On the Sustainability of Web Systems Evolution

Andres Baravalle
ACE – University of East London
London, UK
a.baravalle@uel.ac.uk

Cornelia Boldyreff
Greenwich University
London, UK
c.boldyreff@gre.ac.uk

Andrea Capiluppi and Renata Marques
DISC – Brunel University
Greater London, UK
andrea.capiluppi@brunel.ac.uk
remarques@gmail.com

***Abstract*—In the last twenty years, the evolution of web systems has been driven along three dimensions: the processes used to develop, evolve, maintain and re-engineer the systems themselves; the end products (the pages, content and links) of such processes; and finally the people dimension, with the extraordinary shift in how developers and users shape, interact and maintain the code and content that they put online. This paper reviews the questions that each of these dimensions has addressed in the past, and indicates which ones will need to be addressed in the future, in order for web system evolution to be sustainable.**

We show that the study on websites evolution has shifted from server- to client-side, focusing on better technologies and processes, and that the users becoming creators of content open several open questions, in particular the issue of credibility of the content created and the sustainability of such resources in the long term.

I. Introduction

The study of web systems and their evolution, researched for many years in the WSE series and elsewhere, has grown along three main phases: the first depended heavily on the technologies and infrastructures that web systems were based upon, so servers, their performance [15] and how to optimise their workload [2].

The second phase shifted the research from the issues of server-based development to client-side applications: in this phase, the earlier research focused on the evolution of web pages, and how to develop sites by reusing existing technologies and content, or to re-architecture the structure and links of the pages [14]. Further research confirmed the need to study which processes were most effective when developing new content: it was reported that websites were mostly expanded and eventually replaced, which created an issue with search engine spiders; and that web accessibility has to play a major role in the development of websites [Holger et al 2011]. Finally, the testing of websites played a major role in focusing the researchers attention on the static and evolving features of online content.

In the third phase, the study of web systems was complemented considering the openness of such systems and how collaboration is achieved in web system development. It is only recently that research on website evolution has focused on "users" as "content creators": the developer and user communities, the stakeholders and the whole web system ecosystem have an influence on the axes of products, processes and people.

In this paper we describe three axis along which the recent research on web evolution has developed, focusing on the "products" (Section II), "processes" (Section III) and "people" (Section IV) axes, and discussing the open issues that each of them present to researchers.

II. Products

It is hard to imagine what would be the web today without some technologies born without much fanfare but which have had a very in-depth penetration. While Netscape and Microsoft and later on the Mozilla Foundation and Microsoft were fighting for browser predominance, technologies arose to change the face of the web and shape it as we know it now.

With an increasing number of websites to work on, developers started recognising the importance of code reuse and abstraction. A small number of very popular libraries gained prominence in those years - with use across large number of web sites. This included software as Counter (born circa 1995), FormMail (1995) and phplib (1998). The issues related to maintainability were also reflected in the rise of software as WML (Website Meta Language, created in 1996), the use of Server Side Includes (included in Apache in 1995) and the use of content management systems (Slashdot was created in 1997 and Slash shortly after). Web technologies as $mod\_rewrite$ (circa 1989) and session-based URLs were also popularised in the 1990s, and allowed to move away from the static relationship between files and URLs.

On the web applications side, Geocities (1995) and Hotmail (1996) were born in 1995 and phpNuke and SquirrellMail in 1999; companies of any size were started offering complex services on top of Open Source stacks.

Popularity and reachability were important issues and developers were striving to find solutions. RSS (1995) allowed blogs and web sites in general to increase their popularity through syndication. LinkExchange and LinkTrade were founded in 1996 and 1997, allowing websites to increase their popularity; at his peak link exchange reached more than half of Internet-enabled households.

Because of their public nature, web systems raised the importance of accessibility in web systems. Early research emphasized the readability of web pages [4] and need for "plain" language as well as the needs of minority language users [11]. Usability and accessibility research predated research in human factors and ergonomics, from which it derived most of the early research methods (Nielsen, 1993).

A. Open Issues: Security, Innovation, and Sustainability

At the time of writing, Open Source software dominates the web. Netcraft statistics[1] show that Apache is still the most popular web server (from the time in which they started running their survey in 1995. As other players, as Nginix, start growing their market share, proprietary web servers seem more and more relegated to niche roles. PHP outperforms all other programming languages as number of websites[2]; Drupal and Wordpress have a combined market share of 51%.

This means that the bio-diversity that characterised the early days of the web is disappearing. Most web sites are building on top of the same set of technologies and using similar software stacks – which has implications when it comes to security. Even more dangerously, anecdotal evidence seems to indicate that most web sites do not update regularly their software stack – opening themselves to hackers.

There are indications that innovation is well alive; niche technologies keep mushrooming and the last few years have seen the popularisation of concepts as "cloud hosting", "virtual servers", "no-sql databases" and "Content Delivery Networks". In the coming years, well be probably seeing more progress in this direction - with further separation between the large majority of the web sites using the same software stack and innovators testing and popularising new paradigms.

III. Processes

Early on it was clear to some that traditional software development processes and methods whilst relevant to the development of web systems needed re-thinking in the light of early web system development. Metrics used to access web products and processes needed some adaptation to be applicable to web systems; and new metrics such as "depth of linking" were devised [19]. Earlier research had already suggested that study of metrics obtained from web sites over time provided a means to predict when a web site might be re-structured and also analysis of web content could be fruitful to identity replicated content similar to code clones [18], [5].

Other studies by Boldyreff with Dalton and Kyaw, highlighted the need for systematic development and maintenance processes for web systems [9], [10], [13], [6]. Kyaws survey found that web hypermedia design methods being promoted were derived largely from Object Oriented approaches and Database design approaches. Already at this time it was clear that web systems presented many of the challenges of more traditional large software systems and that existing methods and tools could be employed in their development and maintenance albeit with some adaptation. A fuller summary of this pioneering research on web system engineering can be found in the abstract quoted below from [3]:

Since the mid 90s, we have been studying Web development and maintenance at Durham. One aspect of the research has been concerned with characterizing the evolution of Web sites over time. Another has been concerned with determining appropriate process models of Web development and maintenance. Overlaying this research as been work on Web metrics: firstly to evaluate web application quality, secondly to describe change, and thirdly to guide our research on process models for web engineering. Working with small to medium enterprises developing commercial web-based applications has given our research a firmly based practical emphasis and empirical grounding.

All of our research has been leading towards the establishment of web site engineering as a new discipline within Software Engineering. An overriding concern has been determining how to build the foundations of web site engineering based on the lessons that can be learnt from traditional Software Engineering, the identification of the new challenges introduced by taking an engineering approach to web site evolution, both development and maintenance aspects, and the unique problems posed by the web and its associated technologies and innovative applications.

What we failed to recognise at the time was the impact that open source software engineering and open on-line collaborative communities developing web system content as well as code would have on our current web systems and their evolution.

A. Open Research Issues

The growth of distributed development and maintenance in recent decades has posed challenges for many large software systems and web systems are no exception. Global software development has brought not only a heterogeneity of system components but also of methods and processes used in a systems development and maintenance. The result has been multi-dimensional evolution and this is exhibited by many web based systems in the extreme as they are often composed of different technologies, multiple components and web services developed and maintained by diverse teams, each component and service manifesting its own evolutionary path presenting the overall system manager with a multitude of paths to understand and attempt to manage.

Few web systems today have been developed by a single team following common processes. This heterogeneity of processes and its impact on development and maintenance as manifest through multi-dimensional evolution poses the single greatest challenge to understanding web systems which of course is crucial for their continued and successful overall evolution. This links up with the people issue as more or less any one can contribute to open source software and content on the web.

IV. People

One of the most unforeseeable effects that the evolution of web systems underwent was opening the content creation to anyone, which culminated in the so called Web 2.0 revolution. This change of perspective put the users as the creators of the code and content in many different venues. During the last few years, the content created and maintained online by users (and termed *User-Generated Content*, or UGC [20], [1]) has become substantial in quantity, relevant in quality [8], and centered around major topics and websites: multimedia (YouTube, Flickr, DeviantArt, etc), expert knowledge on specific topics (Street maps [12], Slashdot.org, IMDB, Wikipedia, etc), let alone all the source code released with open licenses through

[1] http://news.netcraft.com/archives/2013/04/02/april-2013-web-server-survey.html

[2] http://trends.builtwith.com/framework

well known OSS repositories (SourceForge, Google Code, etc). The majority of such UGC's faces an "*uncontrolled, self-organized community of volunteer contributors, without the traditional tight conditions and organizational policies imposed in industrial production environments*" [16].

The majority of such UGCs faces an uncontrolled, self-organized community of volunteer contributors, without the traditional tight conditions and organizational policies imposed in industrial production environments [5]. From the research standpoint, some of these UGCs (for instance, the Wikipedia pages and the source code of OSS projects) are based on measurable effort of volunteers and produce measurable output with a determined productivity that an be tracked throughout the contents evolution, by parsing the recorded history of changes.

A. Open Issues

The two main issues that user-generated online content have to face are

- the credibility of the published resources, and
- the sustainability issue.

On the one hand, the creators of online content do not have an objective way to describe how credible or accurate is the information that they are putting into the web. One such example are the sites that collect the reviews and comments by customers on products or services, to be used by other customers to form a judgement on the quality of such products or services.

So, as visible in the Figure 1, even if a site like TripAdvisor shows a sustained intake of reviews by customers, and the judgements of customers (rating between 1 and 5) depict a nearly perfect experience, the number of reviews is still much lower than the potentially much larger number of customers who "physically" visited the premises of a restaurant. This raises an issue of credibility and bias of the reviews posted, since they could represent the views of a sample of customers much smaller (and biased) than the real user base.

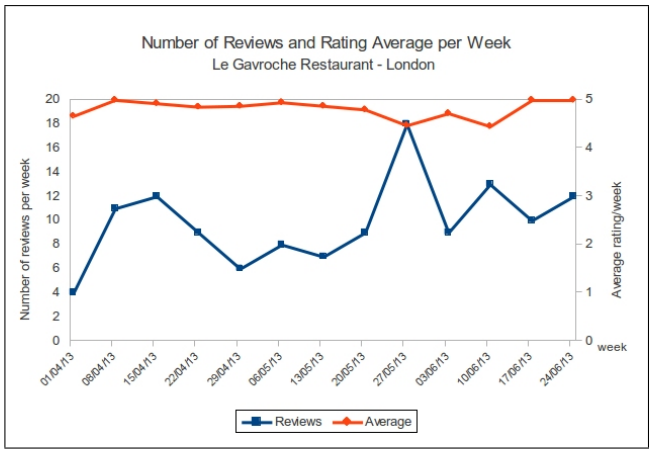

Fig. 1. Number of reviews and average grade per week of a well known London restaurant

Another such example is the collaboratively edited Wikipedia project, where pages are created and maintained by hundreds of online editors: most Wikipedia pages are accessed for information, and their utility or credibility can be measured by the number of views that they have received overall[7]; while another way of determining its credibility is at the level of individual and subjective effort[17].

On the other hand, the online user-generated content faces the issue of sustainability: user-generated content shows very common and recurring patterns where an initial phase of sustained growth is followed by a various descending phases, which eventually decline into zero, as visible in the Figure 2.

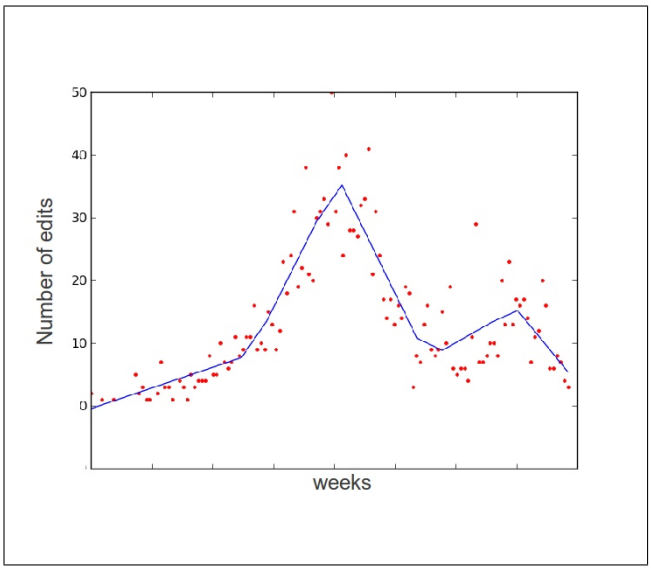

Fig. 2. Number of reviews and average grade per week of a well known London restaurant

V. Conclusions

The evolution of web systems over the last twenty years from humble hypertext systems to sophisticated large distributed systems has been an amazing achievement form both the technological as well as the societal standpoints. Here web systems evolution has been considered along the dimensions of products, processes and people.

The public nature of the web and its openness have been significant factors in its evolution. Tim Berners Lee in "Weaving the Web" envisaged the web as both readable and writable, and this has definitely come to pass in the last decade. Sustaining the future evolution of web systems will be as much a technological and engineering problem as a social problem, as we gain a better understanding of how open on-line collaborative communities develop and function effectively and their role in driving future web evolution.

References

[1] G. M. Alluvatti, A. Capiluppi, G. De Ruvo, and M. Molfetta. User generated (web) content: trash or treasure. In *Proceedings of the 12th International Workshop on Principles of Software Evolution and the 7th annual ERCIM Workshop on Software Evolution*, IWPSE-EVOL '11, pages 81–90, New York, NY, USA, 2011. ACM.

[2] M. F. Arlitt and C. L. Williamson. Web server workload characterization: The search for invariants. In *ACM SIGMETRICS Performance Evaluation Review*, volume 24, pages 126–137. ACM, 1996.

[3] C. Boldyreff, E. Burd, and J. Lavery. Towards the engineering of commercial web-based applications. In *International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*. Citeseer, 2001.

[4] C. Boldyreff, L. Burd, J. Donkin, and S. Marshall. The case for plain english to increase web accessibility. In *Web Site Evolution, 2001. Proceedings. 3rd International Workshop on*, pages 42–48. IEEE, 2001.

[5] C. Boldyreff and R. Kewish. Reverse engineering to achieve maintainable www sites. In *Reverse Engineering, 2001. Proceedings. Eighth Working Conference on*, pages 249–257. IEEE, 2001.

[6] C. Boldyreff and P. Kyaw. A survey of hypermedia design methods in the context of world wide web design.

[7] A. Capiluppi. Similarities, challenges and opportunities of wikipedia content and open source projects. *Journal of Software: Evolution and Process*, 2012.

[8] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 1–14, New York, NY, USA, 2007. ACM.

[9] S. Dalton. A workbench to support development and maintenance of world-wide web documents, department of computer science, university of durham, 1996.

[10] S. Dalton and C. Boldyreff. Web maintenance – the new frontier. In *Proceedings of Durham Maintenance Workshop, Sept.1996.*, 1996.

[11] J. Donkin, C. Boldyreff, L. Burd, and S. Marshall. Supporting sign language users of web-based applications: a feasibility study. In *HCI*, pages 291–295, 2001.

[12] M. M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, Oct. 2008.

[13] P. Kyaw. *An investigation of web-based hypermedia design support: methods and tools*. PhD thesis, Durham University, 1998.

[14] J. Martin and L. Martin. Web site maintenance with software-engineering tools. In *Web Site Evolution, 2001. Proceedings. 3rd International Workshop on*, pages 126–131. IEEE, 2001.

[15] D. Mosberger and T. Jin. httperf – a tool for measuring web server performance. *ACM SIGMETRICS Performance Evaluation Review*, 26(3):31–37, 1998.

[16] F. Ortega. *Wikipedia: A quantitative analysis*. PhD thesis, Universidad Rey Juan Carlos – Escuela Técnica Superior De Ingeniería De Telecomunicación, 2009.

[17] P. Pirolli, E. Wollny, and B. Suh. So you know you're getting the best possible information: a tool that increases wikipedia credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1505–1508. ACM, 2009.

[18] P. Warren, C. Boldyreff, and M. Munro. The evolution of websites. In *Program Comprehension, 1999. Proceedings. Seventh International Workshop on*, pages 178–185. IEEE, 1999.

[19] P. Warren, C. Gaskell, and C. Boldyreff. Preparing the ground for website metrics research. In *Web Site Evolution, 2001. Proceedings. 3rd International Workshop on*, pages 75–85. IEEE, 2001.

[20] S. Wunsch-Vincent and G. Vickery. Participative web: User-created content. *Working Party on the Information Economy*, page 74, 2007.