

**SEVENTH FRAMEWORK PROGRAMME
FP7-ICT-2009-6**

BlogForever
Grant agreement no.: 269963

BlogForever: D2.1 Survey Implementation Report

Editor:	Silvia Arango-Docio
Revision:	Draft V1.0
Dissemination Level:	WP2
Author(s):	S. Arango-Docio, P. Sleeman, H. Kalb
Due date of deliverable:	31/08/2011
Actual submission date:	
Start date of project:	01 March 2011
Duration:	30 months
Lead Beneficiary name:	University of London

Abstract:

This report outlines a principal study aiming to inform the development of preservation and dissemination solutions for blogs. To achieve this, the study encompassed: [i] a user survey exploring the aspects of blog preservation and blogging practices in general; [ii] an investigation into the use of tools and technologies within the Blogosphere; and finally [iii] an inquiry into the recent theoretical and technological advances for analysing blogs and their networks.

The results of the study, as summarised in this report, enable addressing the objectives pursued as part of the D2.1 deliverable of the BlogForever project. More specifically, this report comments on: [a] common weblog authoring practices; [b] important aspects and types of blog data that should be preserved; [c] the patterns in weblogs structure and data; [d] the technology adopted by current blogs; and finally [e] the developments and prospects for analysing blog networks and weblog dynamics. As an account for the conducted work this report includes implementation details and adopted ethical guidelines. The results of the inquiry are discussed within the context of BlogForever – offering directions for further development of the project.

The **BlogForever** Consortium consists of:

Aristotle University of Thessaloniki (AUTH)	Greece
European Organisation for Nuclear Research (CERN)	Switzerland
University of Glasgow (UG)	UK
The University of Warwick (UW)	UK
University of London (UL)	UK
Technische Universitat Berlin (TUB)	Germany
Cyberwatcher	Norway
SRDC Yazilim Arastrirma ve Gelistrirme ve Danismanlik Ticaret Limited Sirketi (SRDC)	Turkey
Tero Ltd (Tero)	Greece
Mokono GMBH	Germany
Phaistos SA (Phaistos)	Greece
Altec Software Development S.A. (Altec)	Greece

History

<i>Version</i>	<i>Date</i>	<i>Modification reason</i>	<i>Modified by</i>
0.1	18 June 2011	First Draft	S.Arango-Docio
0.2	24 July 2011	Updated First Draft	S. Arango-Docio, P. Sleeman
0.3	27 July 2011	Table of Contents	S. Arango-Docio
0.4	27 July 2011	Comments received	S. Arango-Docio
0.5	30 July 2011	Second Draft	S. Arango-Docio, H. Kalb
0.5.1	03 August 2011	Adaptations to the format, Addition of literature references, Extension of the SmartPLS chapter	H. Kalb
0.5.2	03 August 2011	Comments on second draft	P. Sleeman
0.5.3	08 August 2011	Merged versions 0.5.1 & 0.5.2	S. Arango-Docio
0.5.4	18 August 2011	Updates to 0.5.3	S. Arango-Docio
0.5.5	24 August 2011	Inclusion of the chapter 4.1.1 and 4.3.3	H. Kalb
0.5.6	24 August 2011	Integration of the Technical Survey	K. Stepanyan
0.5.7	25 August 2011	Expansion, review, copy-editing and commenting	M. Joy, K Stepanyan
1.0	27 August 2011	Pre-final formatting	H. Kalb, K. Stepanyan
1.1	30 August 2011	Review and editing	M. Joy

Table of Contents

1	EXECUTIVE SUMMARY	8
2	INTRODUCTION	10
3	BLOG PRESERVATION AND STUDIES REVIEW	11
4	BLOGFOREVER SURVEY.....	13
4.1	QUESTIONNAIRE DEVELOPMENT	13
4.1.1	<i>Influence Factors for Author and Reader Behaviour</i>	<i>14</i>
4.1.1.1	Blog author intention to contribute to a blog archive.....	14
4.1.1.2	Influence factors for the search strategy in the archive	17
4.1.2	<i>Survey Pilot Testing</i>	<i>18</i>
4.1.3	<i>Survey Translations</i>	<i>20</i>
4.2	SURVEY IMPLEMENTATION.....	21
4.2.1	<i>Survey Population and Sampling.....</i>	<i>21</i>
4.2.1.1	Sampling strategy	21
4.2.1.2	Internet sampling.....	22
4.2.1.3	Sample size.....	25
4.2.2	<i>Survey software.....</i>	<i>26</i>
4.2.3	<i>Survey Implementation and Promotion</i>	<i>31</i>
4.3	DATA ANALYSES AND RESULTS.....	33
4.3.1	<i>From IProbe to SPSS and SmartPLS</i>	<i>33</i>
4.3.2	<i>SPSS and Excel Analyses</i>	<i>34</i>
4.3.2.1	Authors survey respondents summary	34
4.3.2.2	Common blog authoring practices	37
4.3.2.3	Patterns in blog structure and data	39
4.3.2.4	Network-based metrics	41
4.3.2.5	Blog lifecycle examination	43
4.3.2.4	Aspects of blogs for preservation	45
4.3.3	<i>Blog Author Intentions for Contributing to Blog Archives</i>	<i>47</i>
4.3.3.1	Measurement model	48
4.3.3.2	Structural model	49
4.3.4	<i>Influence Factors for Search Strategies within Archives</i>	<i>50</i>
5	TECHNOLOGY USED BY CURRENT BLOGS	53
5.1	OBJECTIVES, DATA COLLECTION METHODS AND DATASETS	53
5.1.1	<i>Data Collection Methods and Datasets.....</i>	<i>54</i>
5.1.2	<i>Evaluation Method.....</i>	<i>55</i>
5.2	EVALUATION RESULTS.....	56
5.2.1	<i>Platforms and Software Used</i>	<i>56</i>
5.2.2	<i>Document Character Sets.....</i>	<i>58</i>
5.2.3	<i>Use of CSS, Images, HTML5 and Flash</i>	<i>60</i>
5.2.4	<i>Semantic Markup: Microformats, Microdata and Metadata</i>	<i>62</i>
5.2.5	<i>RSS and Atom Feeds.....</i>	<i>65</i>
5.2.6	<i>APIs and Libraries.....</i>	<i>65</i>
5.2.7	<i>Social Media.....</i>	<i>67</i>
5.2.8	<i>Media Types and Common File Formats</i>	<i>68</i>
5.2.9	<i>Single Posts versus Websites.....</i>	<i>69</i>
5.2.10	<i>Differences between the Blogosphere and Web</i>	<i>70</i>
5.3	SUMMARY	71
6	ANALYSIS OF INTER-BLOG RELATIONSHIPS	73
6.1	SOCIAL NETWORK ANALYSIS.....	74
6.2	INTER-BLOG RELATIONSHIPS.....	76

6.2.1	<i>Actors</i>	76
6.2.2	<i>Relations</i>	78
6.2.3	<i>Time</i>	80
6.3	REQUIREMENTS FOR THE SPIDER AND DATA MODEL.....	81
6.3.1	<i>Weblog Spider</i>	81
6.3.2	<i>Data Model</i>	84
6.4	BENEFITS AND USE CASES.....	85
7	LIFE CYCLES OF BLOGS	87
7.1	DYNAMICS ON THE LEVEL OF BLOGS	87
7.2	DYNAMICS OF POSTS	89
7.3	SUMMARY	90
8	CONCLUSIONS	91
9	REFERENCES	94
A.	APPENDIX A – OFFLINE QUESTIONNAIRES	99
A.1	FINAL OFFLINE DESIGN FOR AUTHORS QUESTIONNAIRE.....	99
A.2	FINAL OFFLINE DESIGN FOR READERS QUESTIONNAIRE.....	113
B.	APPENDIX B - BLOGFOREVER SURVEY IPROBE SCREENSHOTS	127
C.	APPENDIX C - BLOGFOREVER SURVEY PROMOTION SCREENSHOTS	133
D.	APPENDIX D – READERS SURVEY DATA SUMMARY	134
E.	APPENDIX E - FACEBOOK, DELICIOUS, AND TWITTER	139
F.	APPENDIX F – BLOG TECHNOLOGY SURVEY SOFTWARE	141

Index of Tables

Table 1 – Items to measure the constructs in the author questionnaire	16
Table 2 – Items to measure the constructs in the reader questionnaire	17
Table 3 – Excerpt from the distribution of countries by IP data	33
Table 4 – IProbe data extraction example	34
Table 5 – Author responses by survey language	34
Table 6 – Authors by education & employment.....	35
Table 7 – Authors by country of residence	35
Table 8 – Authors by nationality	36
Table 9 – Authors by age group	36
Table 10 – Frequency of posting and editing	37
Table 11 – Frequency of mashup activities	37
Table 12 – Frequency of design activities	37
Table 13 – Frequency of dialogue activities.....	38
Table 14 – Main audience of blogs	38
Table 15 – Single authors and collaborators	38
Table 16 – Blogging service providers.....	39
Table 17 – Type of media used	40
Table 18 – Blog content creation.....	40
Table 19 – Availability of a list of links.....	40
Table 20 – Blog traffic monitoring methods	41
Table 21 – User interaction with blog content	42
Table 22 - Ranking analysis tools	43
Table 23 – Hits per day	43
Table 24 – Motivations for maintaining blogs	44
Table 25 – Meaning of blogs.....	44
Table 26 – Impact of losing a blog	45
Table 27 – Use of external services for blog preservation	45
Table 28 – Attitudes towards blog preservation in a trusted archive	46
Table 29 – Interest towards blog archiving tools for increasing readership.....	46
Table 30 – Importance of preserving blog data.....	46
Table 31 – Summary of factor loadings, average variance extracted and composite reliability	48
Table 32 – Cross loadings of the items	49
Table 33 – Average variance extracted and latent variable correlations	49
Table 34 – Summary of factor loadings, average variance extracted and composite reliability	51
Table 35 – List of technologies considered in the evaluation (<i>+count indicates that number of identified occurrences were counted</i>).....	55
Table 36 – File types and frequency of their occurrences.....	68
Table 37 – Comparison between the Web and the Blogosphere.....	70
Table 38 – Important elements and annotations for the spider development	82

Index of Figures

Figure 1 – Initial BlogForever survey structure	13
Figure 2 – Framework for author intention to contribute to the archive	16
Figure 3 – Screenshot of statistics of author survey testing	20
Figure 4 – Screenshot of statistics of reader survey testing.....	20
Figure 5 – BlogForever Facebook entry.....	23
Figure 6 – Survey promotion at http://dablog.ulcc.ac.uk	24
Figure 7 – Survey promotion at http://blogforever.eu/blog	24
Figure 8 – Example of BlogForever survey visitors at getclicky.com	26
Figure 9 – Language selection front page	27
Figure 10 – BlogForever survey introduction	28
Figure 11 – IProbe survey details.....	29
Figure 12 – Example of an iProbe alert.....	29
Figure 13 – Example of the German survey introduction	30
Figure 14 – Example of Spanish survey questions.....	31
Figure 15 - Blog traffic monitoring methods.....	42
Figure 16 - Importance of preservation of blog elements	47
Figure 17 – PLS path analysis model (** p < .01, *** p < .001).....	50
Figure 18 – HTTP response codes registered during the data-collection stage.....	54
Figure 19 – Frequency of weblog-powering software platforms	57
Figure 20 – Variations in versions of adopted software.....	58
Figure 21 – Content type of the evaluated resources.....	59
Figure 22 – Encoding of the evaluated resources.....	59
Figure 23 – Break down of the other 6% (see Figure 22) of charset attributes.....	60
Figure 24 – Average number of images identified	61
Figure 25 – Average use of BMP, SVG, TIFF, WBMP and WEBP formats.....	61
Figure 26 – Distribution of images for pages with less than 20 images only.....	62
Figure 27 – Number of Flash instances detected.....	62
Figure 28 – Summary of metadata use	63
Figure 29 – Histogram of Open Graph references	64
Figure 30 – Use of web feeds by type	65
Figure 31 – Number of JavaScript instances identified.....	66
Figure 32 – Number of identified library/framework instances	66
Figure 33 – Frequency of embedded YouTube videos.....	68
Figure 34 – Differences in use of technology on the level of posts/pages and websites.....	70
Figure 35 – Examples for blog relations	77
Figure 36 – Examples for explicit relationships with blogs	79
Figure 37 – Comments refer to the blog post or previous comments.....	82

1 Executive Summary

This report outlines a principal investigation into: [i] the common practices of blogging and attitudes towards preservation of blogs; [ii] the use of technologies, standards and tools within blogs; and finally, [iii] the recent theoretical and technological advances for analysing blogs and their networks. This investigation aims to inform the development of preservation and dissemination solutions for blogs within the context of BlogForever.

The objectives pursued in this study enabled discussion of: [a] common weblog authoring practices; [b] important aspects and types of blog data that should be preserved; [c] the patterns in weblogs structure and data; [d] the technology adopted by current blogs; and finally [e] the developments and prospects for analysing blog networks and [f] weblog dynamics.

To achieve the aims and objectives of this investigation, a set of review and evaluation exercises were conducted. The members of the BlogForever consortium jointly designed and implemented:

- ✓ An online survey involving 900 blog authors and readers;
- ✓ An evaluation of technologies and tools used in more than 200 thousand active blogs;
- ✓ A review of recent advances in theoretical and empirical research for analysing networks of blogs; and
- ✓ A review of empirical literature discussing dynamic aspects of blogs and blog posts

The methods, implementation details and results emerging from each of these exercises are reported in this document.

The online survey component, conducted as part of this study, included two distinct questionnaires intended for blog authors and readers. The questionnaires were piloted and then made available for four weeks during July/August 2011. The questionnaires were translated into six languages (English, French, German, Greek, Russian and Spanish) and promoted across various online channels. The analysis was conducted by the University of London (UL) and the Technische Universität Berlin (TUB) using SPSS, Excel and SmartPLS software. The results show that the majority of respondents rarely consider archiving their blogs. This fact illuminates the potential for irretrievable loss of blogs and their data and justifies efforts towards development of independent archiving and preservation solutions. Furthermore, the results indicate a considerable interest of readers towards a central source of blog discovery and searching services that could be provided by blog archives.

Further analysis of the survey was conducted using a structural equation modelling technique – Partial Least Squares (PLS). The results suggest that the perception of collective benefits has a stronger influence on the blog authors' intentions to contribute their blogs to archives than the perception of individual benefits. Additionally, the expectation of new or stronger relationships with other people as well as the perception of being a part of a group of bloggers influences the perception of collective benefits. These initial but insightful results should be expanded in future research.

The evaluation of technologies and tools used in blogs was conducted to extend and corroborate the self-reported measures of the online survey. The evaluation looked into the use of third-party libraries, external services, semantic mark-up, metadata, web feeds, and various media formats in the Blogosphere. The sample of evaluated blogs was compiled using: [a] the Weblogs.com ping server; [b] highly ranked blogs from Technorati and Blogpulse; and [c] blogs contributed by the respondents of the online survey. Data collection was conducted by using an internally developed application for accessing blogs, parsing their source code and identifying technologies used. The results of the data analysis suggest a considerable variation within the studied blogs. More specifically, there is a large number of software platforms, encoding standards, third party services and libraries used. Yet, despite a large number of established and widely used technologies, such as

web feeds, cascading style sheets and JavaScript, the results suggest inconsistencies in approaches to adopting meta-data standards and third party services.

The inquiry into the recent theoretical and empirical research encompassed a wide range of interdisciplinary papers and publications. The literature found relevant for the required analysis of inter-blog relationships has been discussed as part of this report. Application of social network analysis for studying blogs is concluded to be a method sufficient for addressing project requirements. Social network analysis can be performed on various types of relationships, for instance, link-exchange between a set of blogs as well as co-citation between individual web pages. The review suggests that network analysis can be beneficial for identifying relevant blogs and communities and understanding the life cycles of blogs in general. Given the availability of timestamps, social network analysis can also be used for studying and visualising the dynamics within the Blogosphere. Some requirements informing the development of the BlogForever spider component and the data model are formulated in this report. Information, collected for providing greater efficiencies in any future analysis (e.g. reduction in the number of surveys and the numbers of translations) is also being reported.

The final component of the report constitutes a review that highlights current understanding of blog dynamics and user online behaviour as discussed in the relevant literature. The primary focus of the review is on collecting evidence related to the life cycle of blogs and posts, and discussing it in the context of BlogForever. It suggests necessary adaptation in approaches used for crawling and archiving blogs with different dynamics, reputation or state.

Regular studies, similar to those described in this report, are being planned by the BlogForever consortium. Future investigations are being considered to eliminate the limitations of the current study, and in particular, ensuring a more even distribution of countries represented in the survey, wider accessibility with additional European languages, and a greater number of respondents.

2 Introduction

The explosive growth of social media has created new ways for individuals to express their opinions online. Bloggers across the globe are writing daily to produce one of the most comprehensive resources of information, making the so called Blogosphere increasingly richer.

This report presents the results of deliverable D2.1 and contains descriptive information about the survey design, management, deployment and dissemination of its data. The aim is to understand blog structures, the types of data presented in blogs, common blog authoring practices and their preservation. The document also covers an inquiry into adopted technologies, inter-blog relationships and blog dynamics - offering design proposals and reporting blog metrics in general.

Initially, the survey was designed to cover authors and readers in the same questionnaire but after discussion between University of London staff and other partners within the BlogForever project the design changed to two sub-surveys that were clearly focusing in authors and readers separately. These two sub-surveys were designed in six languages and available for participation online for twenty eight days. The final survey was used to gather feedback from all participating partners through a week of pilot testing.

Several project partners produced metrics and ranking results related to common blog authoring practices, patterns in blog structure and data, network based metrics, aspects of blogs for preservation, blog lifecycles examination, and technologies used by current blogs. The survey data variables were formed using two different software packages - SPSS and SmartPLS. The resulting frequency and blogging patterns are documented. The target of this report is to collate these findings so that they can be used to drive the requirements of a future blog preservation archive system.

3 Blog Preservation and Studies Review

Blogs and their preservation have similar requirements to other digital assets but as an emerging and ever evolving digital asset, specific considerations need to be taken. These considerations include planning, assessment, selection, rights management, monitoring technologies, size, benefits, and implementation of preservation strategy. Of all these considerations, planning is crucial.

Considerations for the preservation of a blog should be designed ideally prior to its creation. Selection criteria for the preservation of blogs is essential as Caplan [1] notes that not all blogs are equal. Assessing blogs in terms of their size, readership, comments and interactivity are key points to be considered in relation to their selection and, of course, crucially their content. Rights management and identification of ownership are essential prior to any action taken. Being aware of which technologies are being used and which are emerging in the Blogosphere is also important, if it is decided that this environment is to be preserved. Understanding the value of the blog is also important.

The survey is a thorough approach to analysing both the needs of authors and readers in these contexts of preservation and their interest in preservation of their blogs. All these approaches can then feed into decision making about the appropriate archiving strategy for the blogs being selected. Appropriate strategies and formats for preservation are extremely important but one size does not necessarily fit all as there will be differing needs according to different types of blogs. The survey was designed to focus on what content, features and aspects of blogs could be considered for preservation, interest in utilisation of a blog archive, as well as who will be the audience of the preserved data.

As one of the primary considerations of the actual preservation of blogs is precisely that – for whom are we preserving blogs? Of course, we invest in blogs and their preservation for the benefit of society now and in the future, and as we know blogs are as diverse as society itself. Preservation is not solely driven by the technology, and contextual information about blogs is essential. In addition the selection of blogs and good selection criteria are required. Not all blogs are equal and a variety of contexts drive their creation, either self-motivated or the requirements of a working environment. These considerations greatly alter the content aspect of blogs, and as such the preservation of the context of blogs is to be considered and is assessed in the survey. Blogs are versatile and dynamic, and feed into the interactive records of the online community. They produce information and communication interactions worth considering now and in the future. Their fast changing nature affects the preservation questions reviewed in this report. Before the BlogForever survey was designed, several examples of bloggers and their work were available. National or international preservation policies for blogs have not been extended, and it seems that different national libraries, many institutions and researchers are indicating the real need for selection, preservation and access to blogs in a specific blog archive.

As Caplan [1] has noted, the considerations when looking at digital preservation in general are: availability, identity, understandability, fixity, authenticity, viability and renderability. These issues are extremely relevant for the preservation of blogs. As Caplan (*ibid.*) noted, availability requires having the right to acquire (and sometimes modify, e.g. migrate to a different format) to preserve the blog. Identity raises the issue of ownership. The survey tries to assess these considerations and how to easily preserve blogs in a cost efficient manner safeguarding their authenticity and integrity. Authenticity and integrity are essential if the blogs are to be treated with confidence as research objects or as a reliable record of this age

Prior to designing and implementing this study, a number of earlier studies related to blog preservation have been identified and reviewed. Some of these initiatives are discussed below.

The Blogger Callback Survey [2], conducted by the Pew Internet and American Life Project (PIALP) in 2006, described blogging as:

“a personal, and somewhat private, hobby and a smaller group who view their blogs as more time-consuming, and more public, endeavors. For both groups, the primary motivations to blog were to express themselves creatively and to record their personal experiences” (p. 7).

However, this description may no longer be accurate given the extensive adoption of blogs and rapid change in their use by small and large companies and organisations. Understanding what blogs are today and what is valued is a primary concern for those wishing to preserve them.

PANDORA¹ is the Web Archive of the National Library of Australia. Pandora is considered the first to make a step towards blog preservation in 2004, yet, it managed to archive only a single blog. Although, later on the library increased the number of blogs preserved, the gain was not substantial, with only 12 blogs preserved by April 2011.

More recent attempts by ArchivePress² were more successful. Presented at IPRES 2009 Conference, the solutions developed by Pennock and others [3] provided a mechanism for institutions to collectively harvest blog content. They used WordPress Open Source software and RSS feeds to archive parts of blogs believed to be of primary importance and re-use value. The problems with using feeds for preserving the content were many.

The later work by Davis [4] outlined some of these problems in the paper called “**Moving Targets: Web Preservation and Reference Management**” and highlighted some of the institutional and technological challenges related to preservation of blogs and other dynamic web applications.

The challenges of blog preservation are discussed by Hank and her colleagues [5-8] who stress a range of issues that may affect blog preservation practices. The decision-making for designing and implementing preservation programs would benefit from understanding the characteristics and behaviour of blog authors, readers and blogging service providers.

The survey conducted as part of the BlogForever project, as summarised in this report, attempts to address these challenges and develop a greater understanding of blogs, stakeholders and their interaction in general.

¹ <http://pandora.nla.gov.au/>

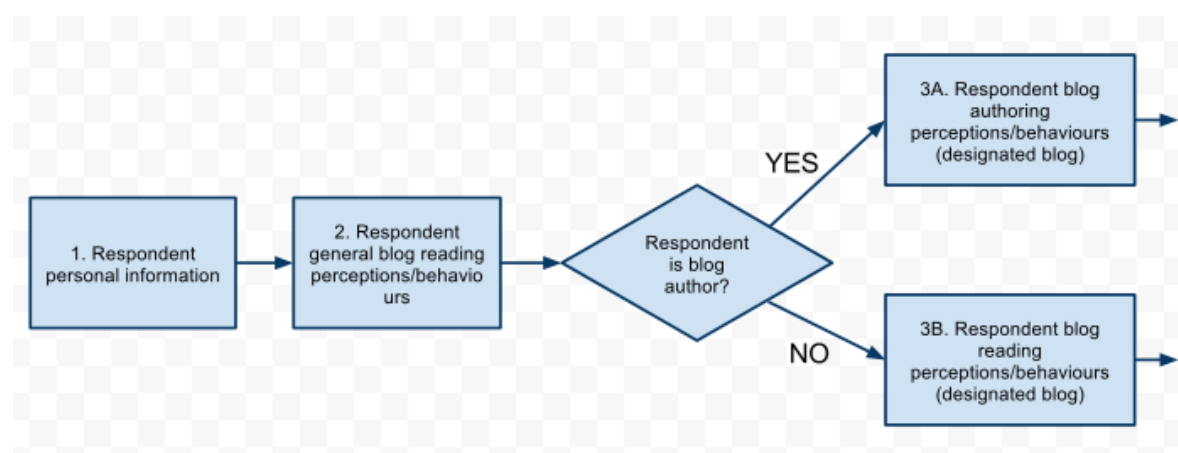
² <http://archivepress.ulcc.ac.uk/>

4 BlogForever Survey

4.1 Questionnaire Development

For a period of four months, the BlogForever consortium, led by the University of London, worked on the design of the different questions that created the final BlogForever survey. Initially, the survey for authors and readers of blogs was just one long questionnaire with several options for branching throughout the survey (Figure 1). After several trials, the decision was taken to divide the survey into two questionnaires, one that covered the authors of blogs and their practices and another which queried blog readers' opinions about blogging, reading patterns and preservation.

Figure 1 – Initial BlogForever survey structure



The areas of research we considered for the final design of the questionnaires were as follows.

Analysis of inter-blog relationships. The survey tried to summarise the importance of inter-blog relationships as a social network phenomenon. Some questions available in the questionnaires related to this area were:

- ✓ How important are communication and networking possibilities (with co-authors, blog owners or other readers) on the blogs you read?
- ✓ How often do you leave a comment(s) on the blogs you read?
- ✓ Does your blog include a list of links like a blogroll?
- ✓ How many other blogs link to your site?

Aspects of weblogs for preservation. What content, features and aspects of blogs should be preserved and who will be the audience of the preserved data? This area looked into opinions of the type of authority and respondents' expectations about preserved material. We aimed to understand what parts of the blogs the authors and users would prefer to preserve. Examples of questions available in the questionnaires included:

- ✓ What reasons can you imagine for using a central blog archive or blog preservation system?
- ✓ What elements of the blog are the most important for you to be preserved?

Common weblog authoring practices. This formed an overview on how authors approached blogging design, editing and posting. Questions covered ways to create content, users reviews about this content, post creation and time spent on authoring activities. Examples of questions available in the questionnaires include:

- ✓ How frequently do you perform: authoring and editing; mashup; design and dialogue activities?
- ✓ Which group do you feel represents the main audience for your blog?

Patterns in weblog structure. The aim was to identify sets of common patterns in structures and data available for future preservation strategies. Examples available in the questionnaires are:

- ✓ What media does your blog contain?
- ✓ How is the media in your blog created?
- ✓ How important for you is the graphical layout or visual appearance of a blog?
- ✓ How important is the availability of rich media (i.e. audio, video, images) for conveying your message?
- ✓ How often do you use a blog's widgets (e.g. News feeds, Flickr, RSS, DIGG, YouTube, Twitter, Skype...)?

Technology. For preservation purposes we wanted to know what type of technology blogs used so an archive strategy based on more common structures could be implemented. Some examples available in the questionnaires are:

- ✓ Do you use a blog provider?
- ✓ Why do you use that platform?

Blog lifecycles and ranking examination. The survey investigated the current roles of blog metrics, popularity indices, decay of blogs, blogs' dynamics within organisational contexts and user ranking. Some examples include:

- ✓ How often do you try to catch up with the top ranked blogs?
- ✓ Is your designated blog connected to an organisation?
- ✓ What kind of organisation?
- ✓ Are you expected or required to read this blog?
- ✓ What does your blog mean to you?
- ✓ What impact would the loss of your blog have on you?
- ✓ Do you have backups for your blog?

4.1.1 Influence Factors for Author and Reader Behaviour

The Technische Universität Berlin (TUB) added the following research objectives that were linked to the survey.

- ✓ A study of the blog authors' intention to contribute to a blog archive repository (as a central and standardised transparent means of preserving and accessing blog contents).
- ✓ A study of blog readers' requirements and preferences for accessing a blog archive repository (as a central and standardised transparent means of preserving and accessing blog contents).

4.1.1.1 Blog author intention to contribute to a blog archive

One of the crucial questions for the success of an archive is to understand what the motivation of blog authors is for contributing their blogs to the archive. Therefore, we aimed to examine factors influencing blog authors' intention to use the archive. Thereby, we see the archive as a central and standardised transparent means of preserving and accessing blog contents. Focussing on the intention instead of the actual use is an appropriate method to inquire about the acceptance of a system that is not yet available. Extensive research in technology acceptance has shown that the intention to behave is the main predictor of the behaviour [9-11].

Writing a blog can be seen as a single activity of a blog author and for many blogs with a diary character this might be true even if they are public. But like other kinds of social software, blogs are also a technology to communicate and collaborate with other people. They allow citing other websites and blogs, commenting on blog posts, and referring to other interesting blogs via blogroll. Aspects of interconnectivity and publicity of blogs indicate that blog authors like to contribute to a common effort of the blog community. Therefore, we assume that the intention to contribute to a

blog archive is also dependent on the expectations of the individual as well as the collective benefit. Thus, we hypothesise that:

- ✓ H1: Perceived individual benefit positively influences the intention to contribute to the blog archive.
- ✓ H2: Perceived collective benefit positively influences the intention to contribute to the blog archive.

Furthermore, we assume that both kinds of perception of benefit are influenced by multiple beliefs and attitudes. In our theoretical model we concentrate on four factors (reputation, self-efficacy, relationship management, and social identity) that we assume to be most influential for blog authors.

Writing a blog is a type of knowledge sharing, and the possibility to increase reputation is a strong influence factor for knowledge sharing [12]. We propose such influence for blog authors because many of them use their blogs to disseminate their opinions and appraisals based on their knowledge and experience. Therefore, it can be assumed that blog authors like to gain a reputation in the subjects they are blogging about. Such reputation belongs to an individual. Thus, we hypothesise that:

- ✓ H3: Expected reputation positively influences the perceived individual benefit.

A blog author shares knowledge, experiences, opinions, etc. to the public. Thus, their statements are open for critique by other people. If the blog author is aware of the potential of critical feedback, they have to be confident about what is published in the blog. Nevertheless, the perceived self-efficacy could vary depending on the status or personality of the author. Self-efficacy with respect to knowledge describes the confidence of an individual in their ability to share useful knowledge and can encourage the intention of open knowledge transfer [13]. Therefore, we suggest that the belief of an author that they can provide useful knowledge to the public has an influence on the expectation of an individual benefit. Thus, we hypothesise that:

- ✓ H4: Perceived self-efficacy positively influences the perceived individual benefit.

If a blog author perceives the blogging activities as a contribution to the effort of team or community then the author probably has relationships at least with some others of this community. Often, the relationships are mainly maintained through virtual activities because the bloggers are geographically distributed. Therefore, blogging can be seen as an activity to establish and to maintain relationships with others. Thus, we assume that if a person perceives the possibility to network to manage relationships through blogging then the blog author will as well perceive blogging as a common benefit. Therefore, we hypothesise that:

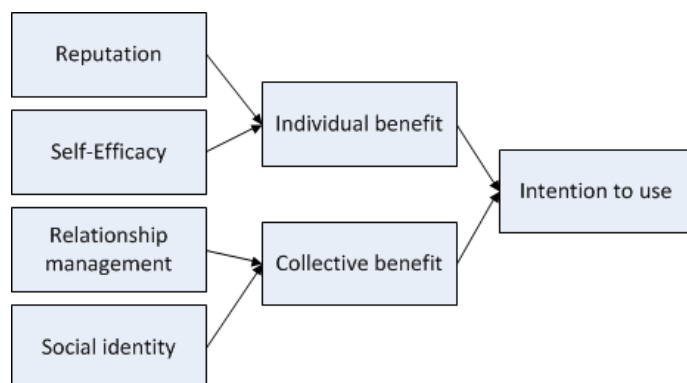
- ✓ H5: Relationship management positively influences the perceived collective benefit.

Additionally, a person who feels embedded in a group or community is more likely to perceive the individual activities as contributions to a common effort. Therefore, the social identity of an individual has an influence on the perception that blogging leads to a collective benefit. Thus, we hypothesise that:

- ✓ H6: Social identity positively influences the perceived collective benefit.

The theorised constructs and their influences are represented in the theoretical framework that is shown in Figure 2.

Figure 2 – Framework for author intention to contribute to the archive



The questions to measure the constructs were mainly taken from previously published research and adapted to the context of this survey. Table 1 shows the constructs and the related questionnaire items.

Table 1 – Items to measure the constructs in the author questionnaire

Variable	Survey items	Comment
Intention to use/adopt a centralised blog archive	I intend to contribute my blog to a central blog archive. If there is/would be a central blog archive, I predict that I would contribute my blog. I plan to contribute my blog to a central archive.	Adapted from Venkatesh & Bala [14]
Perceived individual benefit of blogging like informational value/feedback	I write blogs to get some feedback (advice or criticism) about my blogs. I get to learn other people’s views on my blogs.	Adapted from Lu & Hsiao [15], Dholakia et al. [16]
Perceived reputation through blogging	Writing a blog enhances personal reputation. I earn respect from others by writing a blog.	Adapted from Hsu & Lin [17]
Self-Efficacy in writing a blog	I think I am competent to create a good and well-received blog. I feel confident in my ability to create blogs that are interesting for others.	Adapted from Kankanhalli et al. [18]
Perceived collective benefit	Writing a blog advances the overall Blogosphere. Generally, writing a good blog enhances the relevance of Blogosphere. The Blogosphere is a growing and persistent body of knowledge for Internet users.	TUB proposals
Perceived social benefit of relationship management	Blogging strengthens ties with other bloggers. Blogging creates new relationships with other bloggers.	Adapted from Hsu and Lin [17]

through blogging	I want to stay in touch with other Internet users.	Adapted from Dholakia et al. [16]
Social identity	I think my personal identity overlaps with the other bloggers' identities. I feel part of the group of bloggers.	Adapted from Dholakia et al. [16], Bagozzi and Dholakia [19]

All questions were measured with a five-point Likert scale from “Strongly disagree” to “Strongly agree”. The applicability of the questions was enhanced by the review during the survey testing (see chapter Survey Pilot Testing).

4.1.1.2 Influence factors for the search strategy in the archive

We conceptualise that the blog reader will either search the blog archive or might want to explore it (possibly interactively). Thereby, blog readers can be, but do not have to be, blog authors as well. We further conceptualise that this choice is determined by the influence factors of:

- ✓ Topic relevance instead of author relevance,
- ✓ Demand to assess credibility of the source,
- ✓ Perception of learning based on the search process,
- ✓ Demand to understand complexity,
- ✓ Demand to navigate connected resources, and
- ✓ Attitude regarding the search interface.

Questions for each construct were developed and refined during the testing phase. The number of questions has to be limited due to the fact that the inquiry of influence factors for archive searching was just a (small) part of the survey. Therefore, possible results should be considered with caution because they have strong limitations as well. Table 2 shows the constructs and the related questionnaire items.

Table 2 – Items to measure the constructs in the reader questionnaire

Variable	Survey items
Preference of exploration (Search vs. Exploration)	I think that for comprehensive searching of blogs in a central blog archive: a sorted list would be an effective way to explore a domain (like Google results). I think that for comprehensive searching of blogs in a central blog archive: a visual and interactive map would be an effective way to explore a domain.
Topic vs. author relevance	When I search I go for the topic and the author is usually of secondary importance.
Credibility of the source	I spend a lot of time determining whether or not a source of information is credible. How important is for you to know how credible the blogs' sources of information are?
Learning	Learning is an important aspect of my blog searching and reading.
Complexity	I am often interested in how multiple blogs relate to each other.
Deep Search	It is important for me to trace the links of blogs to find out more.

Rich Interface	<p>I would prefer a comprehensive search over a simple search interface to retrieve relevant pages.</p> <p>A simple search interface with only few options facilitates me best to find relevant blogs.</p> <p>A complex search interface with many options and different views facilitates me best to find relevant blogs.</p>
-----------------------	--

All questions were measured with a five-point Likert scale from “Strongly disagree” to “Strongly agree” or from “Very unimportant” to “Very important”. The first question (preference of exploration) and the second question of the construct (rich interface) were coded reversely.

4.1.2 Survey Pilot Testing

The pilot testing was done during the third week of June 2011, using volunteers connected with the different partners working for the BlogForever Project. The pilot survey, to test the completion time and number of users participating, lasted a week. This pilot reported a few questions that needed some clarification from the pilot respondents. We eliminated some questions due to the overall length of the original questionnaires to make the implementation more popular with potential respondents. Records of the time taking the questionnaires were obtained so identifying other issues was possible. Some question dependencies were added during the testing process.

Some relevant examples of queries and feedback from partners and pilot respondents are listed below:

- ✓ One respondent to the pilot study requested that question “*How important is a comprehensive interface over a simple one to search for relevant blogs?*” from the readerquestionnaire be made more precise and explicit to indicate whether we were asking about “searching for blogs“, “searching for blog posts across many blogs“ or “searching for blog posts within one blog”.
- ✓ Clarifications were requested for the question “*How important to you is the communication and networking on the blogs you read?*” The answer provided was that “I would prefer a comprehensive over a simple interface to search for relevant blogs” is to measure the preference of readers regarding the richness of an interface during a search process. If readers will only use keyword search to look for relevant or interesting blog post and if they search only occasionally then they may prefer a very simple interface like Google, but if they explore the Blogosphere regularly and work with their search results then they may prefer a more comprehensive interface. The question was formulated clearer and we provided some examples.
- ✓ For the question “*What is the software and the version you use for your blog?* ” it was decided that maybe it was a bit hard to answer for a non-technical person and it was not used in the final version of the survey.
- ✓ For the question “*What is your preferred method of accessing a blog post?* ” it was only possible to tick one answer initially so we allowed multiple responses to support the respondent possible multiple answers.
- ✓ The question “*What kind of organisation?* ” allowed the user to tick only one of the multiple answers. The feedback was that if the organisation was a university meant that it was “Academic / Research”, “Scientific” and “Public sector”. This question was updated with multiple choice responses.
- ✓ The following feedback was provided for two questions that were concurrent formulated: “*A sorted list would be the most effective way to explore a domain*” with “strongly agree” then they would have to answer “*A visual and interactive map would be the most effective way to explore a domain*” with “strongly disagree” because there can be only one “most effective way”. The questions were reformulated and only “*an effective way*” was included as both could be perceived as effective but maybe one was more effective than the other.

After gathering a comprehensive list of remarks, all the queries, updates and feedback were resolved. After this pilot test period the author questionnaire was finalised with forty-three questions from the seventy initially presented to the pilot. For the reader questionnaire, the final total was thirty-six questions from an initial number of forty-three³.

The two screen-shots below (Figure 3 and Figure 4) present the averages for overall time, idle time and working time spent by the pilot users. These statistics were produced by iProbe during the pilot testing period for the author and reader questionnaires.

³ For details of the final questionnaires in text format, see Appendix A.1, and Appendix A.2 for details of all the questions available online.

Figure 3 – Screenshot of statistics of author survey testing

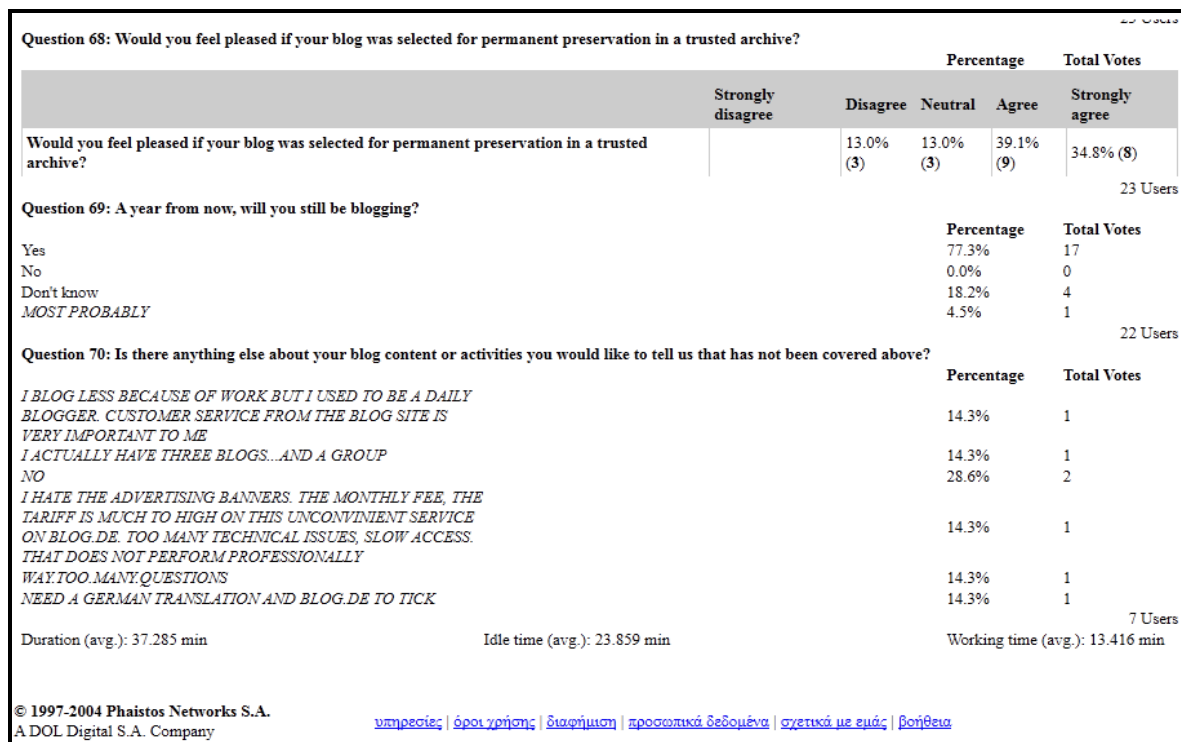
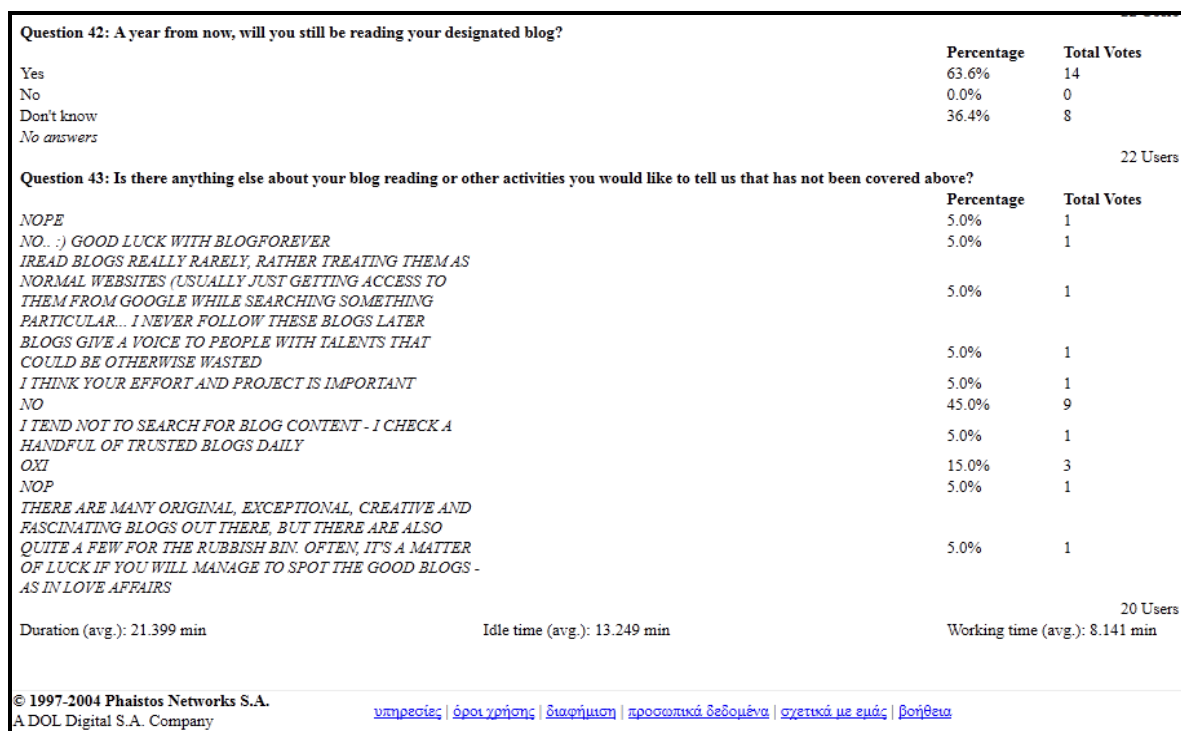


Figure 4 – Screenshot of statistics of reader survey testing



4.1.3 Survey Translations

After the pilot test was completed, the iProbe software was modified in order to support multiple languages and the survey was translated into Greek, Spanish, Russian, French and German. The

translators were within the project and their target was to match the original text and structure of the English version.

The aim of providing the questionnaire in different languages was to reach a wider distribution of native speakers of different mother tongues. Thereby we can evaluate if there were significant differences regarding attitudes and behaviour for different countries and languages. Even if we cannot provide every spoken language in the European Union, we cover six of twenty-three official EU languages. Additionally, English, German, French, Spanish, and Russian are the most common foreign languages in the EU⁴. Thus, the questionnaire should be understandable to the majority of EU citizens.

4.2 Survey Implementation

4.2.1 Survey Population and Sampling

One of the targets of this study is to reach a representative sample of the blogging population. The participants must be selected with reference to a clearly defined population in terms of nature, region and time, and the people to be selected must be approached individually using a clearly defined selection procedure with reference to the selection criteria.

The constant change in the size of the blogging community and the social media channels to spread the survey makes the estimation of the actual size of population sample difficult. It has been considered that the representativeness can only be achieved in online surveys with respect to Internet users as a whole, or with respect to specific groups of Internet users or users of specific websites as the target group of the study.

The participants could have been selected both offline and online. It was decided to only select online participants due to the existing time and resource limitations but mainly due to the online nature of the material overviewed.

4.2.1.1 Sampling strategy

The methods used for sample screening fall in the active approach category. Participants of this online survey who are representative of specific groups of Internet users can, in addition, be selected by means of lists of email addresses providing the direct URL that leads to the BlogForever survey. The participants of online surveys can be actively selected on the basis of typical criteria for statistical random selection, or by means of quotas for certain socio-demographic groups. This selection means that the BlogForever partners decide for themselves whom to approach and ask to participate in the survey.

In the future annual blog analysis, an active as well as a more passive approach are recommended. The usage of a systematic approach, where every nth visitor is invited to participate using a popup or a banner, might add representativeness to the sample. In the current study all the visitors are invited to participate via a banner or a news entry or a blog post. There is no random number generator which at random intervals asks visitors to the website to take part, although this algorithm design could create a biased sample as a result of possible issues with self-selection.

For this study, the online selection did not follow a statistical random procedure. The survey software used allowed multiple participations by the same respondent. If the person wanted to participate in both the author and reader questionnaires, they could do it without any restrictions. The ultimate target was to promote a high rate of responses. Participation in this online survey was

⁴ http://ec.europa.eu/education/languages/pdf/doc3275_en.pdf

possible independently from the design and programming of the questionnaire. This meant that the survey was intended to be processed without problems using different operating systems, types and versions of browsers.

4.2.1.2 Internet sampling

The survey was restricted to people with easy access to their email accounts and the Internet. At any stage, due to time limitations and resources, there was no intention to use other offline methods of sampling.

The survey was available on several partners' web pages as a banner and a hyperlink but the pop-up questionnaire option was not used. Additionally, the survey link was advertised in all the different languages options by the BlogForever partners.

The listserv method was used to provide quality and usefulness for the sample and we depended on the quality of the specific mailing lists used, although the majority were connected with the archival community. Specific mailing lists members' numbers were initially available but some snowball techniques were implemented, hence the list members recommended to their own mailing lists and the total population for the listserv sampling grew without us knowing the final totals.

Blog forums were used but we were not sure what population rates were attached to those forums. Twitter and Facebook professional and personal releases were available but again the final number targeted was inconclusive. For example, The Library of Congress National Digital Information Infrastructure & Preservation Program Facebook page (Figure 5) promoted the survey link via the CERN bulletin article but we do not know the real numbers of followers and the actual respondents who came via this path.

Figure 5 – BlogForever Facebook entry

National Digital Information Infrastructure & Preservation Program
A new blog preservation project in Europe

BlogForever: Intelligent Blog Preservation - CERN Bulletin
cdsweb.cern.ch

A new EU co-funded project, BlogForever, has set its sights on a developing region of the Internet: the blogosphere. With society growing ever more online-oriented, blogs have become rich repositories of cultural, scientific and social information. The BlogForever software platform is designed to ma

25 de julio a las 13:13 · Ya no me gusta · Comentar · Compartir

A ti, Richard Mojibake Davis y otras 4 personas más personas más les gusta esto.

Escribe un comentario...

Another way to reach the blog population was through different blog posts. Below are two examples (

Figure 6 and Figure 7):

Figure 6 – Survey promotion at <http://dablog.ulcc.ac.uk>

The BlogForever survey is live!

By [Silvia Arango-Docio](#) <http://dablog.ulcc.ac.uk/2011/07/11/the-blogforever-survey-is-live/>

After weeks of design work, the BlogForever survey is live, available in 6 languages and running for 28 days.

This survey is part of [BlogForever](#), an EU-funded collaborative project that ULCC collaborates through the Digital Archives department.

The results of the survey, available at the end of the summer, will help to develop digital preservation, management and dissemination facilities for weblogs. Hence, we are keen to gather information about the content, context and usage patterns of current weblogs, so we could identify blogs users' views on their long-term preservation, management, analysis, access and use. If you would like to take part on the survey please use the following link:


Figure 7 – Survey promotion at <http://blogforever.eu/blog>


blog forever Home Activity Work Packages Partners News Events

e > Blog > The BlogForever survey is live!

The BlogForever survey is live!

3:17 pm in [Blog, News](#) by [Silvia Arango-Docio](#)

by [Silvia Arango-Docio](#)

After weeks of design work, the BlogForever survey is live, available in 6 languages and running for 28 days. The results of the survey, available at the end of the summer, will help us to develop digital preservation, management and dissemination facilities for weblogs within the BlogForever project. Hence, we are keen to gather information from you about blog content, context and usage patterns of current weblogs, so we could identify your views on the long-term preservation, management, analysis, access and future use of the BlogForever Archive. We would appreciate if you could take part on the survey using the following link:



All the sampling methods mentioned above lacked information on the total population those samples represented. Apart from their connection with Internet access and blogging, the number of users for each of the chosen methods varied on a daily basis and the recommendations via the social media channels could multiply the audience without having much control or information on that added population.

It is clear that unbiased access to the Internet would have meant obtaining a random population sample through non-Internet ways like Knowledge Networks case with their KNOWLEDGEPANEL tool or the case of The Blogger Callback Survey, sponsored by The Pew Internet and American Life Project (PIALP), which conducted telephone interviews with 233 self-identified bloggers from previous surveys.

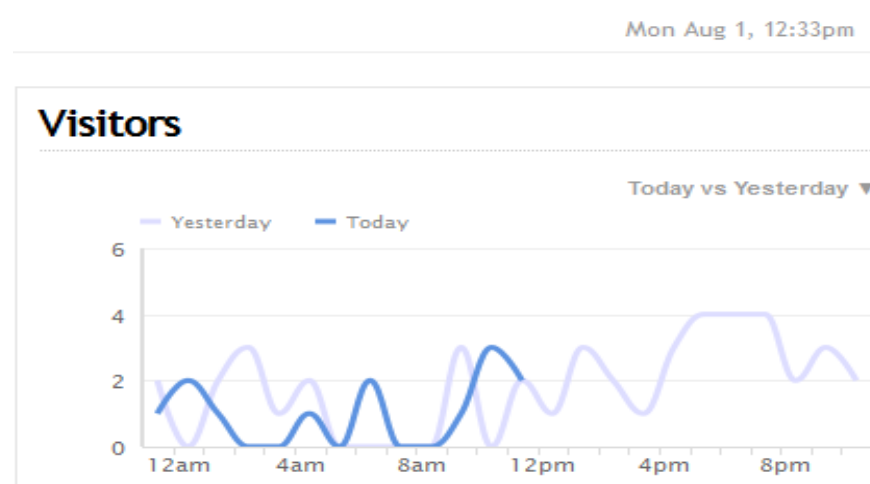
Given the intentions of conducting the BlogForever survey annually, Internet sampling and its representativeness should be addressed by mixing other non-Internet methods of questionnaire administration like telephone or postal methods if resources are available, with the Internet sampling techniques used for the BlogForever survey 2011.

At the current stage of the project, we believe that Internet sampling has been very useful to reach specific samples within organisations which all staff has accessed to the Internet and listserv members who have constant email access.

4.2.1.3 Sample size

The results analysis method was considered in advance and one of the main goals was to make sure the sample was a sufficiently large so then when it was broken down into subgroups (e.g. age and gender) there were sufficient elements in each subgroup. Our main target was to achieve accuracy but this was constrained by limited time and difficult real access to respondents. We ensured the survey was accessed by sufficient numbers for meaningful subgroup analysis. The final sample size was compromised by time and by not knowing exact numbers of the total blogging population, hence the final population size was not determined. The concentration of the survey work was to produce a clean dataset from the author and reader questionnaires within the participation obtained for the twenty-eight days the survey was running.

The respondents who abandoned the survey were not recorded within the software used, information that needs to be available for the future annual surveys. We used <http://getclicky.com/> (Figure 8) to get a regular overview of web analytics for <http://iprobe.gr/Surveys/BlogForever/> but those statistics did not provide any non-response actual data.

Figure 8 – Example of BlogForever survey visitors at getclicky.com

4.2.2 Survey software

Partner Phaistos Networks provided the survey software iProbe, an online platform that empowers small and larger organisations with the ability to easily conduct online surveys, providing real-time monitoring and analysis of the data. iProbe has been used by research companies (e.g. AGB Nielsen Media Research, Metron Analysis), corporate clients (e.g. Vodafone) and advertising firms (e.g. Bold Ogilvy, Tempo OMD) for conducting research about shopping behaviour, user satisfaction, polls, etc. Some of the iProbe surveys have been running continuously for many years (e.g. E-metrics survey in collaboration with Nielsen).

Features of iProbe:

- ✓ Supports several types of questions like open-ended and closed-ended questions (multiple choice, categorical, Likert scale, numerical, etc.).
- ✓ Supports branching: the flow can jump from one question to another. Depending on the answers to one or more questions another question can either appear or disappear. These dependencies can be the conjunction of many questions, specifically if the defined answers are satisfied, then the question appears.
- ✓ Each of the questions can be either optional or compulsory.
- ✓ If a question accepts multiple answers, the limit of the selected answers can be defined and can vary from one selection to the number of answers.
- ✓ Each survey can be set for several languages.
- ✓ The number of submissions per user can be limited based on their IP address.
- ✓ When a user completes a survey, the questions are validated and the user is informed if they did not answer any compulsory questions. The compulsory questions are presented with a different colour from the normal text.

The BlogForever survey link <http://iprobe.gr/Surveys/BlogForever/> was available from iProbe from the 8th of July running for four weeks.

Firstly, the potential interviewee arrived at the following menu and selected their preferred choice of language (Figure 9):

Figure 9 – Language selection front page

Secondly, an introduction text appeared (Figure 10) in the language selected and informed the user about the BlogForever project and the purpose of the survey. The user then had to choose if they wanted to participate as a blog author or reader in order to proceed with the corresponding questionnaire.

Figure 10 – BlogForever survey introduction

blog forever

BlogForever is a collaborative EU funded project. Its key objective is to develop robust digital preservation, management and dissemination facilities for weblogs. These facilities will be able to capture the dynamic and continuously evolving nature of weblogs, their network and social structure, and the exchange of concepts and ideas that they foster; pieces of information omitted by current web archiving methods and solutions.

The purpose of this survey is to gather information about the content, use and context of weblogs that are currently in use in society. The results of the survey will be used to help with the long-term preservation, management, analysis, access and use of weblogs, and thus better serve the community.

Learn more about [BlogForever](#).

Please select if you want to participate as a *blog author* or as a *blog reader*:

  BlogForever (ICT No. 269963) is funded by the European Commission under Framework Programme 7 (FP7) ICT Programme

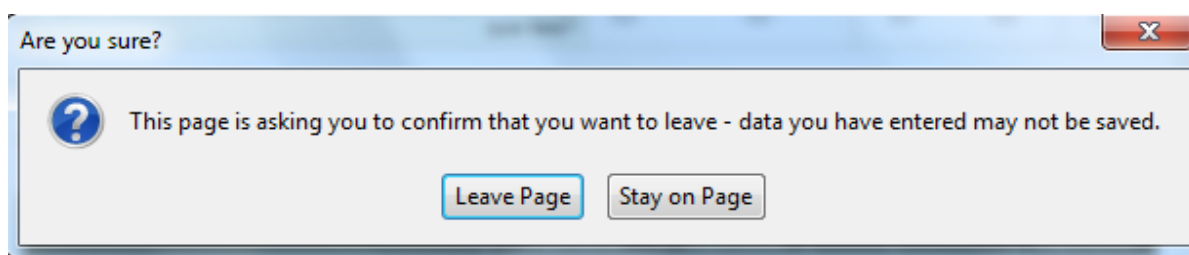
At the top of each page and to facilitate the process of completing the survey, the following details were available (Figure 11):

- ✓ Name of the survey
- ✓ “Thank you for participating” message
- ✓ An estimation of how much time it will take to complete the questionnaire
- ✓ A progress bar that showed the percentage of completion, the percentage of the pages that the user had proceeded.
- ✓ Each questionnaire was separated into sections and each section was in a different page
- ✓ In order for the user to proceed from one page to the next, the respondent had to answer all the mandatory questions of that current page. Otherwise, an alert message informed the user with the number of questions that were not answered, marked in red and the page would scroll to the first unanswered question.

Figure 11 – IProbe survey details

After the completion of the questionnaire, a “thank you” message was available to thank the user for their participation. Moreover, at the end of the author questionnaire, there was a link to connect to the reader questionnaire which prompted the users to participate in the reader questionnaire.

During the survey, the user was informed with alert messages when they performed actions that affected the process of the survey. For example, if the user tried to refresh, close the page, move backwards or mandatory questions were uncompleted, the alerts were displayed (Figure 12).

Figure 12 – Example of an iProbe alert

For the purposes of this survey, Phaistos Networks incorporated new features in order to support multiple languages and to produce estimates of time for completing the questionnaires. Initially, iProbe was setup to support two languages. The software was updated so the questionnaires in languages other than English were incorporated in separate “mapping” files to the English survey file (Figure 13 and Figure 14). The language that the user chose to participate in the questionnaire was recorded.

Also, the *overall time* that the user spent completing each questionnaire was recorded. The overall time was separated into the *idle time* and the *working time*. The working time was estimated as the time that the user spent either using the mouse or the keyboard. The idle time started 5 seconds after the user stopped using the mouse or the keyboard.

Finally, a control tool to validate the blog URL was added to ensure that the URL matched a correct structure and each user was able to participate several times using different blog URLs with a maximum of 100 submissions. The blog URL was attached to a unique identifier for each response.

The survey was hosted on the servers of Phaistos Networks.

Figure 13 – Example of the German survey introduction



BlogForever ist ein EU-gefördertes, kollaboratives Projekt, dass es sich zur Aufgabe gemacht hat, eine Lösung für die "robuste, digitale Bewahrung, Verwaltung und Verbreitung von Weblogs" zu entwickeln. Diese Lösung wird in der Lage sein, die dynamische und kontinuierliche (Weiter-)Entwicklung von Weblogs, ihren Netzwerken und sozialen Strukturen zu erfassen sowie den Austausch von Konzepten und Ideen zu fördern.

Mit dieser Umfrage möchten wir Informationen über den Inhalt, die Nutzung und den Kontext von bereits bestehenden Weblogs erheben. Die Ergebnisse werden dazu genutzt, um bei der Langzeitarchivierung, der Verwaltung, der Analyse, dem Zugriff und der Nutzung von Weblogs zu helfen und somit die Community besser zu unterstützen.

Erfahren Sie mehr über [BlogForever](#).

Bitte wählen Sie aus, ob Sie als *Blog-Autor* oder als *Blog-Leser* befragt werden möchten:

Blog-Autor

Blog-Leser

Weiter zum Fragebogen. »



BlogForever (ICT No. 269963) is funded by the European Commission under Framework Programme 7 (FP7) ICT Programme

Figure 14 – Example of Spanish survey questions

9. ¿Qué importancia tienen las posibilidades de comunicación y trabajo en red (con co-autores, los autores de blogs u otros lectores) en los blogs que lees?

	Muy poco importante	Poco importante	Neutral	Importante	Muy importante
¿Qué importancia tienen las posibilidades de comunicación y trabajo en red (con co-autores, los autores de blogs u otros lectores) en los blogs que lees?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

10. ¿Con qué frecuencia dejas un comentario(s) en los blogs que lees?

	Nunca	En raras ocasiones	A veces	A menudo	Siempre
¿Con qué frecuencia dejas un comentario(s) en los blogs que lees?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

11. ¿Qué importante es el diseño gráfico o la apariencia visual de un blog para ti?

	Muy poco importante	Poco importante	Neutral	Importante	Muy importante
¿Qué importante es el diseño gráfico o la apariencia visual de un blog para ti?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

12. ¿Cómo evalúas las siguientes afirmaciones?

	Muy poco importante	Poco importante	Neutral	Importante	Muy importante
Un sistema de búsqueda sencillo con pocas opciones y maneras de ver los resultados me facilita la mejor manera de encontrar blogs interesantes.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Un sistema de búsqueda complejo con muchas opciones y maneras de ver los resultados me facilita la mejor manera de encontrar blogs interesantes.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

4.2.3 Survey Implementation and Promotion

The survey was launched on 8 July 2011 and respondents were informed about the purpose of the online survey within the introduction. The participation was voluntary and the responses were used in an anonymous form and only for the research purposes explained in this document. The duration of fieldwork for this survey needed to be sufficiently long to allow the target group to participate and to avoid biased samples. It was suggested to run the survey for four weeks during July-August 2011. Ideally, the fieldwork period should have been much longer but the need to produce the results by the end of August affected that decision. For future annual intakes of the survey, a longer running period should be available and possibly after the summer vacation to gather more participants like university students, professionals etc.

The BlogForever survey email (blogforever_survey@blogforever.eu) was provided so respondents were able to obtain information about this project or feedback on any issues, experiences or errors encountered during the survey participation.

On the other hand, to avoid respondents who could click through the questionnaire, many questions were declared compulsory. This action could create invalid data, so after an initial testing process a review of the compulsory questions was done. The survey software did not allow breaking off the questionnaire and continuing later but the time of filling the questionnaires was reduced to balance this and to make the whole survey fieldwork process as efficient as possible. It could have been beneficial to have the breaking off feature in place, so we could have allowed respondents to resume the survey at the point where it was interrupted.

The promotion of the survey was done using different mechanisms:

- ✓ Banners with links to the survey were provided on various partners' networks and websites for dissemination purposes.
- ✓ Individual emails were sent to specific professional lists. Other relevant EU project partners, archival and international mailing lists were contacted.
- ✓ Newsletter items, articles, website press releases and blog posts were published.
- ✓ Social media channels like personal Facebook entries, BlogForever- Linked-in- Group information and individual Twitter accounts were used to disseminate the link of the survey.
- ✓ Internal university students and staff intranets news items were published.
- ✓ Forum posts promoting the survey were initiated when possible.

There was an initial analysis of the country distribution based on the users' IP addresses checked against the RIPE Database while the survey was still live to see which countries were highly represented and from those figures, more efforts were implemented to promote other country and language participation (Table 3). The content of the table is sorted in a descending order - showing most common values first.

Table 3 – Excerpt from the distribution of countries by IP data

Country Code	English Name	Total Users	Percentage
GR	Greece	197	54.6%
DE	Germany	40	11.1%
GB	United Kingdom	39	10.8%
US	USA	22	6.1%
unresolved		17	4.7%
TR	Turkey	15	4.2%
ES	Spain	9	2.5%
AM	Armenia	4	1.1%
IE	Ireland	3	0.8%
NL	The Netherlands	2	0.6%
FR	France	2	0.6%
CH	Switzerland	2	0.6%
CY	Cyprus	2	0.6%
AU	Australia	1	0.3%
IT	Italy	1	0.3%
IN	India	1	0.3%
LK	Sri Lanka	1	0.3%
CA	Canada	1	0.3%
CR	Costa Rica	1	0.3%

4.3 Data Analyses and Results

The data metrics and analyses of the survey aimed to know in depth about blogging patterns so this knowledge could be related to the requirements of a future blog preservation archive system. Some of the multiple initial questions that were in the background of this study were:

- ✓ Why will some blog users aim to use a preservation system?
- ✓ Will their blogging attitudes change if they know their content will be archived?
- ✓ What are the future main objectives of the creators of blogs?
- ✓ How will the blog preservation system be used?

One of the logics behind some of the survey analyses was to know if the variation of using a blog archive matched variations in other variables attached to blogging practices. The first action was to represent the data with percentages using tabulations trying to detect associations between the variables. Excel, SPSS and SmartPLS were used to analyse the final data records transferred from the survey software.

4.3.1 From IProbe to SPSS and SmartPLS

The survey results were extracted into CSV and XLSX file formats. The exported CSV file was a semicolon separated file. The first row of this file had the questions of the survey. The first column of the data file corresponded to the unique user identifier (USER_ID) (Table 4).

Table 4 – IProbe data extraction example

User ID	1 About yourself (Tick which one describes you best)	2 In which country do you live?	3 What is your nationality?	4 What is your gender?	5 Select your Age Group
139818	In paid employment	Greece	Greek	Male	25 - 34
139819	In full-time education	Greece	Greek	Male	25 - 34
139820	Freelancer	Greece	Greek	Male	25 - 34
139821	Freelancer	Greece	Greek	Male	25 - 34
139822	In paid employment	Greece	Greek	Female	25 - 34
139823	In paid employment	Greece	Greek	Male	35 - 44
139825	In paid employment	Greece	Greek	Female	50 - 54
139826	In paid employment	France	Italian	Male	25 - 34
139827	In paid employment	Switzerland	Spanish	Female	18 - 24
139828	In paid employment	Switzerland	British (UK)	Female	45 - 49
139831	In paid employment	Greece	Greek	Male	25 - 34
139834	ACADEMIC, LECTURER	Greece	Greek	Male	35 - 44
139837	Self-employed	Greece	Greek	Female	25 - 34
139838	In paid employment	France	French	Male	35 - 44
139839	In paid employment	France	Polish	Male	25 - 34
139840	In paid employment	Germany	German	Male	25 - 34
139843	In paid employment	Greece	Greek	Male	35 - 44
139844	In paid employment	Switzerland	Malagasy	Female	25 - 34
139853	In paid employment	Greece	Greek	Male	35 - 44
139854	In paid employment	Switzerland	Italian	Male	45 - 49
139857	In paid employment	United Kingdom	Irish	Male	25 - 34

In the CSV file, the questions with multiple answers were enclosed in double quotes (e.g. "English\nGreek") and separated with the new line character (“\n”). The matrix type questions were represented in a separate column for each matrix question. The column title was the combination of the matrix title and a matrix question, separated with the colon symbol (“:”). Field “Blog_URL” was also represented in a separate file with two columns, User_ID and the corresponding Blog_URL.

The total number of responses to the author questionnaire was 517. After an initial data check four records were identified as duplicates and deleted. An extra record seemed to have duplicate answers and it was considered corrupted, leaving the total responses to 512. For the case of the reader questionnaire, we obtained a total of 430 records with one record that was duplicated and corrupted so the data analyses were performed for a total of 428 records.

4.3.2 SPSS and Excel Analyses

The SPSS case or unit of analysis for this survey was the blog URL and its unique ID number. We established data frequencies with counts and percentages using SPSS and Excel. Excel was used to identify strings of responses that were not translated back into in English due to time limitations. The respondents added plenty of feedback that was taken into consideration and added to this report when relevant. The overall data analyses are overviewed in the following sections. Please note that the content presented in the tables is sorted in a descending order - showing most common values first.

4.3.2.1 Authors survey respondents summary⁵

From the frequency analyses performed to the author’s dataset the following general results were found:

- ✓ The majority of the questionnaires were answered in German and English, covering a 73% of the total authors’ responses (Table 5).
- ✓ Of all the participants for the author survey, 40% were employed and nearly 41% were based in Germany and were German nationals (Table 6, Table 7 and Table 8).
- ✓ More than 42% (217 responses) of the authors interviewed used German as their blog language.
- ✓ 27.9% (143 responses) of the total respondents used English to design their blog.

Table 5 – Author responses by survey language

SurveyLanguage	Responses	%
----------------	-----------	---

⁵ Readers Survey Respondents Summary available at Appendix D.

German	220	43
English	154	30.1
Greek	78	15.2
Spanish	41	8
Russian	13	2.5
French	5	1
Blank	1	0.2
<i>Total</i>	<i>512</i>	<i>100</i>

Table 6 – Authors by education & employment

EducationEmployment	Responses	%
In paid employment	207	40.4
In full-time education	122	23.8
Freelancer	55	10.7
Self-employed	49	9.6
Home carer	27	5.3
Other	52	10.2
<i>Total</i>	<i>512</i>	<i>100</i>

Table 7 – Authors by country of residence

CountryResidence	Responses	%
Germany	209	40.8
United Kingdom	79	15.4
Greece	77	15.0
Spain	34	6.6
Armenia	23	4.5
United States	20	3.9
Canada	10	2.0
France	4	0.8
Austria	4	0.8
Italy	3	0.6
Australia	3	0.6
Switzerland	3	0.6
Russian Federation	3	0.6
Indonesia	3	0.6
Costa Rica	3	0.6
Argentina	3	0.6
South Africa	2	0.4
India	2	0.4
New Zealand	2	0.4
Netherlands	2	0.4
Brazil	2	0.4
Peru	2	0.4
Other	19	3.7

<i>Total</i>	<i>512</i>	<i>100</i>
--------------	------------	------------

Table 8 – Authors by nationality

Nationality	Responses	%
German	206	40.2
British	74	14.5
Greek	73	14.3
Spanish	33	6.4
Armenian	27	5.3
US American	21	4.1
Canadian	7	1.4
Austrian	5	1.0
Croatian	4	0.8
Italian	4	0.8
Australian	3	0.6
Indian	3	0.6
Swiss	3	0.6
Polish	3	0.6
Indonesian	3	0.6
French	3	0.6
Irish	3	0.6
Costa Rican	3	0.6
Argentinean	3	0.6
South African	2	0.4
Portuguese	2	0.4
New Zealander	2	0.4
Dutch	2	0.4
Peruvian	2	0.4
Other	21	4.1
<i>Total</i>	<i>512</i>	<i>100</i>

The predominant author age group was 25-34 with 25% of the total, followed by the 35-44 age group. The 18-24 group accounted for only 16% of the total. The rest of the age groups were under 10% apart from the Over 65 with 3.5% (Table 9).

Table 9 – Authors by age group

AgeGroup	Responses	%
25 - 34	131	25.6
35 - 44	106	20.7
18 - 24	82	16
45 - 49	49	9.6
Under 18	47	9.2
55 - 64	42	8.2
50 - 54	37	7.2
Over 65	18	3.5

<i>Total</i>	<i>512</i>	<i>100</i>
--------------	------------	------------

4.3.2.2 Common blog authoring practices

The survey tried to gather as much information as possible over a short period of time, how blog authors performed their most relevant authoring activities. For this, we started asking how frequent blog posts were added, writing was done, and content was uploaded and embedded. The results are summarised below⁶.

Table 10 – Frequency of posting and editing

AuthoringEditing	Responses	%
Weekly	232	45.3
Once a day	105	20.5
Monthly	85	16.6
Several times a day	60	11.7
Rarely	27	5.3
Never / Not at all	3	0.6
<i>Total</i>	<i>512</i>	<i>100</i>

More than 45% of the responses (Table 10) indicated that authoring and editing activities were done on a weekly basis. For the case of mixing, quoting and reusing content from other sources, the results were that more than 30% of the survey participants rarely mixed, quoted or reused content from others and it was never done by 25% of the respondents (Table 11).

Table 11 – Frequency of mashup activities

MashupActivities	Responses	%
Rarely	175	34.2
Never / Not at all	128	25.0
Weekly	84	16.4
Monthly	71	13.9
Several times a day	28	5.5
Once a day	26	5.1
<i>Total</i>	<i>512</i>	<i>100</i>

In terms of blog design activities, like changing the appearance or the feel of the blog, nearly 60% (Table 12) declared that rarely applied those changes. More than 16% of the total who answered the questionnaire performed blog style design activities monthly, followed by more than 12% who responded that never applied blog design changes. Only 5% applied those changes daily.

Table 12 – Frequency of design activities

DesignActivities	Responses	%
Rarely	306	59.8
Monthly	84	16.4

⁶ For details of blog readership from the readers' dataset see Appendix D Table 5 and 6.

Never / Not at all	64	12.5
Once a day	26	5.1
Weekly	20	3.9
Several times a day	12	2.3
<i>Total</i>	<i>512</i>	<i>100</i>

For the case of blog dialogue activities like blog community and comments responses, moderating, linking to other sites and search engine optimisation activities, more than 26% (Table 13) did those activities in a weekly basis and around 20% did them daily.

Table 13 – Frequency of dialogue activities

DialogueActivities	Responses	%
Weekly	135	26.4
Once a day	94	18.4
Rarely	94	18.4
Monthly	77	15
Several times a day	72	14.1
Never / Not at all	40	7.8
<i>Total</i>	<i>512</i>	<i>100</i>

Respondents were asked about their blog audience. They were able to select multiple groups that matched their feelings of who represented these audiences. The highest percentage (Table 14) was for the group “General Public” followed by “Family and Friends”, but many records showed a combination of audiences rather than just a unique group. More than 8% of the responses were from the choice “Other” where the blog audience fell outside the list given in the question.

Table 14 – Main audience of blogs

MainAudience	Responses	%
General Public	306	59.8
Family and Friends	207	40.4
Myself	165	32.2
Colleagues and Professional Peers	164	32.0
Students	91	17.8
OTHER BLOGGERS	6	0.11
Other	45	8.8

The questionnaire tried to see how authors and collaborators considered themselves as a group and the majority designated themselves as the only author, although some feedback was available in some responses regarding how respondents were collaborators and no designated authors. Only 13.5% worked as a group (Table 15).

Table 15 – Single authors and collaborators

SingleAuthorMultipleAuthors	Responses	%
Only author	443	86.5
Multiple authors	69	13.5
<i>Total</i>	<i>512</i>	<i>100</i>

In terms of work related blogs and their authoring, only 11.3% of the 512 respondents were required to blog by their organisation. 9.6% did not know if they were expected to blog at their work. The rest were not expected to blog.

Regarding blog providers, 61.3% were using one, more than 36% were not with any blog provider and the rest did not answer the question. The results gathered about which blog providers were using are in Table 16.

Table 16 – Blogging service providers

BlogProvider	Responses	%
BLOG.DE	90	17.6
WORDPRESS	48	9.4
BLOGGER	43	8.4
BLOG.CO.UK	20	3.9
BLOGSPOT	20	3.9
PATHFINDER	13	2.5
LIVEJOURNAL.COM	13	2.5
PATHFINDER.GR	6	1.2
TUMBLR	4	0.8
GOOGLE	3	0.6
PATHFINDER BLOGS	2	0.4
PBLOGS.GR	2	0.4
MOKONO	2	0.4
BLOGGER/BLOGSPOT	2	0.4
BLOG.CA	2	0.4
BLOG.COM.ES	2	0.4
Other	39	7.6
Blank	201	39.3
<i>Total</i>	<i>512</i>	<i>100</i>

In conjunction with the blog provider questions, there was a need for data about how authors backed up their work, not only to know a bit more of their normal practices but to understand how much consideration they had about maintaining their blog alive if ever their blog providers were not able to do so. The results showed that more than 64% did not have any backups and from the 35.7% who backed up their work, only 7.4% did it daily, 7.8% did it weekly and nearly 20% had backed up their work monthly. 65% of authors, who backed up their work, left this question unanswered. Only 18% retained a copy of their backed up work.

4.3.2.3 Patterns in blog structure and data

The question of blog data was introduced to the respondents as “What media does your blog contain?”. Different arrays of data (Table 17) showed that most of the blogs contained textual data. Photographs and moving images closely followed the predominant type of data. The results showed that 83% of the authors had photographs and 43.2% had moving images. These different arrays of blog structures showed the variety of blog types⁷.

⁷ For details of blog data interaction from the readers' dataset, see Appendix D Tables 7 – 11.

Table 17 – Type of media used

MediaInBlog	Responses	%
Text	503	98.2
Photographs	425	83
Moving images	221	43.2
Images other than photographs	207	40.4
Audio	147	28.7
LINKS	5	1
GIFS	2	0.4
Other	23	4.5

It is necessary to note that some respondents used other types of media not included in the given list of choices. The additions specified by respondents included published papers in PDF format, teaching materials and lists of references.

After knowing a bit more of the predominant types of data available, respondents referred to how these objects were created, showing that more than 90% was self-created and 28.9% was remixed from an original blog object (Table 18). However, users were allowed to choose more than one answer from a list of available options. This explains the value of cumulative percentage exceeding 100%. It appears that a combination of techniques for preparing their posts is being used by authors.

Table 18 – Blog content creation

MediaInBlogCreation	Responses	%
Self-created	462	90.2
From specific websites	166	32.4
Reused/Remixed from original	148	28.9
Search engine results	133	26
From other blogs	98	19.1
You Tube	7	1.4
Other	23	4.5

The data gathered showed multiple combinations of how blog objects were created, but these combinations did not show specific predominant patterns apart from the results above. Moving into querying about the importance of rich media in their blog, results were that nearly 60% found rich media important or very important to their blog and nearly 24% were neutral about it.

In terms of blogrolls or lists of links available, nearly 60% reported (Table 19) that those lists were not in their blog. More than 40% used them and when asked how many links were available, nearly 60% of those with lists of links, left the question unanswered and only 16.4% knew that their list contained between 10 and 49 links.

Table 19 – Availability of a list of links

ListofLinks	Responses	%
No	299	58.4
Yes	212	41.4
Blank	1	0.2
<i>Total</i>	<i>512</i>	<i>100</i>

When authors were asked about how many blogs linked to their site, more than 40% did not know, and 35% responded that “Fewer than 10”.

4.3.2.4 Network-based metrics

Authors were asked about the way their traffic was monitored, how their content was used, and if there was awareness of daily hits and any ranking analysis applied. These metrics showed some indications of how blog authors had records about their networks and were interested to know more about the trends of popularity and established some ranking attributes within their connections⁸.

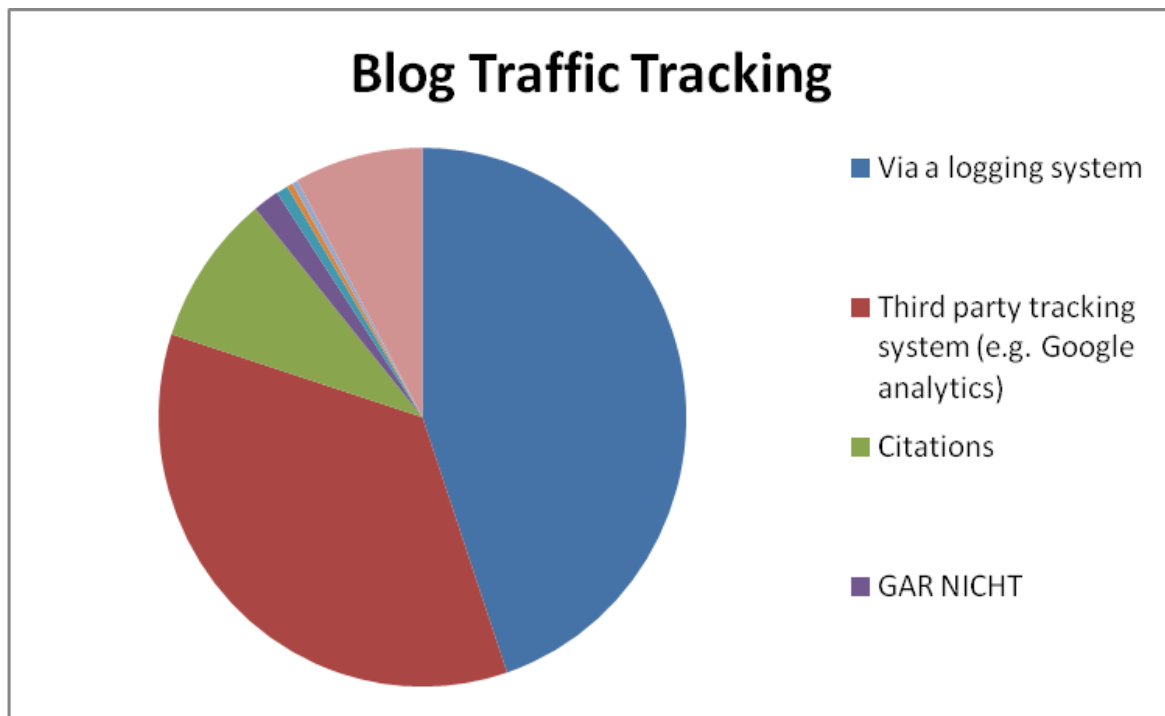
When authors were asked about the ways they tracked their traffic, the following results showed that nearly 50% (Table 20 and Figure 15) used a logging system for these purposes.

Table 20 – Blog traffic monitoring methods

TrackBlogTraffic	Responses	%
Via a logging system	251	49
Third party tracking system	197	38.5
Citations	51	10
GAR NICHT	9	1.8
BLOG.DE	4	0.8
DO NOT	2	0.4
GOOGLE	2	0.4
Other	44	8.6

⁸ For details of blog networking from the readers’ dataset, see Appendix D. Tables 12 - 13.

Figure 15 - Blog traffic monitoring methods



8.6% quoted “Other” as a method of tracking represented a diverse bundle of answers with lots of specific information on how their blog providers transferred data about their blog traffic via newsletters or their in-built statistics option. A low percentage within the “Other” group represented authors who were not interested in knowing anything about their blog traffic.

Table 21 – User interaction with blog content

UsersInteractionBlogContent	Responses	%
Comments	443	86.5
Subscribe or follow	219	42.8
Feeds (e.g. RSS, Atom...)	143	27.9
Linking to blogpost	135	26.4
Adding to blogroll	86	16.8
Acknowledging trackbacks	76	14.8
Contributing posts	53	10.4
EMAIL	3	0.6
NACHRICHTEN (News)	2	0.4
TWITTER	2	0.4
Other	12	2.3

Table 21 represents data about methods used to interact with authors’ blog content. The respondents had the opportunity to mix and match the choices available in the question. “Comments” were the most popular choice of blog content interaction (86.5%) followed by “Subscriptions” and “Feeds” (60%).

Table 22 - Ranking analysis tools

RankingAnalysisTools	Responses	%
None of the above	234	45.7
Comments	184	35.9
Subscribers	108	21.1
WordPress stats	76	14.8
Citations	51	10
Onsite Audience Growth	41	8
Technorati Rank	34	6.6
Offsite Audience Growth via Feeds	31	6.1
BlogPulse	23	4.5
GOOGLE ANALYTICS	3	0.6
STATCOUNTER	3	0.6
WIKIO	3	0.6
Other	17	3.3

The ranking analysis tools query showed that authors used many ways to know about their ranking (Table 22). “Comments” was one of the most popular ways to evaluate their own ranking. The interesting result was that respondents used several methods of ranking and the combinations of these methods were varied. Only 5.1% of the respondents specified details about methods of ranking used that were not available in the given choices.

Table 23 – Hits per day

HitsDay	Responses	%
Between 10 and 49	181	35.4
50 or more	144	28.1
Fewer than 10	115	22.5
Don\'t know	71	13.9
Blank	1	0.2
<i>Total</i>	<i>512</i>	<i>100</i>

Awareness of hits a day on authors’ blogs is represented above. This awareness was spread over the choices but the majority (35.4%) opted for “Between 10 and 49” hits a day.

Overall ranking methods were popular within authors. They used extensive ranges of tools to assess blog networking and its success.

4.3.2.5 Blog lifecycle examination

The analysis of the blog lifecycle in conjunction with blog preservation will be taken further with the establishment of the annual BlogForever Survey. The blog lifecycle assessment was covered by querying the reasons behind starting to blog, the motivations for maintaining blogs, the meaning of these blogs for their authors and the impact of losing them. The blog technology lifecycle was not overviewed and there were no patterns from blog creation to its death assessed.

The lifecycle context was connected with feelings from the respondents about blog creation. When the participants were asked “Why did you start blogging?”, 392 out of the 512 total respondents for the author questionnaire did not answer the question. The rest explained many different reasons and some of the responses referred to actual dates instead of explanations. A summary of those responses are listed:

- ✓ Looking for a way to express ideas, interests, feelings, hobbies, thoughts and judgements
- ✓ To promote specific subjects
- ✓ Share writing skills and research
- ✓ Curiosity
- ✓ As a way to keep an e-calendar system or a record of activities
- ✓ To meet people with similar interests
- ✓ For dissemination and engagement
- ✓ A resource to refer to in the future
- ✓ Useful mean of communication and information sharing

Once thoughts of blog origin were queried, motivations for maintaining blogs were asked. Those motivations were similar to the reasons for creating blogs so consistency was implied. Respondents were able to choose several reasons for maintaining their blogs and the resulting combinations (Table 24) were wide and overall “personal” choice was the predominant one (79.7%), followed by “information sharing” (60.7%) and the “discussion of topics” (48.8%).

Table 24 – Motivations for maintaining blogs

Values	Responses	%
Personal	408	79.7
Information sharing	311	60.7
Discussion of topics	250	48.8
Mostly for myself	218	42.6
Create an online presence	195	38.1
Professional	181	35.4
Entertainment	165	32.2
Record of activities or events	158	30.9
Mostly for my audience	139	27.1
Organise / promote / support an activity	119	23.2
Promote teaching and learning	108	21.1
Commercial	47	9.2
Manage a project	39	7.6
To target markets or communities	39	7.6
Manage a conference	12	2.3
Other	33	6.4

The range of motivations for maintaining blogs for each respondent represented a wide variety of patterns, difficult to analyse in this comprehensive survey but a starting point for further future analysis (Table 24).

Table 25 – Meaning of blogs

MeaningofYourBlog	Responses	%
An enjoyable hobby	240	46.9
Very important part of my life	171	33.4

Don't spend a lot of time on it	41	8
Other	38	7.4

Regarding the meaning of blogs for authors participating in the BlogForever survey, 46.9% considered their blog as a hobby and more than 30% defined the blog as a very important part of their life (Table 25). Some specifications of the choice "Other" provided details about their blog professional meaning or reason for their blog existence.

Table 26 – Impact of losing a blog

ImpactBlogLost	Responses	%
Very important	181	35.4
Important	210	41
Neutral	92	18
Unimportant	23	4.5
Very unimportant	5	1

As a final insight into the blog lifecycle, the survey asked about the impact of losing their blog and the responses reiterated the importance of maintaining it with more than 75% agreeing with how relevant losing their blog was (Table 26). There was no specific question about previous blogs and their decay, or if they had any issues with their current blogs future loss.

4.3.2.4 Aspects of blogs for preservation

The survey was designed to obtain insights into current and future preservation practices for blogs. Considerations of blog selection and its elements for preservation were asked about. Blog archive readership information was gathered.

Of the total of responses for the author questionnaire, only 33.4% self-archived their blog. When respondents were asked about using an external service to preserve their blog, the majority (85.7%) never used one (Table 27). Several answers under "Other" implied that they relied on their blog provider for archival practices. Some feedback pointed the following examples as awareness of blog archival activities:

- ✓ PDF archive at Blog.de
- ✓ Preserved by the people who run the blog
- ✓ Maybe University Digital Archive

Table 27 – Use of external services for blog preservation

UsedExternalServicePreservation	Responses	%
Never used one	439	85.7
Web-archiving service	25	4.9
Digital archive	16	3.1
Archiving service	13	2.5
Institutional repository	2	0.4
Other	17	3.3

Authors were asked about blog selection for preservation purposes in a trusted archived. The majority of the responses indicated a positive attitude towards blog archival selection covering nearly 80% of the total as "Agree" and "Strongly Agree". See Table 28 for details.

Table 28 – Attitudes towards blog preservation in a trusted archive

BlogSelectedTrustedArchived	Responses	%
Strongly agree	195	38.1
Agree	199	38.9
Neutral	99	19.3
Disagree	9	1.8
Strongly disagree	9	1.8

Information about blog readership once they were archived was gathered, asking the respondents about the channels authors found most useful to promote their blog readership within a blog archive. Multiple combinations of the tools given were provided by the respondents, and making “A blog community” and “Sharing and rating” the most popular choices. These popular tools were followed by “Blog news portals” and “Blog marketing tools”. There were results about not being interested in blog sharing and increasing blog readership covering around 20% of the total interviewed. Details of the results are presented in Table 29.

Table 29 – Interest towards blog archiving tools for increasing readership

Values	Responses	%
A blog community	242	47.3
Sharing and rating	242	47.3
Blog news portals	167	32.6
Blog marketing tools	116	22.7
Not interested in increasing my readership	61	11.9
Not interested in sharing beyond my blog	55	10.7

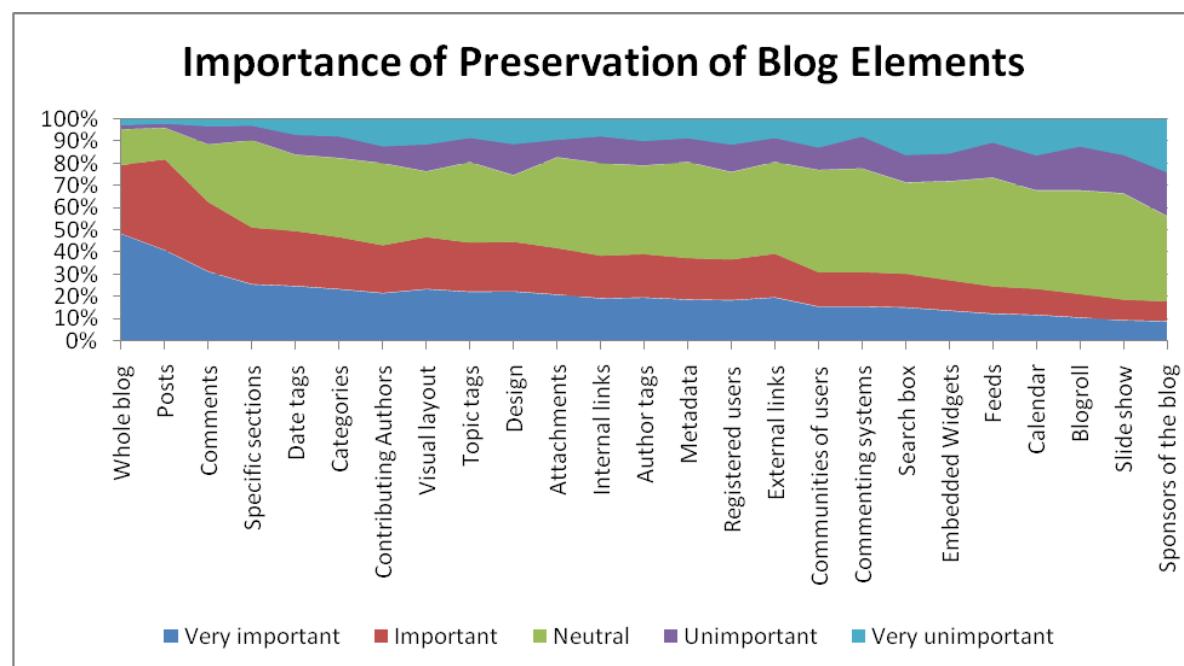
Details of the importance of preserving blog data and which blog components were relevant for the authors to select for long term preservation showed that the whole blog as an entity had the highest outcome (46.3%) followed but posts (45.7%) and comments (25.4%). Specific details of other elements and their relevant role in blog preservation are specified in Table 30 and Figure 16.

Table 30 – Importance of preserving blog data

PreservedElement	Very Important (%)	Important (%)	Neutral (%)	Very Unimportant (%)	Unimportant (%)
Whole blog	46.3	29.7	15.4	2.5	2
Posts	45.7	45.7	15.8	2.1	2.3
Comments	25.4	25.4	21.1	2.5	6.8
Specific sections	20.7	20.7	31.8	2.3	5.5
Date tags	20.7	20.7	28.7	5.9	7.6
Categories	18.4	18.4	27.9	6.1	7.8
Contributing Authors	18.4	18.4	31.4	10.4	6.6
Visual layout	17.8	17.8	22.5	8.6	9.4
Topic tags	17.6	17.6	28.7	6.6	8.8
Design	17.2	17.2	23	8.6	10.9

Attachments	16.8	16.8	32.8	7.4	6.4
Internal links	15.6	15.6	33.8	6.3	10
Author tags	15.6	15.6	31.8	7.8	9
Metadata	15.4	15.4	35.7	7	9
Registered users	15.4	15.4	33	9.6	10.5
External links	15	15	31.6	6.4	8.4
Communities of users	12.9	12.9	38.3	10.5	8.6
Commenting systems	12.5	12.5	37.5	6.3	11.7
Search box	12.5	12.5	34	13.3	10.4
Embedded Widgets	11.1	11.1	36.1	12.5	10.4
Feeds	10.4	10.4	41.2	8.8	13.5
Calendar	9.8	9.8	36.7	13.5	13.3
Blogroll	8.6	8.6	37.9	10	16.2
Slide show	7.8	7.8	40	13.5	14.6
Sponsors of the blog	7.6	7.6	32.6	20.3	17

Figure 16 - Importance of preservation of blog elements



Hence, archiving the whole blog as an overall object with its posts and comments seemed the most important choice for the authors participating in this survey.

4.3.3 Blog Author Intentions for Contributing to Blog Archives

The analysis of the framework described in this was conducted using the Partial Least Squares (PLS) approach. PLS is a structural equation modelling technique that supports confirmatory and exploratory research [20]. The calculations were done with the SmartPLS software. With 512 responses our sample fulfils the required sample size of “at least 10 times the number of items in the most complex construct” [20, p.9].

PLS analysis comprises the measurement model that shows the mapping of measures onto theoretical constructs and the structural model that explains the causal and correlational links between the latent variables.

4.3.3.1 Measurement model

To validate the measurement model we assessed content validity, construct validity, and discriminant validity. To establish content validity, we ensure consistency between measurement items and existing literature. As shown in Table 1 in section 4.1.1, most of the items that we have used were adapted from previously validated work. Additional items were developed based on our experience and evaluated during the pilot testing of the questionnaire.

Construct validity is composed of convergent and discriminant validity. Convergent validity is assessed by examining the average variance extracted (AVE), the composite reliability (CR), and the item loadings. Table 31 shows the constructs, related items, loadings of the items, AVE of the constructs, and CR of the constructs.

Table 31 – Summary of factor loadings, average variance extracted and composite reliability

Construct	Item	Loading	AVE	CR
Intention to use/adopt	ITU1	0.934	0.839	0.94
	ITU2	0.898		
	ITU3	0.915		
Perceived individual benefit	IB1	0.910	0.772	0.87
	IB2	0.847		
Perceived reputation	REP1	0.843	0.768	0.87
	REP2	0.908		
Self-Efficacy	SE1	0.880	0.804	0.89
	SE2	0.913		
Perceived collective benefit	CB1	0.821	0.655	0.85
	CB2	0.811		
	CB3	0.796		
Relationship management	RM1	0.898	0.753	0.90
	RM2	0.890		
	RM3	0.812		
Social identity	SI1	0.582	0.595	0.74
	SI2	0.923		

Factor loadings should be higher than 0.7 but lower loadings can also be acceptable in practice. Factors with loadings lower than 0.5 should be eliminated [21]. Almost all of our items load high on their constructs. Only the item SI1 has a loading lower than 0.7 but higher than 0.5. Therefore, we included the item SI1 as well. The AVE values should be greater than 0.5 and the CR values should be greater than 0.7 [22]. These thresholds were exceeded for every construct.

Table 32 – Cross loadings of the items

	CB	IB	ITU	REP	RM	SE	SI
CB1	0.821	0.272	0.254	0.390	0.340	0.387	0.427
CB2	0.811	0.358	0.233	0.347	0.376	0.356	0.327
CB3	0.796	0.166	0.329	0.375	0.361	0.452	0.342
IB1	0.311	0.910	0.201	0.320	0.434	0.277	0.301
IB2	0.255	0.847	0.228	0.229	0.322	0.167	0.307
ITU1	0.310	0.205	0.934	0.199	0.126	0.190	0.130
ITU2	0.329	0.221	0.898	0.215	0.124	0.235	0.129
ITU3	0.289	0.237	0.915	0.245	0.151	0.201	0.140
REP1	0.438	0.241	0.242	0.843	0.368	0.379	0.235
REP2	0.377	0.309	0.187	0.908	0.374	0.444	0.347
RM1	0.402	0.430	0.100	0.315	0.898	0.221	0.530
RM2	0.385	0.333	0.105	0.361	0.890	0.263	0.427
RM3	0.365	0.370	0.180	0.429	0.812	0.277	0.390
SE1	0.468	0.214	0.233	0.406	0.242	0.880	0.170
SE2	0.422	0.249	0.182	0.440	0.279	0.913	0.302
SI1	0.210	0.178	0.114	0.205	0.217	0.104	0.582
SI2	0.443	0.328	0.120	0.308	0.520	0.272	0.923

For discriminant validity each of the items should load higher on the theoretically assigned construct than on any other construct [23] and the average variance of a construct should be higher than the square of a correlation with any other construct [24]. The first is measured by the cross loadings that are shown in Table 32. The latter comparison is shown in Table 33. Both criteria are satisfied.

Table 33 – Average variance extracted and latent variable correlations

	AVE	CB	IB	ITU	REP	RM	SE	SI
CB	0.655	1.000						
IB	0.772	0.324	1.000					
ITU	0.839	0.338	0.241	1.000				
REP	0.768	0.459	0.318	0.240	1.000			
RM	0.753	0.443	0.436	0.146	0.422	1.000		
SE	0.804	0.494	0.259	0.229	0.472	0.292	1.000	
SI	0.595	0.452	0.344	0.146	0.338	0.520	0.269	1.000

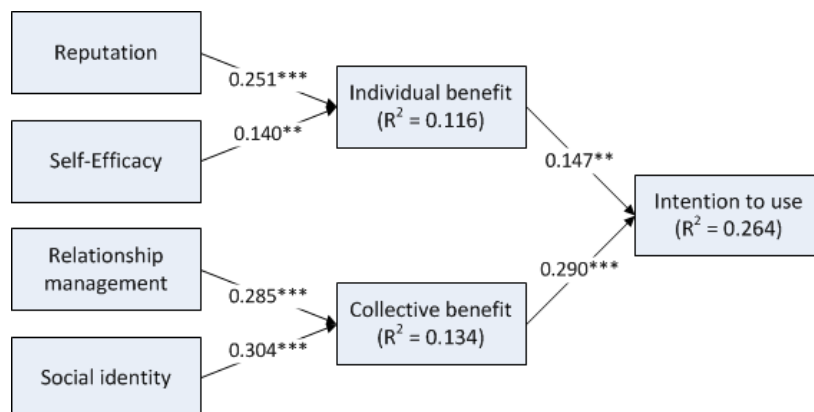
As we have shown in the analysis of the measurement model, all scales in this study are measuring the theoretical constructs of our model sufficiently. Therefore, we can proceed with the analysis of the structural model.

4.3.3.2 Structural model

The structural model allows the testing of the proposed hypotheses. To determine the significance of the paths among the constructs, the bootstrap re-sampling method was used with the option of 1000 re-samples. Figure 17 shows path coefficients and significance for the proposed relationships as well as the R^2 values of the endogenous variables.

All proposed relationships are highly significant at the significance level of at least 0.01. The path coefficients represent the strength of the influence. It can be seen that the intention to use is much more influenced by the perceived collective benefit (0.290) than by the perceived individual benefit (0.147). As well, an expected reputation (0.251) has more influence on a perceived individual benefit than the self-efficacy of the author (0.140). The influence of relationship management and social identity on the perceived collective benefit is almost the same.

Figure 17 – PLS path analysis model (** $p < .01$, *** $p < .001$)



The R^2 values of the dependent constructs indicate the explanatory power of the structural model. It means how many percent of the variance is accounted by the according predictors. Chin [25] denoted a substantial level ($R^2=0.67$), a moderate level ($R^2=0.33$) and a weak level ($R^2=0.19$). It has to be stated that all the dependent variables can be ranked only at a weak level. Thereby, intention to use has the highest R^2 value. 26 % of the variance of the intention to use the archive is accounted by the predictors of that construct. This is in the middle between a weak and a moderate level. But only 12 % of the variance of individual benefit and 13 % of the variance of collective benefit are explained by their proposed influence factors.

All hypotheses for the influence factors of authors' intention to contribute to the archive are supported by the data. But it has to be stated that the influence is only weak or moderate. This indicates that there are possibly other important influence factors that should be considered for a further development of the model. Furthermore, the measurement instruments should be improved in following studies.

Summarising, the study found that the influence on the intention to contribute to the archive is much higher by the perception of a collective benefit than by the perception of individual benefits. This should be taken into account for the development and promotion of the blog archive. Thereby, a support of relationship management and social identity can be seen as starting points.

4.3.4 Influence Factors for Search Strategies within Archives

The analysis of the framework described in chapter "4.1.1.2 Influence factors for the search strategy in the archive" was conducted using the Partial Least Squares (PLS) approach. PLS is a structural equation modelling technique that supports confirmatory and exploratory research [20]. The calculations were done with the SmartPLS software [26]. With 429 responses our sample fulfils the required sample size of "at least 10 times the number of items in the most complex construct" [20, p.9].

PLS analysis comprises a measurement model that shows the mapping of measures onto theoretical constructs and a structural model that explains the causal and correlational links between the latent variables.

In this analysis, the measurement model cannot be fully evaluated because four constructs were measured with only one reflective indicator. Thus, only some estimations and assumptions are possible. A first analysis revealed very low factor loadings for the reverse coded items (RI2 and SE1). Therefore, the item RI2 was eliminated and the item SE1 was not coded reversely anymore. New calculations were conducted and the resulting factor loadings can be seen in Table 34. The loadings are higher than 0.7 for the constructs of credibility of the source and rich interface. Also the construct of preference of exploration has high loadings even if the item SE2 barely misses the threshold of 0.7. Therefore, it can be stated that the respondents did not perceive the questions for the preference of exploration as contrary statements. Furthermore, the AVE and the CR of this construct are below the thresholds of 0.5 (for the AVE) and 0.7 (for the CR). Therefore, the convergent validity of SE1 and SE2 has to be assessed as insufficient. That means that the respondents answered the question differently and both questions are not measuring the same construct. Additional research is needed to examine if there is (or is not) an interdependence between the preference of a simple search and the preference of a more sophisticated exploration of the archive. Similarly, the respondents did not perceive the support by a simple search interface (RI2) as the opposite of the support by a complex search interface (RI1, RI3).

Table 34 – Summary of factor loadings, average variance extracted and composite reliability

Construct	Item	Loading	AVE	CR
Preference of exploration (Search vs. Exploration)	SE1	0.719	0.498	0.665
	SE2	0.693		
Credibility of the source	SC1	0.843	0.676	0.807
	SC2	0.801		
Rich Interface	RI1	0.860	0.664	0.798
	RI2	Eliminated		
	RI3	0.768		
Topic vs. author relevance	AR1	1	1	1
Complexity	C1	1	1	1
Deep Search	DS1	1	1	1
Learning	L1	1	1	1

Discriminant validity was estimated by the cross loadings and a comparison of AVE and latent variable correlations (see chapter “4.3.3.1 Measurement ” for more background information about the method). Both criteria are satisfied and do not provide any additional insight.

Additionally, we evaluated the structural model even if the measurement model cannot be assessed as sufficient. A R^2 value of 0.14 indicates that only 14% of the variance of the preference of exploration is explained by the other variables. This is just a weak level. Furthermore, only two of the six proposed hypotheses are significant. Thereby, the preference of a rich interface has a stronger influence (path coefficient = 0.26, $p < 0.001$) on the preference of exploration than the preference of a deep search (path coefficient = 0.12, $p < 0.05$).

Summarising, only few implications can be made. First, if there is interdependence between the preference of a simple search and a sophisticated exploration then it is not like that one is the opposite of the other. But an understanding of such interdependencies will facilitate the provision of adequate interfaces and functionalities for archive users. Therefore, additional research is needed.

On the other hand, the theoretical model should be reconsidered because the independent variables do not seem to be strong influence factors for the construct of preference of exploration. Additionally, the measurement instruments have to be improved because only proper measures will lead to stable findings.

5 Technology Used by Current Blogs

The main aim of this section is to review the technological foundations of the current Blogosphere. The review is primarily based on a large-scale evaluation of active blogs. This evaluation provides the necessary grounds for extending and corroborating the self-reported measures of the above-mentioned user survey. Furthermore, the extensive list of examined technologies enables commenting on the widely adopted standards and potential trends in the Blogosphere. The evaluation has been conducted in the following stages:

- ✓ Accessing and parsing a large set of blogs
- ✓ Identifying and quantifying the use of technologies such as, standards, adopted services, file formats and platforms.
- ✓ Analysing collected data and reporting the results
- ✓ Discussing the implications in line with the objectives of BlogForever

The detailed account on the data collection methods and analysis is discussed below.

5.1 Objectives, Data Collection Methods and Datasets

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.com⁹ ping server.

Weblogs.com receives notifications when new content is being 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 recognised 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 recognised 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 [27-29].

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.com¹⁰, Top 40 blogs published by Blogpulse.com¹¹ and a collection of blogs acquired from the blog user survey described above.

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.

⁹ Weblogs.com (<http://weblogs.com>) intends to provide a free, open access ping server.

¹⁰ Technorati.com (<http://www.technorati.com>)

¹¹ Blogpulse.com (<http://www.blogpulse.com>)

5.1.1 Data Collection Methods and Datasets

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.

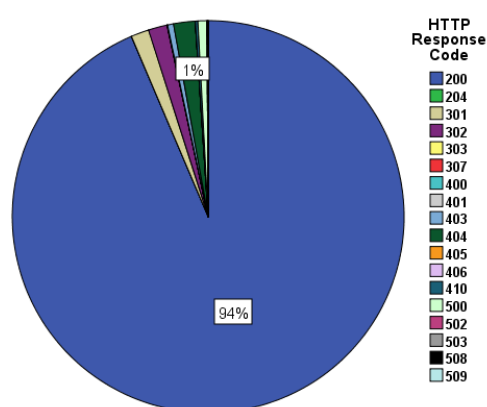
To implement the data collection, custom software was implemented using a combination of PHP¹² and Bash¹³. More specifically, PHP5.3 was used to implement the core of the application. The CURL network library was used to implement communication with the blogs via HTTP and regular expressions were utilised in order to parse the blog source code and evaluate the use of certain technologies. Furthermore, 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 any case, the software retrieves the specific URL HTML code and proceeds with the analysis and results. (See Appendix F for more details).

The datasets were generated on 16 August 2011 between 12:00 to 19:00 GMT for Weblogs.com. The rest of the datasets were generated on the same day between 19:00 and 20:00 GMT. A Linux powered machine with i7 Intel 2.8 MHz CPU and 12GB RAM was used for the development and data collection.

The overall number of data entries collected was 259,930. HTTP response codes have been recorded. Items where status code was not retrieved 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 18.

Figure 18 – HTTP response codes registered during the data-collection stage



The datasets are available for download from the BlogForever project website¹⁴.

Weblogs.com Dataset:

¹²<http://www.php.net>

¹³<http://www.gnu.org/s/bash/>

¹⁴<http://blogforever.eu>

- 1) 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. The XML file was parsed and URL entries were extracted for further processing.
- 2) The URL entries have been filtered to distinguish between updated resources and their hosted websites. Duplicate entries have also been removed.
- 3) 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

Coma-delimited Dataset Files:

- blogs-weblogs-results.csv
- posts-weblogs-results.csv

Technorati and Blogpulse Datasets:

- 1) The list of top 100 and top 40 blogs ranked by Technorati and Blogpulse respectively has been acquired on 12 August 2011.
- 2) 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

Coma-delimited Dataset Files:

- blogpulse-results.csv
- technorati-top100-results.csv

Contributed Blogs:

- 1) The URL entries of all the contributed blogs were made available after processing the results of the survey.

Total number of accessed resources: 504

Total number of valid records: 145

Coma-delimited Dataset Files:

- survey-blogs-results.csv

Total Data Corpus:

Overall total of accessed resources: 259,930

Overall total of valid records: 209,830

5.1.2 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 35:

Table 35 – List of technologies considered in the evaluation (*+count indicates that number of identified occurrences were 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

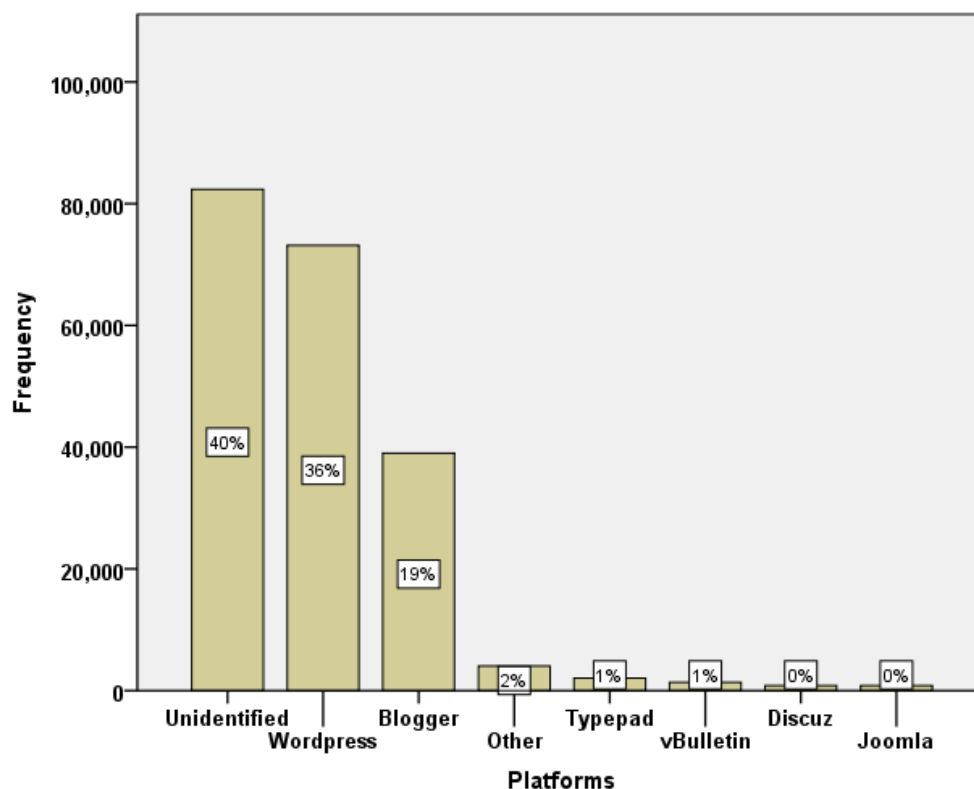
Dojo.js	JQuery.js	SIOC
Dublin Core (+count)	JQueryUI.js	Software/Platform
ExtCore.js	Microdata	Twitter
Facebook	Microformat-hCard	YouTube embedded video
Flash (+count)	Microformat-XFN	YUI.js
FOAF	MooTools.js	XHTML
Google+	Open Graph Protocol (+count)	Other MIME Types (see 5.2.8)
img-PNG (+count)	Open Search	
img-GIF (+count)	Pingback	

5.2 Evaluation Results

5.2.1 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%). However, in 40% of the cases, information about the platform remained unknown. 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 (Figure 19). 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.

Figure 19 – Frequency of weblog-powering software platforms

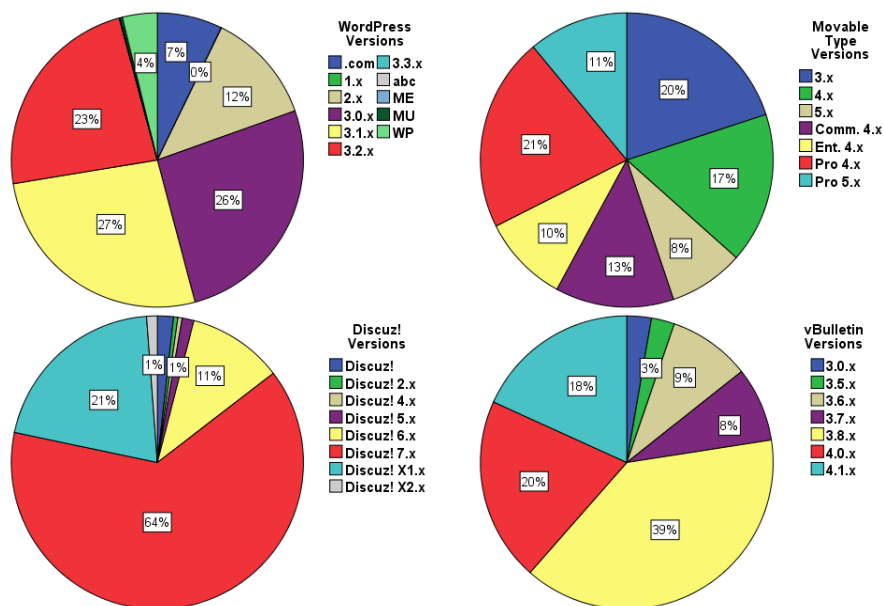


There is a considerable variation across most popular software platforms used. The consistency in specifying versions of adopted software varies too. However, it is still possible to identify the extent of adoption and noticeable patterns within the studied corpus.

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 20, 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 20 – Variations in versions of adopted software



5.2.2 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, recommended¹⁵ 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">
```

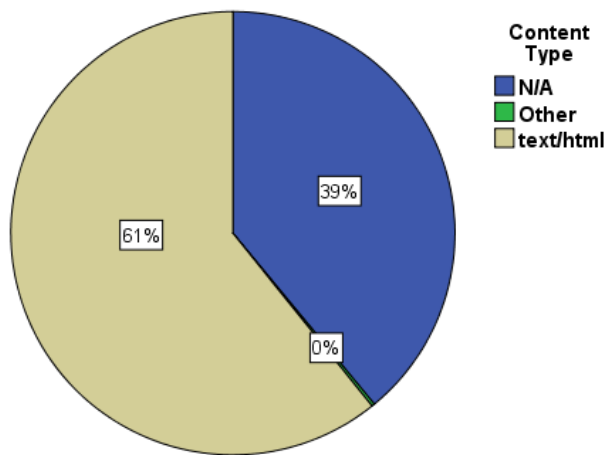
User agents are expected to work with any character encoding registered with IANA¹⁶, however, the support of an encoding is bound to the implementation of a specific user agent.

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 labelling. Information about the types of documents distributed by blogs was also collected.

¹⁵ <http://www.w3.org/TR/html4/charset.html>

¹⁶ <http://www.iana.org/assignments/character-sets>

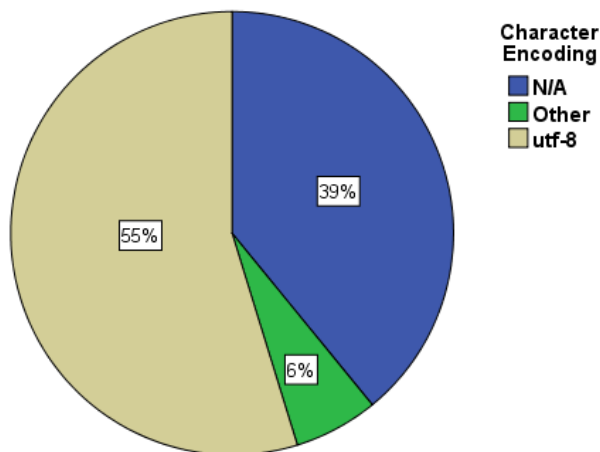
Figure 21 – Content type of the evaluated resources.



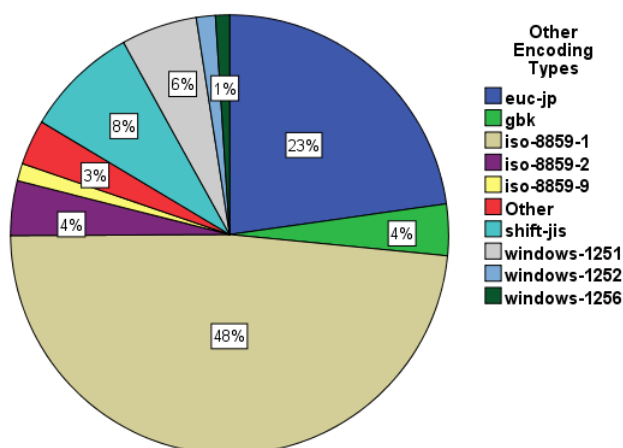
The results suggest that text/html is the most widely (61%) specified content type within the studied corpus. Other types constitute to less than 1% and include: application/xhtml+xml; /xml; /xhtml+xml; /vnd.wap.xhtml+xml, as well as text/xml; / javascript; / php1; / shtml; and / html+javascript. A considerable 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 22).

Figure 22 – Encoding of the evaluated resources.



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 23 for more details.

Figure 23 – Break down of the other 6% (see Figure 22) of charset 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 charsets other than UTF-8.

5.2.3 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)¹⁷ 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.

HTML5¹⁸ 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.

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

¹⁷ <http://www.w3.org/Style/CSS/>

¹⁸ <http://dev.w3.org/html5/spec/Overview.html>

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 24 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 25. 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 26 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 24 – Average number of images identified

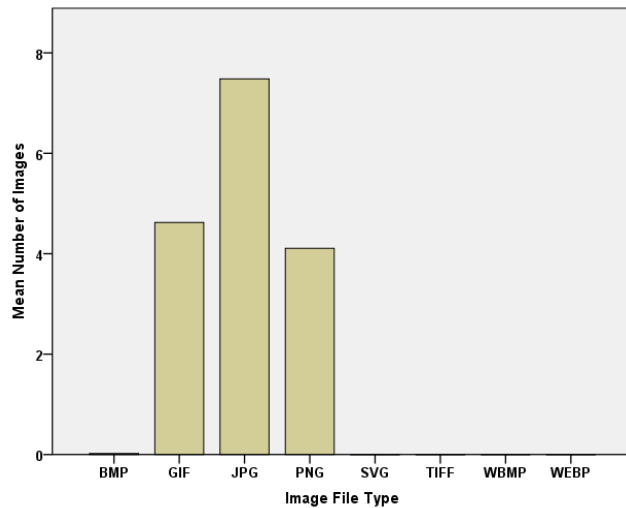


Figure 25 – Average use of BMP, SVG, TIFF, WBMP and WEBP formats

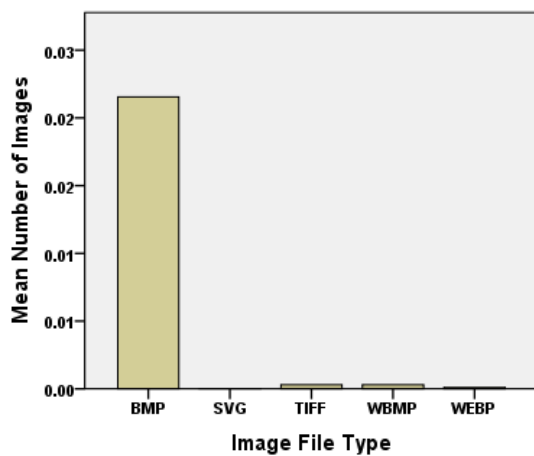
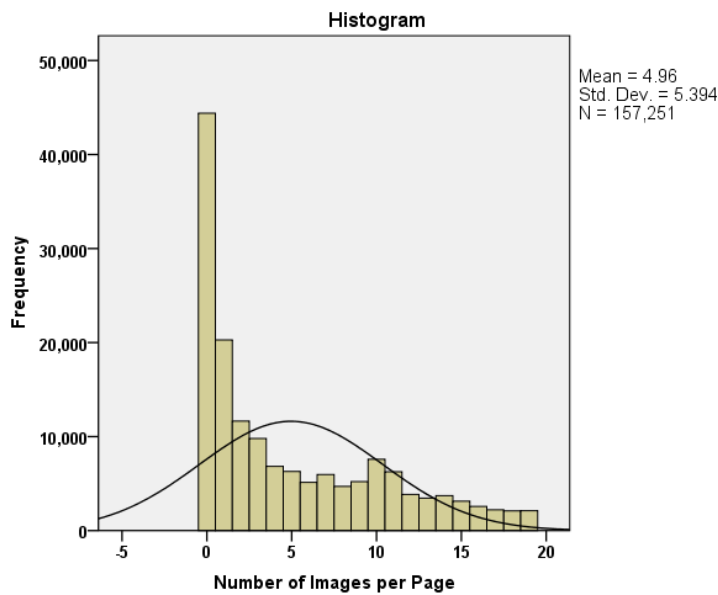


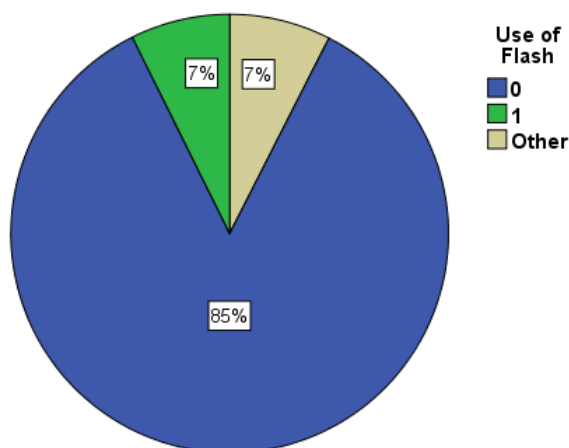
Figure 26 – Distribution of images for pages with less than 20 images only.



Flash¹⁹, 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 resources 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 resources did not include any Flash content (see Figure 27). 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.

Figure 27 – Number of Flash instances detected.



5.2.4 Semantic Markup: Microformats, Microdata and Metadata

¹⁹<http://www.adobe.com/products/flash.html>

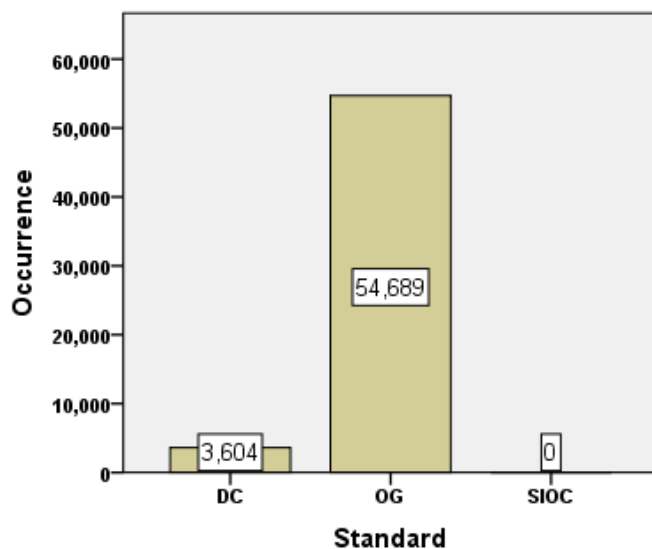
One of the objectives of this evaluation was to evaluate the adoption of semantic mark-up within the Blogosphere. To address this objective this investigation looks into the use of metadata formats and associated technologies. This section discussed the use of:

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

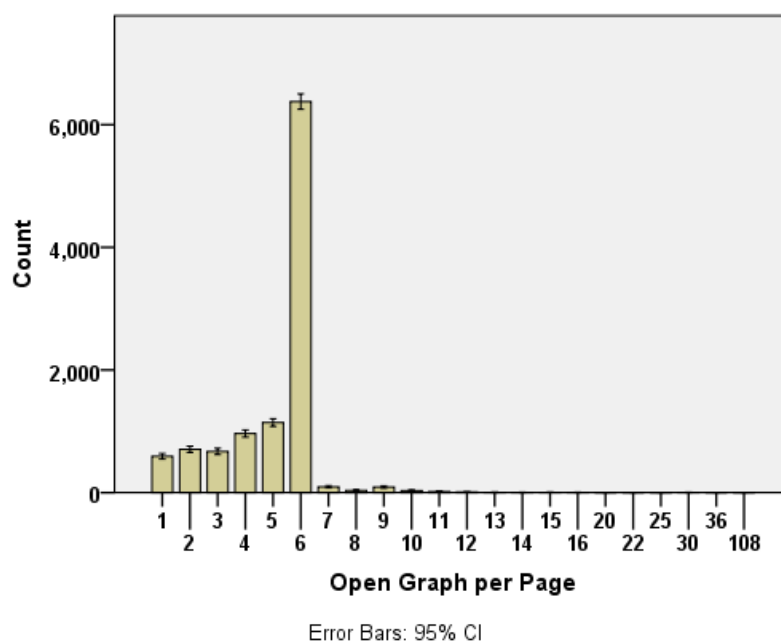
Metadata are commonly defined as data about data. Within the context of the Web, metadata are commonly referred to as the descriptive text used alongside web content. Examples of metadata can include keywords, associations or various content mapping. It is often required to standardise these descriptions 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 28). 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 28 – Summary of metadata use



The histogram of OG occurrences is shown in Figure 29. 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 29 – Histogram of Open Graph references

Microdata²⁰ and **Microformats**²¹ 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" >`

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%) of using RDF. No references to Open Search were identified.

²⁰ <http://www.w3.org/TR/microdata>

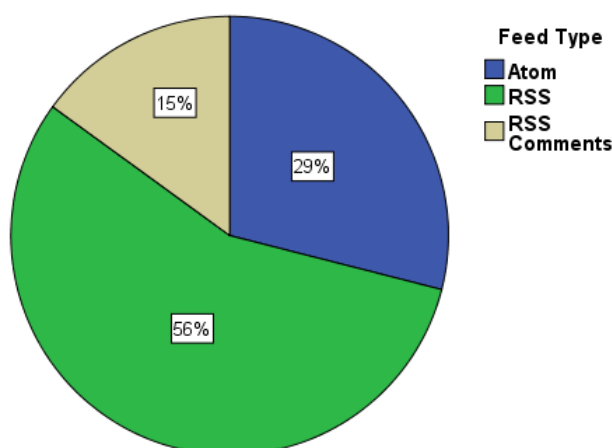
²¹ <http://microformats.org/about>

5.2.5 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 [30]. 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 30. 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 30 – Use of web feeds by type



5.2.6 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:

- ✓ Dojo²²
- ✓ Ext Core²³
- ✓ JQuery²⁴
- ✓ JQuery UI²⁵
- ✓ MooTools²⁶
- ✓ Prototype²⁷
- ✓ YUI Library²⁸

²² <http://dojotoolkit.org/>

²³ <http://www.sencha.com/blog/ext-core-30-beta-released/>

²⁴ <http://jquery.com/>

²⁵ <http://jqueryui.com/>

²⁶ <http://mootools.net/>

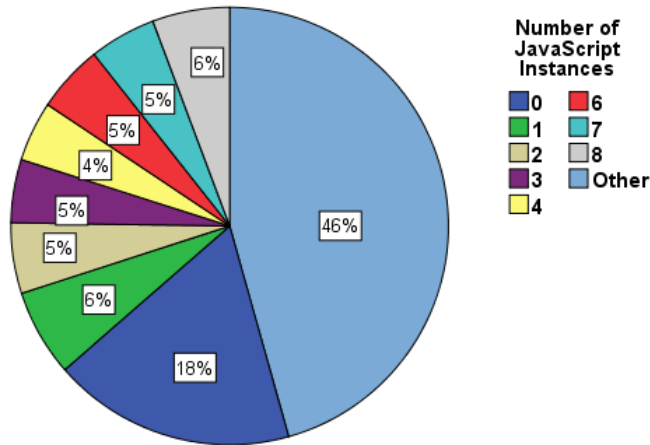
²⁷ <http://www.prototypejs.org/>

²⁸ <http://developer.yahoo.com/yui/>

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

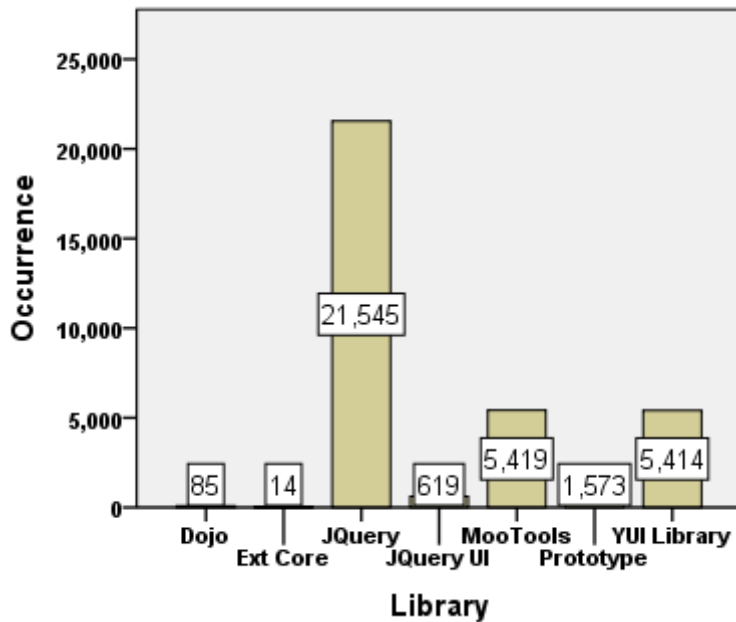
To use of JavaScript by each of the accessed resource has been quantified based on the number of *.js files linked or segments of JavaScript code embedded within the accessed document. The results suggest a wide adoption of JavaScript with 82% of the entire studied corpus having at least one reference to JavaScript. The average number of JavaScript instances is large too – 12.5 instances per resource (Figure 31).

Figure 31 – Number of JavaScript instances identified.



Within the identified instances of JavaScript code, there are references to specific libraries and frameworks. Their 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 32).

Figure 32 – Number of identified library/framework instances



Last, but not least, this sections summarises the use of Pingback²⁹ 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 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.

5.2.7 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:

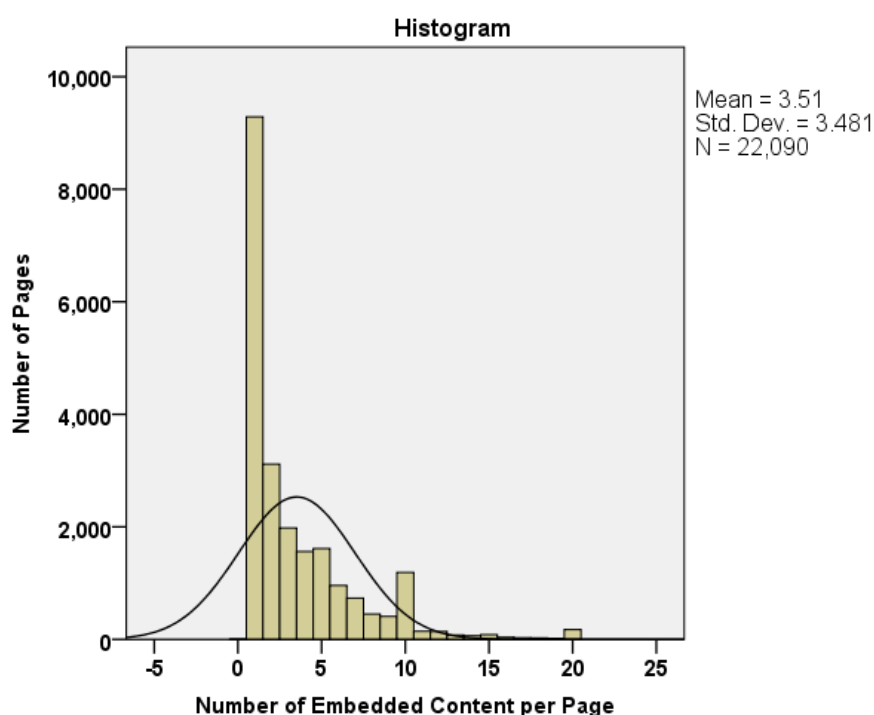
- ✓ Twitter
- ✓ Facebook
- ✓ Google+
- ✓ YouTube

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 – totalling 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 33 (for convenience, the 0.6% of outliers were reduced to 20).

²⁹ <http://hixie.ch/specs/pingback/pingback-1.0>

Figure 33 – Frequency of embedded YouTube videos.



5.2.8 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)³⁰. 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 36.

Table 36 – File 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

³⁰ <http://www.iana.org/assignments/media-types/index.html>

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.

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.

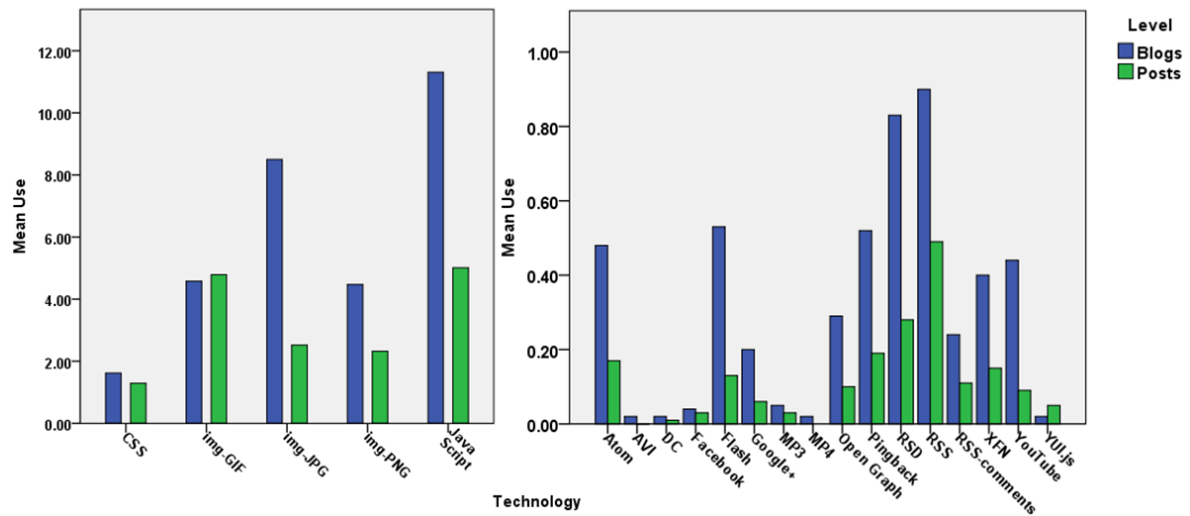
5.2.9 Single Posts versus Websites

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. This section reports the most prevalent differences recorded (Figure 34).

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.

Figure 34 – Differences in use of technology on the level of posts/pages and websites.



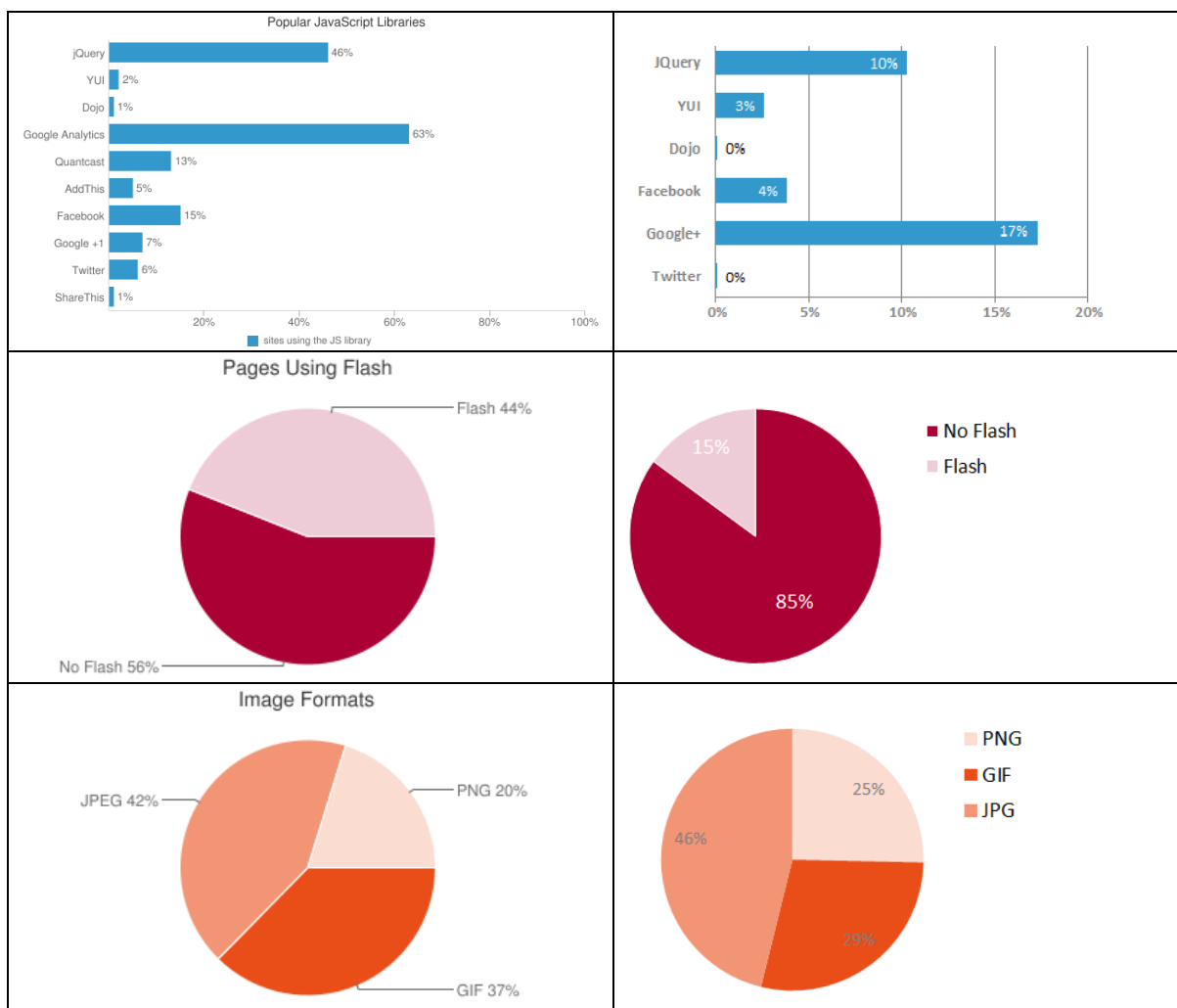
5.2.10 Differences between the Blogosphere and Web

This section of the report compares the data published by HTTP Archive³¹ with the data obtained from Weblogs.com and used in this evaluation. The justification for comparing the sources of information is to discuss the differences between the Blogosphere and the entire Web. The data used from the HTTP Archive corresponds to the timeframe of the data obtained from Weblogs.com. The results are summarised in the following Table 37.

Table 37 – Comparison between the Web and the Blogosphere

The Web	The Blogosphere
<p>Pages with Errors (4xx, 5xx)</p> <p>No Errors 75%</p> <p>Errors 25%</p>	<p>HTTP Response Codes</p> <p>200 94%</p> <p>3xx-5xx 6%</p>

³¹ <http://httparchive.org>



5.3 Summary

This chapter outlines the results of an evaluation that investigated the use of various technologies. The key message emerging from the study argues for the diversity of the Blogosphere. More specifically, there is a large number 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, this diversity exhibited in the Blogosphere may require additional efforts for avoiding data loss or distortion when aggregating, preserving and disseminating blogs.

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 will require fine alignment and intricate tuning of the solutions developed by BlogForever.

On the other side, however, there are a large number of established and widely used technologies and standards used 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. Other file formats are less frequently used. Due to marginal use of some file formats their support within the context of BlogForever may be considered to be non-cost-effective.

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 and aggregation of blogs conducted by the anticipated BlogForever tools.

Consequently, and more generally, this evaluation measures and reports the technological foundations used in the Blogosphere. The results of this evaluation can, therefore, inform the strategies for crawling and preserving blog data. Within the context of BlogForever, this evaluation can particularly inform the development of the required data model and the range of applications for developing the anticipated services and solutions.

6 Analysis of Inter-Blog Relationships

The aim of this chapter is to facilitate and to prepare the application of social network analysis on weblog data gathered by the weblog spider and stored in the BlogForever archive.

Weblogs are web pages written by one or more authors that cite or refer to other web content by linking to other web pages and blogs. Thereby, a network of interconnected blogs emerges – the so-called Blogosphere. Social network analysis (SNA) can help to better understand the structure and dynamics in the Blogosphere.

The BlogForever project has formulated various objectives for the application of SNA. In an initial step we can distinguish between analyses of the structure of the Blogosphere (or parts thereof) and analyses of dynamics in the Blogosphere.

The first aim is the identification of structural particularities, e.g. highly connected blogs, subgroups or clusters in the Blogosphere. Such information can facilitate the deployment of rankings, the selection of related blog posts, and the aggregation of clusters in the Blogosphere. Thereby, the capabilities concerning the knowledge discovery process of the BlogForever platform users can be improved. An analysis regarding structural aspects is done at one given point in time. Of course, the analysis can be repeated but it applies always to a snapshot.

In contrast, the analyses of dynamics in the Blogosphere explore changes during a given time period. This aims on a better understanding of how structures in the Blogosphere evolve and vanish as well as why changes occur. Therefore, the identification and explanation of the behaviour of blogs and their authors is in the focus of dynamic social network analysis. Dynamic analyses on longitudinal network data try to overcome the weaknesses of classical static network models [cp. 31, p. 730]. Thus, dynamic social network analyses enable the inquiry of additional research questions, e.g. information and innovation diffusion [cp. 32, 33].

Referring to the description of work, the following objectives with respect to the structure of the Blogosphere should be supported by the application of social network analysis:

1. Establishing a ranking among weblogs in order to assign a priority to a user's request^{32,33}
2. Determination of central weblogs and boundaries together with the according frequencies to be stored in the archive³⁴

Additionally the following objectives regarding the dynamics in the Blogosphere should be supported:

3. Examination of weblog lifecycles, emerging weblogs and peripheral weblogs moving to a central position^{35,36}
4. Studying the role of group formation among online actors and the creation of shared meaning across group of bloggers, further of the role of prominence and collective filtering³⁷

Some research has already been conducted in the area of social network analysis and blogs. The following seven examples give a first insight in interesting related research.

³²Task 2.1 Weblogs survey, Part A, Description of Work, p. 7

³³Task 2.1 Weblogs survey, Section B1.3.1.1 Detailed work description, Part B, Description of Work, p. 23

³⁴Task 2.1 Weblogs survey, Section B1.3.1.1 Detailed work description, Part B, Description of Work, p. 23

³⁵Task 2.1 Weblogs survey, Part A, Description of Work, p. 7

³⁶Task 2.1 Weblogs survey, Section B1.3.1.1 Detailed work description, Part B, Description of Work, p. 23

³⁷Task 2.1 Weblogs survey, Section B1.3.1.1 Detailed work description, Part B, Description of Work, p. 23

Yang and Lin [34] combined information retrieval techniques with network analysis to improve recommender systems for blogs. Therefore, they developed four approaches based on post citation network, blog-based social network, and post content. The approaches were evaluated with collected blog data.

Goncalves et al. [35] aim on the enhancement of blog rankings and search engines for blogs through the inclusion of popularity measures. Popularity means that a significant portion of a collective or a group “like, approves, or finds the object suitable in some given context”. Proposed indicators for the measurement of popularity are the number of visits, number of downloads or number of social annotations as well as especially for blogs the number of subscribers (to the RSS-feed), relative click-through ratio and number of times the blog appeared in “top lists.

Dolinska [36] examined the use of centrality measures of social network analysis to enhance blog searching. Therefore, she applied the measurement of in degree centrality, out degree centrality, and all degree centrality on the network that emerges by capturing the links from blogrolls.

Xiaoguang Wang et al. [37] perform a social network analysis on a Chinese blog community to enquire blog-supported scientific communication. They applied their analyses on blogroll links and limited their enquiry to the blog provider Csdn.net. In the examined network, they found a lot of star-like loose clusters. Thereby, they identified a few central bloggers with a lot of inbound social hyperlinks, and many ordinary bloggers with mostly only one inbound social hyperlink. Furthermore, the link density is high between central bloggers and ordinary bloggers, and low among ordinary bloggers.

Agarwal et al. [38] aim on the identification of influential bloggers. Therefore, they propose a model to quantify the influence of bloggers (not blogs). The model estimates the influence of a blog post by using the number of inlinks, outlinks, comments, and the length of a blog post to measure blog posts recognition, novelty, activity generation, and eloquence. The influence of a blogger is determined by her or his most influential blog post.

Adar & Adamic [39] examined the spread of information among blogs. They applied link inference techniques to find non-explicit links because not every blog cites the source of the information. The inference techniques relate two blogs based on the number of common blogs explicitly linked to, the number of shared non-blog links, text similarity, order and frequency of repeated infections, and in-link and out-link counts.

Kumar et al. [40] examined the evolution of blogs and blog communities from 1999 to 2003. Thereby, they enquired the dynamics of connectivity in the blogosphere. As one finding, they observed an explosive increase in connectedness, and in local-scale community structure around the end of 2001.

The remainder is organised as follows. Section 6.1 gives a short introduction into the elements and different types of SNA. Section 6.3 identifies the elements and relationships of weblogs that are accessible for enquiries via SNA. To facilitate the application of SNA for weblogs, section 6.3 deduces requirements for the data model and for the weblog spider.

6.1 Social Network Analysis

In the following, we introduce social network analysis (SNA), the main concepts, different types of SNA, and some important measures. This overview helps to understand how different types of social network analyses can be applied on blogs and the Blogosphere.

SNA aims to enquire about social behaviour by focussing on relationships. These relationships can occur between individuals as well as groups. The individuals and their relationships can be represented as a network of nodes and edges. SNA provides various methods and measures to investigate and to assess such networks.

The fundamental concepts in SNA are [31, pp. 17-21]:

- ✓ **Actors.** An actor represents a social entity. Such an entity could be an individual as well as a collective social unit. Therefore, examples are people, departments or nation-states. The actors in an inquired network can be all of the same type (one-mode networks) or of different types. In a virtual environment, actors can be as well digital objects like software agents.
- ✓ **(Relational) Ties.** A relationship between actors is called social tie. A tie links two actors and can be of different types, e.g. expressed friendship, sending messages or business transactions. A special type of ties can be created on common behaviour of two actors (co-occurrence or co-citation), e.g. if they refer to the same book.
- ✓ **Relation.** A relation is a collection of ties of a specific kind. A relation can be seen as a class for relationships between the actors and the ties are the instances of that class. There can be different relations in a network, e.g. diplomatic ties and trade activities among nations. Furthermore, special relations can be aggregated to a general relation, e.g. different types of trade activities to a general “trade activity” relation.
- ✓ **Dyad, Triad and Subgroups.** Dyads are pairwise relationships. They consist of a pair of actors and the (possible) tie(s) between them. The triad consists of three actors and the (possible) tie(s) between them. The assessment of triads allows additional analyses like transitivity and balance. A Subgroup is any subset of actors and all ties among them.
- ✓ **Group.** A group consists of a finite set of actors. It must be reasoned by theoretical, empirical or conceptual criteria why these actors belong together. The set of actors has to be finite to analyse the data of the network.
- ✓ **Social Network.** The social network consists of a finite set of actors and the relation(s) defined on them.

Social network analyses can be distinguished between whole-network analyses and egocentric analyses. Whole-network analyses examine a finite set of actors and the relations among them. Relationships to actors outside the network are not taken into consideration. In contrast, egocentric analysis focus on a single actor(s) (ego) and the relationships of this actor(s) to other actors (alters). In this case, the related actors can be inside the network (the set of actors) as well as outside the network [41, 42].

Another distinction of social network analyses can be made between static and dynamic analyses. Static analyses usually aggregate the ties between actors to create a network representation. Thereby, the social network can be examined with respect to different relations, actors, and subgroups. This leads to propositions, e.g. about the impact of actors in a network or about the closeness of actors. The static analyses have their limitations if the evolution of the network is of interest. Therefore, dynamic analyses disaggregate the ties and can visualise change processes in the network over time [33]. Disaggregation means that the dynamic analyses view the single ties in defined time slots or at points in time, and thereby, how the relationships among the actors are changing, while the static analyses assesses the strength of relationships by aggregating the ties between two actors.

SNA provides a number of measures that can be interpreted in various contexts. Some of the central measures are [31, 33]:

- ✓ **Network size.** Number of nodes (or actors) in a network.

- ✓ **Relationship strength, tie strength.** The strength of the relationship in a dyad. Dependent on the context, the relationship strength can represent e.g. the frequency of interactions, count actual interactions, etc..
- ✓ **Degree.** The degree measures the number of direct contacts of an actor. If the relations are directed then it can be differentiated between out-degree (number of outgoing relationships with other actors) and in-degree (number of incoming relationships from other actors).
- ✓ **Network roles.** Actors in the network can have different roles. For example, a broker (Gatekeeper) connects two cliques or subgroups in a network. The broker has an important (or powerful) role because the subgroups are disconnected without the broker. A transmitter, on the other hand, has only outgoing relationships but no incoming relationships. In contrast, a receiver has only ingoing but no outgoing relationships. Actors with a pulsetaker role have a small degree but connect to actors with a high degree.
- ✓ **Density.** The number of realised relationships in the network divided by the number of theoretically possible relationships.
- ✓ **Shortest path (geodesic).** The shortest path between two nodes depends on the number of steps (or actors in between) that are needed to transfer an information or object from actor A to actor B. For example, the shortest path between A and B is 1 if actor A and actor B have a direct relationship. The shortest path between A and B is 2 if actor A can reach actor B only with at least one other actor in between.
- ✓ **Centrality betweenness.** The number of shortest paths between pairs of actors, which run through the observed actor. A high betweenness represents a control position in the network. A broker (see network roles above) has the highest betweenness in the network.
- ✓ **Centrality closeness.** Average shortest path length of an actor to all other actors in the network. Closeness is a distance measure that represents how many steps are necessary in average to reach each other actor.
- ✓ **Diameter.** Diameter is a distance measure that represents the longest shortest path in the network. The shortest paths between each pair of actors are compared and the longest of them represents the diameter for the network.
- ✓ **Reciprocity.** Symmetry of relationships. A relationship between two actors is called symmetric or reciprocal if the relationship from actor A to actor B also exists from actor B to actor A.

An interesting alternative next to direct relationships, like the exchange of objects between two actors, is the identification or modelling of relations based on co-occurrence or co-citation. A co-citation between two actors exists if both actors refer to the same object, e.g. if two scientific articles cite the same literature. The more citations these two articles have in common the stronger is the relation between the articles. Such relations based on co-citations can be as well very useful in the context of blogs.

As a summary, it can be stated that the network of linked actors is the central aspect of SNA. To perform a SNA in the BlogForever project, a prerequisite is to identify and to define the actors and relations that should be examined as well as their properties. Therefore, the next chapter investigates blogs and their interrelationships.

6.2 Inter-blog Relationships

In the following, we identify the central elements in the context of blogs that can be examined by social network analyses. The elements are distinguished in terms of actors and relations. Additionally, examples of properties for the elements are listed.

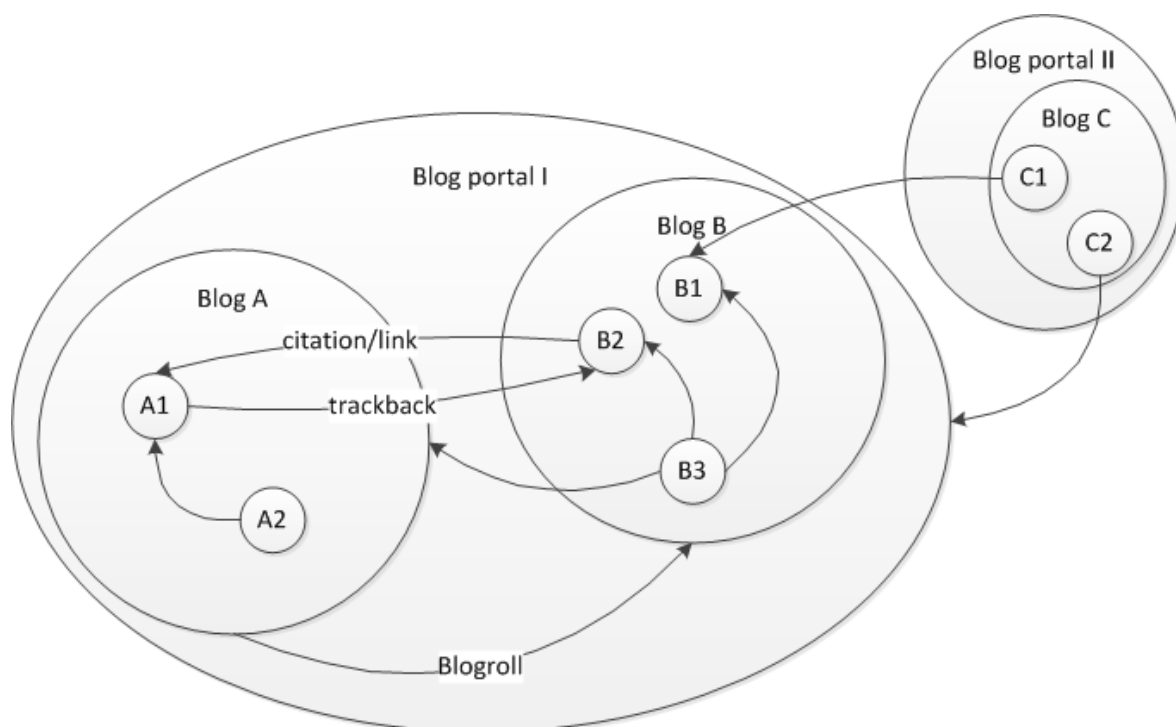
6.2.1 Actors

Actors are the nodes in a social network graph. An actor can have or maintain various ties to other actors in the network. As stated above, an actor can be an individual as well as a collective. In the

context of blogs, the most obvious individuals are *blog authors* and *blog readers*. They are not disjunctive because a blog author can be a reader and vice versa. Normally, a blog author will be a *person* but can also be a *software agent* that generates blog posts (e.g. the daily usage protocol of a system). Persons have various attributes that can be (more or less) permanent like name, gender, date of birth, nationality as well as temporary attributes like geographical position or mood. However, only some people will provide such information publicly in a blog. Therefore, further research has to be performed to examine which attributes could be captured from blogs (e.g. from a “About” page of the blog), and in how many cases detailed information about actors are available.

Next to authors and readers further entities are possible to represent actors in the network. One possible actor is the blog itself. A blog can be seen as an entity that establishes relationships by linking to other blogs as well as other web resources. A blog has various attributes, e.g. a URL, a language, and date of creation. Another possible entity is a blog portal that hosts or aggregates various blogs, e.g. ScienceBlogs³⁸. To use such aggregation of blogs can help to enquire and visualise relations on a higher level, e.g. to inquire if blogs and blogger from different blog portals interact or to navigate from an aggregated level to a more specific level and vice versa. A prerequisite for such aggregation has to be that the affiliation of blogs to a portal is distinctive. Figure 35 shows a simple example for the aggregation of blog posts to blogs, and blogs in blog portals.

Figure 35 – Examples for blog relations



The attributes of actors can be used for two purposes. First, a group of actors can be aggregated based on the attributes of authors, readers or blogs. This is similar to the aggregation in portals stated above. Such a group can appear as one actor (or node) in the network. Relationships of elements of this group to other actors in the network appear as relationships between the group entity and the other actors. Such aggregation can help to examine and visualise the network in an appropriate way, especially if there is a huge amount of actors in the network. The second purpose of the attributes of actors is to identify subgroups that can be examined or compared regarding their structure. For example, a possible study could identify in the group of scientific blogs different subgroups with respect to the language of the blogs. Network measurements can be used to

³⁸ <http://scienceblogs.com/>

compare these subgroups, e.g. if the density of English speaking scientific blogs is different to the density of Greek speaking scientific blogs. Central actors of the subgroups can be identified by other network measurements like betweenness and closeness, e.g. who is a central actor in the Greek speaking scientific Blogosphere even if it is not a central actor in the overall scientific Blogosphere. Therefore, the identification of subgroups based on attributes of the actors could lead to interesting insights about the network structure or network dynamics in different blogger communities.

6.2.2 Relations

Relations are the edges in the social network graph. A relation connects a pair of actors and represents a symmetric or asymmetric relationship between the actors. If we look at the Blogosphere, relations can be explicit and public available (e.g. links), explicit but just internal available (e.g. views that are tracked in the log file), and existent but not available in the blogs themselves (e.g. friendships between authors).

Four relationship types can be identified that are explicit and public in the Blogosphere [43, 44]

- ✓ **Citation/Link.** Blog A cites blog B if one entry (blog post) of blog A contains a hyperlink to blog B. A citation can direct on another blog as a whole or on individual blog posts in the blog. As well citations of parts of blog posts (like contained images or documents) are possible if they can be addressed by a link.
- ✓ **Blogroll.** A blogroll is a collection of links to other blogs that are placed prominently on the start page or as part of the layout of the blog. Therefore, a blogroll relation between two blogs represents a link in the blogroll of blog A that directs to blog B.
- ✓ **Linkback.** A linkback is a back-reference in a cited blog. If blog A cites blog B, a back-reference will be created on blog B that indicates the citation in blog A. The linkback is performed automatically as long as linkback is activated in both blogs respectively the software of the blogs. The three methods of reback, trackback, and pingback implement the linkback mechanism. Semantically, they mean the same but differ in how they implement a linkback technically.
- ✓ **Comment.** Blog posts can be commented if the comments functionality is enabled. The comment relation represents a relationship between the person who creates a comment and the blog (or blog author) where the comment occurs.

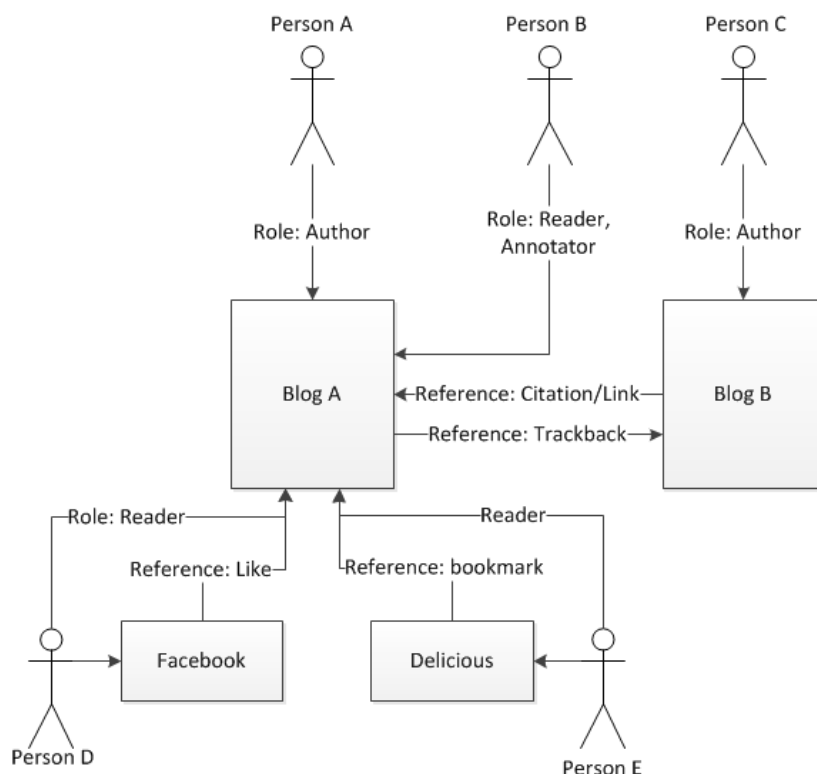
Additional relationships between blogs can be found in the metadata. There are already several metadata standards available that describe blogs (or web pages) semantically and indicate additional relationships to other blogs or web pages. Examples are XHTML Friends Network (XFN) and Friend of a Friend (FOAF). Information about the actual adoption of the standards in blogs is necessary to estimate their potential for social network analysis.

Examples for the relations of citation/link, blogroll and linkback (e.g. trackback) are visualised in Figure 35. There are, as well, other explicit and public relationships, such as bookmarks, and adding to tweet/retweet or Facebook. In contrast to the above mentioned relationships, these establish connections to platforms outside the blogosphere, e.g. delicious.com or facebook.com. Therefore, a gathering of additional platforms other than blogs will be necessary if these relationships are to be taken into consideration for analyses.

To facilitate a better understanding Figure 36 visualises an example with several types of explicit relationships. Person A is the author of blog A. Four other persons (B, C, D, E) read this blog and perform different feedback actions on it. Person B just comments on the blog posts. These comments can simply be captured from the blog itself. Person C runs another blog and refers to blog A by linking from a specific blog post in blog B. Thus, a linkback information will be shown on blog A that indicate that blog B is citing blog A. Both explicit relationships (link and linkback)

can be captured from the two blogs. Person D reads blog A and pushes the “Like” button for one blog post. Thereby, a notification is created in the Facebook profile of person D that links to the blog post. Blog A contains only the information on how many persons like this post, not who they are. Therefore, the identity of the senders of the “like” relationships has to be captured from Facebook. Similarly, Person E creates a delicious bookmark to a blog post of blog A. Again, this relationship cannot be captured solely from the data of the blog itself.

Figure 36 – Examples for explicit relationships with blogs



There exist many applications like Facebook, Delicious or Twitter which allow a person to express a relationship to a blog or blog post. The three examples of Facebook, Delicious, and Twitter are explained more in detail in the appendix. It is impossible to describe a complete list of such applications because there are already a lot and the numbers increasing even as we write this report (e.g. Google+ is a well-known new service).

Blogs contain as well links to other web pages and resources other than blogs. These links can be used for the identification of relations even if they are not direct relationships between two blogs inside the Blogosphere. In these cases, the assumption is made that two blogs relate to each other if they cite the same web page or resource (e.g. the website of a newspaper). The strength of the relation is stronger the more often such co-citation occurs between these blogs. Depending of the aim, an analysis of co-citations would consider only those relationships with a strength that exceeds a specific threshold. Thereby, random connections could be avoided and a better identification of patterns is supported.

The target of an external link can be a web page or other resource (e.g. database). For the analysis of co-citations of web pages, it can be useful to disaggregate the target address into different levels. Thereby, the target address can indicate the overall website (e.g. <http://en.wikipedia.org/wiki/>), a specific page inside the website (e.g. <http://en.wikipedia.org/wiki/Blog>) and a specific part inside the specific page (e.g. <http://en.wikipedia.org/wiki/Blog#Behavior>). Depending on the aim of the analysis and the structure of the network, an analysis of co-citations based on higher levels (e.g.

relate people who cite a Wikipedia) can be as useful as the analysis on more specific levels (e.g. relate people who cite the Blog article in Wikipedia).

Embedded or syndicated objects (e.g. a YouTube video) in a blog can be seen similar to external links. Thereby, the blog embeds an object from another page and shows this object as part of the blog. A relationship between two blogs can be identified based on co-citations if both blogs embed the same object (e.g. the same YouTube video) respectively embed an object from the same platform (e.g. YouTube).

Furthermore, the relationships between linked pages or embedded objects can be lead to further insights, e.g. if cited news articles are from the same author or if embed YouTube videos are from the same user in YouTube. A relation between two blogs could be created that indicate that two blog authors cite the same news author or the same YouTube user. However, a capturing of the external websites (e.g. the news website or YouTube) would be necessary if these relationships should be taken as well into consideration for an analysis.

Another opportunity to identify relationships between blogs can be possibly based on tags or categories. Thereby, a tie between two blogs is created if they use the same tag. The more common tags they use the more ties will be created and the stronger will be the relationship between these two blogs. However, tags are just terms that are formulated by the author and without any restriction regarding a common vocabulary or thesaurus. Therefore, such approach can cause some problems that are typically when dealing with tags, e.g. problems of synonyms or homonyms. Nevertheless, it is an interesting opportunity in BlogForever because the tags and categories can be captured directly from the blog itself.

The relation between the reader of a blog and the blog itself cannot be captured by publicly available data but only from internal blog log files. Additionally, blog hosting services can indicate a “reading relation” [43]. The validity of such indication depends on the internal data. It can be assumed as highly valid if the view of a user account on a blog can be tracked. But this is only possible in a closed scenario where the reader has to sign in for reading. More often, the logs will only track the IP address of the reader and therefore, the validity of the relation can be questioned. An additional reader tracking can be performed in the BlogForever project by logging the activities of archive users. However, the behaviour of blog readers in the archive could be different to blog readers in the original blogosphere.

Besides the relations described above, many relations among blog users (authors and readers) are often not directly extractable, e.g. family relationships, organisational structures, and friendships in the real life. Nevertheless, they can be very beneficial for a social network analysis dependent on the research aim. Therefore, it can become necessary to gather such information from other digital sources (e.g. social networking services like Facebook) or via survey techniques (e.g. questionnaire or interview).

Adding properties can increase the expressiveness of relations. For example, a blog post can link to another blog post because it agrees or disagrees. If such information is added to the ties, networks for agreement and disagreement can be compared as well as interdependences between agreement and disagreement can be examined [cp. 45].

6.2.3 Time

An important property for social network analyses on the Blogosphere will be time information because the Blogosphere is a living network that is changing continuously. Dynamic analyses can examine and visualise the changes as long as the information about the time of the events are available. Therefore, the data for event-based analyses has to include the timing of network events. [cp. 33].

Network events can concern actors as well as relations. Actors (persons or blogs) can appear or vanish in the network at a certain time. In a similar way, a tie between two actors will be created at a certain time, e.g. a link between blog A and blog B occurs when the blog posts that includes the link is published. Furthermore, the properties of actors (e.g. amount of blog posts of a blog) and relations (e.g. click rate of a link) can change over time.

Dynamic analyses will provide considerably better insights into blog related behaviour and evolutions in the Blogosphere. Therefore, as many network events as possible should be captured. This can be done if time information for appearance, change, and disappearance of elements (actors, relations, properties) are explicitly available on the blog or if the blog has to be scanned in intervals because of a missing real time indication of blog changes. If time information is explicitly available, data about time has simply to be captured by the spider. If the blog has to be scanned in intervals to get time information, the difference between two snapshots constitutes changes in the blog or the occurrence of network events. Thereby, it is problematic to estimate an adequate time span for the scan intervals because it will depend on the rate of change with respect to the target blog. An example for scanning a blog in intervals would be if the links in the blogroll are changed from time to time but the blog does not indicate these changes via ping or similar things that allow a real time tracking. Therefore, the blogroll has to be scanned at a time A and at a time B. The difference between both scans constitutes the changes in the blogroll (e.g. adding or deleting of a link). The timestamp for the changes would be the time of the second scan. Nevertheless, it is problematic to estimate how often the blogroll will be changed in this blog and therefore, how often it should be scanned.

6.3 Requirements for the spider and data model

In the following, requirements for the spider and for data model of BlogForever are deduced. The spider is responsible for capturing blog data that should be stored in the archive. The data model defines the structure of stored data.

6.3.1 Weblog Spider

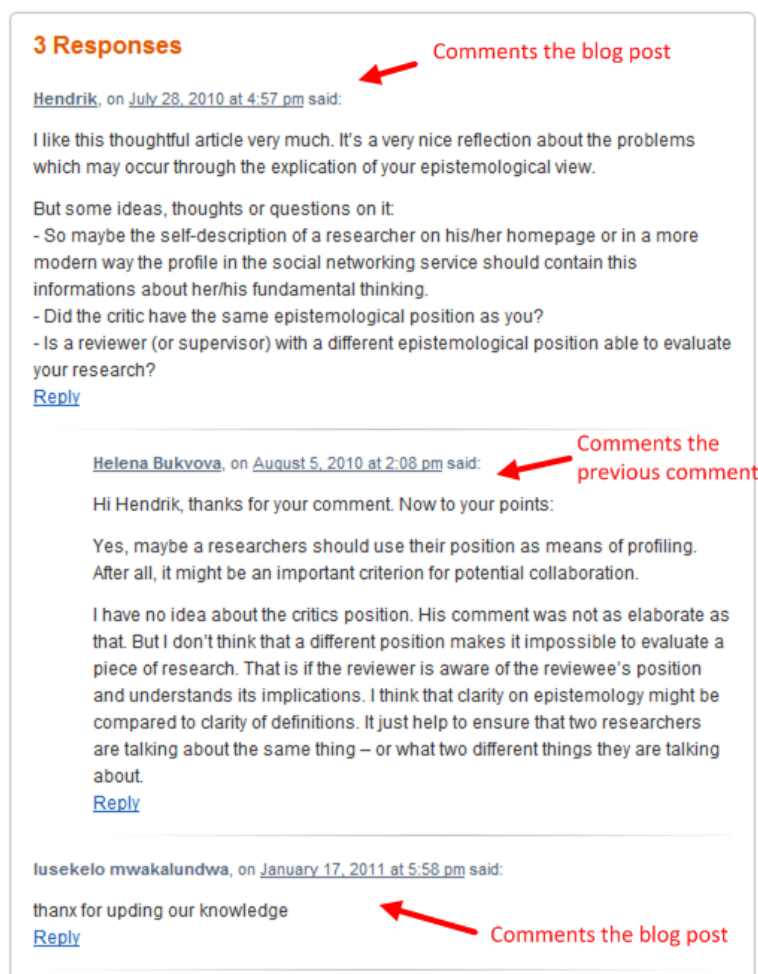
The weblog spider is the software component in the archive that will be accessing the original blog data from outwith the blog server. It is the gateway between the Blogosphere and the blog archive, and, therefore, makes the decision what is captured and what is not. Table 38 shows the elements that are important to perform network analysis for blogs and the Blogosphere, and annotations that should be taken into consideration for the development of the spider.

In the first step, the spider should understand a blog as a container that has a unique identifier (the URL) and contains several other elements like blog posts, comments, blogroll, etc. A blog post has as well a URI that is composed of the blog URL and a unique identification for the post (e.g. date and title of the post). Both URIs have to be stored because they are potential destinations of references in other blogs.

Blogs are interconnected through links. The spider should provide a first differentiation between the different types of links: citation, linkback and blogroll. Additionally, the spider should recognise if the destination of a link is another blog (or part of) or a different resource.

Comments are another kind of relationships. They appear only on the blog post in response to which they are commenting. The content of a comment can address the blog post or other comments. The latter are replies to something that the comment before has stated. Often, it can be identified from the layout or the structure if a comment references to the blog post or a former comment (see Figure 37).

Figure 37 – Comments refer to the blog post or previous comments



To analyse changes in the structure of the Blogosphere, it is essential to gather information about the time when things happen. At least the time of creation of blog posts, comments, and linkbacks should be captured from the blog page because they are often shown explicitly. If additional time information is available, they should be gathered as well, but it has to be accompanied by an analysis of what it is they are addressing. Additionally, blogs should be analysed in intervals to track possible changes e.g. in the blogroll. Thereby, the spider should record the time at which the scan of the blog was performed.

Table 38 – Important elements and annotations for the spider development

Element	Annotation
Blog	A blog is a container element that contains several other elements. A blog has a unique and permanent identifier (URL). A blog has at least one author but can be a corporate blog with multiple authors as well.
Blog post	A blog post has a unique and permanent identifier (URL) that is normally composed of the blog URL with a specific extension. Blog posts have normally just one author.

Links	Distinguish between link, linkback and blogroll. Distinguish the link destination between a blog, blog post, other part of a blog or other resource. Is the link part of a blog post then the date of creation of the blog post can be assumed as the date of creation of the link. It should be captured for external link if they direct to a website (homepage), single page inside the website or specific part on a page.
Embedded objects	The address of embedded/syndicated objects (e.g. embedded YouTube videos) should be captured.
Comments	Does a comment reference to a blog post or to a previous comment?
Author	Can the author of a post or comment be identified by an URI (e.g. URL)? Are there any additional information about the Author and his or her relationships to others (e.g. FOAF data)? Is the author a real person or a software bot?
Time/Date	Are information about the time of creation available for blog posts, comments and linkbacks? Are there other time information available? Information about the time of capturing data from a blog should be stored.
Tags and Categories	Tags and Categories should be captured where available. The hierarchy of categories should be captured if available.
Meta data	Are there any relational meta data (e.g. XFN)? Are there any other semantic meta data that provide additional information? Are there any meta data that help to identify the author or other Internet representations of the author?
Context/Affiliation	Does the blog belong to a bigger entity (e.g. blog portal, organisation)?

Normally, the author of a blog post can be identified or at least marked with an ID so that blog posts of the same author can be found. For social network analysis, it is more important to know which activities belong to the same author than the identification of the real name or real life identity of an author. Therefore, aliases or pseudonyms can be used for identification as well, e.g. to distinguish the blog posts of different authors in a corporate blog. The identification of authors of comments is often more complicated because some blogs allow anonymous comments or guest comments. In these cases the author of a comment cannot be tracked. In some cases the blog author has a name or label but it cannot be used as a unique id. And in some cases the name or the picture of the user is connected to a login of a blog host or to a URL. Therefore, the spider should describe the identity of authors of blog posts and comments as accurately as possible.

Blogs (like any other web pages) can contain various metadata. These should be captured because they can contain relational information (e.g. “known” element in FOAF) or information about the author that facilitate the identification of other relevant digital representations (e.g. Twitter or Facebook representation of the author).

Additionally, other information about the context are important as well and should be captured by the spider. Aggregation of blogs to a bigger entity (e.g. a blog portal) and analysis regarding the influence of affiliations are possible as long as the spider can identify and capture such contextual information. For example, the URL of the blog could be analysed whether the domain indicate that the blog is hosted by a blog provider (e.g. the URL is a subdomain of “blog.de”).

6.3.2 Data Model

The data model defines the structure of the stored data. It is crucial for social network analysis that necessary data are adequately accessible. Proper data structures allow direct access to the data for network structure analyses and enable us to avoid additional effort to extract these data later.

For the inner structure of blogs, we can distinguish between blogs, blog posts and comments. Blogs are more or less container elements that contain the blog posts, the related comments and some other content (e.g. blog roll or additional pages). Blogs and blog posts possess a URL and therefore can be linked by others. Normally, comments cannot be linked from outside the blog but internally, comments can refer to previous comments. Due to the relevance of these elements for social network analysis, we can formulate the following requirements for the data model:

1. It should be differentiated between blogs, blog posts, and comments.
2. Blogs and blog posts have a URL. Additionally, comments could have a URI.
3. Comments can address a blog post or a previous comment.

As stated in section 6.2, there are various types of relations from blogs to other blogs or to other resources (e.g. web pages). To differentiate the relationships associated with a blog is highly important in facilitating a meaningful network analysis. Therefore, we can formulate the following requirements:

4. Links should be made explicit in the data model.
5. Links should be differentiated by several well defined types. At least, we might make a distinction between blogroll, link to another blog, link to another blog post, link to another web page, and link to another resource.
6. Links should have a sender and a recipient.
7. For external links that have a web page or other resource as link target, it should be possible to differentiate between the whole website, a single page on the website and a specific part of a single page. A flexible opportunity is needed because other differentiations are possible as well.

Blogs, blog posts and comments are written by people. The names of these authors often appear next to the blog post or comment. In some cases, the author has a URI (e.g. the account in the blogging software). Nevertheless, a lot of people use aliases or guest accounts. The following requirement can be formulated:

8. Authors of blogs, blog posts and comments should be stored if available.

Authors, blogs, blog posts, comments and relations can have various attributes or properties. These attributes allow a categorisation or aggregation of the elements in the analysis. For example, an analysis can focus on blog posts written in English. In this case, the necessary property would be the language of the blog post. There are many possible properties. Some can be extracted directly from the blog because they appear explicitly (e.g. category of a blog post). Other properties have to be derived from the content through additional analysis (e.g. derivation of the subject of a blog post through text analysis). Therefore, we can formulate as a requirement:

9. There should be the opportunity to add various properties to the elements in the data model.

The affiliation of a blog or an author can facilitate the identification and understanding of groups and subgroups in a social network. For example the blogs that belong to a specific blog portal or the authors that belong to the same organisation can be in the interests of network analysis. Therefore, the following requirement can be formulated:

10. There should be an opportunity to add affiliations to authors and blogs.

Often, blogs provide additional functions for their readers to give feedback or to connect blog posts with other platforms. Examples are Facebook, Delicious, and Twitter (see the appendix for more details) but there are many more. These external connections are useful for analyses that take connections into account which relate to resources outside the Blogosphere. Especially as the authors of these connections (e.g. a Twitter tweet) can be bloggers as well, it can increase the expressiveness of social network analysis. Therefore, the following requirement can be formulated:

11. There should be an opportunity to add relationships to a blog or blog post which describe the external relations from other platforms to the blog or blog post. The character of such relationships can vary from the indication how many people have created a connection (e.g. how many “likes” the page) to meaningful statements about the blog post (e.g. a comment about the blog post via Twitter).

Blogs and the Blogosphere are changing permanently even if the intervals of changing vary greatly among the blogs. Information about time are needed to analyse dynamics in the Blogosphere and therefore, should be considered in the data model. Three types of events – appearance (or creation), adaption, and disappearance – can occur for an element and can be related with a timestamp. Therefore, the following requirements are formulated:

12. Elements of the data model like blogs, blog posts, comments, authors, links, etc. should have a property for their initial appearance and disappearance.
13. Elements that can change over time like blog or blog posts should have an additional property indicating the time of their changing. This property would belong to the version of the element, e.g. the version of a blog post (if versioning is supported). For some elements like simple links it is unlikely that they change. Therefore, they do not need such property or versioning.

6.4 Benefits and use cases

Conducting different social network analyses in BlogForever can lead to various benefits for the project and for the intended archive.

A social network analysis on the blogs in the archive can produce several measures that facilitate the evaluation of the stored Blogosphere. For example the measure “density” provides a statement on how strongly the blogs in the archive are connected. This measure can facilitate the assessment of knowledge exchange inside an organisation (e.g. the CERN) if the archive stores the blogs of this organisation. The measurements can be applied as well on subgroups like a division in the organisation.

As another example there are blogs that are not connected to other pages or not connected to other blogs [46]. In combination with a lack of new blog posts in the blog, it can be estimated that the blog is dead or only placeholder for the domain. For blogs of an organisation, that can trigger further examination if the blog subject is not relevant anymore, if it is an orphaned blog because the author has left the organisation or if there are other reasons.

Furthermore, the capabilities of social network analysis can provide additional support for users of the blog archive to identify relevant blogs or blog posts. The first of these opportunities includes the implementation of a PageRank algorithm or similar algorithms [47]. Thereby, a ranking of blogs can be generated that estimates the relevance of a blog by the links from other pages. Another opportunity is the identification of similar or related blogs or blog posts based on the analysis of co-citations. Co-citation means that it is analysed if several blogs cite the same resource. Thereby, the similarity of these blogs or blog posts is estimated just by network information and independently

of the content of the blogs. This could be advantageously if the keyword or text-based analyses do not provide sufficient results.

The available time information can be used for further opportunities of rankings. E.g. the data that are used for the analysis can be limited to a specific time span, e.g. the last month. Thereby, a ranking would be adapted to that time span. This can be very beneficial especially in the Blogosphere where changes occur very often. Examples are the identification of the most influential blogs in the last month or in a specific historical time period.

The behaviour of different actors in the blogosphere (e.g. authors and persons who comment on blog posts) can be analysed and compared by social network analysis. Thereby, similarities could be identified and heuristics could lead to the proposition that two actors are the same person. However, it has to be researched how valid the results of such heuristics could be.

Additionally, the capability to support social network analysis is highly relevant to scientists with various research interests. Blog life cycles and their impact in the Blogosphere can be explored. Models and theories about the structure and evolution of communities can be evaluated or tested. Other examples for research interests are information diffusion or communication behaviour in digital environments. This list can be continued endlessly. Therefore, the provision of the data in the archive in a way that facilitates social network analysis will make the archive a very useful resource for researchers in this area.

7 Life Cycles of Blogs

This section extends the report by highlighting current understanding of blog dynamics and user online behaviour as discussed in the relevant literature. The primary focus of this section is on collecting and collating evidence related to the life cycle of blogs and posts, and discussing it in the context of BlogForever.

Blogs are dynamic and versatile web spaces. As one of the Web 2.0 technologies, blogs are user-centred web applications that promote social connectedness, sharing, content creation and collaboration. These primary characteristics of blogs constitute a blog platform that sets them apart from the more traditional Web of inert web pages. Understanding the dynamic nature of this platform is necessary for developing suitable blog archiving solutions. More specifically, in the context of BlogForever, an insight into the life-cycle of blogs and user interaction with them (viewed under the umbrella term of blog dynamics) can help identifying the challenges and avoiding potential pitfalls when developing the anticipated archiving solutions.

This section intends to answer questions such as:

Blog Level:

- ✓ How often are blogs abandoned, deleted or migrated?

Blog Entry Level:

- ✓ How often are blog posts added, modified or deleted?
- ✓ How frequently are blog posts re/visited throughout their life-span?

The two levels (Blog and Blog Entry) are in line with the two primary views considered for modelling the Blogosphere. These are Blog Networks and Post Networks. However, since the primary difference between the two is the granularity of information the developed models are often interchangeable [48]. Hence, some of the works that discuss the dynamics of posts, for example may be suitable for describing the dynamics of blogs and vice versa.

The publications for this review have been identified using the ACM, IEEE Xplore, databases, and more general Google and Google Scholar search engines. Keywords used: `life-span blogs`; `life-cycle blogs`, `abandoned blogs`, `blog mortality`, `life-cycle of blogposts`, `life-span of blogposts`.

7.1 Dynamics on the Level of Blogs

One of the frequently appearing statistics describing the Blogosphere is the number of abandoned blogs. It is commonly accepted that abandoned blogs constitute a large share of the Blogosphere. Although, decisions on whether a blog is abandoned or not are multilateral, there is a common view that blogs are ephemeral by nature. Some of the early inquiries into the issue of mortality were conducted by Perseus. The study reports that 66.0% of blogs remain without updates for at least two months. This accounts to a large number of blogs (i.e. 2.72 million) that are being considered permanently or temporarily abandoned [27]. Other studies have shown similar results. The study of the blogs, which are powered by the LiveJournal service, suggests that only half of all blogs are being active [28]. More recently, the Danish blogging portal overskrift.dk reported [29] that more than a half (90,000) of its blogs established by late 2009 did not have any updates for the last three months.

The average life-span of abandoned blogs has been reported to be 126 days, with large number of those remaining 'one-day wonders'. These statistics highlight the scope of the ephemeral characteristics that some blogs may exhibit. It is not surprising, therefore, that blogs are sometimes considered to have a 'life-span of a fruitfly' [49].

More interestingly, there are several studies that attempt to model the life-span of blogs and discuss the factors that contribute to abandoning blogs. Gurzick and Lutters [50] noted the variation in timeframes of deserted blogs and tried to understand why some blogs live longer. Although the majority of blogs are abandoned within a few days, there are still many blogs that die after being active for more than a year (including a blog noted to be abandoned after 923 days). They (*ibid.*) conducted a set of semi-structured interviews to understand the issues. They model the life-cycle of a blog that consists of four consecutive states: [a] Non-Directed Personal Storage; [b] Growth and Aggregation; [c] A Personal Voice; and [d] Established and Interactive. They argue that the death of a blog in the initial stages is more likely due to a discovery of alternative tools and services. More established blogs, on the contrary, are more likely to be abandoned due to changes in bloggers' employment or family circumstances. This model, however, does not capture or investigate the role of the community. Some studies attempt to address this gap by discussing the issue. Miura and Yamashita [51] suggest that established communication with the readership greatly encourages bloggers to continue writing. However, they do not give clear answers on what happens to the community once a blog is abandoned.

A different approach, but an interesting one nonetheless, is proposed by Venolia [28]. She models the life of a blog mathematically by taking into account its 'decay' rate – the probability of a blog being abandoned. Although, the study is limited to a set of weak assumptions (i.e. expecting constant rates of posting and mortality of blogs), the proposed simple model deserves credit for extending the discussion on understanding the life-span of blogs. Similarly, but taking into account that interval between the posts may vary considerably from one blog to another, Kramer and Rodden [52] developed a mathematical model that is not confined to constant rates of postings. Spotting that the majority of academic works disregards the frequency of posts and define abandoned blogs based on averaged numbers, they developed a model that estimates the timing of the next post based on the previous posting habits. They analysed a large set of blogs (approximately 1.1 million) obtained from Blogger, 2% of which have been reclaimed as 'active' despite having no updates within a commonly used measure of 30 days.

The solutions developed as part of BlogForever should reflect the life-span statistics and patterns of dynamics within the Blogosphere. For instance, the methods for crawling and archiving frequently updated blogs may not be suitable for working with abandoned or deleted ones. The large numbers of abandoned blogs can justify the required efforts from the BlogForever team. Additionally, considerations should be given to differences between the new and established blogs when considering their archiving and preservation.

Life-Span Trends in the Blogosphere:

The concerns about the large number of abandoned blogs and their mortality rates may raise questions about the current and future states of the Blogosphere.

Technorati, in its annual report on the state of the Blogosphere [53], highlights the continued, though more stable growth of the Blogosphere. Yet, still a considerable number of respondents of Technorati's survey are reported to blog less due to their devotion to microblogging (30%) and social networks (28%). The New York Times [54] notes similar trends with the younger generation choosing Twitter, Tumblr and Facebook over blogging. However, while dead blogs are plentiful, the trend of increasing number of active and durable blogs persists [28].

In line with these changes, Kopytoff (*ibid.*) notes, referring to the director of the Internet and American Life Project, Lee Rainie, that blogging is not dying, but shifting with the times. Features popularised by blogging are being weaved into other kinds of services. The Technorati Report suggests that a large number of bloggers adopt social media such as Twitter and Facebook to distribute information about a published blog post and extend their audience [53]. More recently, for instance, Mashable [55] announced that Blogger, Google's blogging service that recorded a decline in numbers of unique visitors from the US by 2% in 2011 [54], will be rebranded by the end

of August 2011. It is believed that the rebranding will allow integration of Blogger with the recently started Google+ communication platform. Some users already suggest Google to follow the example of the Posterous blogging platforms and enable an easy mechanism for posting from Blogger onto other social networking sites. To which extent these changes may affect blogging experience or the evolution of blogs as a whole is impossible to tell. However, to ensure that the solutions developed by BlogForever remain adequate within a perpetually changing Web environment, their design should allow a certain level of flexibility. Some of the trends highlighted here may be considered when developing the data model, crawling and archiving tools.

7.2 Dynamics of Posts

One of the major differences between the blogs and more conventional websites is the temporal nature of the published content. Blog authors are willing to make the published content available to their readers immediately, while readers are often interested in receiving notifications about the published entry.

Composing and publishing a blog post initiates a life cycle that extends beyond the blog. Once published, a blog post almost instantaneously slips into a vast network of agents. It is being crawled, indexed, mined, scraped, republished and distributed throughout the Web. Software agents and robots pick up and ensure its distribution among potentially interested parties from fellow bloggers to corporate marketers. Frank Rose, in his article in Wired Magazine [56] vividly depicts the life cycle of a post in an interrelated diagram of agents and processes among them. The role of individual agents in the system is certainly well-understood and is primarily ubiquitous in the Blogosphere. Yet, the patterns of accessing, distributing and republishing blog posts vary from post to post and across the Blogosphere in general. There is no doubt, of course, that the life cycle of blogs is not trivial. It has already been demonstrated. What continues to be pursued by researchers is the desire to understand the role of agents and the community in that network. The following section highlights some of the identified works in this direction.

User behaviour associated with certain posts is at times considered bursty. Patterns similar to power law distribution are common for both topological and temporal characteristics of blog posts.

Early attempts to characterise the nature of blog posts is made by Gruhl and co-authors [57]. They crawled about 12 thousand blogs by using RSS and collected around 400 thousand blog postings. Their observations suggested existence of distinct patterns of user activities around certain topics. They distinguish three topic related activities: [a] Mostly Chatter – where topics are discussed continuously at relatively moderate levels; [b] Spiky Chatter – where topics react quickly and strongly to external events, but are maintained at a significant level; and [c] Just Spike – where topics turn from inactive to very active during a short period of time. Classification of users based on their posting behaviour has got the characteristics of power law distribution. They also observe that individuals with more than four recorded posts are contributing during the spike or within the middle 25% of registered posts.

It is important to highlight, however, that the average number of comments throughout the Blogosphere is relatively small. Only 10%-20% of the entire Blogosphere is attributed to comments, with an average of 0.3 comments to a post. The number of comments in influential blogs is significantly bigger compared to less-influential ones. The number of comments in top-ranked blogs can exceed the volume of posts [58]. This phenomenon continues to remain under the spotlight of researchers who are trying to understand or model the dynamics of the Blogosphere.

The work of Götz and her colleagues [59] propose and validate a model that can be used for generating a synthetic Blogosphere according to ‘what if’ scenarios and exploring blog dynamics. As a foundation for this model the authors (*ibid.*) referred to an earlier study [60] that explored 45

thousand blogs with 2.2 million blog posts. The results of this study suggested that blogs have a weekly periodicity (rather than a bursty behaviour), where the popularity of posts drops with a power law, instead of exponentially. Even more, almost every measurements followed a power law, including the size distribution of cascades (= number of involved posts). The observations suggested posts with a large number of links (i.e. 'stars') to be most popular, yet, these posts themselves remain un-cited.

Furthermore, the temporal characteristics of blog posts have been proposed to inform blog ranking mechanisms [61]. Menezes and his colleagues analysed the dynamics of blog topic discussion, initiation and participation. Arguing that the temporal characteristics are not currently used by search engine ranking algorithms, they propose an algorithm for their improvement. They suggest assigning a score to the blogs according to their precursor or laggard behaviour. They argue that the score will enable distinguishing between the blogs that have similar structural properties and, therefore, have similar ranking.

These recurring patterns suggest that the solutions designed by BlogForever should take into account the uneven distribution of posts and comments registered in blogs. Some posts may be discussed intensely during the initial period of time and attract only fewer comments later in their life-time. Hence, consideration should be given to the *timing* and the *amounts of traffic* needed for archiving the posts. Additionally, the ranking mechanisms incorporated into BlogForever could consider the temporal properties of blogs and offer an alternative ranking system to the mainstream structure based approach.

7.3 Summary

Referring to the earlier studies this section discusses the life-span of blogs and blog posts; this section outlines the results of empirical research and highlights the points for potential consideration within the boundaries of BlogForever. More specifically:

- ✓ Different approaches may be necessary for crawling and archiving the large number of abandoned blogs compared to the blogs that are actively maintained.
- ✓ Consideration should be given (e.g. prioritising some of the blogs) to differentiate between archiving established blogs versus recent (prone to dying) blogs. The existing readership and community around the blogs may also be taken into account.
- ✓ Design of the archiving solution should allow a certain level of flexibility to accommodate the frequently changing nature of the Web.
- ✓ Consideration should be given to the power law distribution pattern frequently observed with many measures (e.g. frequency of posts and comments). Variations in popularity of blogs and their generated data may require different crawling or preservation strategies.

8 Conclusions

This report propelled the process of informing the development of preservation and dissemination solutions for blogs. It started by summarising the online survey that explored the aspects of blog preservation and blogging practices in general. It continued by evaluating the use of tools and technologies within the Blogosphere. The report was further extended by including an inquiry into the theoretical and technological advances of analysing blogs and their networks.

Discussing the online survey this report outlines the prominent patterns and user behaviour in the Blogosphere and discusses user attitudes towards archiving and preservation. The survey was designed by giving careful consideration to the feedback received from all BlogForever partners. This allowed achieving one of the main targets and obtaining a considerable number of responses. The representative sample of internationally diverse participants enabled commenting on blogging practices beyond a single community or a nation.

German and Greek languages were most prevalent, following the highest number of responses completed in English. Other languages did not initially receive considerable attention. However, after putting additional effort on promoting questionnaires in other languages a greater balance has been achieved. Monitoring the preference for languages and the followed additional promotion improved (though still not significantly) the numbers of completions in Spanish, Russian and French languages. Providing conclusive explanations to the observed variation was not feasible. Further research is needed to identify the reasons behind the low numbers of completions for the less popular languages. Speculations, however, may include a lack of motivation, lower proportions of population or selected timing of the study (e.g. holiday season).

The number of responses collected from survey participants both, authors and readers, totals 900. While, this sample was not the largest possible it was fairly comprehensive – covering various areas and subjects of blogging. The data collected from the survey was rich, since many respondents provided additional and detailed feedback via provided options (i.e. ‘Other’ text boxes) for open feedback. The bloggers were found to be open to tell about their work within a blogging community and their blogging practices. The results showed that a large number of bloggers do not normally archive or preserve their work. Many of them, however, expressed willingness to deposit their blogs into archives. However, providing answers on whether archiving and preservation activities will become commonly adopted in the future remained beyond the scope of this study.

In general, interesting results were obtained about the design of blogs, blog audiences and their communication. The role of collaborators versus main authors was discussed. Possible future research directions were identified for exploring how the anticipated blog archives could address the role of collaborators in blog communities. A large number of blogs were found to use a variety of media objects, but most of them used textual data. The use of photographs and moving images was also reported to be frequent. Nearly 90% of all the blogs used self-created content, while 28.9% used remixed data. The importance of rich media, links and citations was found to be important – having direct implications for blog preservation strategies. It was shown that blog users frequently relied on monitoring blog traffic, comments, subscriptions and feeds as measures of popularity. The use of ranking methods varied widely. Motivations for maintaining blogs were primarily personal – for sharing information and promoting discussion topics.

When asked about the types of data that blog users would like to preserve in an archive, the majority expected their entire blogs, with posts and comments, to be preserved. Preserving blogs in their entirety rather than their selective sections was found to be a prevalent preference. Inbuilt archival functions for increasing readership of the archived blogs were also found to be important. Multiple ranking tools within a blog archive were also considered important. Nearly 90% of the authors interviewed never used an external service to preserve their blog and they mainly relied on their blog provider for these activities.

Furthermore, some hypotheses regarding the intention of blog authors to contribute their blogs to a central blog archive were tested. Thereby, the analysis shows that the perception of a collective benefit has a stronger influence as the perception of an individual benefit. Furthermore, the perception of a collective benefit is influenced by the expectation of better or more relationships to other people and by the identification of a blogger as part of a bigger community or a group. These findings support the proposition that blogging is not seen by the authors as an individual activity even if the most blogs have only just one author. Instead it actually seems that bloggers are aware of the Blogosphere and intend to contribute to it. Therefore, it can be assumed that it is more promising for a wide adoption to establish the archive as a service that supports the Blogosphere instead as a service that supports an individual.

Nevertheless, it has to be stated that the measurement instruments for the influence factors need improvements. Only few indicators could be used due to the length of the questionnaire. Therefore, the results had to be interpreted patiently due to the limitations. Additionally, the theory work should be strengthened to stabilise and to expand the theoretical framework.

In addition to analysing the survey, the report outlined a large-scale investigation into the technological backbone of the Blogosphere. It studied more than 200 thousand blogs and evaluated the technologies, tools and standards adopted by currently active and most popular blogs. The results showed a considerable variation in the ways the technology is being adopted. More specifically, the study revealed a large number of software platforms used for powering blogs and a variety in the versions and subsystems adopted. It identified more commonly used technologies among which, RSS and Atom feeds, CSS and JavaScript. The study identified similarities between the Blogosphere and the Web in using various image formats – highlighting the parallels from the preservation perspectives where possible. Last, but not least, a consideration was given to the use of metadata standards and the integration of social media. The study revealed less frequent and inconsistent approaches to adopting these standards and services – signposting possible challenges and pitfalls in developing the anticipated archiving solutions.

Conceptual work was conducted towards the analysis of inter-blog relationships. Thereby, it was shown that different social network analysis can be conducted on various types of blog data. Direct relationships like citations or blogroll were described as well as the use of co-citations, e.g. based on common citations of the same web page. Benefits like ranking, community detection or the support of visual exploration of the Blogosphere were identified. Next to the specific network capabilities of blogs, it was taken into account that the Blogosphere is a continuously and frequently changing network. Therefore, a special attention was pointed on the consideration of dynamics. Dynamic social network analysis of blogs can reveal more details about the evolution of communities in the Blogosphere. Furthermore, rankings can be adjusted with respect to long term or short term evolutions. Requirements for the weblog spider and the data model were formulated to prepare the analyses.

Finally, the report extended to include a discussion on the life-span of blogs and blog posts. It outlined the results of empirical research and highlighted the points for consideration within the boundaries of BlogForever. More specifically, the report pointed to the variations in the requirements for crawling and archiving the large number of abandoned and active blogs, and maintaining the deleted ones. It suggested differentiating and prioritising blogs before choosing appropriate strategies for their aggregation and archiving.

Some of the limitations of this inquiry were related to managing a multi-lingual survey with more than one questionnaire. This was shown to be challenging. It required additional time for translation of the survey as well as the received feedback. Translation of the results, however, was a prerequisite for starting data analysis. String match frequencies were performed to work with this type of data due to time limitations. Future research should allow extra time for efforts against potential data loss when translating or for using additional tools for analysis. Furthermore, a

considerably larger sample with a more even distribution of languages and participant countries should be expected from future studies. Finally, and most importantly, qualitative methods should be introduced for analysing open-ended questions and feedback.

Inquiries, similar to those outlined in this report, will be conducted on a regular basis. BlogForever survey will be implemented annually to obtain up-to date information on the dynamics, trends and changes of the Blogosphere over time. Implementation details for the regular studies are already being discussed – defining strategies for making the survey faster and more efficient and focusing on the use of archives, internal blog ranking and user satisfaction.

9 References

- [1] P. Caplan, "The Preservation of Digital Materials," *Library Technology Reports*, vol. 44, 2008.
- [2] A. Lenhart and S. Fox, *Bloggers: A portrait of the internet's new storytellers*: Pew Internet & American Life Project, 2006.
- [3] M. Pennock and R. Davis, "ArchivePress: A really simple solution to archiving blog content," 2009.
- [4] R. Davis, "Moving targets: web preservation and reference management," *Ariadne*, 2010.
- [5] L. Sheble, S. Choemprayong, and C. Hank, "Surveying bloggers' perspectives on digital preservation: Methodological issues," 2007, p. 2010.
- [6] C. Hank. (2008). *Considerations for the Preservation of Blogs*. Available: http://www.digitalpreservationeurope.eu/publications/briefs/preservartion_blogs.pdf
- [7] C. Hank, "Science and scholarship in the blogosphere: Blog characteristics, blogger behaviours and implications for digital curation.," in *Poster presented at the 5th International Digital Curation Conference*, 2009.
- [8] C. Hank, "Blogger perspectives on digital preservation: Attributes, behaviors, and preferences," 2009, p. 3.
- [9] F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: theory and results," *Sloan School of Management, Massachusetts Institute of Technology*, 1986.
- [10] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319-340, 1989.
- [11] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User acceptance of computer technology: a comparison of two theoretical models," *Management science*, pp. 982-1003, 1989.
- [12] M. M. L. Wasko and S. Faraj, "Why should I share? Examining social capital and knowledge contribution in electronic networks of practice," *MIS quarterly*, pp. 35-57, 2005.
- [13] M. Kang, Y. G. Kim, and G. W. Bock, "Identifying different antecedents for closed vs open knowledge transfer," *Journal of Information Science*, vol. 36, p. 585, 2010.
- [14] V. Venkatesh and H. Bala, "Technology acceptance model 3 and a research agenda on interventions," *Decision Sciences*, vol. 39, pp. 273-315, 2008.

- [15] H. P. Lu and K. L. Hsiao, "Understanding intention to continuously share information on weblogs," *Internet Research*, vol. 17, pp. 345-361, 2007.
- [16] U. M. Dholakia, R. P. Bagozzi, and L. K. Pearo, "A social influence model of consumer participation in network-and small-group-based virtual communities," *International Journal of Research in Marketing*, vol. 21, pp. 241-263, 2004.
- [17] C. L. Hsu and J. C. C. Lin, "Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation," *Information & Management*, vol. 45, pp. 65-74, 2008.
- [18] A. Kankanhalli, B. C. Y. Tan, and K. K. Wei, "Contributing knowledge to electronic knowledge repositories: An empirical investigation," *MIS quarterly*, pp. 113-143, 2005.
- [19] R. P. Bagozzi and U. M. Dholakia, "Intentional social action in virtual communities," *Journal of Interactive Marketing*, vol. 16, pp. 2-21, 2002.
- [20] D. Gefen, "Structural Equation Modeling and Regression: Guidelines for Research Practice," *Communications of the Association for Information Systems*, vol. 4, 2000.
- [21] J. Hulland, "Use of partial least squares (PLS) in strategic management research: a review of four recent studies," *Strategic Management Journal*, vol. 20, pp. 195-204, 1999.
- [22] W. W. Chin, Ed., *The partial least squares approach for structural equation modeling* (Modern Methods for Business Research. Lawrence Erlbaum, 1998, p.^pp. Pages.
- [23] D. Gefen, D. Straub, and M. C. Boudreau, "Structural equation modeling and regression: Guidelines for research practice," *Communications of the Association for Information Systems*, vol. 4, p. 7, 2000.
- [24] C. Fornell and D. F. Larcker, "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, vol. 18, pp. 39--50, 1981.
- [25] W. W. Chin, "The Partial Least Squares Approach to Structural Equation Modeling," presented at the Modern Methods for Business Research, Mahwah, New Jersey, 1998.
- [26] C. M. Ringle, S. Wende, and A. Will, "SmartPLS 2.0 (beta)," ed. Hamburg: Softwarepaket, 2005.
- [27] D. Li and G. Walejko, "Splogs and abandoned blogs: The perils of sampling bloggers and their blogs," *Information, Communication & Society*, vol. 11, pp. 279-296, 2008.
- [28] G. Venolia, "A matter of life or death: Modeling blog mortality," *Unveröffentlicher Forschungsbericht. Redmond. Online verfügbar: research.microsoft.com/~ginav/ljmodeling.pdf*, 2007.

- [29] S. Bogh-Andersen. (2009). *Den danske blogosfære ved udgangen af årtiet/The Danish blogosphere at the end of the decade*. Available: <http://blog.overskrift.dk/2009/12/31/den-danske-blogosf%C3%A6re-ved-udgangen-af-artiet/>
- [30] B. Hammersley, *Developing feeds with RSS and Atom*: O'Reilly, 2005.
- [31] S. Wassermann and K. Faust, *Social Network Analysis: Methods and Applications*. New York, NY, USA: Cambridge University Press, 1994.
- [32] M. D. Choudhury, H. Sundaram, A. John, and D. D. Seligmann, "Analyzing the Dynamics of Communication in Online Social Networks," B. Furht, Ed., ed New York, Dordrecht, Heidelberg, London: Springer, 2010, pp. 59-94.
- [33] M. Trier, "Towards dynamic visualization for understanding evolution of digital communication networks," *Information Systems Research*, vol. 19, pp. 335-350, 2008.
- [34] W.-S. Yang and Y.-R. Lin, "An Analysis of Network Structure and Post Content for Blog Post Recommendation." vol. 6637, J. Xu, G. Yu, S. Zhou, and R. Unland, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 276-286.
- [35] M. A. Goncalves, J. Almeida, L. Santos, A. H. F. Laender, and V. Almeida, "On Popularity in the Blogosphere," *IEEE Internet Computing*, 2010.
- [36] I. Dolinska, "Simple Blog Searching Framework Based on Social Network Analysis," pp. 611-617.
- [37] W. Xiaoguang, J. Tingting, and M. Feicheng, "Blog-supported scientific communication: An exploratory analysis based on social hyperlinks in a Chinese blog community," *Journal of Information Science*, vol. 36, pp. 690-704, 2010.
- [38] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," New York, NY, USA, pp. 207-218.
- [39] E. Adar and L. A. Adamic, "Tracking Information Epidemics in Blogspace," pp. 207-214.
- [40] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the Bursty Evolution of Blogspace," *World Wide Web*, vol. 8, pp. 159-178, 2005.
- [41] P. V. Marsden, "Recent developments in network measurement," *Models and Methods in Social Network Analysis*, vol. 8, pp. 8-30, 2005.
- [42] M. Schnegg and H. Lang, "Die Netzwerkanalyse: Eine praxisorientierte Einführung," H. Lang and M. Schnegg, Eds., ed, 2001, pp. 1-55.
- [43] T. Furukawa, Y. Matsuo, I. Ohmukai, K. Uchiyama, and M. Ishizuka, "Social Networks and Reading Behavior in the Blogosphere."

- [44] C. Marlow, "Audience, structure and authority in the weblog community," 2004.
- [45] E. Gilbert, T. Bergstrom, and K. Karahalios, "Blogs are Echo Chambers: Blogs Are Echo Chambers," 2009, pp. 1-10.
- [46] S. C. Herring, L. A. Scheidt, S. Bonus, and E. Wright, "Bridging the gap: A genre analysis of weblogs," p. p. 11, 2004.
- [47] A. Kritikopoulos, M. Sideri, and I. Varlamis, "BlogRank: ranking weblogs based on connectivity and similarity features," Pisa, Italy, 2006, pp. 8-es.
- [48] N. Agarwal and H. Liu, "Modeling and Data Mining in Blogosphere," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 1, pp. 1-109, 2009.
- [49] B. Arnold. (2009, 15.07.2011). *Caslon Analytics: Blogging*. Available: <http://www.caslon.com.au/weblogprofile1.htm>
- [50] D. Gurzick and W. G. Lutters, "From the personal to the profound: understanding the blog life cycle," 2006, pp. 827-832.
- [51] A. Miura and K. Yamashita, "Psychological and social influences on blog writing: An online survey of blog authors in Japan," *Journal of Computer Mediated Communication*, vol. 12, pp. 1452-1471, 2007.
- [52] A. D. I. Kramer and K. Rodden, "Applying a user-centered metric to identify active blogs," 2007, pp. 2525-2530.
- [53] J. Sobel. (2010, 28.04.2011). *State of the Blogosphere 2010*. Available: <http://technorati.com/blogging/article/state-of-the-blogosphere-2010-introduction/>
- [54] V. Kopytoff, "Blogs Wane as the Young Drift to Sites Like Twitter," in *New York Times*, 20.02.2011 ed. New York Edition, 2011, p. B1.
- [55] B. Parr. (2011, 24.07.2011). *EXCLUSIVE: Google To Retire Blogger & Picasa Brands in Google+ Push*. Available: <http://mashable.com/2011/07/05/google-blogger-picasa-rebranding/>
- [56] F. Rose, "The life cycle of a blog post, from servers to spiders to suits-to you," *Beskikbaar by: http://www.wired.com/special_multimedia/2008/ff_secretlife_1602*, 2007.
- [57] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," 2004, pp. 491-501.
- [58] G. Mishne and N. Glance, "Leave a reply: An analysis of weblog comments," 2006.
- [59] M. Götz, J. Leskovec, M. McGlohon, and C. Faloutsos, "Modeling blog dynamics," 2009.

- [60] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs," *Arxiv preprint arXiv:0704.2803*, 2007.
- [61] T. Menezes, C. Roth, and J. P. Cointet, "Precursors and Laggards: An Analysis of Semantic Temporal Relationships on a Blog Network," 2010, pp. 120-127.

A. Appendix A – Offline Questionnaires

A.1 Final Offline Design for Authors Questionnaire

[Note: Questions marked (*) are mandatory]

Title: BlogForever - Blog Author Survey

Page #1: Section 1. Your Personal Profile

(*)Question 1: About yourself (Tick which one describes you best)

Answers:

- 1: In full-time education
- 2: In paid employment
- 3: Self-employed
- 4: Freelancer
- 5: Home carer
- 6: Other (please specify) [User Input]

(*)Question 2: In which country do you live?

Answers:

- Choice 1: Afghanistan
- Choice 2: Albania
- Choice 3: Algeria
- Choice 4: Andorra
- Choice 5: Angola
- Choice 6: Argentina
- Choice 7: Armenia
- Choice 8: Australia
- Choice 9: Austria
- Choice 10: Azerbaijan
- Choice 11: Bahamas
- Choice 12: Bahrain
- Choice 13: Bangladesh
- Choice 14: Barbados
- Choice 15: Belarus
- Choice 16: Belgium
- Choice 17: Belize
- Choice 18: Benin
- Choice 19: Bhutan
- Choice 20: Bolivia
- Choice 21: Bosnia Herzegovina
- Choice 22: Botswana
- Choice 23: Brazil
- Choice 24: Brunei
- Choice 25: Bulgaria
- Choice 26: Burkina
- Choice 27: Burundi
- Choice 28: Cambodia
- Choice 29: Cameroon
- Choice 30: Canada
- Choice 31: Cape Verde
- Choice 32: Central African Republic
- Choice 33: Chad

Choice 34: Chile
Choice 35: China
Choice 36: Colombia
Choice 37: Comoros
Choice 38: Congo
Choice 39: Congo Democratic Republic
Choice 40: Costa Rica
Choice 41: Croatia
Choice 42: Cuba
Choice 43: Cyprus
Choice 44: Czech Republic
Choice 45: Denmark
Choice 46: Djibouti
Choice 47: Dominica
Choice 48: Dominican Republic
Choice 49: Ecuador
Choice 50: Egypt
Choice 51: El Salvador
Choice 52: Equatorial Guinea
Choice 53: Eritrea
Choice 54: Estonia
Choice 55: Ethiopia
Choice 56: Fiji
Choice 57: Finland
Choice 58: France
Choice 59: Gabon
Choice 60: Gambia
Choice 61: Georgia
Choice 62: Germany
Choice 63: Ghana
Choice 64: Greece
Choice 65: Grenada
Choice 66: Guatemala
Choice 67: Guinea
Choice 68: Guinea-Bissau
Choice 69: Guyana
Choice 70: Haiti
Choice 71: Honduras
Choice 72: Hungary
Choice 73: Iceland
Choice 74: India
Choice 75: Indonesia
Choice 76: Iran
Choice 77: Iraq
Choice 78: Ireland
Choice 79: Israel
Choice 80: Italy
Choice 81: Jamaica
Choice 82: Japan
Choice 83: Jordan
Choice 84: Kazakhstan
Choice 85: Kenya
Choice 86: Kiribati
Choice 87: Korea North
Choice 88: Korea South

Choice 89: Kosovo
Choice 90: Kuwait
Choice 91: Kyrgyzstan
Choice 92: Latvia
Choice 93: Lebanon
Choice 94: Lesotho
Choice 95: Liberia
Choice 96: Libya
Choice 97: Liechtenstein
Choice 98: Lithuania
Choice 99: Luxembourg
Choice 100: Madagascar
Choice 101: Malawi
Choice 102: Malaysia
Choice 103: Maldives
Choice 104: Mali
Choice 105: Malta
Choice 106: Marshall Islands
Choice 107: Mauritania
Choice 108: Mauritius
Choice 109: Mexico
Choice 110: Micronesia
Choice 111: Moldova
Choice 112: Monaco
Choice 113: Mongolia
Choice 114: Montenegro
Choice 115: Morocco
Choice 116: Mozambique
Choice 117: Myanmar
Choice 118: Namibia
Choice 119: Nauru
Choice 120: Nepal
Choice 121: Netherlands
Choice 122: New Zealand
Choice 123: Nicaragua
Choice 124: Niger
Choice 125: Nigeria
Choice 126: Norway
Choice 127: Oman
Choice 128: Pakistan
Choice 129: Palau
Choice 130: Panama
Choice 131: Papua New Guinea
Choice 132: Paraguay
Choice 133: Peru
Choice 134: Philippines
Choice 135: Poland
Choice 136: Portugal
Choice 137: Qatar
Choice 138: Romania
Choice 139: Russian Federation
Choice 140: Rwanda
Choice 141: Samoa
Choice 142: San Marino
Choice 143: Sao Tome & Principe

- Choice 144: Saudi Arabia
- Choice 145: Senegal
- Choice 146: Serbia
- Choice 147: Seychelles
- Choice 148: Sierra Leone
- Choice 149: Singapore
- Choice 150: Slovakia
- Choice 151: Slovenia
- Choice 152: Solomon Islands
- Choice 153: Somalia
- Choice 154: South Africa
- Choice 155: Spain
- Choice 156: Sri Lanka
- Choice 157: Sudan
- Choice 158: Suriname
- Choice 159: Swaziland
- Choice 160: Sweden
- Choice 161: Switzerland
- Choice 162: Syria
- Choice 163: Taiwan
- Choice 164: Tajikistan
- Choice 165: Tanzania
- Choice 166: Thailand
- Choice 167: Togo
- Choice 168: Tonga
- Choice 169: Trinidad & Tobago
- Choice 170: Tunisia
- Choice 171: Turkey
- Choice 172: Turkmenistan
- Choice 173: Tuvalu
- Choice 174: Uganda
- Choice 175: Ukraine
- Choice 176: United Arab Emirates
- Choice 177: United Kingdom
- Choice 178: United States
- Choice 179: Uruguay
- Choice 180: Uzbekistan
- Choice 181: Vatican City
- Choice 182: Venezuela
- Choice 183: Vietnam
- Choice 184: Yemen
- Choice 185: Zambia
- Choice 186: Zimbabwe

(*)Question 3: What is your nationality?

Answers:

- Choice 1: Afghan
- Choice 2: Albanian
- Choice 3: Algerian
- Choice 4: US American
- Choice 5: Andorran
- Choice 6: Angolan
- Choice 7: Argentinean
- Choice 8: Armenian
- Choice 9: Australian

Choice 10: Austrian
Choice 11: Azerbaijani
Choice 12: Bahamian
Choice 13: Bahraini
Choice 14: Bangladeshi
Choice 15: Barbadian
Choice 16: Belarusian
Choice 17: Belgian
Choice 18: Belizean
Choice 19: Beninese
Choice 20: Bhutanese
Choice 21: Bolivian
Choice 22: Bosnian
Choice 23: Brazilian
Choice 24: British (UK)
Choice 25: Bruneian
Choice 26: Bulgarian
Choice 27: Burkinabe
Choice 28: Burmese
Choice 29: Burundian
Choice 30: Cambodian
Choice 31: Cameroonian
Choice 32: Canadian
Choice 33: Cape Verdean
Choice 34: Central African
Choice 35: Chadian
Choice 36: Chilean
Choice 37: Chinese
Choice 38: Colombian
Choice 39: Comoran
Choice 40: Congolese
Choice 41: Costa Rican
Choice 42: Croatian
Choice 43: Cuban
Choice 44: Cypriot
Choice 45: Czech
Choice 46: Danish
Choice 47: Djibouti
Choice 48: Dominican
Choice 49: Dutch
Choice 50: Ecuadorean
Choice 51: Egyptian
Choice 52: Emirati
Choice 53: Equatorial Guinean
Choice 54: Eritrean
Choice 55: Estonian
Choice 56: Ethiopian
Choice 57: Fijian
Choice 58: Filipino
Choice 59: Finnish
Choice 60: French
Choice 61: Gabonese
Choice 62: Gambian
Choice 63: Georgian
Choice 64: German

Choice 65: Ghanaian
Choice 66: Greek
Choice 67: Grenadian
Choice 68: Guatemalan
Choice 69: Guinea-Bissauan
Choice 70: Guinean
Choice 71: Guyanese
Choice 72: Haitian
Choice 73: Honduran
Choice 74: Hungarian
Choice 75: I-Kiribati
Choice 76: Icelander
Choice 77: Indian
Choice 78: Indonesian
Choice 79: Iranian
Choice 80: Iraqi
Choice 81: Irish
Choice 82: Israeli
Choice 83: Italian
Choice 84: Jamaican
Choice 85: Japanese
Choice 86: Jordanian
Choice 87: Kazakhstani
Choice 88: Kenyan
Choice 89: Kosovar
Choice 90: Kuwaiti
Choice 91: Kyrgyz
Choice 92: Latvian
Choice 93: Lebanese
Choice 94: Liberian
Choice 95: Libyan
Choice 96: Liechtensteiner
Choice 97: Lithuanian
Choice 98: Luxembourger
Choice 99: Malagasy
Choice 100: Malawian
Choice 101: Malaysian
Choice 102: Maldivan
Choice 103: Malian
Choice 104: Maltese
Choice 105: Marshallese
Choice 106: Mauritanian
Choice 107: Mauritian
Choice 108: Mexican
Choice 109: Micronesian
Choice 110: Moldovan
Choice 111: Monacan
Choice 112: Mongolian
Choice 113: Montenegrin
Choice 114: Moroccan
Choice 115: Mosotho
Choice 116: Motswana
Choice 117: Mozambican
Choice 118: Namibian
Choice 119: Nauruan

Choice 120: Nepalese
Choice 121: New Zealander
Choice 122: Nicaraguan
Choice 123: Nigerian
Choice 124: Nigerien
Choice 125: North Korean
Choice 126: Norwegian
Choice 127: Omani
Choice 128: Pakistani
Choice 129: Palauan
Choice 130: Panamanian
Choice 131: Papua New Guinean
Choice 132: Paraguayan
Choice 133: Peruvian
Choice 134: Polish
Choice 135: Portuguese
Choice 136: Qatari
Choice 137: Romanian
Choice 138: Russian
Choice 139: Rwandan
Choice 140: Salvadoran
Choice 141: Samoan
Choice 142: San Marinese
Choice 143: Sao Tomean
Choice 144: Saudi
Choice 145: Senegalese
Choice 146: Serbian
Choice 147: Seychellois
Choice 148: Sierra Leonean
Choice 149: Singaporean
Choice 150: Slovakian
Choice 151: Slovenian
Choice 152: Solomon Islander
Choice 153: Somali
Choice 154: South African
Choice 155: South Korean
Choice 156: Spanish
Choice 157: Sri Lankan
Choice 158: Sudanese
Choice 159: Surinamer
Choice 160: Swazi
Choice 161: Swedish
Choice 162: Swiss
Choice 163: Syrian
Choice 164: Taiwanese
Choice 165: Tajik
Choice 166: Tanzanian
Choice 167: Thai
Choice 168: Togolese
Choice 169: Tongan
Choice 170: Trinidadian or Tobagonian
Choice 171: Tunisian
Choice 172: Turkish
Choice 173: Turkmen
Choice 174: Tuvaluan

Choice 175: Ugandan
Choice 176: Ukrainian
Choice 177: Uruguayan
Choice 178: Uzbekistani
Choice 179: Venezuelan
Choice 180: Vietnamese
Choice 181: Yemenite
Choice 182: Zambian
Choice 183: Zimbabwean

(*)Question 4: What is your gender?

Answers:

- 1: Female
- 2: Male
- 3: Rather not say

(*)Question 5: Select your Age Group

Answers:

- 1: Under 18
- 2: 18 - 24
- 3: 25 - 34
- 4: 35 - 44
- 5: 45 - 49
- 6: 50 - 54
- 7: 55 - 64
- 8: Over 65

Page #2: Your Current or Main Blog

(*)Question 6: Blog URL

Answers:

- 1: [User Input]

Question 7: Short description of content

Answers:

- 1: [User Input]

(*)Question 8: Are you the only author on this blog. or are there multiple authors?

Answers:

- 1: I'm the only author
- 2: Multiple authors

(*)Question 9: What language is your blog written in?

Answers:

- Choice 1: Bulgarian
- Choice 2: Czech
- Choice 3: Danish
- Choice 4: Dutch
- Choice 5: English
- Choice 6: Estonian
- Choice 7: Finnish
- Choice 8: French
- Choice 9: German
- Choice 10: Greek
- Choice 11: Hungarian

- Choice 12: Irish
- Choice 13: Italian
- Choice 14: Latvian
- Choice 15: Lithuanian
- Choice 16: Maltese
- Choice 17: Polish
- Choice 18: Portuguese
- Choice 19: Romanian
- Choice 20: Russian
- Choice 21: Slovak
- Choice 22: Slovene
- Choice 23: Spanish
- Choice 24: Swedish
- Choice 25: Other

Question 10: Do you use a blog provider?

Answers:

- 1: Yes
- 2: No

Question 11: If yes. which one?

Answers:

- 1: [User Input]

Question 12: Why do you use that platform?

Answers:

- 1: [User Input]

(*Question 13: What media does your blog contain? (Tick all that apply)

Answers (Allow multiple options):

- 1: Text
- 2: Photographs
- 3: Images other than photographs (e.g. drawings. graphs. clipart...)
- 4: Audio
- 5: Moving images (e.g. video. animation...)
- 6: Other (Please specify) [User Input]

(*Question 14: How is the media in your blog created? (Tick all that apply)

Answers (Allow multiple options):

- 1: Self-created
- 2: Reused/Remixed from original
- 3: From other blogs
- 4: From specific websites
- 5: Search engine results
- 6: Other (Please specify) [User Input]

Question 15: How important is the availability of rich media (i.e.audio. video. images) for conveying your message?

Answers:

Row 1: How important is the availability of rich media (i.e.audio. video. images) for conveying your message?

Column 1: Very unimportant

Column 2: Unimportant

Column 3: Neutral

Column 4: Important

Column 5: Very important

Page #3: Section 2. Your Activity

Question 16: Why did you start blogging?

Answers:

1: [User Input]

(*)Question 17: What is your motivation for maintaining a blog? (Tick all that apply)

Answers (Allow multiple options):

1: Professional

2: Personal

3: Entertainment

4: Commercial

5: Organise / promote / support an activity

6: Discussion of topics

7: Promote teaching and learning

8: Information sharing

9: Keep a record of activities or events

10: Manage a conference

11: Manage a project

12: Create an online presence

13: To target markets or communities

14: Mostly for myself

15: Mostly for my audience

16: Other (please specify) [User Input]

(*)Question 18: How frequently do you perform these activities?

Answers:

Row 1: Authoring and editing activities (adding new blog post. writing. uploading content. embedding content)

Row 2: Mashup activities (making a mix of various content from other sources. quotes from other sources. reusing content from other authors)

Row 3: Design activities (changing look and feel of blog. new skin or theme. modifying style or appearance)

Row 4: Dialogue activities (community response. responses to comments. moderating. search engine optimisation. adding links to other sites)

Column 1: Never / Not at all

Column 2: Rarely

Column 3: Monthly

Column 4: Weekly

Column 5: Once a day

Column 6: Several times a day

Question 19: How do you assess the following statements?

Answers:

Row 1: Writing a blog enhances personal reputation.

Row 2: I want to stay in touch with other Internet users.

Row 3: I earn respect from others by writing a blog.

Row 4: Blogging creates new relationships with other bloggers.

Row 5: I write blogs to get some feedback (advice or criticism) about my blogs.

Row 6: Blogging strengthens ties with other bloggers.

Row 7: I write blogs to learn other people's views on my blogs.

Row 8: Generally. writing a good blog enhances the relevance of blogosphere.

Row 9: I think I am competent to create a good and well-received blog.
Row 10: I think my personal identity overlaps with the other bloggers' identities.
Row 11: Writing a blog advances the overall blogosphere.
Row 12: I feel part of the group of bloggers.
Row 13: I feel confident in my ability to create blogs that are interesting for others.
Row 14: The blogosphere is a growing and persistent body of knowledge for Internet users.
Row 15: The blogosphere has long-term value for Internet users.
Column 1: Strongly disagree
Column 2: Disagree
Column 3: Neutral
Column 4: Agree
Column 5: Strongly agree

(*)Question 20: Are you expected or required by your organisation to blog?

Answers:

- 1: Yes
- 2: No
- 3: Don't Know

Question 21: What kind of organisation?

Answers (Allow multiple options):

- 1: Commercial
- 2: Academic / Research
- 3: Media
- 4: Public sector
- 5: Other (Please specify) [User Input]

Page #4: Section 3. Your Users

(*)Question 22: Which group do you feel best represents the main audience for your blog? (Tick all that apply)

Answers (Allow multiple options):

- 1: Colleagues and Professional Peers
- 2: Family and Friends
- 3: Myself
- 4: General Public
- 5: Students
- 6: None of the above (please describe) [User Input]

(*)Question 23: How do you track your blog traffic?

Answers (Allow multiple options):

- 1: Via a logging system
- 2: Third party tracking system (e.g. Google analytics)
- 3: Specific users' comments
- 4: Citations
- 5: Other (please specify) [User Input]

(*)Question 24: How many hits do you receive on your blog on a typical day (approx.)?

Answers:

- 1: Fewer than 10
- 2: Between 10 and 49
- 3: 50 or more
- 4: Don't know

(*)Question 25: Do you use any of the following ranking analysis tools for your blog success?

Answers (Allow multiple options):

- 1: BlogPulse
- 2: Technorati Rank
- 3: Wordpress stats
- 4: Onsite Audience Growth
- 5: Offsite Audience Growth via Feeds
- 6: Subscribers
- 7: Comments
- 8: Citations
- 9: Other (please specify) [User Input]
- 10: None of the above

(*)Question 26: How do users interact with your blog and its content?

Answers (Allow multiple options):

- 1: Comments
- 2: 'Like' buttons
- 3: 'Post' buttons or bookmarking (e.g: Facebook. Twitter...)
- 4: 'Send to' option
- 5: Subscribe or follow
- 6: Adding a user as a 'Friend'
- 7: Linking to blogpost
- 8: Adding to blogroll
- 9: Acknowledging trackbacks
- 10: Feeds (e.g. RSS. Atom...)
- 11: Contributing posts
- 12: Don't know
- 13: Other (please specify) [User Input]

(*)Question 27: Does your blog include a list of links like a blogroll?

Answers:

- 1: Yes
- 2: No

(*)Question 28: If yes. how many links do you have?

Answers:

- 1: Fewer than 10
- 2: Between 10 and 49
- 3: 50 or more
- 4: Don't know

(*)Question 29: How many other blogs link to your site?

Answers:

- 1: Fewer than 10
- 2: Between 10 and 49
- 3: 50 or more
- 4: Don't know

Page #5: Section 4. A Central Blog Archive

Question 30: What reasons can you imagine for using a central blog archive or blog preservation system?

Answers:

- 1: [User Input]

(*)Question 31: In a blog archive. what channels would be most useful to you to increase readership of your blog? (Tick all that apply)

Answers (Allow multiple options):

- 1: A blog community
- 2: Sharing and rating
- 3: Blog news portals
- 4: Blog marketing tools
- 5: Not interested in increasing my readership
- 6: Not interested in sharing beyond my blog
- 7: I don't know

(*)Question 32: How do you assess the following statements?

Answers:

Row 1: If there is/would be a central blog archive. I predict that I would contribute my blog.

Row 2: I intend to contribute my blog to a central blog archive.

Row 3: I plan to contribute my blog in a central archive.

Column 1: Strongly disagree

Column 2: Disagree

Column 3: Neutral

Column 4: Agree

Column 5: Strongly agree

Page #6: Section 5. Preservation

(*)Question 33: Do you have backups for your blog?

Answers:

- 1: Yes
- 2: No

Question 34: Approximately. how often do you make backups?

Answers:

- 1: Monthly
- 2: Weekly
- 3: Daily

(*)Question 35: Do you retain a copy of each backup?

Answers:

- 1: Yes
- 2: No
- 3: Only selected versions (Please specify) [User Input]

(*)Question 36: Have you ever self-archived your blog?

Answers:

- 1: Yes
- 2: No

(*)Question 37: If so. how?

Answers:

- 1: Exporting to server
- 2: Backing up database
- 3: Service provider does it automatically
- 4: Other (please specify) [User Input]

(*)Question 38: Have you ever used an external service to preserve your blog?

Answers:

- 1: Web-archiving service
- 2: Archiving service
- 3: Institutional repository
- 4: Digital archive
- 5: Never used one
- 6: Other (specify) [User Input]

Question 39: What elements of the blog are the most important for you to be preserved?

Answers:

- Row 1: Whole blog
- Row 2: Specific sections
- Row 3: Posts
- Row 4: Comments
- Row 5: Commenting systems
- Row 6: Feeds
- Row 7: Internal links
- Row 8: Blogroll
- Row 9: Sponsors of the blog
- Row 10: Design
- Row 11: Visual layout
- Row 12: Topic tags
- Row 13: Author tags
- Row 14: Date tags
- Row 15: Categories
- Row 16: Metadata
- Row 17: Attachments
- Row 18: External links
- Row 19: Calendar
- Row 20: Slide show
- Row 21: Search box
- Row 22: Embedded Widgets
- Row 23: Registered users
- Row 24: Communities of users
- Row 25: Contributing Authors' Profiles
- Column 1: Very unimportant
- Column 2: Unimportant
- Column 3: Neutral
- Column 4: Important
- Column 5: Very important

Page #7: Section 6. Your Perceptions

Question 40: What does your blog mean to you?

Answers:

- 1: Very important part of my life
- 2: An enjoyable hobby
- 3: Don't spend a lot of time on it
- 4: Other (please specify) [User Input]

(*)Question 41: What impact would the loss of your blog have on you?

Answers:

- Row 1: What impact would the loss of your blog have on you?
- Column 1: Very unimportant
- Column 2: Unimportant
- Column 3: Neutral

Column 4: Important

Column 5: Very important

(*)Question 42: I would feel pleased if my blog was selected for permanent preservation in a trusted archive.

Answers:

Row 1: I would feel pleased if my blog was selected for permanent preservation in a trusted archive?

Column 1: Strongly disagree

Column 2: Disagree

Column 3: Neutral

Column 4: Agree

Column 5: Strongly agree

Question 43: Is there anything else about your blog content or activities you would like to tell us that has not been covered above?

Answers:

1: [User Input]

A.2 Final Offline Design for Readers Questionnaire

[Note: Questions marked (*) are mandatory]

Title: BlogForever - Blog Reader Survey

Page #1: Section 1. Your personal profile

(*)Question 1: About yourself (Tick which one describes you best)

Answers:

1: In full-time education

2: In paid employment

3: Self-employed

4: Freelancer

5: Home carer

6: Other (please specify) [User Input]

(*)Question 2: In which country do you live?

Answers:

Choice 1: Afghanistan

Choice 2: Albania

Choice 3: Algeria

Choice 4: Andorra

Choice 5: Angola

Choice 6: Argentina

Choice 7: Armenia

Choice 8: Australia

Choice 9: Austria

Choice 10: Azerbaijan

Choice 11: Bahamas

Choice 12: Bahrain

Choice 13: Bangladesh

Choice 14: Barbados

Choice 15: Belarus

Choice 16: Belgium

Choice 17: Belize
Choice 18: Benin
Choice 19: Bhutan
Choice 20: Bolivia
Choice 21: Bosnia Herzegovina
Choice 22: Botswana
Choice 23: Brazil
Choice 24: Brunei
Choice 25: Bulgaria
Choice 26: Burkina
Choice 27: Burundi
Choice 28: Cambodia
Choice 29: Cameroon
Choice 30: Canada
Choice 31: Cape Verde
Choice 32: Central African Republic
Choice 33: Chad
Choice 34: Chile
Choice 35: China
Choice 36: Colombia
Choice 37: Comoros
Choice 38: Congo
Choice 39: Congo Democratic Republic
Choice 40: Costa Rica
Choice 41: Croatia
Choice 42: Cuba
Choice 43: Cyprus
Choice 44: Czech Republic
Choice 45: Denmark
Choice 46: Djibouti
Choice 47: Dominica
Choice 48: Dominican Republic
Choice 49: Ecuador
Choice 50: Egypt
Choice 51: El Salvador
Choice 52: Equatorial Guinea
Choice 53: Eritrea
Choice 54: Estonia
Choice 55: Ethiopia
Choice 56: Fiji
Choice 57: Finland
Choice 58: France
Choice 59: Gabon
Choice 60: Gambia
Choice 61: Georgia
Choice 62: Germany
Choice 63: Ghana
Choice 64: Greece
Choice 65: Grenada
Choice 66: Guatemala
Choice 67: Guinea
Choice 68: Guinea-Bissau
Choice 69: Guyana
Choice 70: Haiti
Choice 71: Honduras

Choice 72: Hungary
Choice 73: Iceland
Choice 74: India
Choice 75: Indonesia
Choice 76: Iran
Choice 77: Iraq
Choice 78: Ireland
Choice 79: Israel
Choice 80: Italy
Choice 81: Jamaica
Choice 82: Japan
Choice 83: Jordan
Choice 84: Kazakhstan
Choice 85: Kenya
Choice 86: Kiribati
Choice 87: Korea North
Choice 88: Korea South
Choice 89: Kosovo
Choice 90: Kuwait
Choice 91: Kyrgyzstan
Choice 92: Latvia
Choice 93: Lebanon
Choice 94: Lesotho
Choice 95: Liberia
Choice 96: Libya
Choice 97: Liechtenstein
Choice 98: Lithuania
Choice 99: Luxembourg
Choice 100: Madagascar
Choice 101: Malawi
Choice 102: Malaysia
Choice 103: Maldives
Choice 104: Mali
Choice 105: Malta
Choice 106: Marshall Islands
Choice 107: Mauritania
Choice 108: Mauritius
Choice 109: Mexico
Choice 110: Micronesia
Choice 111: Moldova
Choice 112: Monaco
Choice 113: Mongolia
Choice 114: Montenegro
Choice 115: Morocco
Choice 116: Mozambique
Choice 117: Myanmar
Choice 118: Namibia
Choice 119: Nauru
Choice 120: Nepal
Choice 121: Netherlands
Choice 122: New Zealand
Choice 123: Nicaragua
Choice 124: Niger
Choice 125: Nigeria
Choice 126: Norway

Choice 127: Oman
Choice 128: Pakistan
Choice 129: Palau
Choice 130: Panama
Choice 131: Papua New Guinea
Choice 132: Paraguay
Choice 133: Peru
Choice 134: Philippines
Choice 135: Poland
Choice 136: Portugal
Choice 137: Qatar
Choice 138: Romania
Choice 139: Russian Federation
Choice 140: Rwanda
Choice 141: Samoa
Choice 142: San Marino
Choice 143: Sao Tome & Principe
Choice 144: Saudi Arabia
Choice 145: Senegal
Choice 146: Serbia
Choice 147: Seychelles
Choice 148: Sierra Leone
Choice 149: Singapore
Choice 150: Slovakia
Choice 151: Slovenia
Choice 152: Solomon Islands
Choice 153: Somalia
Choice 154: South Africa
Choice 155: Spain
Choice 156: Sri Lanka
Choice 157: Sudan
Choice 158: Suriname
Choice 159: Swaziland
Choice 160: Sweden
Choice 161: Switzerland
Choice 162: Syria
Choice 163: Taiwan
Choice 164: Tajikistan
Choice 165: Tanzania
Choice 166: Thailand
Choice 167: Togo
Choice 168: Tonga
Choice 169: Trinidad & Tobago
Choice 170: Tunisia
Choice 171: Turkey
Choice 172: Turkmenistan
Choice 173: Tuvalu
Choice 174: Uganda
Choice 175: Ukraine
Choice 176: United Arab Emirates
Choice 177: United Kingdom
Choice 178: United States
Choice 179: Uruguay
Choice 180: Uzbekistan
Choice 181: Vatican City

Choice 182: Venezuela
Choice 183: Vietnam
Choice 184: Yemen
Choice 185: Zambia
Choice 186: Zimbabwe

(*)Question 3: What is your nationality?

Answers:

Choice 1: Afghan
Choice 2: Albanian
Choice 3: Algerian
Choice 4: US American
Choice 5: Andorran
Choice 6: Angolan
Choice 7: Argentinean
Choice 8: Armenian
Choice 9: Australian
Choice 10: Austrian
Choice 11: Azerbaijani
Choice 12: Bahamian
Choice 13: Bahraini
Choice 14: Bangladeshi
Choice 15: Barbadian
Choice 16: Belarusian
Choice 17: Belgian
Choice 18: Belizean
Choice 19: Beninese
Choice 20: Bhutanese
Choice 21: Bolivian
Choice 22: Bosnian
Choice 23: Brazilian
Choice 24: British (UK)
Choice 25: Bruneian
Choice 26: Bulgarian
Choice 27: Burkinabe
Choice 28: Burmese
Choice 29: Burundian
Choice 30: Cambodian
Choice 31: Cameroonian
Choice 32: Canadian
Choice 33: Cape Verdean
Choice 34: Central African
Choice 35: Chadian
Choice 36: Chilean
Choice 37: Chinese
Choice 38: Colombian
Choice 39: Comoran
Choice 40: Congolese
Choice 41: Costa Rican
Choice 42: Croatian
Choice 43: Cuban
Choice 44: Cypriot
Choice 45: Czech
Choice 46: Danish
Choice 47: Djibouti

Choice 48: Dominican
Choice 49: Dutch
Choice 50: Ecuadorean
Choice 51: Egyptian
Choice 52: Emirati
Choice 53: Equatorial Guinean
Choice 54: Eritrean
Choice 55: Estonian
Choice 56: Ethiopian
Choice 57: Fijian
Choice 58: Filipino
Choice 59: Finnish
Choice 60: French
Choice 61: Gabonese
Choice 62: Gambian
Choice 63: Georgian
Choice 64: German
Choice 65: Ghanaian
Choice 66: Greek
Choice 67: Grenadian
Choice 68: Guatemalan
Choice 69: Guinea-Bissauan
Choice 70: Guinean
Choice 71: Guyanese
Choice 72: Haitian
Choice 73: Honduran
Choice 74: Hungarian
Choice 75: I-Kiribati
Choice 76: Icelander
Choice 77: Indian
Choice 78: Indonesian
Choice 79: Iranian
Choice 80: Iraqi
Choice 81: Irish
Choice 82: Israeli
Choice 83: Italian
Choice 84: Jamaican
Choice 85: Japanese
Choice 86: Jordanian
Choice 87: Kazakhstani
Choice 88: Kenyan
Choice 89: Kosovar
Choice 90: Kuwaiti
Choice 91: Kyrgyz
Choice 92: Latvian
Choice 93: Lebanese
Choice 94: Liberian
Choice 95: Libyan
Choice 96: Liechtensteiner
Choice 97: Lithuanian
Choice 98: Luxembourger
Choice 99: Malagasy
Choice 100: Malawian
Choice 101: Malaysian
Choice 102: Maldivan

Choice 103: Malian
Choice 104: Maltese
Choice 105: Marshallese
Choice 106: Mauritanian
Choice 107: Mauritian
Choice 108: Mexican
Choice 109: Micronesian
Choice 110: Moldovan
Choice 111: Monacan
Choice 112: Mongolian
Choice 113: Montenegrin
Choice 114: Moroccan
Choice 115: Mosotho
Choice 116: Motswana
Choice 117: Mozambican
Choice 118: Namibian
Choice 119: Nauruan
Choice 120: Nepalese
Choice 121: New Zealander
Choice 122: Nicaraguan
Choice 123: Nigerian
Choice 124: Nigerien
Choice 125: North Korean
Choice 126: Norwegian
Choice 127: Omani
Choice 128: Pakistani
Choice 129: Palauan
Choice 130: Panamanian
Choice 131: Papua New Guinean
Choice 132: Paraguayan
Choice 133: Peruvian
Choice 134: Polish
Choice 135: Portuguese
Choice 136: Qatari
Choice 137: Romanian
Choice 138: Russian
Choice 139: Rwandan
Choice 140: Salvadoran
Choice 141: Samoan
Choice 142: San Marinese
Choice 143: Sao Tomean
Choice 144: Saudi
Choice 145: Senegalese
Choice 146: Serbian
Choice 147: Seychellois
Choice 148: Sierra Leonean
Choice 149: Singaporean
Choice 150: Slovakian
Choice 151: Slovenian
Choice 152: Solomon Islander
Choice 153: Somali
Choice 154: South African
Choice 155: South Korean
Choice 156: Spanish
Choice 157: Sri Lankan

Choice 158: Sudanese
Choice 159: Surinamer
Choice 160: Swazi
Choice 161: Swedish
Choice 162: Swiss
Choice 163: Syrian
Choice 164: Taiwanese
Choice 165: Tajik
Choice 166: Tanzanian
Choice 167: Thai
Choice 168: Togolese
Choice 169: Tongan
Choice 170: Trinidadian or Tobagonian
Choice 171: Tunisian
Choice 172: Turkish
Choice 173: Turkmen
Choice 174: Tuvaluan
Choice 175: Ugandan
Choice 176: Ukrainian
Choice 177: Uruguayan
Choice 178: Uzbekistani
Choice 179: Venezuelan
Choice 180: Vietnamese
Choice 181: Yemenite
Choice 182: Zambian
Choice 183: Zimbabwean

(*)Question 4: What is your gender?

Answers:

- 1: Female
- 2: Male
- 3: Rather not say

(*)Question 5: Select your Age Group

Answers:

- 1: Under 18
- 2: 18 - 24
- 3: 25 - 34
- 4: 35 - 44
- 5: 45 - 49
- 6: 50 - 54
- 7: 55 - 64
- 8: Over 65

Page #2: Section 2. Reading blogs

(*)Question 6: How often do you read other people's blogs?

Answers:

Row 1: How often do you read other people's blogs?

- Column 1: Never
Column 2: Rarely
Column 3: Sometimes
Column 4: Often
Column 5: Always

(*Question 7: What languages are used in the blogs you read?

Answers (Allow multiple options):

- 1: Bulgarian
- 2: Czech
- 3: Danish
- 4: Dutch
- 5: English
- 6: Estonian
- 7: Finnish
- 8: French
- 9: German
- 10: Greek
- 11: Hungarian
- 12: Irish
- 13: Italian
- 14: Latvian
- 15: Lithuanian
- 16: Maltese
- 17: Polish
- 18: Portuguese
- 19: Romanian
- 20: Russian
- 21: Slovak
- 22: Slovene
- 23: Spanish
- 24: Swedish
- 25: Other (please specify) [User Input]

(*Question 8: What would you consider the main reasons you read blogs? (Tick all that apply)

Answers (Allow multiple options):

- 1: To scan news
- 2: To scan comments on news
- 3: To scan general content
- 4: For professional research
- 5: As a personal interest
- 6: To be up to date with blog trends
- 7: Other reasons (Please specify) [User Input]

(*Question 9: How important are communication and networking possibilities (with co-authors, blog owners or other readers) on the blogs you read?

Answers:

Row 1: How important are communication and networking possibilities (with co-authors, blog owners or other readers) on the blogs you read?

Column 1: Very unimportant

Column 2: Unimportant

Column 3: Neutral

Column 4: Important

Column 5: Very important

(*Question 10: How often do you leave a comment(s) on the blogs you read?

Answers:

Row 1: How often do you leave a comment(s) on the blogs you read?

Column 1: Never

Column 2: Rarely

Column 3: Sometimes

Column 4: Often
Column 5: Always

(*)Question 11: How important for you is the graphical layout or visual appearance of a blog?

Answers:

Row 1: How important for you is the graphical layout or visual appearance of a blog?

Column 1: Very unimportant

Column 2: Unimportant

Column 3: Neutral

Column 4: Important

Column 5: Very important

(*)Question 12: How do you assess the following statements?

Answers:

Row 1: A simple search interface with only few options facilitates me best to find relevant blogs.

Row 2: A complex search interface with many options and different views facilitates me best to find relevant blogs.

Column 1: Very unimportant

Column 2: Unimportant

Column 3: Neutral

Column 4: Important

Column 5: Very important

(*)Question 13: How often do you try to catch up with the top ranked blogs?

Answers:

Row 1: How often do you try to catch up with the top ranked blogs?

Column 1: Never

Column 2: Rarely

Column 3: Sometimes

Column 4: Often

Column 5: Always

(*)Question 14: How important is for you to know how credible the blogs' sources of information are?

Answers:

Row 1: How important is for you to know how credible the blogs' sources of information are?

Column 1: Very unimportant

Column 2: Unimportant

Column 3: Neutral

Column 4: Important

Column 5: Very important

(*)Question 15: Learning is an important aspect of my blog searching and reading

Answers:

Row 1: Learning is an important aspect of my blog searching and reading

Column 1: Strongly disagree

Column 2: Disagree

Column 3: Neutral

Column 4: Agree

Column 5: Strongly agree

(*)Question 16: I am often interested in how multiple blogs relate to each other

Answers:

Row 1: I am often interested in how multiple blogs relate to each other

Column 1: Strongly disagree

Column 2: Disagree
Column 3: Neutral
Column 4: Agree
Column 5: Strongly agree

(*)Question 17: What is your preferred method of accessing a blog post?

Answers (Allow multiple options):

- 1: Keyword search
- 2: Tag or tag cloud
- 3: Category
- 4: Searching for recent updates by date
- 5: Searching by author
- 6: Other (please specify) [User Input]

Question 18: How often do you access static web pages (e.g. About. Contacts) in a blog?

Answers:

Row 1: How often do you access static web pages (e.g. About. Contacts) in a blog?

- Column 1: Never
Column 2: Rarely
Column 3: Sometimes
Column 4: Often
Column 5: Always

(*)Question 19: How often do you use a blog's widgets (e.g. News feeds. Flickr. RSS. DIGG. YouTube. Twitter. Skype)?

Answers:

Row 1: How often do you use a blog's widgets (e.g. News feeds. Flickr. RSS. DIGG. YouTube. Twitter. Skype)?

- Column 1: Never
Column 2: Rarely
Column 3: Sometimes
Column 4: Often
Column 5: Always

Page #3: Section 3. Your Designated Blog

(*)Question 20: Blog URL

Answers:

- 1: [User Input]

Question 21: Short description of content

Answers:

- 1: [User Input]

(*)Question 22: What language is your designated blog written in?

Answers:

- Choice 1: Bulgarian
- Choice 2: Czech
- Choice 3: Danish
- Choice 4: Dutch
- Choice 5: English
- Choice 6: Estonian
- Choice 7: Finnish
- Choice 8: French
- Choice 9: German

- Choice 10: Greek
- Choice 11: Hungarian
- Choice 12: Irish
- Choice 13: Italian
- Choice 14: Latvian
- Choice 15: Lithuanian
- Choice 16: Maltese
- Choice 17: Polish
- Choice 18: Portuguese
- Choice 19: Romanian
- Choice 20: Russian
- Choice 21: Slovak
- Choice 22: Slovene
- Choice 23: Spanish
- Choice 24: Swedish
- Choice 25: Other

(*)Question 23: What subject(s) would you use to classify your designated blog? (Tick all that apply)

Answers (Allow multiple options):

- 1: Arts
- 2: Business
- 3: Computers
- 4: Culture
- 5: Economics
- 6: Education
- 7: Entertainment
- 8: Family
- 9: Games
- 10: Health
- 11: Home
- 12: Information and communication
- 13: Life and personal experience (diary. journal)
- 14: News and current affairs
- 15: Politics
- 16: Recreation
- 17: Reference
- 18: Regional
- 19: Religion
- 20: Science
- 21: Shopping
- 22: Society
- 23: Sports
- 24: Other (please specify) [User Input]

(*)Question 24: What attracted you to your designated blog? (Tick all that apply)

Answers (Allow multiple options):

- 1: The content
- 2: The design
- 3: The layout
- 4: Other (please specify) [User Input]

(*)Question 25: How did you find the designated blog?

Answers:

- 1: Recommended by another site

- 2: Recommended by a friend
- 3: Random encounter as result of a search
- 4: Don't know
- 5: Other (please specify) [User Input]

(*Question 26: Which group do you feel best represents the main audience for your designated blog? (Tick all that apply)

Answers (Allow multiple options):

- 1: Colleagues and Professional Peers
- 2: Family and Friends
- 3: General Public
- 4: Students
- 5: None of the above (please describe) [User Input]
- 6: Don't know

(*Question 27: Is your designated blog connected to an organisation?

Answers:

- 1: Yes
- 2: No
- 3: Don't know

Question 28: What kind of organisation?

Answers (Allow multiple options):

- 1: Commercial
- 2: Academic / Research
- 3: Media
- 4: Public sector
- 5: Other (Specify) [User Input]

(*Question 29: Are you expected or required to read this blog?

Answers:

- 1: Yes
- 2: No
- 3: Don't know

Page #4: Section 4. A Central Blog Archive

(*Question 30: What reasons can you imagine for using a blog preservation system?

Answers:

- 1: [User Input]

(*Question 31: Assuming BlogForever were able to provide a service across a large volume of blogs. what would you most like from a service like that? (Tick all that apply)

Answers (Allow multiple options):

- 1: Access
- 2: Search engine
- 3: Information services
- 4: Clustering
- 5: Feeds
- 6: Blogs sorted according to topics or clusters
- 7: Blog content rated in some way
- 8: Other (please specify) [User Input]

(*Question 32: When searching for blog content. what are you looking for? (Tick all that apply)

Answers (Allow multiple options):

- 1: Consumer blogs
- 2: Expert opinions from industry experts or journalists
- 3: Twitter feeds
- 4: Postings relevant to my interests
- 5: Discussion areas and forums
- 6: To participate in discussions
- 7: Review sites
- 8: Communities
- 9: Overviews and aggregations of content
- 10: Other (please specify) [User Input]

(*)Question 33: How do you assess the following statements?

Answers:

Row 1: I would prefer a comprehensive search over a simple search interface to retrieve relevant pages.

Row 2: When I search I go for the topic and the author is usually of secondary importance.

Row 3: I want to identify and locate top ranked blogs.

Row 4: I spend a lot of time determining whether or not a source of information is credible.

Row 5: It is important for me to trace the links of blogs to find out more

Column 1: Strongly disagree

Column 2: Disagree

Column 3: Neutral

Column 4: Agree

Column 5: Strongly agree

(*)Question 34: I think that for comprehensive searching of blogs in a central blog archive:

Answers:

Row 1: A sorted list would be an effective way to explore a domain (like Google results)

Row 2: A visual and interactive map would be an effective way to explore a domain

Column 1: Strongly disagree

Column 2: Disagree

Column 3: Neutral

Column 4: Agree

Column 5: Strongly agree

Page #5: Section 5.Your Perceptions

(*)Question 35: What does your designated blog mean to you?

Answers:

1: Very important part of my life

2: An enjoyable hobby

3: Don't spend a lot of time reading it

4: Other (please specify) [User Input]

Question 36: Is there anything else about your blog reading or other activities you would like to tell us that has not been covered above?

Answers:

1: [User Input]

B. Appendix B - BlogForever Survey IProbe Screenshots

Appendix Figure 1- Authors Survey IProbe Screenshot Section 1

Section 1. Your Personal Profile Progress: 14%

1. About yourself (Tick which one describes you best)

In full-time education

In paid employment

Self-employed

Freelancer

Home carer

Other (please specify)

2. In which country do you live?

select:

3. What is your nationality?

select:

4. What is your gender?

Female

Male

Rather not say

5. Select your Age Group

Under 18

18 - 24

25 - 34

35 - 44

45 - 49

50 - 54

55 - 64

Over 65

[next page](#) Powered by iProbe™

Appendix Figure 2 - Authors Survey IProbe Screenshot


19. How do you assess the following statements?

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Writing a blog enhances personal reputation.	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I want to stay in touch with other internet users.	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I earn respect from others by writing a blog.	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Bloggng creates new relationships with other bloggers.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
I write blogs to get some feedback (advice or criticism) about my blogs.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Bloggng strengthens ties with other bloggers.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I write blogs to learn other people's views on my blogs.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generally, writing a good blog enhances the relevance of blogosphere.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think I am competent to create a good and well-received blog.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think my personal identity overlaps with the other bloggers' identities.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Writing a blog advances the overall blogosphere.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel part of the group of bloggers.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel confident in my ability to create blogs that are interesting for others.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The blogosphere is a growing and persistent body of knowledge for internet users.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The blogosphere has long-term value for internet users.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix Figure 3 - Authors Survey IProbe Screenshot Section 4

BlogForever - Blog Author Survey

Thanks for agreeing to participate. This survey should take no more than 15 minutes to complete.



Progress: 71%

Section 4. A Central Blog Archive

30. What reasons can you imagine for using a central blog archive or blog preservation system?

31. In a blog archive, what channels would be most useful to you to increase readership of your blog? (Tick all that apply)



- A blog community
- Sharing and rating
- Blog news portals
- Blog marketing tools
- Not interested in increasing my readership
- Not interested in sharing beyond my blog
- I don't know

32. How do you assess the following statements?

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
If there is/would be a central blog archive, I predict that I would contribute my blog.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I intend to contribute my blog to a central blog archive.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I plan to contribute my blog in a central archive.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

← previous page next page →

Powered by IProbe™

BlogForever (ICT No. 269963) is funded by the European Commission under Framework Programme 7 (FP7) ICT Programme

Appendix Figure 4 - Readers Survey IProbe Screenshot Section 2

Section 2. Reading blogs

6. How often do you read other people's blogs?

	Never	Rarely	Sometimes	Often	Always
How often do you read other people's blogs?	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

7. What languages are used in the blogs you read?

- Bulgarian
- Czech
- Danish
- Dutch
- English
- Estonian
- Finnish
- French
- German
- Greek
- Hungarian
- Irish
- Italian
- Latvian
- Lithuanian
- Maltese
- Polish
- Portuguese
- Romanian
- Russian
- Slovak
- Slovene
- Spanish
- Swedish
- Other (please specify)

8. What would you consider the main reasons you read blogs? (Tick all that apply)

- To scan news
- To scan comments on news
- To scan general content
- For professional research
- As a personal interest
- To be up to date with blog trends

Appendix Figure 5 - Readers Survey IProbe Screenshot Section 3

Section 3. Your Designated Blog

Please tell us about your favourite blog! Pick ONE BLOG that you regard as your favourite and tell us about it. All your replies should apply to that one blog.

20. Blog URL

21. Short description of content

22. What language is your designated blog written in?


select:

23. What subject(s) would you use to classify your designated blog? (Tick all that apply)

- Arts
- Business
- Computers
- Culture
- Economics
- Education
- Entertainment
- Family
- Games
- Health
- Home
- Information and communication
- Life and personal experience (diary, journal)
- News and current affairs
- Politics
- Recreation
- Reference
- Regional
- Religion
- Science
- Shopping
- Society

Appendix Figure 6 - Readers Survey IProbe Screenshot Section 5

BlogForever - Blog Reader Survey
Thanks for agreeing to participate. This survey should not take more than 10 minutes to complete.



Progress: 100%

Section 5. Your Perceptions

35. What does your designated blog mean to you?

Very important part of my life



An enjoyable hobby

Don't spend a lot of time reading it

Other (please specify)

36. Is there anything else about your blog reading or other activities you would like to tell us that has not been covered above?

[← previous page](#) [Finish survey →](#) Powered by iProbe™



BlogForever (ICT No. 269963) is funded by the European Commission under Framework Programme 7 (FP7) ICT Programme

C. Appendix C - BlogForever Survey Promotion Screenshots

Appendix Figure 7 - Mokono Promotion of BlogForever Survey at blog.co.uk



Appendix Figure 8 - Mokono Promotion of BlogForever Survey at blog.de



D. Appendix D – Readers Survey Data Summary

Appendix Table 1 - Readers Responses by Survey Language

SurveyLanguage	Responses	%
Greek	229	53.5
English	134	31.3
German	49	11.4
Spanish	10	2.3
French	3	0.7
Russian	2	0.5
Blank	1	0.2
<i>Total</i>	<i>428</i>	<i>100</i>

Appendix Table 2 - Readers by Country of Residence

CountryResidence	Responses	%
Greece	231	54
United Kingdom	55	12.9
Germany	49	11.4
United States	20	4.7
Turkey	15	3.5
Armenia	11	2.6
Spain	10	2.3
France	7	1.6
Australia	4	0.9
Canada	4	0.9
India	2	0.5
Ireland	2	0.5
Cyprus	2	0.5
Sweden	2	0.5
Other	14	3.3
<i>Total</i>	<i>428</i>	<i>100</i>

Appendix Table 3 - Readers by Nationality

Nationality	Responses	%
Greek	230	53.7
German	47	11
British (UK)	47	11
US American	25	5.8
Turkish	15	3.5
Armenian	12	2.8
Spanish	10	2.3

Dutch	4	0.9
Irish	3	0.7
Australian	3	0.7
French	3	0.7
Portuguese	3	0.7
Cypriot	3	0.7
Canadian	3	0.7
Indian	2	0.5
Ukrainian	2	0.5
Other	16	3.7
<i>Total</i>	<i>428</i>	<i>100</i>

Appendix Table 4 - Blog Languages Usage

BlogLanguage	Responses	%
English	341	79.7
Greek	229	53.5
German	76	17.8
French	51	11.9
Spanish	28	6.5
Russian	19	4.4
Italian	17	4
TURKISH	13	3
Dutch	8	1.9
ARMENIAN	8	1.9
Blanks	6	1.4
Irish	2	0.5
Polish	2	0.5
PERSIAN	2	0.5

Appendix Table 5 - Frequency of Blogs Readership

ReadershipFrequency	Responses	%
Often	168	39.3%
Sometimes	130	30.4%
Always	71	16.6%
Rarely	53	12.4%
Never	6	1.4%
<i>Total</i>	<i>428</i>	<i>100.0%</i>

Appendix Table 6 - Main Reasons for Reading Blogs

BlogReadershipMainReason	Responses	%
As a personal interest	295	68.9
To scan news	245	57.2
ΠΑΡΑΠΟΜΠΗ ΑΠΟ ΚΑΠΟΙΑ ΙΣΤΟΣΕΛΙΔΑ Η Ε-MAIL	200	46.7
ΝΑ ΜΑΘΑΙΝΩ ΤΗ ΓΝΩΜΗ ΤΟΥΣ ΓΥΡΩ ΑΠΟ ΠΡΟΙΟΝΤΑ ΠΟΥ ΘΕΛΩ ΝΑ ΑΓΟΡΑΣΩ ΚΑΙ ΙΣΩΣ ΤΑ ΕΧΟΥΝ ΑΠΟΚΤΗΣΕΙ ΠΡΩΤΟΙ ΚΑΙ ΓΡΑΦΟΥΝ ΣΧΟΛΙΑ ΠΕΡΙ ΑΥΤΩΝ	200	46.7
ΑΠΟΨΕΙΣ ΓΙΑ ΚΟΙΝΑ ΕΝΔΙΑΦΕΡΟΝΤΑ	183	42.8
ΨΥΧΑΓΩΓΙΑ	49	11.4
Other	29	6.8
Blanks	9	2.1

Appendix Table 7 - Preferred Methods of Accessing a Blog Post

AccessBlogPost	Responses	%
Keyword search	261	61
Category	170	39.7
Searching for recent updates by date	125	29.2
Tag or tag cloud	93	21.7
Searching by author	92	21.5
RSS	32	7.5
Other	10	2.3
Blanks	8	1.9
Twitter	2	0.5

Appendix Table 8 - Access Static Web Pages by Frequency

Access Static Web Pages	Responses	%
Sometimes	137	32
Rarely	132	30.8
Often	93	21.7
Never	39	9.1
Always	18	4.2
Blank	9	2.1
<i>Total</i>	<i>428</i>	<i>100</i>

Appendix Table 9 - Blogs Widgets Usage by Frequency

BlogsWidgetsUsage	Responses	%
Often	116	27.1
Sometimes	109	25.5
Rarely	90	21
Never	54	12.6
Always	53	12.4
Blank	6	1.4
<i>Total</i>	<i>428</i>	<i>100</i>

Appendix Table 10 - Comments by Frequency

LeaveComment	Responses	%
Rarely	192	44.9
Sometimes	118	27.6
Never	79	18.5
Often	32	7.5
Blanks	6	1.4
Always	1	0.2
<i>Total</i>	<i>428</i>	<i>100</i>

Appendix Table 11 – Details of Blog Content Search

BlogContentSearch	Responses	%
Postings relevant to my interests	312	72.9
Discussion areas and forums	154	36
Expert opinions from industry experts or journalists	153	35.7
Overviews and aggregations of content	111	25.9
Review sites	103	24.1
Communities	92	21.5
Consumer blogs	74	17.3
To participate in discussions	55	12.9
Twitter feeds	32	7.5%
Other	16	3.7
Blanks	3	0.7

Appendix Table 12 – Importance of Communication and Networking

CommunicationNetworking	Responses	%
Important	156	36.4
Neutral	146	34.1
Very important	56	13.1
Unimportant	42	9.8
Very unimportant	22	5.1
Blank	6	1.4
<i>Total</i>	<i>428</i>	<i>100</i>

Appendix Table 13 – Top Ranked Blogs Reading by Frequency

TopRankedBlogsCatchUp	Responses	%
Rarely	126	29.4
Always	14	3.3
Never	115	26.9
Sometimes	124	29
Often	43	10
Blanks	6	1.4
<i>Total</i>	<i>428</i>	<i>100</i>

E. Appendix E - Facebook, Delicious, and Twitter

Excursus Facebook “Like”-Button

Facebook provides the functionality of the “Like”-Button. Therefore, the owner or admin of the web page has to have a Facebook account and has to provide the “Like”-Button on her/his web page. When the “Like”-Button is integrated into the web page, a Facebook fan page will be connected to this website. Now, another person can press the like button and thereby, a “like”-connection is created to the fan page. Therefore, the person who likes the page has to have a Facebook account as well because the connection is created between the Facebook account of the person who like the page and the Facebook fan page of the web page.

The number of how many people like the page is public available. It is shown on the fan page in Facebook and can be shown on the web page. Additionally, the owner of the fan page has access to some demographic statistics about the persons who like the page. The view that shows who likes the page is restricted to the Facebook friends of a person. Thus, a person can only see the like-expression of these people that are connected to the person in Facebook.

Further, Facebook provide the “Like Box”³⁹. The box can be integrated into a web page and shows how many people like this web page as well as some people who like the page. The Facebook user that are shown is restricted to the Facebook friends of the person who integrated the Like Box into the web page (the owner respectively admin).

It can be summarised that the people who like a page can only be discovered as long as they are friends of the web page owner or friends of the seeker.

Excursus delicious bookmarks

Delicious⁴⁰ is a social bookmarking tool that allows the user to store bookmarks on an online platform, to manage the bookmarks with tagging, and to share the bookmarks with other users. Thereby, tag clouds are emerging on a huge collection of bookmarks and interesting links can be found through exploring a folksonomy.

If a web page is bookmarked, then it can be assumed that the person who bookmarks

- ✓ has read the page,
- ✓ has assessed the page as valuable to bookmark, and
- ✓ has expressed the her/his subjective view in the form of tags.

The information about who has bookmarked a web page can normally not be found on the web page itself. Therefore, a search must be performed in delicious with the link of the web page. This will show users with their user names in delicious and the tags that they have added to the bookmark of this web page. The list has not to be complete because it is as well possible to add private bookmarks in delicious. These will not be shown to other users.

Excursus Twitter (re)tweets

Twitter⁴¹ is a microblogging service. Microblogging allows a user to publish short messages (twitter restricts the message length to 140 characters) in his or her microblogging stream in reverse chronological order. A twitter message is called a “tweet”. Users can follow the stream of other

³⁹ <https://developers.facebook.com/docs/reference/plugins/like-box/>

⁴⁰ <http://www.delicious.com/>

⁴¹ <http://twitter.com/>

users and thereby, receive the message of interesting people. Common syntactical patterns have emerged in these messages which allow

- ✓ to reference other users by adding an “@” in front of a user name (e.g. @blogforever),
- ✓ to cite other messages (it is called retweet) by adding “RT” and the user reference in front of the message (e.g. RT @blogforever Cited message comes here), and
- ✓ to add tags to a message by adding “#” in front of the tag (e.g. #Blogforever).

A tweet or a retweet that references to a blog or blog post can be seen as a relationship between the twitter account and the blog or between the twitter user and the blog author. Such relationships can be tracked by searching the link to the blog or blog post in twitter. However, tweets are available on twitter only for seven days. Therefore, it would be necessary to perform a search and capture the relationships at least every seven days. Some services are already available that provide a regular archiving⁴² for specific search queries or a social media search⁴³ that includes some older tweets as well.

In summarising it can be stated that the tracking of twitter mentions of blogs and blog posts in the BlogForever archive is theoretical possible but it would require a permanent observation of twitter regarding new appearances of links to a blog or blog post. Therefore, it would require performing a search for every possible link destination at least every seven days. The resulting effort with a growing number of blogs in the archive will not be reasonable, especially because it can be assumed that only a small number of search queries (mainly for newer blog posts) will return results. Therefore, heuristics are needed that indicate the probability of a successful twitter search.

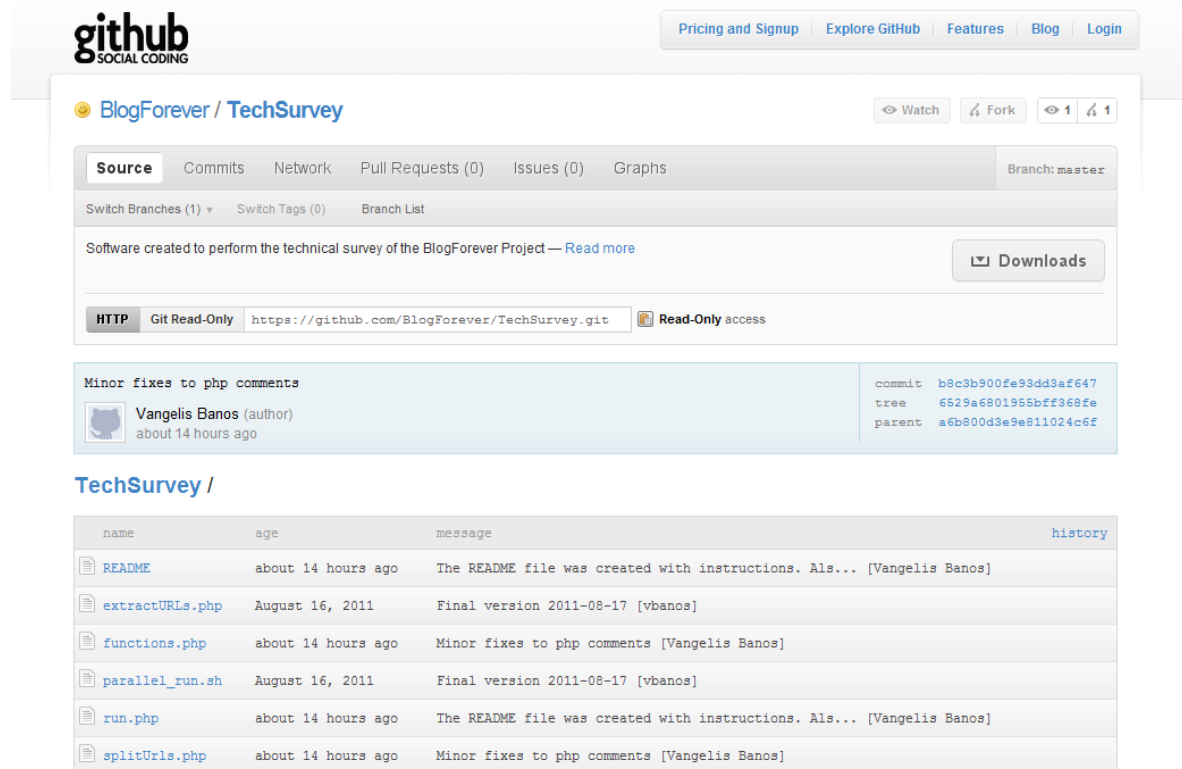
⁴² <http://archivist.visitmix.com/>

⁴³ <http://topsy.com/> and <http://socialmention.com/>

F. Appendix F – Blog Technology Survey Software

The blog technology survey software source code & documentation are publicly available at <https://github.com/BlogForever/TechSurvey>

Appendix Figure 9: github.com/BlogForever application source code repository



The screenshot shows the GitHub repository page for `BlogForever / TechSurvey`. The repository is on the `master` branch. A commit by Vangelis Banos is highlighted, with the message "Minor fixes to php comments". The commit history table below lists the following files and their commit messages:

name	age	message	history
README	about 14 hours ago	The README file was created with instructions. Als... [Vangelis Banos]	
extractURLs.php	August 16, 2011	Final version 2011-08-17 [vbanos]	
functions.php	about 14 hours ago	Minor fixes to php comments [Vangelis Banos]	
parallel_run.sh	August 16, 2011	Final version 2011-08-17 [vbanos]	
run.php	about 14 hours ago	The README file was created with instructions. Als... [Vangelis Banos]	
splitURLs.php	about 14 hours ago	Minor fixes to php comments [Vangelis Banos]	

Application README file:

BlogForever Project - <http://blogforever.eu>

```
@author Vangelis Banos vbanos [at] gmail [dot] com
```

TechSurvey software

One of the main goals of BlogForever 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 software was implemented using a combination of PHP and Bash scripting.

The BlogForever Tech Survey software is capable of analysing a large number of blogs in parallel and detect the use of specific technologies.

Requirements:

- Linux operating system with at least 8GB RAM for running in parallel mode
- Bash
- xargs command line tool (for parallel execution of multiple processes)
- PHP 5.3 or latest
- PHP CURL extension

Usage:

1. Download the code using:

```
git clone git://github.com/BlogForever/TechSurvey.git
```

2. Run in single process mode

```
php run.php url-list-input.txt result-file.csv
```

url-list-input.txt is a text file containing a list of URLs to be analysed

result-file.csv is a CSV formatted text files containing the results of the analysis

3. Run in parallel process mode

```
./parallel_run.sh big-url-list-input.txt output-folder/file
```

big-url-list-input.txt is a text file containing a list of URLs to be analysed

output-folder is the folder which will store the results

```
output-folder/file1.csv
```

```
output-folder/file2.csv
```

```
...
```

```
output-folder/file100.csv
```

When executing the program in parallel mode, 100 processes will initiate simultaneously, increasing considerably the performance of the application. Warning! At least 8 GB of RAM are required in this mode.

After the execution of the program, the output-folder will contain 100 CSV files

containing the results. (one file for each process).