Technological foundations of the current Blogosphere

Vangelis Banos
Department of Informatics
Aristotle University of Thessaloniki
Greece
vbanos@gmail.com

Karen Stepanyan
Computer Science Department
University of Warwick
Coventry, United Kingdom
K.Stepanyan@warwick.ac.uk

Mike Joy
Computer Science Department
University of Warwick
Coventry, United Kingdom
M.S.Joy@warwick.ac.uk

Alexandra I. Cristea
Computer Science Department
University of Warwick
Coventry, United Kingdom
A.I.Cristea@warwick.ac.uk

Yannis Manolopoulos
Department of Informatics
Aristotle University of Thessaloniki
Greece
manolopo@csd.auth.gr

ABSTRACT

In this paper, we review the technological foundations of the current Blogosphere. The review is primarily based on a large-scale evaluation of active blogs. The extensive list of examined technologies enables commenting on a range of widely adopted standards and potential trends in the Blogosphere. The evaluation has been conducted in the following stages:

1. Retrieving and parsing a large set of blogs
2. Identifying and quantifying the use of technologies such as web standards, adopted services, file formats and platforms.
3. Analysing collected data and reporting the results
4. Comparing the results with existing findings from the generic Web to identify similarities and differences in the Blogosphere.

The presented work was performed as part of BlogForever (ICT No. 269963), an EC funded research project aiming to aggregate, preserve, manage and disseminate blogs. The results of this study are relevant within the context weblog preservation and weblog data extraction.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Blogosphere, Blog Technologies, Web Technologies, Data Extraction

1. INTRODUCTION & RELATED WORK

BlogForever is a research project aiming to aggregate, preserve, manage and disseminate blogs. In order to achieve its goals, it is crucial to gain an understanding of the Blogosphere by learning more about the technology used by current weblogs.

It is therefore necessary to explore patterns in weblog structure, data and semantics, blog-specific APIs, social media interconnections and other unique blog characteristics.

There are already a number of important initiatives aiming to identify and record how the web is constructed and what its main ingredients are. The HTTP Archive[1] is a permanent repository of web performance information and technologies utilized. W3Techs[2] also provides information about the usage of various types of technologies on the web. Alexa Internet[3] is a company that is maintaining a database of information about sites, including technical information, since 1996. On the other hand, there are also many initiatives which gather web and especially blog information via user surveys and online questionnaires. Technorati's state of the blogosphere[4] is the most high profile one but there are also others such as The State of Web Development[5]. However, while some of the abovementioned initiatives publish descriptive statistics about the technological foundations of the Blogosphere, the scope and depth of these studies remains limited. For instance, while Technorati [1] may publish basic statistics about the most widely adopted platforms and popular devices for accessing blogs, the use of libraries, formats and tools remain beyond the focus of the review. Consequently, there appear to be no other initiatives, to the best of our knowledge, that conducts technical surveys and evaluate the technological foundations of the Blogosphere. This paper addresses this gap.

WIMS'12, June 13-15, 2012 Craiova, Romania

[1] http://httparchive.org

[2] http://w3techs.com/

[3] http://www.alexa.com

[4] http://technorati.com/state-of-the-blogosphere/

[5] http://www.webdirections.org/sotw10/

The rest of this paper is organized as follows: Section 2 presents the implementation of the technical survey. Section 3 presents and analyses the results of the survey. Section 4 outlines the differences identified between the generic web and the Blogosphere. Section 5 includes conclusions and planned future work.

2. TECHNICAL SURVEY IMPLEMENTATION

2.1 Accessing and parsing a large set of blogs

The main goal of this study is to evaluate the use of third-party libraries, external services, semantic mark-up, metadata, web feeds, and various media formats in the Blogosphere. To achieve this, a considerably large set of blogs has been studied. The sample of blogs has been acquired primarily from the Weblogs.com[6] ping server.

Weblogs.com receives notifications when new content is published on blogs and, subsequently, notifies its subscribers about recent updates. Hence, Weblogs.com is considered a hub between publishers and generally large-scale consumers of content (e.g. search engines). Relying on XML-RPC-based ping mechanism, Weblogs.com provides a quick and efficient interchange service between the two sides. As one of the first recognized ping servers Weblogs.com remains a widely used platform with a large number of daily notifications (around 4 million pings) coming from blogs, news and other information sources. The benefits of using ping update services are widely recognized for supporting the visibility across the Blogosphere and the Web in general.

The choice of using Weblogs.com for this evaluation is justified by two factors. Firstly, Weblogs.com remains a widely accepted and popular service in the Blogosphere, which makes it suitable for conducting a broad survey with a large sample of blogs. Secondly, Weblogs.com publishes a list of resources updated within the last hour. Using a list of recently updated resources can eliminate abandoned or inactive blogs which constitute about the half of all the blogs [2-4]. This enabled keeping the focus of the paper on active blogs only.

In addition to using Weblogs.com, additional resources for accessing weblog data have been considered. More specifically, the study extended its data collection to include the list of Top 100 blogs published by Technorati.com[7], Top 40 blogs published by Blogpulse.com[8] and a collection of blogs acquired from the BlogForever Weblog Survey[5].

The inclusion of additional blogs shared by participants of the survey extends the automatically generated list of blogs with a set of selectively contributed ones. On the other hand, the use of Technorati and Blogpulse provides a potential for enriching the evaluation. Technorati and Blogpulse are among the earlier and established authorities on indexing, ranking and monitoring blogs. Inclusion of top blogs from Technorati and Blogpulse enables a comparative analysis between the more general Weblogs.com cohort and the list of highly ranked blogs.

2.2 Data collection methods

The datasets for this study have been acquired by accessing the list of blogs from the above mentioned sources. The content of the accessed resources (i.e. the source code of the web page acquired via HTTP) was then evaluated for the presence of specific technologies, tools, standards and services.

To implement the data collection, custom software was implemented using a combination of PHP[9] and Bash[10]. More specifically, PHP5.3 was used to implement the core of the application. The CURL[11] network library was used to implement communication with the blogs via HTTP and regular expressions where utilised in order to parse the blog source code and evaluate the use of certain technologies. Finally, Bash was used to implement process management and file I/O.

The software is a Linux command line application which requires a URL list text file as input and generates CSV files with the results. For each URL in the input file, the application performs an HTTP request and retrieves the respective HTML code. Subsequently, a set of regular expressions are executed, one for each technology or digital object type we are trying to detect, and the results are stored in a comma delimited CSV file. It must be noted that input URLs can be blog base URLs but also specific blog post URLs. In either of the cases, the software retrieves the specific URL's HTML code and proceeds to parse and analyse it.

The complete software for implementing this survey is freely available via github[12].

2.3 Datasets

The overall number of accessed blogs was 259,930. HTTP response codes have been recorded. Items where status code was not retreated successfully were discarded. The acquired data was considered valid for analysis only when 200 (OK) status code was received. 94% of all the received status codes were successful. The total number of valid (i.e. Response Status Code: 200) records surveyed was 209,830. The summary of the registered response codes is shown in Figure 1.

[6] *Weblogs.com (http://weblogs.com) intends to provide a free, open access ping server.*

[7] *Technorati.com (http://www.technorati.com)*

[8] *Blogpulse.com (http://www.blogpulse.com)*

[9] *PHP programming language http://www.php.net*

[10] *GNU Bash http://www.gnu.org/s/bash/*

[11] *cURL Network Library http://curl.haxx.se/*

[12] *BlogForever git repository http://github.com/BlogForever/TechSurvey*

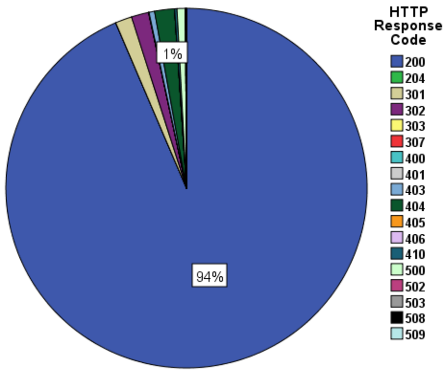


Figure 1, HTTP Status response codes registered during the data-collection stage

The datasets are available for download from the BlogForever project website[13].

Weblogs.com Dataset: An XML file published by Weblogs.com has been downloaded on 12 August 2011 at 13:00 GMT. This file usually contains names and URLs of resources submitted to the ping server within the last hour. Given the earlier studied patterns of activity within the Blogosphere [6] an informed decision on the time of collecting the data was made. The choice of the specified time frame for accessing the data is justified by the anticipated increase in publishing activity of blogs in European and other states within the time zone proximity. The XML file was parsed and URL entries were extracted for further processing.

The URL entries have been filtered to distinguish between updated resources and their hosted websites. Duplicate entries have also been removed.

Two separate datasets for individual pages and hosting websites were generated after accessing and evaluating each of the URLs.

Total number of accessed resources: 259,286

Total number of valid records: 209, 560

The number of accessed resources contains all the URLs that have been extracted and followed by the survey script. The validity of the record is defined by the HTTP response code received when making an attempt to access the resource. The data transmitted along with an HTTP response code other than 200, was stored, but was included in the evaluation.

Technorati and Blogpulse Datasets: The list of top 100 and top 40 blogs ranked by Technorati and Blogpulse respectively has been acquired on 12 August 2011.

The URL entries to top-ranked blogs have been extracted for compiling the datasets.

Total number of accessed resources: 140

Total number of valid records: 125

Contributed Blogs: The URL entries of all contributed blogs were made available after processing BlogForever user survey results.

Total number of accessed resources: 504

Total number of valid records: 145

Total Data Corpus:

Overall total of accessed resources: 259,930

Overall total of valid records: 209,830

The total number of valid records constitutes the studied dataset.

2.4 Evaluation Method

The methods for evaluating the use of certain technologies were limited to parsing the source code of accessed resources and looking for evidence of adopted technologies. The list of technologies that were considered as part of this evaluation are summarised in Table 1.

Table 1, Technologies considered in the evaluation (+count indicates that number of identified occurrences was counted)

HTTP Response Status Code (200, 404, etc.)	img-BMP (+count)	Prototype.js
Atom Feed	img-JPG (+count)	RDF (+count)
Atom Feed Comments	img-WEBP (+count)	RSD
Content Type	HTML5	RSS
CSS (+count)	JavaScript (+count)	RSS-comments
Dojo.js	JQuery.js	SIOC
Dublin Core (+count)	JQueryUI.js	Software/Platform
ExtCore.js	Microdata	Twitter
Facebook	Microformat-hCard	Embedded YouTube video
Flash (+count)	Microformat-XFN	YUI.js
FOAF	MooTools.js	XHTML
Google+	Open Graph Protocol (+count)	Other MIME Types (see Table 2)
img-PNG (+count)	Open Search	
img-GIF (+count)	Pingback	

3. EVALUATION RESULTS

3.1 Platforms and software used

The data, collected from the studied blogs, included some information about the hosting platform that powered the blogs. The analysis in this section is based on the combined dataset that includes the primary source of from Weblogs.com, as well as less extensive sources of Technorati, Blogpulse and list of URLs contributed by the participants the online survey. The information was obtained from the `<meta>` tag that included attributes `generator` and `content`. In addition to the type of software information about its version was also included were available. The most frequent platforms that appear in the studied cohort of the blogs are WordPress (36%) and Blogger (19%).

[13] *BlogForever project http://blogforever.eu*

Technorati, similarly to our findings reported WordPress, followed by Blogger, to be the platform of choice. However, the number of WordPress instances observed within the studied dataset is considerably lower from the 51% reported by Technorati. Similar observation was made in relation to the Blogger platform. These differences may be due to a large number of cases (40%) for which information about the platform remained hidden (Figure 2).

A still considerable number of instances were registered for Typepad, vBulletin Discuz and Joomla. Among other (2%) frequently appearing platforms are: Webnode, PChoc, Posterous, Blogspirit, DataLife Engine and BlueFish. The total number of unique platforms registered however is considerably large – totalling 469 unique platforms. However, even combined together they do not exceed the 19% of the entire list of studied blogs. It remains an open question why a large number of blogs do not exhibit the platforms they are built on. It requires further investigation to identify whether some blogs prefer not to acknowledge the use of a certain blogging engine or whether they are based on custom systems.

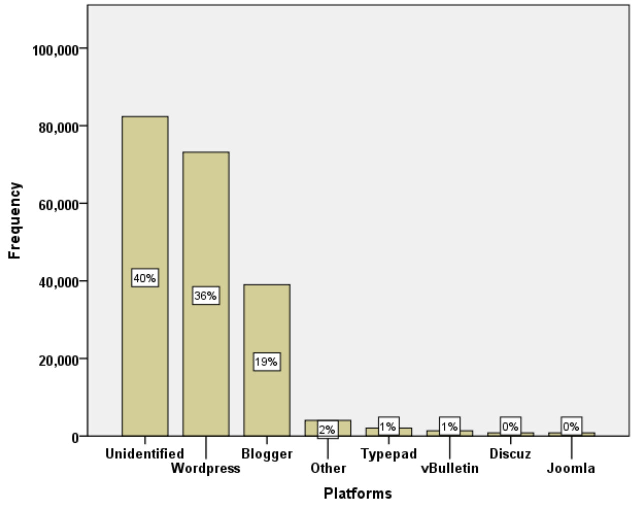


Figure 2, Frequency of weblog-powering software platforms

There is a considerable variation across most popular software platforms used. The consistency in specifying versions of adopted software varies too. However, it is still possible to identify the extent of adoption and noticeable patterns within the studied corpus.

Firstly, and most importantly, it becomes apparent that a large number of websites are maintained without a software upgrade, despite the availability of more recent versions. For instance, 20% of all the Movable type blogs continue using version 3, as shown on Figure 4, despite the availability of versions 4 and 5. There is a similar pattern, with around 13% (and some of the generic 4%) of the WordPress users choosing earlier versions of software released between 2004 and 2009, despite the availability of newer versions. Therefore, from the perspective of blog preservation, and within the context of BlogForever, decisions need to be made on whether anticipated archiving solutions should accommodate blogs that remain active, but are still powered by earlier, and possibly no longer supported software.

While the number of earlier platforms across active blogs remains substantial, the majority of software platforms (with an average of around 75%) use more recent versions. These results are limited to the providers of software packages that do specify their versions. Among the providers that do not specify information about the software version are: Blogger, Typepad and Joomla.

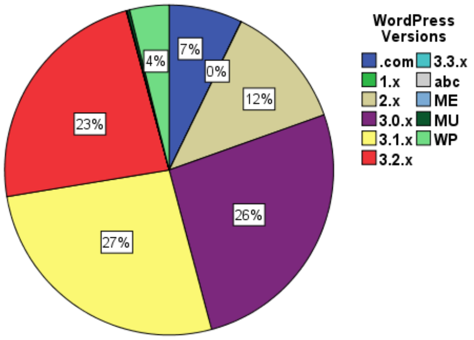


Figure 3, Variation in versions of Wordpress software

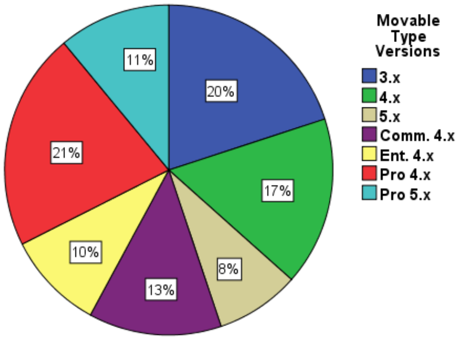


Figure 4, Variation in versions of MovableType software

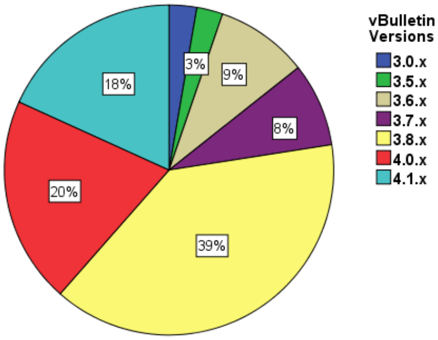


Figure 5, Variation in versions of vBulletin software

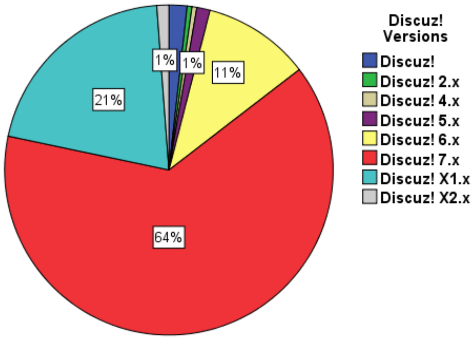


Figure 6, Variations in versions of Discuz! software

3.2 Document Character Sets

Documents transmitted via HTTP are expected to specify their character encoding. Character encoding defines the type text, such as text/html, text/plain, etc. Often referred to as “charset”, it represents a method of converting a sequence of bytes into a sequence of characters. When servers send HTML documents to user agents (e.g. browsers) as a stream of bytes, user agents interpret them as a sequence of characters. Due to a large number of characters throughout written languages, and a variety of ways to represent them, charsets are used to help user agents rendering and representing them.

It is, therefore, recommended by the W3C [7] to label Web documents explicitly by using `<meta>` element as a way of conveying this information. An example of specifying character encoding is given below:

```
<META http-equiv="Content-Type"
content="text/html; charset=EUC-JP">
```

User agents are expected to work with any character encoding registered with IANA[14], however, the support of an encoding is bound to the implementation of a specific user agent.

This evaluation recorded the use of `content` and `charset` attributes across the studied blogs. This enabled commenting on most widely used charsets or the absence of the recommended labeling. Information about the types of documents distributed by blogs was also collected.

The results suggest that text/html is the most widely (61%) specified content type within the studied corpus. Other types constitute to less than 1% and include: application/xhtml; /xml; /xhtml+xml; /vnd.wap.xhtml+xml, as well as text/xml; / javascript; / phpl; / shtml; and / html+javascript. A considerable number of accessed resources were not labelled.

In addition to content type, information about encoding has also been captured and analysed here. UTF-8 is most frequently used encoding. Other identified charsets did not exceed 6%. Encoding information was not specified or remained unidentified in 39% of the cases (Figure 7). The number of blog instances that do not specify charset information are worthy of notice. Within the 6% of other types of charset specifications 48 distinct records were identified. Most common charset specifications included: iso-8859-1 (48%), euc-jp (23%), shift-jis (8%) and windows-1251 (6%). See Figure 8 for more details.

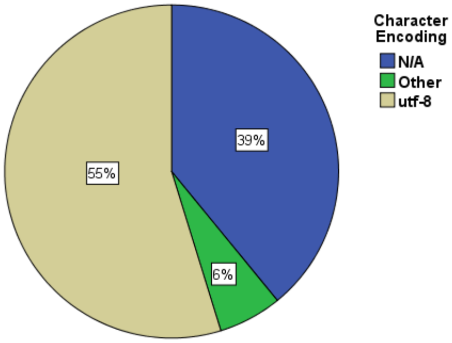


Figure 7, Encoding of evaluated resources

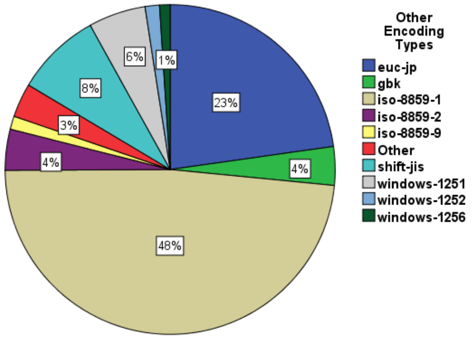


Figure 8, Break down of the other 6% (see Figure 8) of character set attributes

The results demonstrate that the overwhelming majority of studied resources are distributed in Unicode as text/html documents. A still considerable number (6%) of resources are using alternative encoding. It may therefore be required to consider solutions for capturing and preserving the blogs distributed in character sets other than UTF-8.

3.3 Use of CSS, Images, HTML5 and Flash

This section discusses the findings of the study into the use of: CSS, HTML5, Flash and certain image file formats. The dataset includes:

- Number of embedded references to CSS files linked
- Presence of HTML5 based on `<!DOCTYPE>` declaration
- Number of Flash objects used based on references to SWF files
- Number of png, gif, bmp, jpg, webp, wbmp, tiff and svg images used

Cascading Style Sheet (CSS)[15] is a language that enables separation of content from presentation. Used primarily with HTML documents, CSS provides a common mechanism for shared formatting among pages, improved accessibility and greater flexibility and control over the presentation elements of various web documents.

The study demonstrates that most of the accessed resources use CSS elements (without distinguishing between CSS1 and CSS2). The average number of references to CSS is 1.94 – suggesting a frequent use of this technology. 81% of all the studied resources employed CSS.

HTML5[16] is the fifth and (on the day of writing this document) the most recent revision of the HTML language. HTML5 intends to improve its predecessors and define a single markup language for HTML and XHTML. It introduces new syntactical features such as, `<video>`, `<audio>`, `<header>` and `<canvas>` elements, along with the integration of SVG content.

This evaluation looked into adoption of HTML5 within the studied corpus. The results suggest that only 25% (53,546) of all the considered resources are using HTML5. However, it is important to specify here that identification and count of native to HTML5 elements was not performed as part of this study.

[14] *http://www.iana.org/assignments/character-sets*

[15] *http://www.w3.org/Style/CSS/*

[16] *http://dev.w3.org/html5/spec/Overview.html*

Image File Formats that are primarily used on the Web vary widely. Graphical elements displayed on websites are primarily divided into raster and vector images. Raster images, however, are more widely used across the web. This study identified and quantified the number of images used within each of the accessed resource. The raster formats used here include: png, gif, bmp, jpg/jpeg, webp, tiff and wbmp. SVG graphics were considered from the range of vector formats. Figure 9 outlines the use of file formats in the studied corpus of resources. Most frequently used formats are JPG, GIF and PNG images. The average number of these graphic types per web page is between 4 and 8.

The overview of the less frequently used images is shown in Figure 11. The largest number for (and the only instance of) SVG images identified within the dataset is 5. This explains the low value of the averages. The average number of BMP images is the largest with 0.02 per accessed resource. The average of other file types does not exceed 0.01.

Interestingly, the average number of resources with no images identified was considerably high (21.2%). Figure 11 illustrates the frequencies of images identified on a single resource. 90% of all the pages exhibited less than 40 images. The long tail of distribution indicates a rapid decline in the number of websites using large numbers of images.

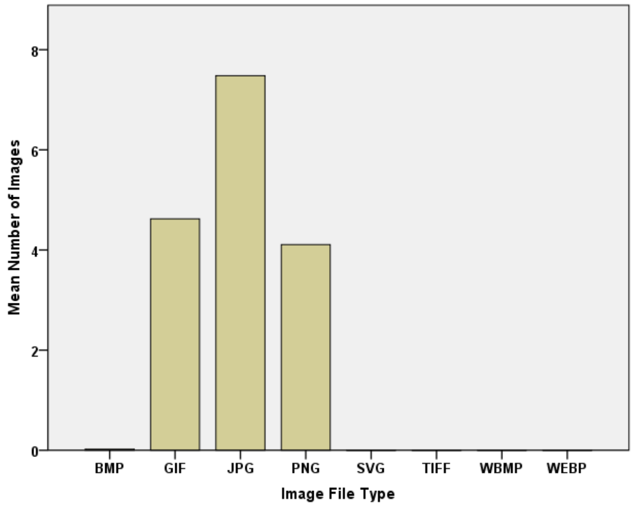


Figure 9, Average number of images identified

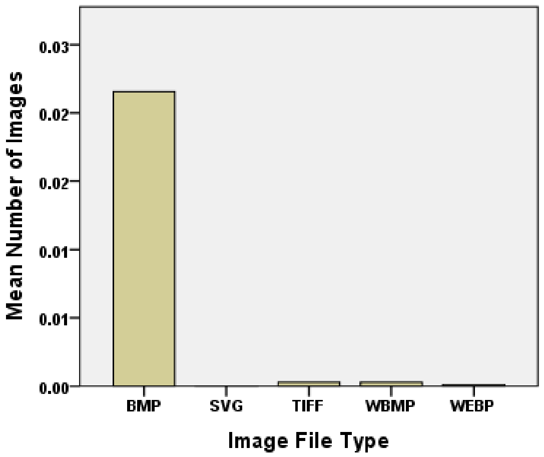


Figure 10, Average use of BMP, SVG, TIFF, WBMP and WEBP formats

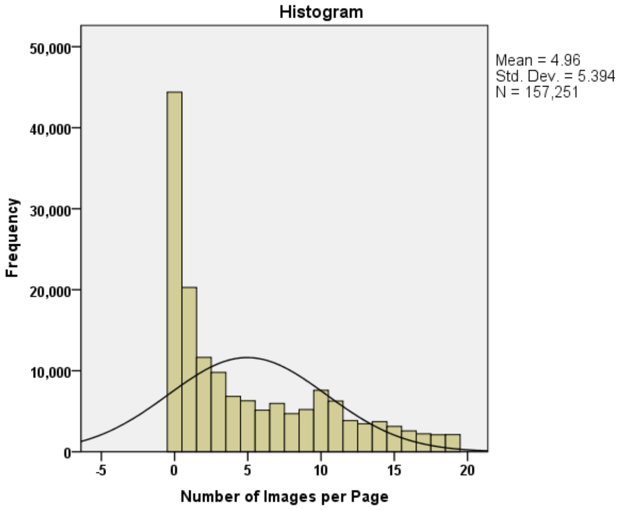


Figure 11, Distribution of images for pages with less than 20 images only

Flash[17]**,** also known as Macromedia/Adobe Flash, is a multimedia platform used for adding interactivity or animation to web documents. It is frequently used for advertisement, games streaming video or audio. Flash is provided by using an object-oriented ActionScript programming language and allows the use of both vector and rasterised graphical content.

The detection of Flash content within the studied resources was based on the use of SWF format. Accessed resourced were searched for `<object>` elements with a source that points to an `*.swf` file. The instances of Flash content were counted as well. The results indicate that the overwhelming majority (85%) of the accessed resourced did not include any Flash content. Around 7% of all the resources were identified as having a single reference to a Flash object. The number of occurrences exceeds double figures only in exceptional cases.

[17] http://www.adobe.com/products/flash.html

3.4 Semantic Markup: Microformats, Microdata and Metadata

One of the objectives of this evaluation was to evaluate the adoption of semantic mark-up within the Blogosphere. To address this objective this investigation looks into the use of metadata formats and associated technologies. This section discussed the use of:

- **Metadata**
  - Dublin Core
  - The Friend of a Friend (FOAF)
  - Open Graph Protocol (OG)
  - Semantically-InterlinkedOnline Communities (SIOC)
- **Micro/data/formats**
  - Microdata
  - hCard (Microformats)
  - XFN (Microformats)
- **Common Semantic Technologies**
  - Resource Description Framework (RDF)
  - Really Simple Discovery (RSD)
  - Open Search

Metadata are commonly defined as data about data. Within the context of the Web, metadata are commonly referred to as the descriptive text used alongside web content. Examples of metadata can include keywords, associations or various content mapping. It is often required to standardize these descriptions for ensuring consistency and interoperability of web content. Referring to Dublin Core, Open Graph, SIOC and FOAF as simply metadata would be inaccurate. However, their use is discussed jointly due to some similarities of their application.

The summary of identified uses of metadata standards is presented in Open Graph (OG) is most frequently used standard (see Figure 12). Each of the instances of OG and DC mark-up has been counted. The average occurrence of OG is 5.7 per page compared to 1.37 for DC.

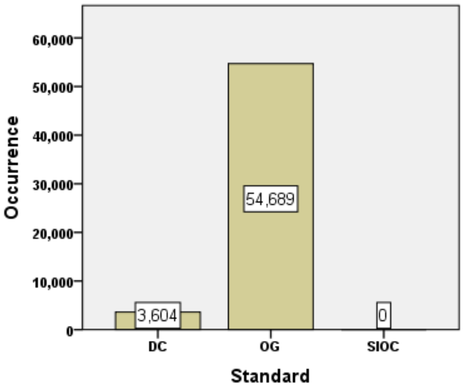


Figure 12, Summary of metadata usage

The histogram of OG occurrences is shown in Figure 13. The use of FOAF has been identified in only 561 cases, which constitutes to less than 0.3% of all the studied pages. The overwhelming majority of evaluated resources did not use FOAF. Across the entire corpus of studied resources no reference to SIOC was identified.

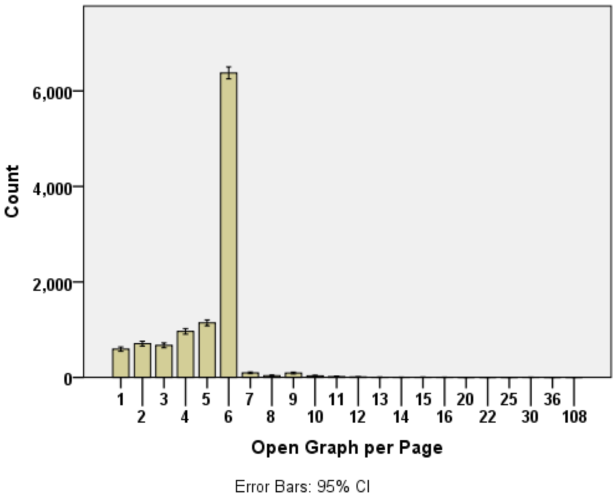


Figure 13, Histogram of Open Graph references

Microdata[18] **and Microformats**[19] are conceptually different approaches to enriching web content with semantic notation. This evaluation counted the number of resources where presence of microdata or microformats has been identified. More specifically, when referring to microformats, the investigation distinguished between XFN, a way of representing human relationships using hyperlinks, and hCard – a simple, distributed format for representing people, companies, organisations, and places. The presence of Microdata within a resource was based on locating `itemscope` and `itemtype="http://schema.org/*"` within a studied page. hCard and XFN microformats were identified, respectively, as class attributes with hcards values and `rel` attributes within `<a>` tags.

To add a property to an item, the itemprop attribute is used on one of the item's descendants. The use of XFN was identified in 74,709 cases, which constitutes to 35.6% of the entire corpus. On the opposite, the use of microdata and hCards was less frequent. Only 27 instances of microdata were identified within the studied resources. The number of identified hCards was limited to 607 (0.3%). A large portion of the studied corpus contained no evidence of either microdata nor microformats.

Common Semantic Technologies considered in this evaluation are limited to the use of: RDF language, Open Search and Really Simple Discovery (RSD) formats. The identification of RDF was based on finding description of resource types – application/rdf+xml. The identification of Open Search format was based on the use of application/ opensearchdescription+xml content type and the use of a relevant namespace declaration:

```
<OpenSearchDescription xmlns=
"http://a9.com/- /spec/opensearch/1.1/">
```

Similarly, the identification of RSD was based on the following namespace declaration:

```
<rsd version="1.0" xmlns=
"http://archipelago.phrasewise.com/rsd" >
```

[18] *MicroData http://www.w3.org/TR/microdata*

[19] *MicroFormats http://microformats.org/about*

The results demonstrate the use of RSD is widespread. About 74% of all the accessed resources were identified as using RSD. On the contrary, only 567 records (0.3%) are using RDF. No references to Open Search were identified.

3.5 RSS and Atom Feeds

Web feeds, like RSS and Atom, have been widely used across weblog platforms and services. Represented in a machine readable format, web feeds enable data sharing among applications. Most common use of web feeds is to provide content syndication and notification of updates from multiple websites into a single application [8]. Aggregators or news readers are commonly used for syndicating the web content by enabling users to subscribe to web feeds. The simple mechanisms for accessing and distributing web content justify the wide adoption of feeds on weblog platforms.

The use of web feed within the studied resources have been identified by the use of the `<link>` tag with `type="application/atom+xml"` for Atom feeds, `type="application/rss +xml"` for standard RSS feeds with an additional distinction to comments where applicable. The results are outlined in Figure 14. RSS feeds are most widely used (56%) feeds. The use of Atom feeds (29%) is still common. 15% of RSS feeds were used distinctly for distributing the content of comments. Yet, no Atom feeds were identified for this purpose.

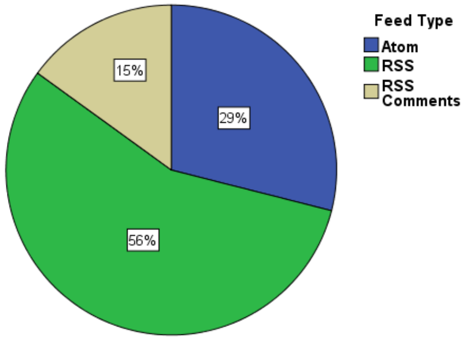


Figure 14, Use of XML feeds by type

3.6 APIs and Libraries

This section discusses the use of JavaScript client-side Object Oriented programming language and a set of libraries adopted by the studied resources. Among the studied libraries and frameworks are:

- Dojo[20]
- Ext Core[21]
- JQuery[22]
- JQuery UI[23]
- MooTools[24]
- Prototype[25]
- YUI Library[26]

In addition to the above mentioned libraries this section discusses the use of Pingback services throughout the studied cohort.

To use of JavaScript by each of the accessed resource has been quantified based on the number of *.js files linked or segments of JavaScript code embedded within the accessed document. The results suggest a wide adoption of JavaScript with 82% of the entire studied corpus having at least one reference to JavaScript. The average number of JavaScript instances is large too – 12.5 instances per resource (Figure 15).

Within the identified instances of JavaScript code, there are references to specific libraries and frameworks. There use is identified by the reference to their name (e.g. dojo.js, jquery.js, etc.). The most frequently used technologies are JQuery, Moo Tools and YUI Library. The cumulative use of Dojo, Ext Core JQuery UI and Prototype constitute to just over 1% of all the accessed resources (Figure 16).

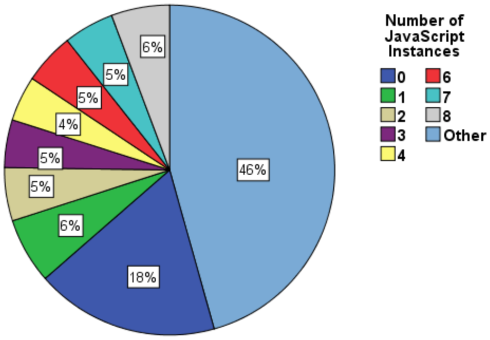


Figure 15, Number of Javascript instances identified

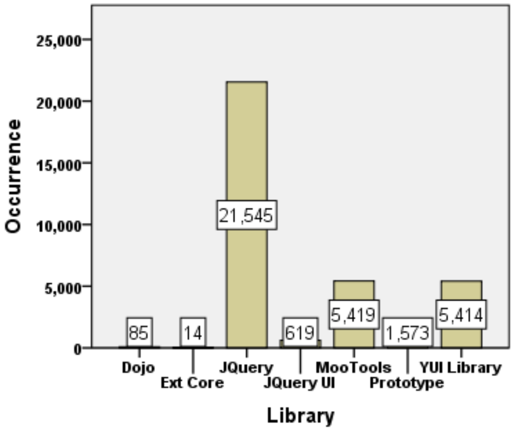


Figure 16, Number of identified library/framework instances

Last, but not least, this sections summarises the use of Pingback[27]APIs. The identification of Pingback is based on the reference of `<link>` tags with `rel="pingback"` attribute within the accessed recourses. The results suggest that 46.4% of all the accessed resources used pingbacks. The use of other

[20] http://dojotoolkit.org/

[21] http://www.sencha.com/blog/ext-core-30-beta-released/

[22] http://jquery.com/

[23] http://jqueryui.com/

[24] http://mootools.net/

[25] http://www.prototypejs.org/

[26] http://developer.yahoo.com/yui/

[27] http://hixie.ch/specs/pingback/pingback-1.0

Linkback mechanisms, including Trackbacks and Refbacks have not been considered in this evaluation. The use of other third party libraries such as Google Analytics were also omitted.

3.7 Social Media

The rise of social media such as Facebook, Twitter and YouTube is believed to have a profound effect on people's blogging behaviour and the Blogosphere in general. A large number of blogs already integrate mechanism for easy distribution of its content on social media websites. Social media are used for promoting and notifying readership about new posts. This section summarises the investigation into the use of social media within the studied corpus of resources. It outlines the extent of adoption of Twitter, Facebook, Google+ and YouTube.

The use of Twitter, Google+ and Facebook were considered integrated with the accessed website when a use of specific JavaScript libraries and XML namespaces with appropriate references to Twitter, Google and Facebook sources are used. The results suggest that almost 4% of all the studied resource indicate an evidence of integration with Facebook. The number of references to Twitter are marginal with only a handful of identified instances. The adoption of Google+, on the other hand, is shown to be considerably higher – totaling 17.2% among the studied resources. This high number of instances is surprising given the announcement of the service less than two months ago from the time of writing this report.

The use of YouTube was studied differently from that of earlier discussed social media. Each of the accessed resources were scanned for occurrences of embedded content from YouTube. The use of `<iFrame>` that points to the source of the hosting site was used to count the number of instances of embedded YouTube content. The results suggest that more than 10% of all the studied resources are using embedded YouTube videos. However, the number of references to embedded content within each of the resources is fairly large. The results demonstrating the use of YouTube is shown in Figure 17(for convenience, the 0.6% of outliers was reduced to 20).

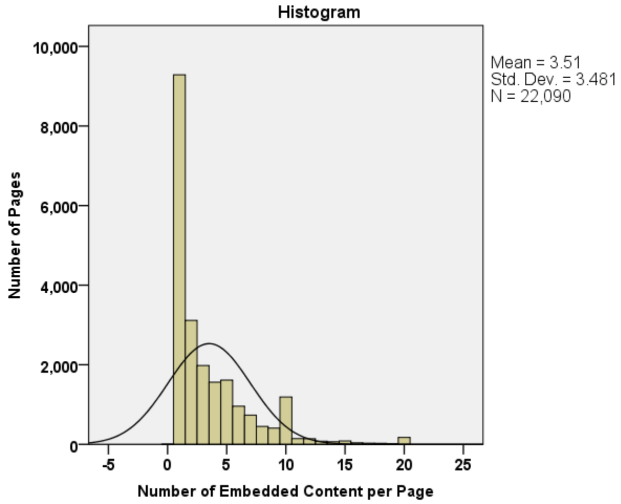


Figure 17, Frequency of embedded YouTube videos

3.8 Media Types and Common File Formats

This evaluation was extended to consider the use of various file formats described as MIME types by Internet Assigned Numbers Authority (IANA)[28]. This evaluation looked into some of the files categorised as audio, video, text, and applications. Originally used to describe email content MIME standard extends further and used along with communication protocols like HTTP. Similarly to email, HTTP requires certain data be transmitted where MIME specification is considered suitable.

The full list of the studied file formats and the frequency of their use as part of the accessed resources is presented in Table 2.

Table 2 – File MIME types and frequency of their occurrences.

File Ext.	Application	Instances
doc	Word Processing	1097
docx	Word Processing	147
odt	Word Processing	51
pdf	Word Processing	13731
txt	Word Processing	641
mp4	Video/Audio	3265
mpeg	Video	36
mpg	Video	613
avi	Video	3265
mov	Video	71
3gpp	Video	1429
xls	Spread Sheet	138
xlsx	Spread Sheet	24
ods	Spread Sheet	722
ppt	Presentation	67
pptx	Presentation	20
odd	Presentation	618
odf	Math Formulas	63
odg	Graphics	4
mdb	Database	0
ccbd	Database	0
odb	Database	153
vCard	Card	14
mp3	Audio	10231
wav	Audio	13
vrml	3D	0

The results suggest that the most frequently (13,731) used file type across the studied corpus is PDF. Slightly less frequent (10,231) occurrences were recorded for mp3 Audio files. The use of MS Word documents, AVI and MP4 videos is between 1,097 and 3,265. No database or 3D reality files were identified within the studied corpus.

[28] *http://www.iana.org/assignments/media-types/index.html*

Given the large number of resources studied as part of this evaluation, even most frequently used file types constitute to a small proportion. The use of MS Word and PDF documents is between 4.9-6.4% of all the studied resources. The combined use of all audio and video files constitutes 9% of all the studied resources.

3.9 Single Posts versus Websites

The data contained in the dataset published by Welogs.com contain both URLs that refer to single posts/pages as well as general domains. The distinction between the two was introduced during the data collection stage. This enables discussing the differences between the use of technologies on the levels of single posts/pages and larger websites..

The results suggest that the average number of technologies used on the website level is approximately twice as large as that on a single page/post level. This does not hold for every element studied here. For instance the use of YUI JavaScript library is 2.5 times more frequently used on the post/page level than on a website level. Almost twice more FOAF references were recorded on the post/page level compared to general websites. The number of GIF images used is also slightly higher on the post/page level compared to their use on the home page.

On the contrary the number of JPG images used on the website level is 3.4 times higher than on the post/page level. A similar pattern holds for embedded YouTube videos with 5 times more videos used on a website level. These results are not surprising since, posts and pages contain more focused content compared to homepages that may include listings with excerpts from a set of posts.

4. DIFFERENCES BETWEEN THE BLOGOSPHERE AND THE WEB

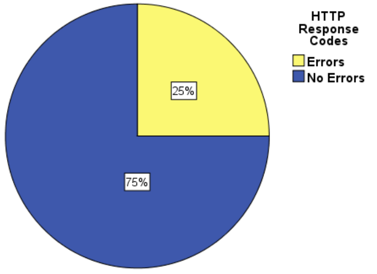
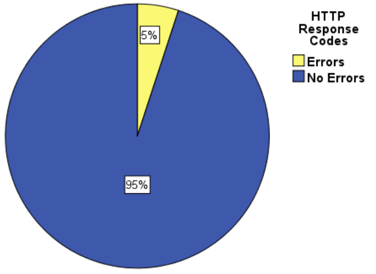
This section questions whether prominent technological difference exists within the domain of blogs and the Web in general. It compares the data published by HTTP Archive with the data obtained from Weblogs.com and discussed in this evaluation.

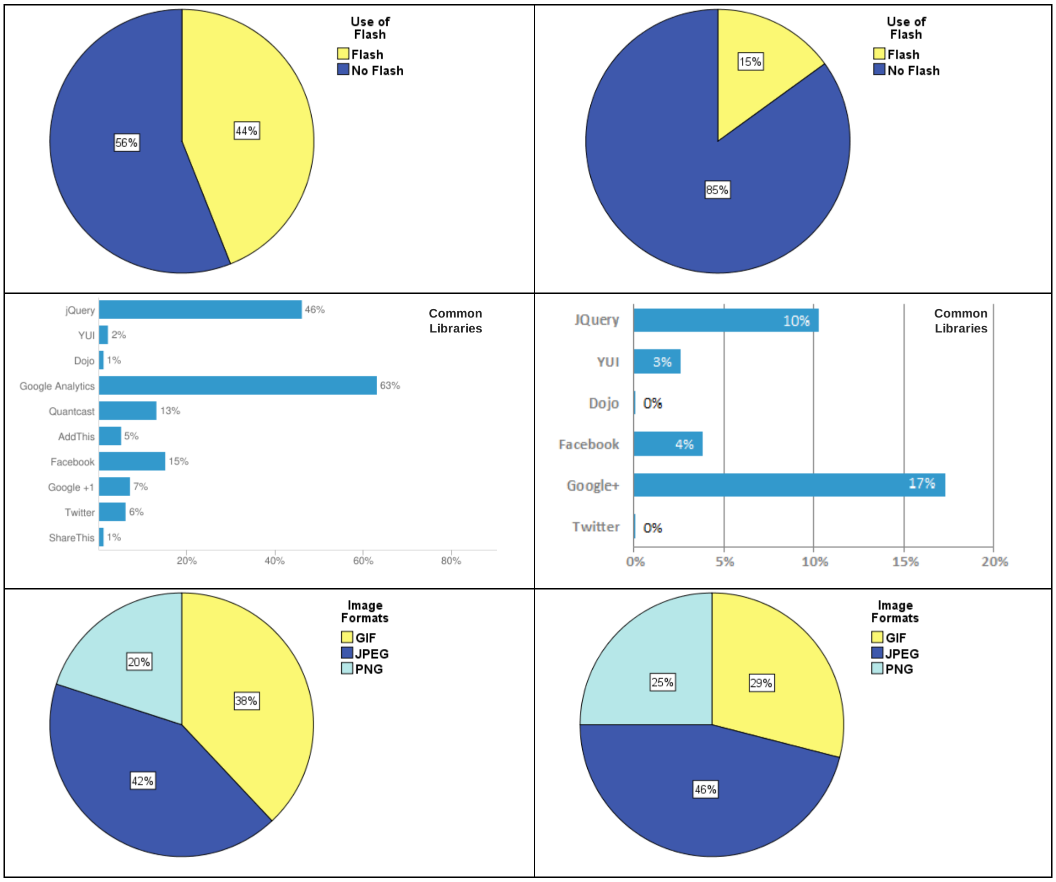
Http Archive attempts to record the changes that take place on the Internet by collecting and storing Web content. Furthermore, HTTP Archive attempts to preserve the ways in which Web content is being constructed and served including information about: size of pages, failed requests and technologies utilized. We aim to compare the commonly occurring technologies with those in the Blogosphere. The rationale for comparing these domains is to identify whether major differences between the Blogosphere and the generic Web exist and whether different strategies may be necessary for blog preservation. Some differences between the Blogosphere and the rest of the Web have already been demonstrated. The spread of information and influence dynamics within blog networks are among common perspectives that distinguish blogs from other websites. For instance, emerging patterns of information propagation and content sharing within the Blogosphere have already been highlighted [9]. However, more work is needed to capture technological differences between the two domains.

The data used from the HTTP Archive corresponds to the timeframe of the data obtained from Weblogs.com. Identical period sourcing the data ensures the comparability of the datasets and eliminates the possibility of technological changes that may affect the results.

The results indicate the presence of considerable differences between the two domains, particularly in relation to the use of Flash, Gif images and JavaScript libraries. Side-by-side comparison is presented in in Table 3.The results indicate a considerable variance in using Flash elements generally as part of the Web (44%), compared to only 15% within the Blogosphere. It is also apparent that the adoption of PNG images is higher (25%) within the Blogosphere compared to the general Web (20%). There is a greater number of GIF images used within the general compared to the blog domain. Furthermore, a considerable difference was observed in the number of recorded HTTP response errors. While conclusive evaluation will requires expanding the study beyond the recently active blogs, the descriptive statistics presented here highlights the problem of no longer accessible web resources.

Table 3, Comparison between generic web and the Blogosphere

The Web	The Blogosphere
	



5. SUMMARY AND FUTURE WORK

This paper outlines the results of an evaluation that investigated the use of various technologies in the Blogosphere. The key message emerging from the study argues for the diversity of the Blogosphere. More specifically, there are large numbers of *software platforms*, *encoding standards*, *third party services* and *libraries* used. There are considerable differences in the ways the standards are being adopted. In the context of BlogForever and preservation of blogs in general, this diversity exhibited in the Blogosphere may require additional efforts for avoiding data loss or distortion when aggregating, preserving and disseminating blogs. For example, given the empirical evidence that indicates limited use of certain file types and popularity of others, informed decisions can be made for focusing on specific file formats and omitting others. Informed trade-offs can enable conserving the resources and contribute to the greater sustainability of the archive.

Firstly, and most importantly, the evaluation suggested existence of around 470 platforms in addition to the dominating WordPress and Blogger. Furthermore, there is a wide variety in the versions and subsystems adopted. The wide variety of content types in addition to the 61% of text/html published in a wide range of encoding standards showcase this fact.

On the other side, however, there are a large number of established and widely used technologies and standards applied consistently throughout the Blogosphere. The use of RSS and Atom feeds, along with CSS and JavaScript are among those technologies. The frequency of images used and their formats are very similar to the ways they are used within the entire Web.

There is a wide variation on the adoption of third party libraries and services. The use of social media APIs is not consistent throughout the studied corpus. However, support for Google+, a service announced within the recent 2 months is considerably large. The adoption of metadata such as Dublin Core, Open Graph, FOAF and SIOC is not consistently spread either. This may have direct implications for crawling, data extraction and aggregation.

Consequently, and more generally, this evaluation measures and reports the technological foundations used in the Blogosphere. The results of this evaluation can, therefore, be used for many purposes by services and solutions geared towards the Blogosphere. In particular, BlogForever is currently using these results to develop strategies for crawling and preserving blog data.

Future work includes the annual implementation of the technical survey to obtain up-to date information on the dynamics, trends and changes of the Blogosphere over time. Improvements in the design of the survey could also be introduced. Randomisation of the dataset and selection of a wider range of blogs should be considered to avoid the current limitations of this study that include only a relevant, though, specific time zone. Future studies, can also take into account the size and popularity of the blog. Comparison of future surveys with current results can help identify evolutionary changes within the Blogosphere. Implementation details for the regular studies are already being discussed – defining strategies for making the technical survey faster and more efficient, while enabling the evaluation of more technologies and larger data sets.

6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission Framework Programme 7 (FP7), BlogForever project, grant agreement No.269963.

7. REFERENCES

[1] Technirati. *State of the Blogosphere*. 2011.

[2] Li, D. and Walejko, G. Splogs and abandoned blogs: The perils of sampling bloggers and their blogs. *Information, Communication & Society*, 11, 2 2008), 279-296.

[3] Venolia, G. A matter of life or death: Modeling blog mortality. *Unveröffentlichter Forschungsbericht. Redmond. Online verfügbar: research. microsoft. com/~ ginav/ljmodeling. pdf*2007).

[4] Bogh-Andersen, S. *The Danish blogosphere at the end of the decade*. City, 2009.

[5] Arango-Docio, S., Sleeman, P. and Kalb, H. *BlogForever: D2.1 Survey Implementation Report*. 2011.

[6] Duarte, F., Mattos, B., Bestavros, A., Almeida, V. and Almeida, J. *Traffic characteristics and communication patterns in blogosphere*. Boston University Computer Science Department, 2006.

[7] W3C *HTML Document Representation*. City.

[8] Hammersley, B. *Developing feeds with RSS and Atom*. O'Reilly, 2005.

[9] Klamma, R., Cao, Y. and Spaniol, M. *Watching the blogosphere: Knowledge sharing in the Web 2.0*. City, 2007.