Trends in Blog Preservation

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

[1]Department of Informatics, Aristotle University, Thessaloniki 54124, Greece

vbanos@gmail.com, manolopo@csd.auth.gr

[2]Department of Computing, Imperial College, London SW7 2AZ, UK

nb605@doc.ic.ac.uk

Keywords: Blogs, Blog preservation, Web archiving

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 talk, we will 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 will introduce the BlogForever EC funded project, its main motivation and findings towards widening the scope of blog preservation.

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 (Agarwal N.). The importance and the influence of the blogosphere are constantly rising and have become the subject of modeling and research (Java A.). 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” (Sroka T.N.). 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 (Tumblr Numbers); 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.

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 (Ashley K.). The topic of web preservation was initially addressed in a large scale by the Internet Archive in 1996 (The Internet Archive). 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).

Table 1, International Internet Preservation Consortium Members

Organization	Year	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

As digital preservation techniques progress and awareness is raised on the matter, there is a continuous trend towards preserving more complex objects (Billenness C.). 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 (Campbell L.).

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 (Commission, Information and Communications Technologies). A number of EC funded projects pursuing advanced web preservation are listed below:

- **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 (LiWA).
- **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 (Edelstein O.).
- **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 testbeds from diverse application areas: web content, digital repositories, and research data sets (Edelstein O.).
- **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 (LAWA).

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 (PageFreezer.com) is claiming to support web archiving and social media archiving. Another popular service is VaultPress (VaultPress), which provides security, backup and support for Wordpress blogs.

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 will present a number of issues that arise when dealing with blog preservation.

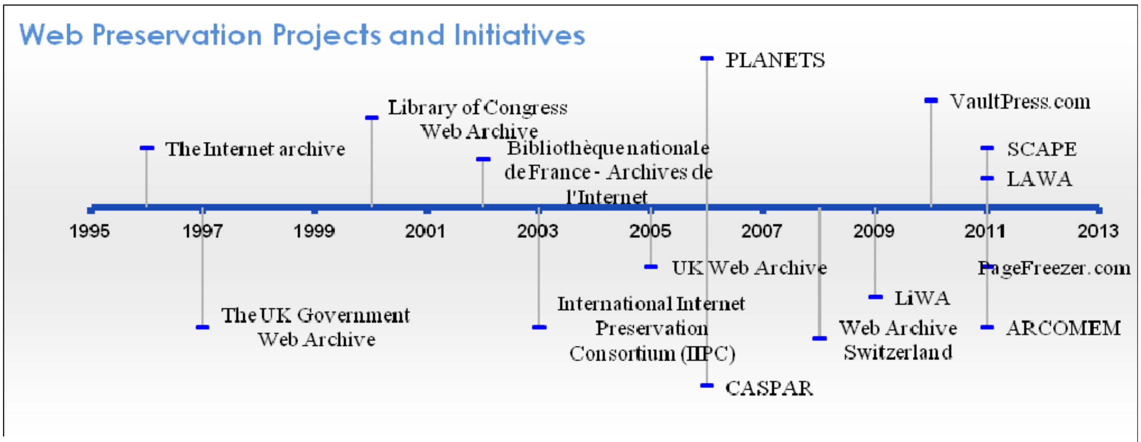

Figure 1, Timeline of important web preservation projects and initiatives

3. BLOG PRESERVATION ISSUES, OVERVIEW AND CONSIDERATIONS

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 (Web Archiving) 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 (Arvidson A.). 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 (Archive-it), which uses the latest Heritrix web crawler (Heritrix), enables harvesting material from the Web as frequently as every 24 hours, once per week, once per month, once per quarter, annually or just once. The same also applies to the popular (Web Curator Tool Project), 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 (Winer D.) and PubSubHub (Bhola S.). 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 (McPhillips S.) 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 (Khare R.) and Microdata (Ronallo J.) 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 to collect, process and provide permanent access to the entire cultural, scientific and general output of Catalonia in digital format (PADICAT).

Alternatively, the Library of Congress has developed online collections for isolated historical events such as September 11, 2001 (Library of Congress). 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 (Occasio News Archive Database). 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.

4.1.1 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.

4.1.2 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.

4.1.3 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:

1. The weblog aggregation component, which will be capable of searching, harvesting and analysing large volumes of weblogs.
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.

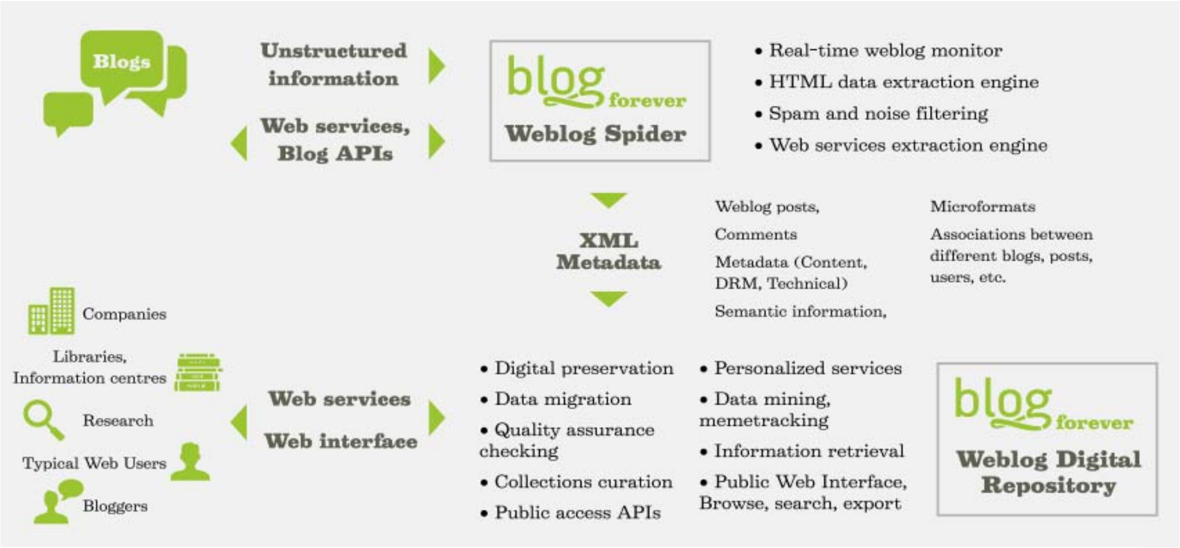
Blogs
Unstructured information
Web services, Blog APIs
blog forever
Weblog Spider
• Real-time weblog monitor
• HTML data extraction engine
• Spam and noise filtering
• Web services extraction engine
XML Metadata
Weblog posts,
Comments
Metadata (Content, DRM, Technical)
Semantic information,
Microformats
Associations between different blogs, posts, users, etc.
Companies
Libraries, Information centres
Research
Typical Web Users
Bloggers
Web services
Web interface
• Digital preservation
• Data migration
• Quality assurance checking
• Collections curation
• Public access APIs
• Personalized services
• Data mining, memetracking
• Information retrieval
• Public Web Interface, Browse, search, export
blog forever
Weblog Digital Repository

Figure 2, BlogForever Platform Architecture

A detailed depiction of the BlogForever platform architecture can be seen on Figure 2.

4.1.4 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.

4.2.1 Weblog Survey

The BlogForever Weblog Survey report (Arango-Docio S.) 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. 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:

- 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.

4.2.2 Weblog data model

Our work on weblog data model (Stepanyan K.) 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.

Pending work on weblog modeling includes an exploration of ontologies' applications in the context of blog preservation.

4.2.3 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 a users' survey (Kalb H.). 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 modelled with the unified modelling 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.

4.3.1 Weblog aggregation prototypes

During our work on weblog aggregation techniques (Rynning M.), we evaluated available weblog data extraction methodologies and technologies. Additionally, 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 will be continued to articulate an optimal set of weblog aggregation techniques.

4.3.2 Spam filtering

Our research on Spam filtering in the context of blog aggregation (Kim Y.) comprises a survey of weblog spam technology 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. It contains a general overview of spam types, ready-made anti-spam APIs available for weblogs, possible methods that have been suggested for preventing the introduction of spam into a blog, and research related to spam focusing on those that appear in the weblog context, concluding in a proposal for a spam detection workflow that might form the basis for the spam detection component of the BlogForever software.

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.

4.4.1 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 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).

4.4.2 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 (Papazoglou M.). 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.

4.4.3 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 section 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 (CERN). 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 (Kalb H.):

- **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 CONCLUSION

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

Agarwal N. and Liu H. (2008). Blogosphere: Research Issues, Tools and Applications. *ACM SIGKDD Explorations*, 10(1):18-31.

Arango-Docio S., Sleeman P. and Kalb H. (2011) BlogForever: D2.1 Survey Implementation Report. BlogForever WP2 Deliverable.

Archive-it, Web Archiving Services. 11 04 2012 http://www.archive-it.org/

Arvidson A. (2001). Kulturarw3. *Proceedings Preserving the Present for the Future.* Copenhagen, pages 101-104.

Ashley K., Davis R., Guy M., Kelly B., Pinsent E. and Farrell S. (2010) *A Guide to Web Preservation*.

Bhola S., Strom R., Bagchi S. and Zhao Y. (2002). Exactly-once Delivery in a Content-based Publish-Subscribe System. *Proceedings International Conference on Dependable Systems and Networks (DNS).* Washington DC, pages 7-16.

Billenness C. (2011). The Future of the Past – Shaping New Visions for EU-research in Digital Preservation. Proceedings Workshop, European Commission, Information Society and Media Directorate-General, Luxemburg.

BlogForever. BlogForever Project. 15 04 2012 http://blogforever.eu

Campbell L. and Dulabahn B. (2010). Digital Preservation: the Twitter Archives and NDIIPP. *Proceedings 7th International Conference Preservation of Digital Objects (iPRES),* Vienna.

CERN. Invenio. 09 04 2012 http://invenio-software.org/

Commission, European. (2011). Information and Communications Technologies.

Edelstein O., Factor M., King R., Risse T., Salant E. and Taylor P. (2011). Evolving Domains, Problems and Solutions for Long Term. *Proceedings 8th International Conference Preservation of Digital Objects (iPRES),* Singapore.

Heritrix (2012). IA Web Crawler. 14 04 2012 https://webarchive.jira.com/wiki/display/Heritrix/

IIPC (2012). International Internet Preservation Consortium. 10 04 2012. http://www.netpreserve.org

ISO 28500:2009 (2009), Information and Documentation – WARC File Format. Geneva: ISO.

Java A., Kolari P., Finin T. and Oates T. (2006). Modeling the Spread of Influence on the Blogosphere. *Proceedings 3rd WWW Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. Edinburgh.

Kalb H., Kasioumis N., García Llopis J., Postaci S and Arango-Docio S. (2011). BlogForever: D4.1 User Requirements and Platform Specifications Report. Blogforever WP4 Deliverable.

Khare R. and Celik T. (2006). Microformats: a Pragmatic Path to the Semantic Web. *Proceedings 15th International Conference on World Wide Web (WWW).* Edinburgh, pages 865-866.

Kim Y. and Ross S. (2012). BlogForever: D2.5 Weblog Spam Filtering Report and Associated Methodology. BlogForever WP2 Report.

LAWA. Longitudinal Analytics of Web Archive Data Project. 15 04 2012 http://www.lawa-project.eu/

Library of Congress, September 11, 2001, Web Archive. 10 04 2012. http://lcweb2.loc.gov/diglib/lcwa/html/sept11/

LiWA. Living Web Archives Project. 15 04 2012 http://liwa-project.eu

McPhillips S. (2012) PANDORA Archive Technical Details. 05 08 2004. 12 04 2012 http://pandora.nla.gov.au/pandoratech.html

Occasio News Archive Database. 10 04 2012 http://newsarchive.occasio.net/

PADICAT: The Digital Heritage of Catalonia. 10 04 2012. http://www.padicat.cat/

PageFreezer.com - Social Media and Website Archiving. 10 04 2012. http://pagefreezer.com

Papazoglou M.P. and Ribbers P.M.A. (2006). *E-business: Organizational and Technical Foundations*. John Wiley.

Ronallo J. (2012). HTML5 Microdata and Schema.org. *code4lib journal* (2012-02-03).

Rynning M., Banos V., Stepanyan K., Joy M. and Gulliksen M. (2011) BlogForever: D2.4 Weblog spider prototype and associated methodology." BlogForever WP2 Deliverable.

Sroka T.N. (2006). Understanding the Political Influence of Blogs: a Study of the Growing.

Stepanyan K., Joy M., Cristea A., Kim Y., Pinsent E. and Kopidaki S. (2011). BlogForever D2.2 Weblog Data Model. BlogForver WP2 Deliverable.

The Internet Archive (1996). http://archive.org

Tumblr Numbers: The Rapid Rise of Social Blogging. 14 04 2012, http://mashable.com/2011/11/14/tumblr-infographic/

VaultPress - Safeguard your site. 10 04 2012. http://www.vaultpress.com

Web Archiving, Library of Congress. 12 04 2012 http://www.loc.gov/webarchiving/

Web Curator Tool Project. 12 04 2012 http://webcurator.sourceforge.net/

Winer D. (2012) Original Announcement of Blog Ping. 12 04 2012 http://xmlrpc.scripting.com/weblogsCom.html