Blog Preservation: Current Challenges and a New Paradigm

Vangelis Banos[1(✉)], Nikos Baltas[2], and Yannis Manolopoulos[1]

[1]Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece
vbanos@gmail.com, manolopo@csd.auth.gr
[2]Department of Computing, Imperial College, London, SW7 2AZ, UK
nb605@doc.ic.ac.uk

Abstract. Blogging is yet another popular and prominent application in the era of Web 2.0. According to recent measurements often considered as conservative, as of now worldwide there are more than 152 million blogs with content spanning over every aspect of life and science, necessitating long term blog preservation and knowledge management. In this work, we present a range of issues that arise when facing the task of blog preservation. We argue that current web archiving solutions are not able to capture the dynamic and continuously evolving nature of blogs, their network and social structure as well as the exchange of concepts and ideas that they foster. Furthermore, we provide directions and objectives that could be reached to realize robust digital preservation, management and dissemination facilities for blogs. Finally, we introduce the BlogForever EC funded project, its main motivation and findings towards widening the scope of blog preservation.

Keywords: Blogs · Blog preservation · Web archiving

1 Introduction

Blogs are types of websites regularly updated and intended for general public consumption. Their structure is defined as a series of pages in reverse chronological order. Blogs have become fairly established as an online communication and web publishing tool. The set of all blogs and their interconnections is referred to as the Blogosphere [1]. The importance and the influence of the blogosphere are constantly rising and have become the subject of modeling and research [13]. For instance, a 2006 study of the importance of blogs in politics, and for US Congress in particular, concluded that blogs play "an increasingly powerful role in framing ideas and issues for legislators and leaders directly" [27]. Blogpulse, a blog trend discovery service, identified 126 million blogs in 2009 and over 152 million blogs in 2010; while Tumblr, a relatively new blogging service, reports that they host over 33 million blogs [28]; statistics which undoubtedly prove the wide acceptance and dynamic evolution of weblogs. Moreover, they underline the importance of this novel electronic publication medium and exert its significance as part of contemporary culture.

J. Cordeiro et al. (Eds.): ICEIS 2012, LNBIP 141, pp. 29–51, 2013.
DOI: 10.1007/978-3-642-40654-6_3,
© Springer-Verlag Berlin Heidelberg 2013

But despite the fast growth of blogosphere, there is still no effective solution for ubiquitous semantic weblog archiving, digital preservation, management and dissemination. Current weblog archiving tools and methods are ineffective and inconsistent, disregarding volatility and content correlation issues, while preservation methods for weblog data have not yet been duly considered. Indeed, existing Web Archiving solutions provide no means of preserving constantly changing content, like the content of weblogs.

Furthermore, to the best of our knowledge, no current Web Archiving effort has ever developed a strategy for effective preservation and meaningful usage of Social Media. The inter-dependence aspect of those media, demonstrated by weblogs featuring shared or adversary opinions, as well as weblogs that support, imitate or revolve around more central ones, is profoundly neglected. Two reasons are mainly responsible for this: firstly, the occasional harvesting of web resources and, secondly, their treatment as unstructured pages, leave little margin for capturing the aforementioned communication perspective of weblogs.

In this work, we present the new challenges that have to be met when facing blog preservation, including information integrity, data management, content dynamics and network analysis. Furthermore, we present the BlogForever EC funded project, its main motivation, objectives and findings towards widening the scope of blog preservation.

2 Related Work

Web preservation is defined as 'the capture, management and preservation of websites and web resources'. Web preservation must be a start-to finish activity, and it should encompass the entire lifecycle of the web resource [5]. The topic of web preservation was initially addressed in a large scale by the Internet Archive in 1996 [29]. Subsequently, many national memory institutions understood the value of web preservation and developed special activities towards this goal. Table 1 displays all major national and international web archiving projects which are part of the International Internet Preservation Consortium (IIPC).

As digital preservation techniques progress and awareness is raised on the matter, there is a continuous trend towards preserving more complex objects [7]. In the scope of web preservation, this means evolving from the preservation of simple web resources (i.e. html documents, images, audio and video files) towards preserving more complex web entities such as complete websites, dynamic web portals and social media. This trend is persisting with more social media content being considered for preservation. For instance, the Library of Congress has started preserving all Twitter content since 2010 [8].

The European Commission has identified the growing need to keep digital resources available and usable over time. To support research in the field, the FP7 ICT Research Programme 2009–2010 and 2011–2012 included specific provisions for digital preservation and web preservation under objectives ICT-2009.4.1: Digital Libraries and Digital Preservation and Objective ICT-2011.4.3 Digital Preservation [10]. A number of EC funded projects pursuing advanced web preservation are listed below:

Table 1. International internet preservation consortium members.

Organization	Years	Access Methods
Bibliotheca Alexandrina's Internet Archive, Egupt	1996	URL Search
Bibliothèque nationale de France - Archives de l'Internet	2002	URL Search, Keyword Search, Full-Text Search, Topical Collections
Government of Canada Web Archive	2005	URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search
Croatian Web Archive (HAW)	2004	URL Search, Keyword Search
The Internet Archive (International)	1996	URL Search, Topical Collections
The Icelandic Web Archive	2004	URL Search
Finnish Web Archive	2006	URL Search, Full-Text Search
Kulturarw3 - The Web Archive of the National Library of Sweden	1997	URL Search
Library of Congress Web Archive, USA	2000	URL Search, Alphabetic Browsing, Subject Browsing, Topical Collections
Royal Library and the State and University Library, Aarhus, Denmark	2005	URL Search
Nettarkivet Norge (WebArchive Norway)	2001	Keyword Search
New Zealand Web Archive	1999	URL Search, Keyword Search, Alphabetic Browsing, Subject Browsing
National Library of Korea	2005	URL Search, Keyword Search, Subject Browsing
PANDORA Australia's Web Archive	1996	URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search, Subject Browsing
Digital Heritage of Catalonia (PADICAT)	2005	URL Search, Keyword Search, Alphabetic Browsing, Subject Browsing, Topical Collections
Webarchive of Slovenia	2007	URL Search, Alphabetic Browsing
The UK Government Web Archive	1997	URL Search, Alphabetic Browsing
UK Web Archive	2005	URL Search, Alphabetic Browsing, Full-Text Search, Subject Browsing, Topical Collections
Web Archiving Project, Japan	2002	Keyword Search, Full-Text Search, Topical Collections
Web archive of The Netherlands	2007	URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search, Topical Collections
WebArchiv - archive of the Czech web	2007	URL Search, Subject Browsing
Web Archive Switzerland	2008	URL Search, Keyword Search, Full-Text Search, Subject Browsing, Topical Collections
Webarchive Austria	2008	URL Search, Topical Collections

- **LiWA** (Living Web Archives) aimed to extend the current state of the art and develop the next generation of Web content capture, preservation, analysis, and enrichment services to improve fidelity, coherence, and interpretability of web archives [20].

- **ARCOMEM** (From Collect-All Archives to Community Memories) is about memory institutions like archives, museums and libraries in the age of the social web. Social media are becoming more and more pervasive in all areas of life. ARCOMEM's aim is to help to transform archives into collective memories that are more tightly integrated with their community of users and to exploit Web 2.0 and the wisdom of crowds to make web archiving a more selective and meaning-based process [11].
- **SCAPE** (Scalable Preservation Environments) project will address scalability of large-scale digital preservation workflows. The project aims to enhance the state of the art in three concrete and significant ways. First, it will develop infrastructure and tools for scalable preservation actions; second, it will provide a framework for automated, quality-assured preservation workflows; and, third, it will integrate these components with a policy-based preservation planning and watch system. These concrete project results will be driven by requirements from, and in turn validated within, three large-scale test beds from diverse application areas: web content, digital repositories, and research data sets [11].
- **LAWA** (Longitudinal Analytics of Web Archive Data) project will build an Internet-based experimental test bed for large-scale data analytics. Its focus is on developing a sustainable infra-structure, scalable methods, and easily usable software tools for aggregating, querying, and analyzing heterogeneous data at Internet scale. Particular emphasis will be given to longitudinal data analysis along the time dimension for Web data that has been crawled over extended time periods [18].
- **PATHS** (Personalised access to cultural heritage spaces) project goals are to provide innovative user-driven personalised access to cultural heritage collections and to support user's knowledge discovery and exploration. The project will create a system that acts as an interactive personalised tour guide through existing digital library collections by extending the state of the art in user-driven information access and by applying language technologies to analyse and enrich online content, with links to related items and background information.

The topic of web preservation in general and blog preservation in particular has been also addressed by a number of private startup companies throughout the world. Pagefreezer [24] is claiming to support web archiving and social media archiving. Another popular service is VaultPress [30], which provides security, backup and support for Wordpress blogs. Figure 1 displays a timeline of important web preservation projects and initiatives since 1995.

Despite the presented activities in the field of web preservation, we argue that there is still no effective solution for ubiquitous semantic weblog archiving, digital preservation and dissemination. Current web archiving tools and methods are not designed for the semantic web era and are ineffective and inconsistent, disregarding volatility and content correlation issues. Additionally, preservation methods for weblog have not yet been duly considered.

In the following section, we present a number of issues that arise when dealing with blog preservation and the current solutions to these issues.

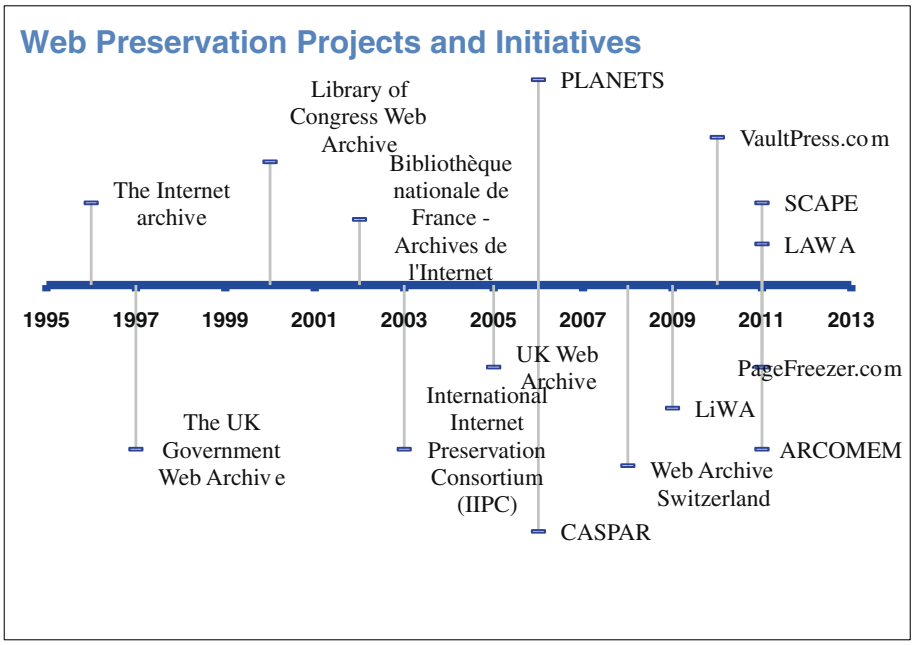

Fig. 1. Timeline of important web preservation projects and initiatives.

3 Blog Preservation Issues, Overview and Consideration

Blog preservation activities can be divided into three main groups: (a) content aggregation, (b) archiving, and (c) management. Here, we present the blocking issues for each one of these groups of activities.

3.1 Blog Content Aggregation

Existing web archiving solutions provide no means of aggregating and preserving constantly changing content, like the content of weblogs. The following two broad technical approaches are usually followed.

Firstly, there are initiatives that select and replicate web sites on an individual basis, an approach exemplified by the Web Capture Initiative [31] and by some projects developed by national archives. A second group of initiatives use crawler programs to automatically gather and store large sets of publicly available web sites. The Internet Archive follows this approach by taking periodic snapshots of the entire web since 1996. Other crawler-based initiatives have focused on national domains, e.g. the pioneering Swedish Royal Library's Kulturarw project, which is now discontinued [4]. A complete list of national web archiving projects is shown on Table 1. These initiatives are usually complemented by deposit approaches, where owners or administrators of websites choose to deposit the web content they are publishing to the repository.

Regardless of the target content, current initiatives employ general purpose web harvesting to collect their material. This approach, although easy to implement, results in problematic and incorrect web archives, especially for highly dynamic site types such as weblogs and wikis, which exhibit special characteristics. More precisely, current weblog content aggregation and digital preservation suffers from the following issues:

- **Web Content Aggregation Scheduling** is a common issue among web archiving projects, since all of them perform this task on regular intervals without considering web site updates. On the other hand, weblogs are extremely volatile and tend to be updated several times during the day, with new content from editors as well as user comments and discussions. As a result, a large amount of weblog content is not preserved, resulting in subsequent information loss and inconsistent web archives. For instance, Internet Archive's latest web preservation project, Archive-it [3], which uses the latest Heritrix web crawler [12], enables harvesting material from the Web as frequently as every 24 h, once per week, once per month, once per quarter, annually or just once. The same also applies to the popular [32], which is used by the Web Capture Initiative of the Library of Congress, the National Library of New Zealand and numerous other institutions worldwide.
- **Web Content Aggregation Performance** is also a major issue for current web preservation initiatives. Most projects use brute-force methods to crawl through a domain or a set of URLs, retrieving each page, extracting links and visiting each one of them recursively, according to a set of predefined rules. This process is performed periodically without taking into account whether the target site has been modified since the previous content aggregation or which components of the weblog have actually been updated. Unlike regular web sites, weblogs support smart content aggregation by notifying third party applications in the event of content submission or modification. Two technologies supporting this are Blog Ping [33] and PubSubHub [6]. Nevertheless, they are not utilized by current web preservation initiatives, resulting in a waste of computing resources.
- **Quality Assurance Checking** is performed manually or in a semi-automatic way for most web preservation projects. The widely used Web Curator Tool requires the administrator to perform a "Quality Review Task" while the PANDORA Archive's quality checking process [21] also requires human supervision.

3.2 Blog Content Preservation

Preservation refers to the long-term storage and access of digital or digitised content. Existing generic web archiving solutions suffer from several preservation-related shortcomings that render them as poor choices for weblog archiving. These relate to both the long-term storage of a weblog as well as to the access and usage of the preserved content.

1. Current web preservation initiatives are geared towards aggregating and preserving **files** and not **information entities**. For instance, the Internet Archive aggregates web pages and stores them into WARC files (ISO 28500:2009),

compressed files similar to zip which are assigned a unique identification number and stored in a distributed file system. Additionally, WARC supports some metadata such as provenance and HTTP protocol metadata. Implicit page elements, such as:

- Page title, headers, content, author information,
- Metadata such as Dublin Core elements,
- RSS feeds and other Semantic Web technologies such as Microformats [16] and Microdata [26] are completely ignored. This impacts greatly the way stored information is managed, reducing the utility of the archive and also hindering the creation of added-value services.

2. **Current Web Archiving Efforts** disregard the preservation of Social Networks and of interrelations between the archived content. However, weblog interdependencies demonstrated by the identification of central actors and peripheral weblogs, as well as by the meme-effect that applies to them, need to be preserved, to provide meaningful features to the weblog repository.
3. **Current Web Archive Scope is Limited** to monolithic regions, subjects or events. There is no generic web archiving solution capable to implement arbitrary subjects and topic hierarchies. For instance, the National Library of Catalonia has initiated a web crawling and access project aiming collect, process and provide permanent access to the entire cultural, scientific and general output of Catalonia in digital format [23].
   Alternatively, the Library of Congress has developed online collections for isolated historical events such as September 11, 2001 [19]. There is an ongoing debate, about benefits or disadvantages of one or another long-term preservation methodology. Many papers have been written and many conferences dedicated to this issue have appeared. It is surprising however, how little has been done at practical level.

3.3 Blog Archive Management

Regardless of the way a weblog is archived, current solutions do not provide users with meaningful management features of the stored information. For example, the Internet Archive stores weblogs as generic documents, listing one post after another, an approach that hinders if not forbids further weblog management. Examining the list of national web archiving initiatives (Table 1) one can see that out of 23 projects, only 8 support Full text search (34 %), 9 support Alphabetic Browsing (39 %) and 8 support Topical Collections (34 %). The most common feature available to all archives is URL Search.

Current solutions completely disregard the social aspect and interrelations of weblogs or other social media. Furthermore, due to the nature of periodic web crawling, users can only view the exact state of their weblog on prefixed dates or times. This solution cannot keep track of the evolving semantics and usage context of highly volatile hypertext pages like weblogs. For example, the Occasio News archive,

which collects sites based on their relevance to social issues, only preserves specific snapshots from a certain newsgroup [22]. Articles do not follow a continuous timeline, a fact that renders their substantial analysis in the future impossible. This results in prolific loss of information with respect to recording the weblog's evolution.

Additionally, current weblog archives cannot preserve the information regarding how posts, relevance links or other weblogs affect the original content and how they led to its propagation or extinction. However, this process must be identified to be of high cultural and sociological value: it is essential to preserve the notions and reactions of contemporary society, the motivations and drives, the interactions between complementary and adversary approaches to certain topics.

Moreover, browsing the preserved Blogosphere through current Web Archiving solutions, like Internet Archive or PANDORA, remains a tentative if not impossible task. For example, within the framework of these solutions, weblog interrelations indicated in the form of Blogrolls are treated as regular hyperlinks of the retrieved Web page with no particular informational value. Not only does this approach lead to the risk of them being omitted during the harvesting stage, especially by domain specific web archives, but it also disregards the value of preserving how thematically correlated weblogs interact with each other.

Finally, though web archived content is generally classified into wide thematic, regional or temporal categories, there exists no robust categorization technique. Weblogs' topic metadata are omitted if they do not fall into the predefined categories. For example, inter-relational authorship information is rarely incorporated into the generic archive model. However, the authorship of electronic publication bears several interesting features, like identification of central actors with authority ranking, person searches and interrelations between authors and the role of anonymity. This has many channels of interest in text mining and the social networking and scientific communities, and would be a stronghold of web archives focusing on social network websites Moreover, the temporal aspect of each Web Archive merely relates to a specific web-snapshot acquired through harvesting. Our methods of real-time harvesting, result into a continuous observation of the lifecycle of a weblog and provide accurate representation for each weblog at any point in time.

As implied by the aforementioned facts, a large fraction of current weblogs lacks digital preservation or it is partially archived. Additionally, digital archives created by means of any of the above mentioned solutions do not guarantee correctness and consistency, thus preventing their effectiveness and their proper usage.

4 Directions Towards Robust and Effective Blog Preservation

In this section, we present our approach towards robust and effective blog preservation. This is a challenge that the BlogForever project (BlogForever) is addressing from four different perspectives: modelling, aggregation, preservation and dissemination. The project's objectives are presented and then each one of the perspectives is outlined.

4.1 Objectives

The project's strategic objective is to provide complete and robust digital preservation, management and dissemination facilities for weblogs. Towards this end, the following scientific and technological objectives have been identified.

Study Weblog Structure and Semantics. BlogForever aims to analyse weblog structure and semantics to understand the unique and complex characteristics of weblogs and develop a generic data model as well as an ontology-based representation of the domain. To achieve this, weblogs are required to be understood and managed in 6 aspects:

1. As physical phenomena,
2. As logical encodings,
3. As conceptual objects with meaning to humans,
4. As structural objects of networked discourse and collaboration for knowledge creation in large groups of humans,
5. As sets of essential elements that must be preserved to offer future users the essence of the object,
6. As ontologies created in a bottom-up manner by communities rather than specialists.

Additionally, weblog aggregation heuristics will be developed to allow us to determine the best practices for efficient data extraction from weblogs.

Define a Robust Digital Preservation Policy for Weblogs. Developing a robust digital preservation policy for weblogs is one of the key objectives. The policy will include the following information:
1. Preservation strategy considerations for assessing risk, requirements for accessing deposited content and long-term accessibility of digital objects, as these factors are deemed to have enduring value. Furthermore, the preservation approach is to be described, including actions that are considered necessary for immediate, intermediate, and long-term preservation. In terms of depositing, it is important to have structures that allow for easy retrieval (and this relates to extracting structures and mapping to them; but also to predicting what and how queries of the future will look like – depending on the amount of flexibility that is required, the data storing can be simpler, or more complex).
2. The Assessment of Interoperability Prospects, which intends to address collaboration issues with existing generic European Web Archiving solutions. Moreover, means for reliable content transfer from the digital archive to other digital repositories, in the event of project termination are to be proposed.
3. The Digital Rights Management Policy, which addresses weblog copyright issues and controls the access level for each item and user in the digital archive.

Implement a Weblog Digital Repository. BlogForever aims to implement a digital repository web application, which will collect, archive, manage and disseminate weblogs. The platform will have the following 2 main components:

Fig. 2. BlogForever platform architecture.

1. The weblog aggregation component, which will be capable of searching, harvesting and analysing large volumes of weblogs. The output of this component will be an information package encoded in XML which will be submitted to the digital repository component.
2. The digital repository component, which will be responsible for weblog data preservation. The digital repository will ensure weblog proliferation, safeguard their integrity, authenticity and long-term accessibility over time, and allow for better sharing and re-using of contained knowledge.

A detailed depiction of the BlogForever platform architecture is depicted in Fig. 2.

Implement Specific Case Studies. BlogForever aims to design and implement specific case studies to apply and test the created infrastructure on extensive and diverse sets of weblogs. The case studies will be both generic (collecting weblogs from a wide array of topics) and domain specific (for example, a case study in academic bloggers community). Thus the case studies will provide the required breadth and depth to validate the developed tools, and guarantee that the project's results could be successfully and widely replicated after the project ends. The impact of the digital repository will be also evaluated by monitoring system usage and gathering user feedback.

The case studies will begin in summer 2012 and are expected to be completed in August 2013. The largest case study will include 500.000 blogs.

4.2 Modelling

Working towards the objectives identified previously, we have already examined a number of tasks regarding modelling the blogosphere.

Weblog Survey. The BlogForever Weblog Survey report [2] outlines a principal investigation into:

1. the common practices of blogging and attitudes towards preservation of blogs;
2. the use of technologies, standards and tools within blogs; and finally,
3. 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:

- common weblog authoring practices;
- important aspects and types of blog data that should be preserved;
- the patterns in weblogs structure and data;
- the technology adopted by current blogs; and finally
- the developments and prospects for analysing blog networks and
- 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.

Some key outcomes of the survey are the following:

- 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.
- 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.
- 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. Figure 3 illustrates the importance of preservation of all blog elements according to the BlogForever Survey.

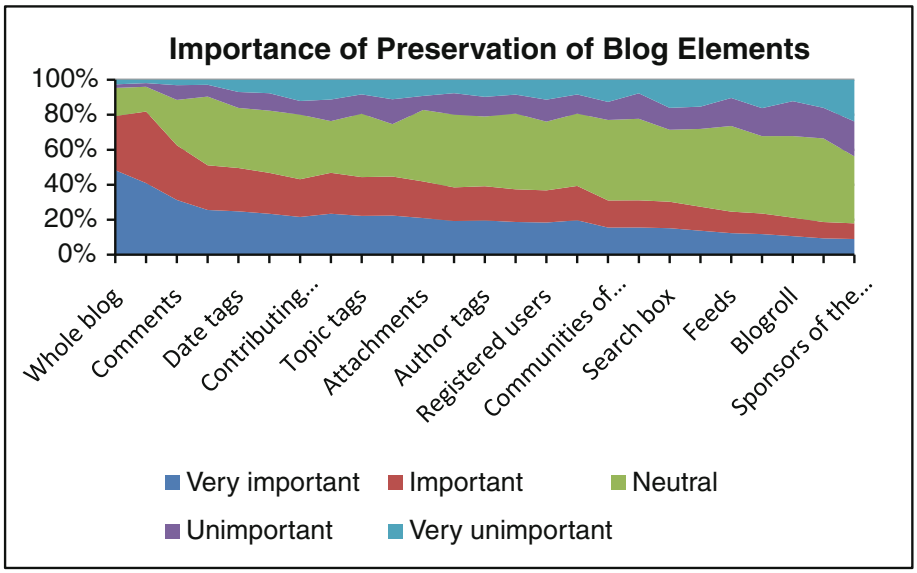

Fig. 3. Importance of preservation of blog elements according to the BlogForever Weblog Survey [2].

- 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.
- Some hypotheses regarding the intention of blog authors to contribute their blogs to a central blog archive were tested the analysis shows that the perception of a collective benefit has a stronger influence as the perception of an individual benefit.
- 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.
- A detailed analysis of the survey can found in BlogForever D2.1 Survey Implementation Report [2].

Weblog Data Model. Our work on weblog data model [28] identifies the data structures considered necessary for preserving blogs by revisiting the earlier inquiry summarised in the BlogForever Weblog Survey. The report includes an inquiry into

- the existing conceptual models of blogs,
- the data models of Open Source blogging systems, and
- data types identified from an empirical study of web feeds.

The report progresses to propose a data model intended to enable preservation of blogs and their individual components. A generic blog data model representation displaying core components and their interconnections is shown in Fig. 4. This basic model can then be extended to ensure the integrity and authenticity of preserved blogs, satisfactory to the requirements of successful preservation and archiving.

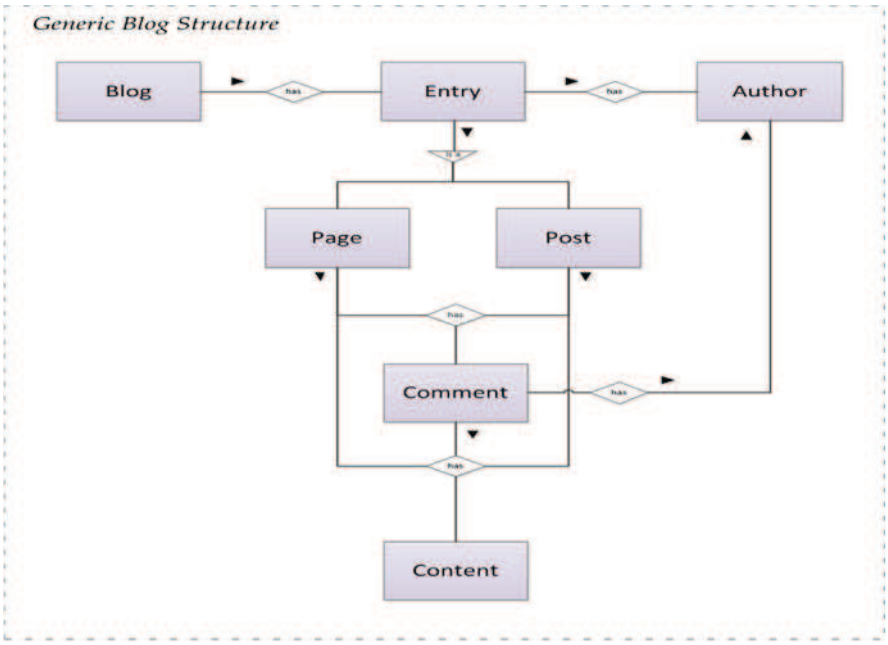

Fig. 4. Generic Blog data model elements and their relationships [28].

In addition to the core data model, a set of extension elements is also defined:

- **Weblog Context:** The entities described as part of the Blog Context component provide descriptive information about the blog and its elements in particular. It includes information about the selected presentation layer of the blog, description and keywords provided by the blogger, or the specific mark-up of individual elements of the blog.
- **Web Feed:** The Web Feed component consists of entities that are necessary to preserve information about web feeds of the blog.
- **Network and Linked Data:** This component contains the necessary associations that may exist across the blogs or their authors.
- **Community:** The Community component enables storing additional information about the active users – authors of posts and comments.
- **Categorised Content:** Categorised Content contains a number of entities that store the content collected from the blogs, which is decomposed into a number of smaller, but 'meaningful' pieces.
- **Standards and Ontology Mapping:** A mechanism for enabling the representation of stored blog data in specific standards, or for mapping it to certain ontologies.
- **Content Semantics:** They provide necessary structures to store the results of some analysis into the semantics of the content. For instance, the results of the sentiment analysis (i.e. sentiment scores) conducted on a specific piece of content can be stored along with additional data describing the algorithm, its version and the status of the results association with the content.

- **Spam Detection:** The Spam Detection component provides a mechanism for storing information about the algorithms and tools used for detecting spam and flagging the content included in the repository.
- **Crawling Information:** This component is intended to store information about the process of crawling. This will allow storage of information about the way crawling was conducted for a specific blog or sets of blogs. Storing information about crawling will make it possible to explain any differences between data along with the development of the crawler.
- **External Widgets:** External widgets make a fairly common appearance on blogs. Widgets are applications embedded into blogs or web pages. Some of the data describing the widget are planned to be stored as part of the preserved blog data.
- **Ranking Category and Similarity:** All of the entities described as part of this component are derived as a result of analysing captured blogs. These structures enable storing information about the ranking of blogs, or assigning them to certain categories.

A complete analysis of the BlogForever Data Model can be found in D2.2 Weblog Data Model Report [28].

Weblog Ontologies. Our work on weblog ontologies outlines an inquiry into the area of ontologies, conducted within the context of blog preservation, management and dissemination [15]. Three different scenarios regarding the application of ontologies are studied:

Semantic Extension of Tags. User generated tags and resulting folksonomies are widespread in blogs. However, while tags can organise blog posts inside a single blog according to the understanding of the blog author(s), it becomes more complicated if posts are aggregated from various blogs with possibly different contexts and topics. Therefore, it is necessary to identify and expose the meaning of the tags to overcome problems that result from the free choice of tags by different users, like homonyms and synonyms, and impair content retrieval.

Interoperability with Linked Open Data (LOD). Facilitation of the interoperability among repositories through the exposure and linking of data including explicit semantics. The application of ontologies can enhance the interoperability by the provision of open standards for describing, accessing, and connecting data. Figure 5 shows how interoperability could be established in this scenario.

The interoperability of the BlogForever environment has to be considered on two levels. First, there should be interoperability among different BlogForever archives. For example, a retrieval process for weblog data could operate on several archives and the results of complex search queries can be merged automatically. The use of shared vocabularies and a common ontology would allow an application to automatically merge the data from both repositories, providing a user of the repository with the means of searching and exploring the data as if they are from one repository.

Furthermore, interoperability with respect to other external repositories could be supported, for example, with other digital libraries. Digital libraries contain endless amounts of data that can be related to the data preserved in a BlogForever archive. Unlike interoperability between two BlogForever archives, the connection with

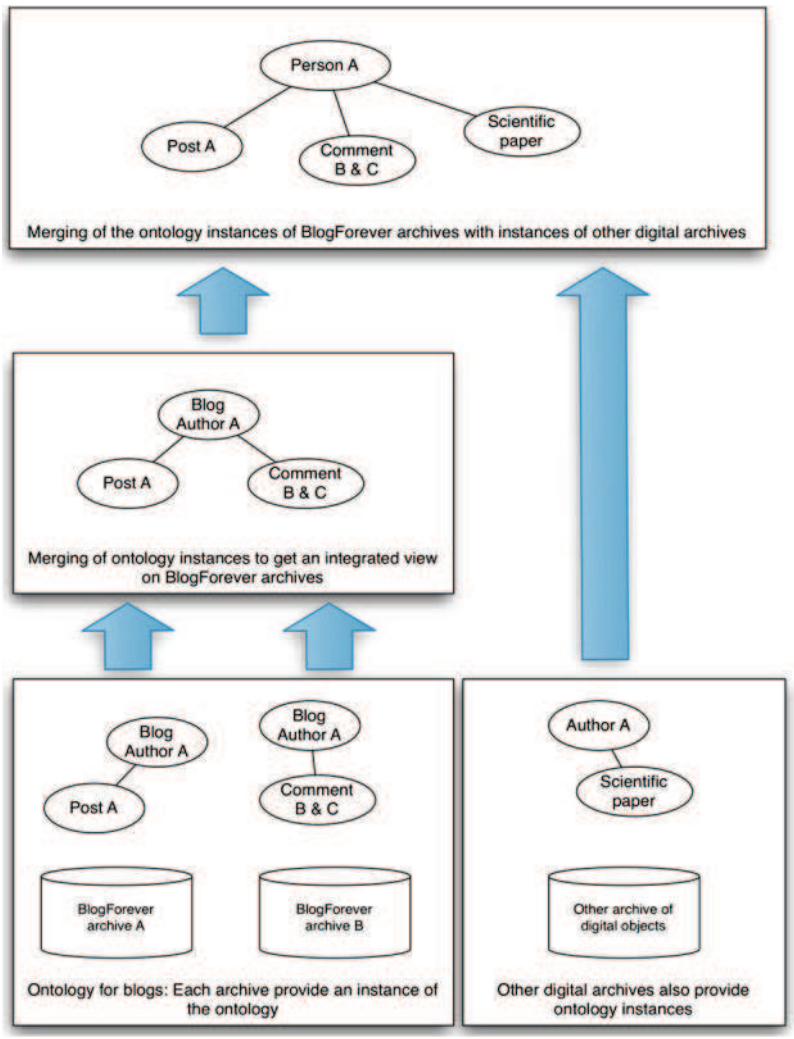

Fig. 5. Blog archive interoperability example using ontologies [15].

another digital library will extend the amount of concepts in the resulting ontology. In other words, two BlogForever archives share a common set of concepts (e.g. blog, post, blog author) and a merging means to merge instances of these concepts. However, another digital library has its own concepts like author, book, newspaper, etc. The relations between the concepts of both repositories have to be expressed (e.g. a blog author is a kind of author). Once the relations between the concepts are expressed formally, a merging of instances of both repositories will be possible.

Microformats, Microdata and RDFa. The utilisation of already available explicit semantics in the web pages that should be preserved can improve the quality of blog crawling. While data extraction on the layout specified by XHTML requires some

heuristics to identify the meaning of data (e.g. to identify the author of a document), it can be obtained directly if microformats are available. Therefore, microformats, microdata, and RDFa will be examined regarding their possible utilization for data aggregation in BlogForever.

User Requirements and Platform Specifications. Requirements descriptions for the BlogForever platform were thoroughly investigated and assembled from several sources including already completed work; semi-structured interviews with relevant stakeholders and users' survey [14]. The report illustrates the method of interview conduction and qualitative analysis. It includes a description of relevant stakeholders and requirement categories.

The identified requirements were specified in a standardised template and modeled with the unified modeling language (UML). Thus, they can be easily explored and utilised by developers. Overall, the requirements are the foundation for the design phase because they represent the perspective of demand.

4.3 Aggregation

The first step to preserve blogs is to manage to achieve effective and complete blog content aggregation. This problem can be split down to two sub-problems, detecting blog updates and retrieving updated blog content.

Weblog Data Extraction Methodologies and Prototypes. The BlogForever Data Extraction Methodology Report [29] outlines an inquiry into the area of web data extraction, conducted within the context of blog preservation. In this work, we review theoretical advances and practical developments for implementing data extraction. The inquiry is extended through an experiment that demonstrates the effectiveness and feasibility of implementing some of the suggested approaches. More specifically, we look into an approach based on unsupervised machine learning that employs the RSS feeds and HTML representations of blogs. It outlines the possibilities of extracting semantics available in blogs and demonstrates the benefits of exploiting available standards such as microformats and microdata.

The detailed workflow presented in Fig. 6 represents the sequential process of data extraction and can be used to inform the design of the data extraction system of the BlogForever platform. The workflow branches depending on a number of conditional checks (depicted as diamonds). The first check looks for an available wrapper. The other branches determine the flows that enable both capturing and updating posts. The loops within the diagram illustrate the processes where more than one entry is subject to extraction. A complete analysis of this work can be found in BlogForever Report D2.6 Data Extraction Methodologies [29].

In addition to the development of a new data extraction methodology, a number of weblog data extraction prototypes were implemented to test the aforementioned techniques and evaluate alternative ways to implement the weblog spider component, one of the two key elements of the BlogForever platform. This work is continued in order to articulate an optimal set of weblog aggregation techniques.

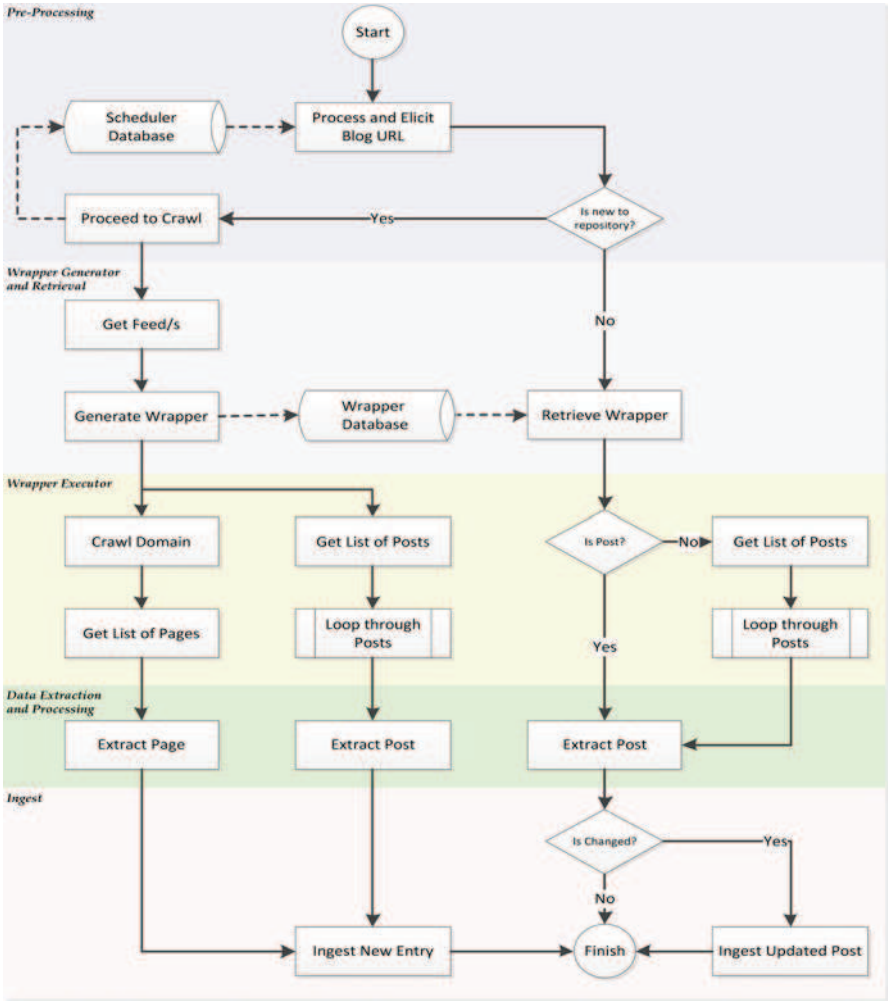

Fig. 6. Wrapper-based data extraction workflow (*Note: Dashed arcs indicate data flow*) [29].

Spam Filtering. Our research on spam filtering in the context of blog aggregation [17] comprises a survey of weblog spam technologies and approaches to their detection. While our work focused on identifying possible approaches to spam detection as a component within the BlogForever software, the discussion has been extended to include observations related to the historical, social and practical value of spam, and proposals of other ways of dealing with spam within the repository without necessarily removing them. We have identified three types of spam specific to blogs:

1. splogs, i.e. blogs that exist to promote affiliated websites, by influencing users to visit a webpage or buy a product, as well as spamdexing to undeservedly improve the ranking of a page in a web search by plagiarising content, stuffing keywords or creating large number of links,

2. blog comments that contain abusive content or are irrelevant to the original post, and,
3. fraudulent pings from non-blogs and/or splogs to attract visitors by misrepresenting content as fresh.

These spam types are analysed in the context of general web spam and various spam detection methods for weblogs are considered, concluding in a proposal for a spam detection workflow that might form the basis for the spam detection component of the BlogForever platform. The proposed methodologies in literature address web spam detection at the time of crawling, indexing and ranking. The types of features that can be used to detect spam can be grouped in spatial and temporal features. Spacial features refer to:

- **URL Pattern Information:** Methods using these type of features are based on the observation that:
  - spammers tend to stuff the URL with combination, mutations, and permutations of context rich keywords to benefit from the search engine ranking that rewards these URLs
  - spammers tend to use hyphens and long length URLs
  - domains that are cheap to acquire, such as ".info" domains, tend to be populated with a higher proportion of spam sites than expensive sites such as ".edu".
- **Home Page Content Information:** Methods using home page are based on the observation that:
  - spammers tend to repeat the same links and keywords
  - spam sites have short life span and grow very quickly
  - spammers employ a high percentage of nouns and only a few pronouns characteristic of expression of opinions
  - coherence of spam content may be lower that authentic content, that is the content would exhibit deviation from a general n-gram language model for $n > 1$
  - HTML templates created by automated blog creation software are repetitive
- **Feed-based Information:** Methods using feeds tend to copy home page content information techniques, and,
- **Link-based Information:** Methods using links are based on the observation that spammers try to exploit search engine ranking algorithms based on links, creating large numbers of fraudulent links.

Temporal features refer to the fact that the keywords, repetitiveness, network structure, size, and density with respect to splogs change at a different rate from that oberved with respect to authentic blogs.

To remedy the spam detection issues of blogs, a number of features and their pros and cons have been evaluated and are summarized in Table 2. Based on this information, we have formulated the proposed BlogForever spam filtering workflow which can be described in three main steps:

Table 2. Approaches to spam blog detection using various features.

Features	Pros	Cons
URL analyser/ template	Low process cost	Limited information – not very adaptive to change
IP/Post frequency	Could be difficult for spammers to manipulate	Must have history of updates and could become quite involved – e.g. where is the threshold for the frequency and how will it adapt to changes in the spam landscape?
Blacklist	Straightforward methodology and thrid party support available	Could lead to exploding blacklists.
RSS/content match	Indicates some level of agreement that the content is what the RSS feed says it is.	This requires that RSS feed is already available. Strictly speaking this is not spam filtering.
Full content analysis	Could be useful for removing duplicates. Difficult for spam to completely confound.	Could be process intensive.
User feedback	High precision	Labour intensive. Low recall because too many items for humans to examine.

1. Apply Ready-made Filters: black list databases (e.g. SplogSpot[1]), URL filtering, words in the content.
2. Apply Adaptive Filters: Ensemble classifier labels new posts, comments and blogs based on features such as temporal change, content, links.
3. Apply Adaptive Ranking Algorithms: Improve ranking performance of search engine by adopting positive crawling policy and implicit user feedback.

A complete analysis of this work can be found in BlogForever Report D2.5 Spam Filtering and Associated Methodologies [17].

4.4 Preservation

The process of digital preservation requires optimal retrieval and interpretation of the information to be preserved. As presented in the previous sections, our modelling and aggregation prototyping work will be the pillars upon which we will build an effective blog preservation platform.

Preservation Strategy. The preservation strategy will include information on assessing risk, requirements for accessing deposited content and long-term accessibility of digital objects, as these factors are deemed to have enduring value. Furthermore, the preservation approach is to be described, including actions that are

[1] http://www.splogspot.com

considered necessary for immediate, intermediate, and long-term preservation. In terms of depositing, it is important to have structures which allow for easy retrieval (and this relates to extracting structures and mapping to them; but also to predicting what and how queries of the future will look like – depending on the amount of flexibility that is required, the data storing can be simpler, or more complex).

Interoperability Strategy. Our planned work on the interoperability prospects of the BlogForever platform intends to analyse the different facets of interoperability: syntactic, semantic and pragmatic [25], by creating an interoperability testing methodology and specific scenarios. Furthermore, we are planning to address collaboration issues with existing platforms as well as libraries, archives, preservation initiatives and businesses that might be in synergistic relationships with BlogForever archives.

Digital Rights Management. Our planned work on Digital Rights Management (DRM) will initially include the identification and analysis of open issues and relevant discussions on the topic of blog preservation. Our aims will be protecting public access to information, content creators and content managers.

4.5 Management and Dissemination

To facilitate weblog digital preservation, management and dissemination, the project will implement a digital repository specially tailored to weblog needs. BlogForever digital repository will have to facilitate not only the weblog content but also the extended metadata and semantics of weblogs, which have been accumulated by the weblog aggregator as presented in Sect. 4.3.

The solution of creating a new software system as the basis of the weblogs repository has been considered and dismissed for this task, since many open-source repository back-ends are freely available on the Internet. In this respect, and taking into account the participation of CERN into the BlogForever consortium, the project will extend and adapt the globally acknowledged and widely used Invenio software [9]. The technology offered Invenio covers all aspects of digital library management. It complies with the Open Archives Initiative metadata harvesting protocol (OAI-PMH) and uses MARC 21 as its underlying bibliographic standard. Its flexibility and performance make it a comprehensive solution for the management of document repositories of large size and render it as an ideal basis for the BlogForever platform.

Long term blog preservation will be one aspect of the BlogForever platform. The other will be providing facilities for various stakeholders [14]:

- Content providers are people or organisations, which maintain one or more blogs and, hence, produce blog content that can or should be preserved in the archive.
- Individual blog authors are people that maintain their own blog.
- Organisations can serve as content providers if they maintain their own corporate blogs.
- Content retrievers are people or organisations which have an interest in the content stored in a blog archive and, therefore, they like to search, read, export, etc. that content.

- Individual blog readers are people who already read blogs for various reasons, e.g. family, hobbies, professional.
- In contrast, libraries operate more as a gatekeeper for individual retrievers. They provide access to various kinds of information sources, e.g. books, journals, movies, etc. Thereby, the access includes value added services like selecting and sorting the sources as well as adding metadata.
- Businesses also offer value added services based on the available information.

Each one of the aforementioned stakeholder has different blog preservation, archiving, management and dissemination requirements which have already been recorded and thoroughly documented, setting the priorities and work plan for the implementation of the BlogForever platform.

5 Conclusions

In this paper, we presented our perspective on the status of blog preservation and the blocking issues that arise when dealing with blog aggregation, preservation and management. Also, we identified a number of open issues that existing web archiving initiatives and platform face when dealing with blogs. Lastly, we presented an outline of the BlogForever EC funded project's current and future work towards creating a modern blog aggregation, preservation, management and dissemination platform.

Acknowledgements. The research leading to these results has received funding from the European Commission Framework Programme 7 (FP7), BlogForever project, grant agreement No.269963. We would also like to thank all BlogForever project partners for their invaluable contributions to the project.

References

1. Agarwal, N., Liu, H.: Blogosphere: research issues, tools and applications. ACM SIGKDD Explor. **10**(1), 18–31 (2008)
2. Arango-Docio, S., Sleeman, P., Kalb, H.: BlogForever: D2.1 survey implementation report. BlogForever WP2 Deliverable (2011)
3. Archive-it. Web Archiving Services. http://www.archive-it.org/. Accessed 11 April 2012
4. Arvidson, A.: Kulturarw3. In: Proceedings Conference on Strategies for the Internet: Preserving the Present for the Future, Copenhagen, pp. 101–104 (2001)
5. Ashley, K., Davis, R., Guy, M., Kelly, B., Pinsent, E., Farrell, S.: A guide to web preservation (2010)
6. Bhola, S., Strom, R., Bagchi, S., Zhao, Y.: Exactly-once delivery in a content-based publish-subscribe system. In: Proceedings International Conference on Dependable Systems and Networks (DNS), Washington, DC, pp. 7–16 (2002)
7. Billenness, C.: The future of the past – shaping new visions for EU-research in digital preservation. In: Proceedings Workshop, European Commission, Information Society and Media Directorate-General, Luxemburg (2011)

8. Campbell, L., Dulabahn, B.: Digital Preservation: the Twitter Archives and NDIIPP. In: Proceedings 7th International Conference Preservation of Digital Objects (iPRES), Vienna (2010)
9. CERN. Invenio. http://invenio-software.org/. Accessed 09 April 2012
10. Commission, European: Information and Communications Technologies (2011)
11. Edelstein O., Factor, M., King, R., Risse, T., Salant, E., Taylor, P.: Evolving domains, problems and solutions for long term. In: Proceedings 8th International Conference Preservation of Digital Objects (iPRES), Singapore (2011)
12. Heritrix.: IA Web Crawler. https://webarchive.jira.com/wiki/display/Heritrix/. (2012). Accessed 14 April 2012
13. Java A., Kolari P., Finin, T., Oates, T.: Modeling the spread of influence on the blogosphere. In: Proceedings 3rd WWW Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Edinburgh (2006)
14. Kalb, H., Kasioumis, N., García Llopis, J., Postaci, S., Arango-Docio, S.: BlogForever: D4.1 User requirements and platform specifications report. Blogforever WP4 Deliverable (2011)
15. Kalb, H., Gkotsis, G, Pincent, E., Banos, V., Davis, R.: BlogForever D2.3 Weblog Ontologies Report. BlogForever WP2 Deliverable (2012)
16. Khare, R., Celik, T.: Microformats: a pragmatic path to the semantic web. In: Proceedings 15th International Conference on World Wide Web (WWW), Edinburgh, pp. 865–866 (2006)
17. Kim, Y., Ross, S.: BlogForever: D2.5 Weblog spam filtering report and associated methodology. BlogForever WP2 Report (2012)
18. LAWA. Longitudinal Analytics of Web Archive Data Project. http://www.lawa-project.eu/ (2012). Accessed 15 April 2012
19. Library of Congress. Web Archive. http://lcweb2.loc.gov/diglib/lcwa/html/sept11/. (2011). Accessed 10 April 2012
20. LiWA. Living Web Archives Project. http://liwa-project.eu. Accessed 15 April 2012
21. McPhillips, S.: PANDORA Archive technical details. http://pandora.nla.gov.au/pandoratech.html. (2012). Accessed 05 Aug 2004
22. Occasio News Archive Database. http://newsarchive.occasio.net/. (2012). Accessed 10 April 2012
23. PADICAT: The Digital Heritage of Catalonia. http://www.padicat.cat/. (2012). Accessed 10 April 2012
24. PageFreezer.com - Social Media and Website Archiving. http://pagefreezer.com. (2012). Accessed 10 April 2012
25. Papazoglou, M.P., Ribbers, P.M.A.: E-business: Organizational and Technical Foundations. Wiley, West Sussex (2006)
26. Rynning, M., Banos, V., Stepanyan, K., Joy, M., Gulliksen, M.: BlogForever: D2.4 Weblog spider prototype and associated methodology. BlogForever WP2 Deliverable (2011)
27. Sroka, T.N.: Understanding the Political Influence of Blogs: A Study of the Growing Importance of the Blogosphere in the US Congress. Institute for Politics, Democracy and the Internet. http://www.ipdi.org/UploadedFiles/PoliticalInfluenceofBlogs.pdf. (2006) Accessed 14 June 2009
28. Stepanyan, K., Gkotsis, G., Pincent, E., Banos, V., Davis, R.: BlogForever D2.6 Data extraction methodology report. BlogForever WP2 Deliverable (2012)
29. The Internet Archive. http://archive.org. (1996)
30. VaultPress - Safeguard your site. http://www.vaultpress.com. (2012). Accessed 10 April 2012

31. Web Archiving, Library of Congress. http://www.loc.gov/webarchiving/. (2012). Accessed 12 April 2012
32. Web Curator Tool Project. http://webcurator.sourceforge.net/. (2012). Accessed 12 April 2012
33. Winer, D.: Original announcement of blog ping. http://xmlrpc.scripting.com/weblogsCom.html. (2012). Accessed 12 April 2012