ for much more on the history of standards and specific information on CSS and (X)HTML
\item\textsuperscript{32} This table-free version can be accessed on (note: the CSS external spreadsheet file does not always seem to load within Wayback Machine properly) http://web.archive.org/web/20071031185621/http://twitter.com/
\item\textsuperscript{33} Dates of some major re-designs for Twitter include – September 2006, November 2006, January 2007 and March 2007.
\end{itemize}
\end{footnotesize}
AJAX was detected within the prototype JavaScript library (http://www.prototypejs.org/) used on Twitter.

For reference these other features were: RSS, Podcast and Wiki, in four, nine and one page entries, respectively. Web 2.0 activities, Share (2), Comment (10), Submit (4), Rate (9), Tag (10), Like (124) and Edit (9).


It seems that the Flickr blog was introduced on the 12th June 2007 with the re-design.

Flickr supports many more interesting features. For example sets, mutually non-exclusive collection of pictures, and
archived pages was necessary since the Flickr website is mostly hidden behind a login / register landing page. Although the design isn’t necessarily the cleanest and despite the fact our conclusions are limited due to limits of access to pages, it seems Flickr has mostly concentrated its efforts at allowing high levels of social engagement from its earliest days.

4.7 Craigslist

Craigslist is an inherently collaborative application which started off as a mailing list and has evolved from a relatively centrally managed system, with Craig Newman submitting a lot of the content, to a fully user generated, and well categorised list of wanted and for-sale items marketplace. During the end of 1998 and until November 1999 the landing page for craigslist was kept more or less unchanged. A minor re-design in November 1999 was followed by a substantial one during March of 2000, in which HTML tables were introduced to play a larger role in page layout. A seemingly cleaner looking interface was introduced during June 200039, which is when some very basic JavaScript was introduced for the first time and new ads categories were also introduced40. During July 2000 a discussion forum was added and later that year in October, community based vetting of ads was introduced through the “flag for review” link-buttons. In January 2001, a wishlist feature was introduced, and the Craigslist Blog in November 2003. By June 2004 Craigslist offered localised ads for the UK (London) and Canada (Montreal, Toronto and Vancouver), in addition to the US. The visual design of Craigslist; however, has not changed much since 2004. Today the design looks surprisingly similar, even tables are still used to layout content, although old and depreciated tags were entirely removed from the page-design.

Craigslist grew organically from a mailing list, into a website, where user contributions played a central role from the beginning. Users contributed mainly in the form of ads and ad vetting. In many ways; however, Craigslist simply resembles an advanced, moderated forum, where participation is driven by the goal of exchanging goods. There isn’t an API that would open up Craigslist as an application platform either. It would also seem that much of Craigslist functionality could be accomplished via a more advanced forum design. An extremely simple table-based visual design was employed. JavaScript was introduced quite early to the page, but the pages on Craigslist have never evolved much in complexity.

5. DISCUSSION

Clearly there was an overall trend within all the investigated websites to re-design old HTML based code. Old tags were phased out in every case, although some websites seemed to have relatively clean designs from the outset, e.g. Digg. In other cases, such as Yahoo, rather bad and old design elements prevailed for longer than expected. As for Web 2.0, Tim Berners Lee did suggest that there is no fundamental change in the technology (see section 2.2); however, this concerns the W3C introduced standards rather than actual adoption of standards by browsers and websites. It wasn’t until 2006 that some of the main browsers became compliant with a substantial set of CSS and HTML W3C standards (Zeldman 2007, Murphy and Persson).


40 While investigating web-design related features on Craigslist we noticed that the ads categories evolved over time, sometime in quite interesting ways. For example in year 1998/99 there were mostly practical job and house related ads (http://web.archive.org/web/20000303014907/http://www.craigslist.com/), which of course has changed considerably over time, with the notorious personals ads and the many international / localised categories introduced (e.g. see http://web.archive.org/web/20000612232944/http://craigslist.com/). One would imagine that a historical study of Craigslist could be interesting for some social sciences research. Unfortunately the scope of this study does not allow us to explore this idea further.
2008, Budd et al. 2009). The problems associated with earlier incompatibilities and inabilities of browsers to render much content simply meant that user-interfaces and rendering across different browsers was buggy and inconsistent (Zeldman 2007). Hence Tim Berners Lee’s argument when put into context transforms into the argument of the issue of technical feasibility of standards into real world implementations, which did not occur more extensively, until the mid 2000s. This is much more important than it would seem. One may argue that the broader public didn’t take the Web seriously at this time and trust was difficult to build, if correct page rendering would be inconsistent. Since full CSS and HTML standards were often not supported, websites had to usually compromise on the web-application design. Ability to manipulate the HTML DOM-tree from client side scripts with proper CSS support, and to integrate this with server side logic via AJAX would have been a major development; however, unlikely to be used if the browsers did not support such implementations consistently.

It was found that interactivity was increasingly being added (in terms of JavaScript). Although Javascript adoption was probably understated, since it was measured on the main landing page, and deeper level pages, where most functionality for sharing and collecting user contributions is located would have been ignored. AJAX was not found over all websites, and maybe AJAX hasn’t been as significant for Web 2.0 as some might suggest.

Also it was found that sites have generally opened up during the later 2000s, with incorporating APIs, RSS feeds or opening up corporate Blogs. An API fulfils an important role for a web-application as it allows for the dataset and application logic to be extended and reused by third parties, and effectively allows for more complex and potentially useful applications to be built. Since data on Web 2.0 websites is generally user contributed, the access and sharing of the data via APIs is to some extent justified. Not every website however has been found to provide an API, e.g. Craigslist. RSS feeds also open up the website as a platform, since Mashups can be built and information consumed in a number of ways. Blogs are a more symbolic way of opening up communication and allowing for a more informal communication channel. It was found that all of the websites discussed, introduced Blogs at some point, most have done so during the mid 2000s.

Finally it must be noted that dynamic interactive features and community dynamics were not investigated at a functional level within this study. For instance how tagging differs and has evolved on the different websites over time, or whether and to what extent Youtube subscription channels or Flickr photo sets are comparable and how do online communities interact with such elements. These are relevant and interesting questions to consider but are clearly beyond the scope of this particular study as the focus of this paper was to predominantly establish the technological development of the website itself, and the aim was not to focus on online communities per se. Nevertheless this type of work would be appropriate for further future study and approaches from virtual excavation, which concerns itself with the systematic acquisition of digital trace data from virtual settlements using the archaeological paradigm, to specifically identify and study activities of active groups of participants may be highly relevant here (Jones, 1997). More recently, for instance Jacovi et al. (2011) looked at Twitter and Blogs, Akoumianakis et al. (2012) explored Youtube, and Akoumianakis et al. (2014) investigated the various community activities of social media websites in such a context, and these types of studies show significant promise in the study of digital artifacts and related community activities. There are some for instance Comrode and Krishnamurthy 2008

5.1 Limitations

There are a number of limiting factors to this study which must be considered. First of all, only the main page (homepage) and referenced JavaScript files for each available date were analysed, as the WM does not usually store related pages, and if it does, then the relative links tend to often be broken, especially if these are composed of dynamic content (i.e. server side generated content based on input parameters, e.g. logged in users, location specific content, etc.). This is a significant limitation to the representativeness of the given website and effectively limits the applicability of this method over analysing websites in their entirety. Nevertheless, the analysis of the main homepage and referenced client-side scripts, as in other studies (see section 3.1), do provide a useful glimpse into the website’s history, although any interpretations of findings must be considered in light of this limitation. A more significant limitation was that pages themselves were not segmented by content, other than based on the HTML mark-up. For example, on a page like Craigslist, some of the user contributed content would also be scanned for words referring to “Wiki” or “Blog”. This can skew the results for Web 2.0 activities and Web 2.0 elements upwards, since this text isn’t part of the Craigslist application. The assumption was that despite some inaccurate readings, overall the accuracy will be
indicative and sufficient, especially due to the manual qualitative inspection that took place over a random sample of instances for each website. Another potential limitation of this work was the sampled time-frame and also that only seven main-stream applications were analysed with the particular choice of websites. Finally since this study is observational in its nature, observing the use of certain technologies does not imply causality for Web 2.0, but rather evidence of correlations.

5.2 Breath of the Study

A limitation of this work was that only seven applications were analysed and out of these, two were behind a login wall for most of the time-period, which limited their analyses further. Although the observations from this study may be interesting and insightful, the coverage of applications relative to the landscape of existing Web 2.0 systems is severely limited. In order to address this criticism of the WM study being too narrow in focus, and partly by a need for baseline features characteristic of Web 2.0 applications; a second study in which a much wider set of Web 2.0 applications is analysed was undertaken, but only of a current-snapshot, i.e. no historical webpages. To keep this study consistent, the same features extracted from websites were used (as described in section 3.2).

A directory of Web 2.0 applications mentioned in Lindmark (2009) was employed to provide a list of 99 top Web 2.0 websites, as ranked by the Alexa Inc. traffic ranking service. In the top 99 list nearly a third of the URL-resources could not be accessed, hence all in all, 59 unique website landing pages were retrieved, with HTML pages and JavaScript files of all items in the list downloaded on the 5th January 2011. The 59 websites included high traffic sites, such as LinkedIn, Facebook, Livejournal, Orkut and others, with a full list of the 59 websites available at http://www.authors-web-server.com/web_evolution_study.html. It was found that 64% (38 out of the 59) websites did not use tables at all, and 49% did not use old depreciated HTML tags either (such as b, i, u, marquee…). This is compatible with the historical study. Also, only on two websites it was found that no JavaScript, and on one website that no CSS were being used. Hence, interactivity and standards based layout was completely prevalent within Web 2.0 applications at this time. It was further found that well known Javascript libraries were referenced within 36 websites (61%), and as many as 40 websites (68%) contained or referenced code with AJAX. Nearly 82% of surveyed pages seemed to contain Blogs, followed by pages containing APIs (37%), RSS (29%), Wiki (19%), Mashups (3%), and one page referred to Podcasts. The following Web 2.0 activities were mentioned across the given websites; tag (72%), like (64%), rate (61%), submit (56%), edit (49%), share (46%), and comment (31%).

Evidently most websites use client-side Javascript and CSS, despite a relatively wide set of pages relying on outdated HTML, i.e. interestingly, 36% of the websites still relied on tables. This could have been expected, and does not contradict the results from the broader historical study. In line with the historical study, most websites now have a Blog, and eventhough potentially not accurate, the content analysis points to some widespread UGC activities. Especially lighter or easier activities, in terms of time and effort needed for generating UGC, such as tagging, like-ing, or rating tended to be widespread.

6. CONCLUSION

Several historical internet studies were undertaken by researchers in the past; however, this study is the first to investigate Web 2.0 adoption related trends, rather than accessibility or generic design issues. Results from this study were presented in the context of historical developments of the Web and largely support the Web 2.0 phenomenon, in that tangible, i.e. measurable, changes associated with Web 2.0 were observed. A unique dataset was employed to present a record of the historical developments of seven major websites over a number of years, with direct links to the cached pages provided throughout for the benefit of the reader. It was shown that increasing standardisation and adoption of social web design elements was a significant trend over the years, more so than AJAX alone. It was further observed that websites opened up their communication and datasets via Blogs and API platforms, which as other academics hinted, helped built more trust with users, and allowed deeper integration with related websites and web-services to effectively build more complex web-applications, rather than just stand-alone websites. These developments occurred gradually, and significant differences in the timings of such trends exist between the analysed pages, although the years around 2005-2007 seemed to have witnessed most of the key changes. Corroborating findings from a second study analysing 56, high traffic social websites were also presented, which lent further validity to
the observed historical trends.

In conclusion observables and measurable evidence in support of the full vision and potential of Web 2.0 as social media was found and presented throughout the 2000s.

REFERENCES


Budd A., Moll C. and Collison S., 2009. CSS Mastery: Advanced Web Standards Solutions, Friends of Ed, USA

Burgess J. E. and Green J. B., 2009, The entrepreneurial vlogger: Participatory culture beyond the professional-amateur divide, The YouTube Reader, pp. 89-107


Leadbeater C., 2009. We-Think: Mass innovation, not mass production, Profile Books Press, USA

Lindmark S., 2009. Web 2.0: Where does Europe stand?, European Comission (JRC) Joint Research Centre - (IPTS) Institute for Prospective Technological Studies Report, Seville (Spain)


Schroth C. and Janner, T, 2007. Web 2.0 and SOA: Converging Concepts Enabling the Internet of Services, IT Professional 9 (3), pp. 36-41


Vosn G and Hagemann S, 2007. Unleashing Web 2.0: From Concepts to Creativity, Morgan Kaufmann, USA


