BlogForever: From Web Archiving to Blog Archiving

Hendrik Kalb, Paraskevi Lazaridou, Vangelis Banos, Nikos Kasioumis, Matthias Trier

hendrik.kalb@tu-berlin.de
paraskevi.lazaridou@tu-berlin.de
vbanos@gmail.com
nikos.kasioumis@cern.ch
mt.itm@cbs.dk

Abstract: In this paper, we introduce blog archiving as a special type of web archiving and present the findings and developments of the BlogForever project. Apart from an overview of other related projects and initiatives that constitute and extend the capabilities of web archiving, we focus on empirical work of the project, a presentation of the BlogForever data model, and the architecture of the BlogForever platform.

1 Introduction

The aim of this paper is to introduce blog archiving as a special type of web archiving.

Web archiving is an important aspect in the preservation of cultural heritage [Mas06] and, therefore, several projects from national and international organisations are working on web preservation activities. The most notable web archiving initiative is the Internet Archive[1] which has been operating since 1996. In national level, there are several remarkable activities, mainly from national libraries, to preserve web resources of their national domain. For example, the British Library announced this spring a project to archive the whole .uk domain [Coo13].

Web archiving is always a selective process, and only parts of the existing web are archived [GMC11, AAS+11]. The selection seems often to be driven by human publicity and search engine discoverability [AAS+11]. Furthermore, contrary to traditional media like printed books, web pages can be highly dynamic. Therefore, the selection of archived information comprises not only the decision of what to archive (e.g. topic or regional focus) but also additional parameters such as the archiving frequency per page, and parameters related to the page request (e.g. browser, user account, language etc.) [Mas06]. Thus, web archiving is a complex task that requires a lot of resources.

All active national web archiving efforts, as well as some academic web archives are members of the International Internet Preservation Consortium[2] (IIPC). Therefore, the web archiving tools[3] developed by the IIPC are widely accepted and used by the majority of

[1] http://archive.org
[2] http://netpreserve.org
[3] http://www.netpreserve.org/web-archiving/tools-and-software

internet archive initiatives [GMC11]. However, the approach inherent in these tools has some major limitations. The archiving of large parts of the web is a highly automated process, and the archiving frequency of a webpage is normally determined by a schedule for harvesting the page. Thus, the life of a website is not recorded appropriately if the page is updated more often than it is crawled [HY11]. Next to the harvesting problem of web archiving, the access of the archived information is inadequate for sophisticated retrieval. Archived information can be accessed only on site or page level according to a URI because analysis and management of current web archiving does not distinguish between different kinds of web pages. Thus, a page with a specific structure like blogs is handled as a black box.

The blogosphere, as part of the web, has an increasing societal impact next to traditional media like press or TV. Prominent examples are the influential blogs in political movements in Egypt [Ish08, Rad08] or Iran [Col05]. But there are also other domains that people engage in blogging, e.g. in the fields of arts or science [WJM10], teaching [TZ09] or leisure activities [Chi10]. The blogosphere as an institution has two connotations: On the one hand it is considered as a place where people build relationships – the blogosphere as a social networking phenomenon [AHA07, Tia13]. This view is emphasizing the activity of relating to others. On the other hand, it is also important to recognize that the numerous contributions yield a joint creation - the blogosphere as a common oeuvre, an institution shared by all bloggers and readers [KT12]. However, blogs as other social media are ephemeral and some that described major historical events of the recent past are already lost [Che10, Ent04]. Also the loss of personal diaries in the form of blogs has implications for our cultural memory [O'S05].

The BlogForever[4] project creates a novel software platform capable of aggregating, preserving, managing and disseminating blogs. Through the specialisation in blog archiving, as a subcategory of web archiving, the specific features of the blog as a medium can be exploited in order to overcome limitations of current web archiving.

2 Related work

In the following section, we review related projects and initiatives in the field of web archiving. Therefore, we inspect the existing solutions of the International Internet Preservation Consortium[5] (IIPC) for web archiving and the ArchivePress[6] blog archiving project. Furthermore, we look into several research projects such as Longitudinal Analytics of Web Archive Data[7] (LAWA), Living Web Archives[8] (LiWA), SCalable Preservation Environments[9] (SCAPE), Collect-All ARchives to COmmunity MEMories[10] (ARCOMEM), and

[4] http://blogforever.eu
[5] http://netpreserve.org
[6] http://archivepress.ulcc.ac.uk/
[7] http://www.lawa-project.eu/
[8] http://liwa-project.eu/
[9] http://www.scape-project.eu/
[10] http://www.arcomem.eu/

the Memento[11] project. Table 1 provides an overview of the related initiatives and projects we examine in this section.

Table 1: Overview of related initiatives and projects

Initiative	Description	Started
ArchivePress	Explore practical issues around the archiving of weblog content, focusing on blogs as records of institutional activity and corporate memory.	2009
ARCOMEM	Leverage the Wisdom of the Crowds for content appraisal, selection and preservation, in order to create and preserve archives that reflect collective memory and social content perception, and are, thus, closer to current and future users.	2011
IIPC projects	Web archiving tools for acquisition, curation, access and search.	1996
LAWA	Development of tools and methods to aggregate, query, and analyse heterogenous Internet data at large scale.	2010
LiWA	Develop and demonstrate web archiving tools able to capture content from a wide variety of sources, to improve archive fidelity and authenticity and to ensure long term interpretability of web content.	2009
Memento	Development of a technical framework that integrates current and past Web.	2009
SCAPE	Developing an infrastructure and tools for scalable preservation actions	2011

The **IIPC**[12] is the leading international organization dedicated to improving the tools, standards and best practices of web archiving. The software they provide as open source comprises tools for

- acquisition (Heritix[13]),
- curation (Web Curator Tool[14] and NetarchiveSuite[15]), and
- access and finding (Wayback[16], NutchWAX[17], and WERA[18]).

They are widely accepted and used by the majority of internet archive initiatives [GMC11].

[11] http://www.mementoweb.org/

[12] http://netpreserve.org/

[13] http://crawler.archive.org; an open-source, extensible, Web-scale, archiving quality Web crawler

[14] http://webcurator.sourceforge.net/; a tool for managing the selective Webharvesting process

[15] https://sbforge.org/display/NAS/Releases+and+downloads; a curator tool allowing librarians to define and control harvests of web material

[16] http://archive-access.sourceforge.net/projects/wayback/; a tool that allows users to see archived versions of web pages across time

[17] http://archive-access.sourceforge.net/projects/nutch/; a tool for indexing and searching Web archives

[18] http://archive-access.sourceforge.net/projects/wera/; a Web archive search and navigation application

The **ArchivePress**[19] project was an initial effort to attack the problem of blog archiving from a different perspective than traditional web crawlers. To the best of our knowledge, it is the only existing open source blog-specific archiving software. ArchivePress utilises XML feeds produced by blog platforms in order to achieve better archiving [PD09]. The scope of the project explicitly excludes the harvesting of the full browser rendering of blog contents (headers, sidebars, advertising and widgets), focusing solely on collecting the marked-up text of blog posts and blog comments (including embedded media). The approach was suggested by the observation that blog content is frequently consumed through automated syndication and aggregation in news reader applications, rather than by navigation of blog websites themselves.

The **LiWA**[20] project aims at the improvement of web archiving technologies. Thereby, it focuses on the areas of archive fidelity [DMSW11, OS10], spam cleansing to filter out fake content [EB11, EGB11], temporal coherence [EB11, BBAW10, MDSW10], semantic evolution of the terminology [TZIR10, TNTR10], archiving of social web material, and archiving of rich media websites [PVM10]. The project aims at the creation of long term web archives, filtering out irrelevant content and trying to facilitate a wide variety of content.

The **ARCOMEM** project focuses mainly on social web driven content appraisal and selection, and intelligent content acquisition. It aims at the transformation of "archives into collective memories that are more tightly integrated with their community of users and to exploit Social Web and the wisdom of crowds to make Web archiving a more selective and meaning-based process" [RP12]. Therefore, methods and tools are developed and research is undertaken in the areas of social web analysis and web mining [MAC11, MCA11], event detection and consolidation [RDM+11], perspective, opinion and sentiment detection [MF11], concise content purging [PINF11], intelligent adaptive decision support [PKTK12], advanced web crawling [DTK11], and approaches for semantic preservation [TRD11].

The **SCAPE** project is aiming to create scalable services for planning and execution of preservation strategies [KSBS12]. They address the problem through the development of infrastructure and tools for scalable preservation actions [SLY+12, Sch12], the provision of a framework for automated, quality-assured preservation workflows [JN12, HMS12], and the integration of these components with a policy-based preservation planning and watch system [BDP+12, CLM11].

The **LAWA** project aims at large-scale data analytics for Internet data. Therefore, it focusses on the development of a sustainable infrastructure, scalable methods, and software tools for aggregating, querying, and analysing heterogeneous data at Internet scale with a particular emphasis on longitudinal data analysis. Research is undertaken in the areas of web scale data provision [SBVW12, WNS+11], web analytics [BB13, PAB13, WDSW12, YBE+12, SW12], distributed access to large scale data sets [SPNT13, YWX+13, SBVW12], and virtual web observatory [SPNT13, YWX+13, ABBS12].

[19] http://archivepress.ulcc.ac.uk/

[20] http://liwa-project.eu/index.php

The **Memento**[21] project aims to provide access to the Web of the past in the way that current Web is accessed. Therefore, it proposes a framework that overcome the lack of temporal capabilities in the HTTP protocol [VdSNS+09]. It is now active Internet-Draft of the Internet Engineering Task Force [VdSNS13].

The aforementioned projects are evidence of various remarkable efforts to improve the harvesting, preservation and archival access of Web content. The BlogForever project, presented in the following, puts the focus on a specific domain of Web, the weblogs.

3 BlogForever project

In the following, we introduce the BlogForever project. In particular, we present three surveys that have been conducted, the BlogForever data model which constitutes a foundation for blog archiving, and the two components of the BlogForever platform.

3.1 Surveys about blogs and blog archiving

Several surveys were conducted in the project to reveal the peculiarities of blogs and the blogosphere, and to identify the specific needs for blog preservation.

Two distinct online questionnaires were disseminated in six language to blog authors and blog readers. The aim was to examine blogging and blog reading behaviour, the perceived importance of blog elements, backup behaviour of bloggers, perceptions and intentions for blog archiving and blog preservation. Complete responses were gathered from 512 blog authors and 428 blog readers. One finding was that the majority of blog authors rarely consider archiving of their blogs. This increases the probability of irretrievable loss of blogs and their data, and, therefore, justifies efforts towards development of independent archiving and preservation solutions. Additionally, the results indicated a considerable interest of readers towards a central source of blog discovery and searching services that could be provided by blog archives [ADSK+11].

A large-scale evaluation of active blogs has been conducted to reveal the adoption of standards and the trends in the blogosphere. Therefore, 259,390 blogs have been accessed and 209,830 retrieved and furhter analysed. The evaluation revealed the existence of around 470 blogging platforms in addition to the dominating WordPress and Blogger. There is also a large number of established and widely used technologies and standards, e.g. RSS, Atom feeds, CSS, and JavaScript. However, the adoption of metadata standards like Dublin Core[22], Open Graph[23], Friend of a Friend[24] (FOAF), and Semantically Interlinked Online Communities[25] (SIOC) varies significantly [BSJ+12, ADSK+11].

[21] http://www.mementoweb.org/
[22] http://dublincore.org/
[23] http://ogp.me/
[24] http://www.foaf-project.org/
[25] http://sioc-project.org/

Another survey, aiming on the identification of specific requirements for a blog archive, comprised 26 semi-structured interviews with representatives of different stakeholder groups. The stakeholder groups included blog authors, blog readers, libraries, businesses, blog provider, and researchers. Through a qualitative analysis of the interviews, 114 requirements were identified in the categories functional, data, interoperability, user interface, performance, legal, security, and operational requirements, and modelled with the unified modelling language (UML). While several of the requirements were specifically for blogs (e.g. comments to a blog may be archived even if they appear outside the blog, for example in Facebook), various requirements can be applied on web archives in general [KKL+11].

3.2 The BlogForever data model

While it seems that it is almost impossible to give an exclusive definition for the nature of blogs [Gar11, Lom09], it is necessary for preservation activities to identifiy blogs' properties [SGK+12]. This is even more crucial for the BlogForever platform which aims on sophisticated access capabilities for the archived blogosphere. Therefore, the different appearances of blogs were examined, and an comprehensive data model was created.

The development of the data model was based on existing conceptual models of blogs, data models of open source blogging systems, an empirical study of web feeds, and the online survey with blogger and blog reader perceptions. Thus, it was possible to identify various entities like [SJC+11]:

- Core blog elements, e.g. blog, post, comments,
- Embedded content, e.g. images, audio, video,
- Links, e.g. embedded links, blogroll, pingback,
- Layout, e.g. css, images,
- Feeds, e.g. RSS, Atom, and
- User profiles and affiliations.

The full model comprises over forty single entities and each entity is subsequently described by several properties, e.g. title, URI, aliases, etc. Figure 1 shows, therefore, the high level view of the blog core. The directions of the relationships between the primary identified entities of a weblog are indicated by small triangles [SJC+11].

Beside the inherent blog properties, additional metadata about archiving and preservation activities are captured, stored, and managed. For example, information regarding the time of harvesting of a blog or the legal rights of the content, have to be documented as well. Furthermore, additional data may emerge as well as annotations from the archive users, like tags or comments.

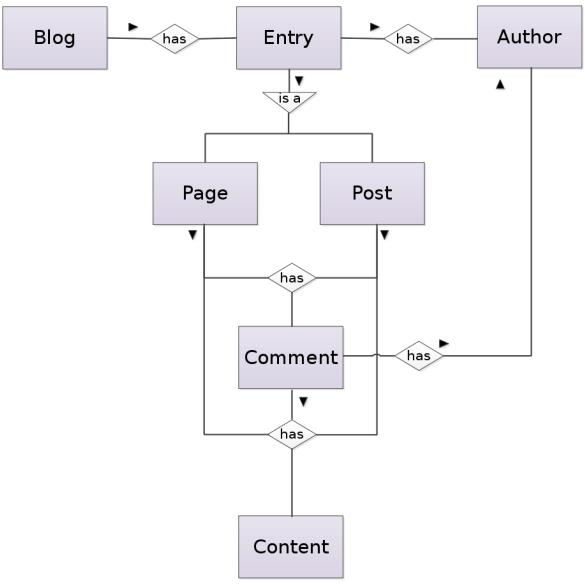

Figure 1: Core of the generic blog data model [SJC+11, p. 45]

3.3 BlogForever platform components

The BlogForever platform consists of the spider component and the repository component. The segmentation into two distinct parts with a well-defined communication interface between them makes the platform more flexible because the components can be developed seperately or even replaced if necessary.

The **spider component** is responsible for harvesting the blogs. It comprises of several subcomponents as shown in figure 2. The *Inputer* is the starting point, where the list of blogs that should be monitored is maintained. The list should be manually defined instead of using ping servers in order to enable the harvesting of qualified blogs and avoid spam blogs (also known as splogs). All blog URLs collected by the Inputer have to pass through the *Host Analyzer*, which approves them or blacklists them as incorrect or inappropriate for harvesting. Therefore, it parses each blog URL, collects information about the blog host and discovers the feeds that the blog may provide. The *System Manager* consists of the source database and the scheduler. While the source database stores all monitored blogs, including various metadata like filtering rules and extraction patterns, the scheduler determines when the blogs are checked for updates. The *Worker* is responsible for the actual harvesting and analysing of the blog content. Therefore, it fetches the feeds of the blogs as well as HTML content. Both are analysed in order to identify distinct blog elements. Further parsing enables the creation of an XML representation of the identified information and entities, and the identification and harvesting of embedded materials.

Finally, the *Exporter* delivers the extracted information together with the original content and embedded objects to the repository component [RBS+11].

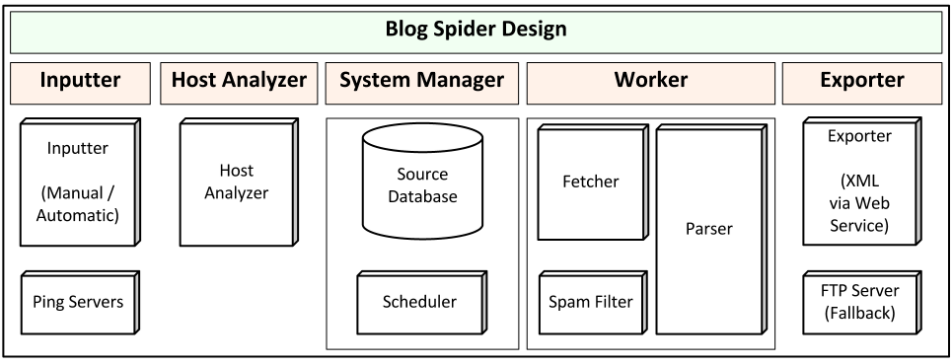

Figure 2: BlogForever spider component design [RBS+11]

The **repository component** represents the actual preservation platform. It facilitates the ingest, management, and dissemination of the harvested blog content and the extracted information. The repository component is based on the open source software suite Invenio[26], and the subcomponents are shown in figure 3.

New blogs for archiving are announced through the *Submission* as single blogs or bulk submissions. Thereby, a topic and a license can be indicated. The repository component informs in turn the spider about changes in the list of blogs to monitor. The submission of new blogs in the repository component enables the management of the archived blog selection through one point. The *Ingest* receives and processes the packages that spider component delivers. It conducts validity checks before the information is transferred to the internal storage. The *Storage* consists of databases and a filesystem. It manages the archived data and is responsible for the replication, incremental backup, and versioning. The latter is necessary to keep every version of an entity, e.g. a post, even if the entity has been updated. The *Core Services* comprise indexing, ranking, digital rights management (DRM), and interoperability. Indexing is performed to enable high speed searching on the archived content. Additionally, the search results can be sorted or ranked, e.g. according to their similarity. The DRM facilitates the access control on the repository's resources. Interoperability is a crucial aspect to facilitate a broader dissemination and integration into other services. Therefore, the repository compotent supports beside others the protocols of the Open Archive Initiative[27] (OAI), the OpenURL format, the Search/Retrieval via URL[28] (SRU), and Digital Object Identifiers (DOI). Finally, the *User Services* provide the functionalities of searching, exporting, personalising, and collaborating to the archive users. Searching can be performed through a search phrase in a single text field but also more enhanced search strategies are possible through the focussing on specific metadata (e.g.

[26] http://invenio-software.org
[27] http://www.openarchives.org/
[28] http://www.loc.gov/standards/sru/

title, author) and the use of regular expressions. The retrieved metadata can be exported in several formats (e.g. Dublin Core[29], MODS[30]) for further processing. Additionally, users can create personal collections and configure notifications that keep them informed about changes in their collection. Collections can also be shared with other users. The possibility to comment and rate any repository content facilitates further collaboration.

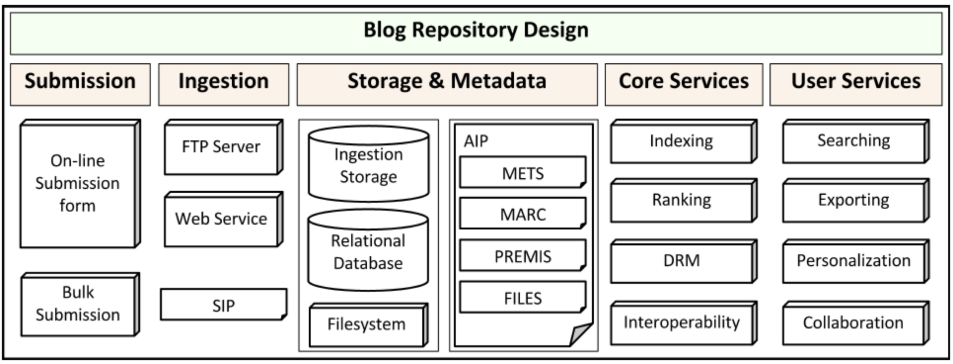

Figure 3: BlogForever repository component design

4 Conclusion

In this paper, we introduced the BlogForever project and blog archiving as a special kind of web archiving. Additionally, we gave an overview about related projects that constitute and extend the capabilities of web archiving. While there are certainly several other aspects to present about the BlogForever project and its findings, we focused on an overview of the empirical work, the presentation of the foundational data model, and the architecture of the BlogForever platform. The software will be available as open source at the end of the project and can be adopted especially by memory institutions (libraries, archives, museums, clearinghouses, electronic databases and data archives), researchers and universities, as well as communities of bloggers. Furthermore, guidelines and recommendations for blog preservation will be provided but could not be introduced in this paper. Two institutions plan already to adopt the BlogForever platform. The European Organization for Nuclear Research (CERN) is going to create a physics blogs archive to maintain blogs related to their research. The Aristotle University of Thessaloniki is going to create an institutional blog archive to preserve university blogs.

The approach of the BlogForever platform is dedicated but not limited to blog archiving. News sites or event calendars have often the same structure and characteristics as blogs (e.g. The Huffington Post). Thus, they could be also archived with BlogForever. However,

[29] http://dublincore.org/
[30] http://www.loc.gov/standards/mods/

it should be also emphasized that blogs are just one type of Web content and social media. Other types may cause different challenges but create also additional opportunities for exploitation. Therefore, additional research should be conducted in the future to further improve, specialise and support the current status of web archiving.

5 Acknowledgments

This work was conducted as part of the BlogForever[31] project co-funded by the European Commission Framework Programme 7 (FP7), grant agreement No.269963.

References

[AAS+11] Scott G Ainsworth, Ahmed Alsum, Hany SalahEldeen, Michele C Weigle, and Michael L Nelson. How much of the web is archived? In *Proceeding of the 11th annual international ACM/IEEE joint conference*, page 133, New York, New York, USA, 2011. ACM Press.

[ABBS12] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. Index maintenance for time-travel text search. In *the 35th international ACM SIGIR conference*, pages 235–243, New York, New York, USA, 2012. ACM Press.

[ADSK+11] Silvia Arango-Docio, Patricia Sleeman, Hendrik Kalb, Karen Stepanyan, Mike Joy, and Vangelis Banos. BlogForever: D2.1 Survey Implementation Report. Technical report, BlogForever Grant agreement no.: 269963, 2011.

[AHA07] Noor Ali-Hasan and Lada A Adamic. Expressing Social Relationships on the Blog through Links and Comments. *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 2007.

[BB13] Klaus Berberich and Srikanta Bedathur. Computing n-Gram Statistics in MapReduce. In *16th International Conference on Extending Database Technology (EDBT '13)*, Genoa, Italy, 2013.

[BBAW10] Klaus Berberich, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. A Language Modeling Approach for Temporal Information Needs. In *32nd European Conference on IR Research (ECIR 2010)*, pages 13–25, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[BDP+12] Christoph Becker, Kresimir Duretec, Petar Petrov, Luis Faria, Miguel Ferreira, and Jose Carlos Ramalho. Preservation Watch: What to monitor and how. In *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES 2012)*, pages 215–222, Toronto, 2012.

[BSJ+12] Vangelis Banos, Karen Stepanyan, Mike Joy, Alexandra I. Cristea, and Yannis Manolopoulos. Technological foundations of the current Blogosphere. In *International Conference on Web Intelligence, Mining and Semantics (WIMS) 2012*, Craiova, Romania, 2012.

[31] http://blogforever.eu/

[Che10] Xiaotian Chen. Blog Archiving Issues: A Look at Blogs on Major Events and Popular Blogs. *Internet Reference Services Quarterly*, 15(1):21–33, February 2010.

[Chi10] Tara Chittenden. Digital dressing up: modelling female teen identity in the discursive spaces of the fashion blogosphere. *Journal of Youth Studies*, 13(4):505–520, August 2010.

[CLM11] Esther Conway, Simon Lambert, and Brian Matthews. Managing Preservation Networks. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011)*, Singapore, 2011.

[Col05] Stephen Coleman. Blogs and the New Politics of Listening. *The Political Quarterly*, 76(2):272–280, April 2005.

[Coo13] Robert Cookson. British Library set to harvest the web, 2013.

[DMSW11] Dimitar Denev, Arturas Mazeika, Marc Spaniol, and Gerhard Weikum. The SHARC framework for data quality in Web archiving. *The VLDB Journal*, 20(2):183–207, March 2011.

[DTK11] Katerina Doka, Dimitrios Tsoumakos, and Nectarios Koziris. KANIS: Preserving k-Anonymity Over Distributed Data. In *Proceedings of the 5th International Workshop on Personalized Access, Profile Management and Context Awareness in Databases (PersDB 2011)*, Seattle, 2011.

[EB11] Miklós Erdélyi and András A Benczúr. Temporal Analysis for Web Spam Detection: An Overview. In *TWAW 2011*, Hyderabad, India, 2011.

[EGB11] Miklós Erdélyi, András Garzó, and András A Benczúr. Web spam classification. In *the 2011 Joint WICOW/AIRWeb Workshop*, pages 27–34, New York, New York, USA, 2011. ACM Press.

[Ent04] Richard Entlich. Blog Today, Gone Tomorrow? Preservation of Weblogs. *RLG DigiNews*, 8(4), 2004.

[Gar11] M Garden. Defining blog: A fool's errand or a necessary undertaking. *Journalism*, September 2011.

[GMC11] Daniel Gomes, João Miranda, and Miguel Costa. A Survey on Web Archiving Initiatives. In Stefan Gradmann, Francesca Borri, Carlo Meghini, and Heiko Schuldt, editors, *Research and Advanced Technology for Digital Libraries*, volume 6966 of *Lecture Notes in Computer Science*, pages 408–420. Springer Berlin / Heidelberg, 2011.

[HMS12] Reinhold Huber-Mörk and Alexander Schindler. Quality Assurance for Document Image Collections in Digital Preservation. In *Proceedings of the 14th International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 108–119, Brno, Czech Republic, 2012. Springer.

[HY11] Helen Hockx-Yu. The Past Issue of the Web. In *Proceedings of the ACM WebSci'11*, Koblenz, Germany, 2011.

[Ish08] Tom Isherwood. A new direction or more of the same? Political blogging in Egypt. *Arab Media & Society*, September 2008.

[JN12] Bolette Ammitzbøll Jurik and Jesper Sindahl Nielsen. Audio Quality Assurance: An Application of Cross Correlation. In *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES 2012)*, pages 144–149, Toronto, 2012.

[KKL+11] Hendrik Kalb, Nikolaos Kasioumis, Jaime García Llopis, Senan Postaci, and Silvia Arango-Docio. BlogForever: D4.1 User Requirements and Platform Specifications. Technical report, BlogForever Grant agreement no.: 269963, 2011.

[KSBS12] Ross King, Rainer Schmidt, Christoph Becker, and Sven Schlarb. SCAPE: Big Data Meets Digital Preservation. *ERCIM NEWS*, 89:30–31, 2012.

[KT12] Hendrik Kalb and Matthias Trier. THE BLOGOSPHERE AS ŒUVRE: INDIVIDUAL AND COLLECTIVE INFLUENCES ON BLOGGERS. In *ECIS 2012 Proceedings*, page Paper 110, 2012.

[Lom09] Stine Lomborg. Navigating the blogosphere: Towards a genre-based typology of weblogs. *First Monday*, 14(5), May 2009.

[MAC11] Silviu Maniu, Talel Abdessalem, and Bogdan Cautis. Casting a Web of Trust over Wikipedia: an Interaction-based Approach. In *Proceedings of the 20th International Conference on World wide web (WWW 2011*, Hyderabad, India, 2011.

[Mas06] Julien Masanès. *Web Archiving*. Springer-Verlag, Berli, Heidelberg, 2006.

[MCA11] Silviu Maniu, Bogdan Cautis, and Talel Abdessalem. Building a signed network from interactions in Wikipedia. In *Databases and Social Networks (DBSocial '11)*, pages 19–24, Athens, Greece, 2011. ACM Press.

[MDSW10] Arturas Mazeika, Dimitar Denev, Marc Spaniol, and Gerhard Weikum. The SOLAR System for Sharp Web Archiving. In *10th International Web Archiving Workshop*, pages 24–30, Vienna, Austria, 2010.

[MF11] Diana Maynard and Adam Funk. Automatic Detection of Political Opinions in Tweets. In *Proceedings of MSM 2011: Making Sense of Microposts. Workshop at 8th Extended Semantic Web Conference (ESWC 2011)*, pages 88–99, Heraklion, Greece, 2011. Springer Berlin Heidelberg.

[O'S05] Catherine O'Sullivan. Diaries, On-line Diaries, and the Future Loss to Archives; or, Blogs and the Blogging Bloggers Who Blog Them. *The American Archivist*, 68(1):53–73, 2005.

[OS10] Marilena Oita and Pierre Senellart. Archiving Data Objects using Web Feeds. In *10th International Web Archiving Workshop*, pages 31–41, Vienna, Austria, 2010.

[PAB13] Bibek Paudel, Avishek Anand, and Klaus Berberich. User-Defined Redundancy in Web Archives. In *Large-Scale and Distributed Systems for Information Retrieval (LSDS-IR '13)*, Rome, Italy, 2013.

[PD09] Maureen Pennock and Richard M. Davis. ArchivePress: A Really Simple Solution to Archiving Blog Content. In *Sixth International Conference on Preservation of Digital Objects (iPRES 2009)*, pages 148–154, San Francisco, USA, 2009. California Digital Library.

[PINF11] George Papadakis, Ekaterini Ioannou, Claudia Niederée, and Peter Fankhauser. Efficient entity resolution for large heterogeneous information spaces. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11)*, pages 535–544, New York, New York, USA, 2011. ACM Press.

[PKTK12] Nikolaos Papailiou, Ioannis Konstantinou, Dimitrios Tsoumakos, and Nectarios Koziris. H2RDF: Adaptive Query Processing on RDF Data in the Cloud. In *Proceedings of the 21st international conference companion on World Wide Web (WWW '12 Companion)*, pages 397–400, New York, New York, USA, 2012. ACM Press.

[PVM10] Radu Pop, Gabriel Vasile, and Julien Masanès. Archiving Web Video. In *10th International Web Archiving Workshop*, pages 42–47, Vienna, Austria, 2010.

[Rad08] Courtney Radsch. Core to Commonplace: The evolution of Egypt's blogosphere. *Arab Media & Society*, September 2008.

[RBS+11] M. Rynning, V. Banos, K. Stepanyan, M. Joy, and M. Gulliksen. BlogForever: D2. 4 Weblog spider prototype and associated methodology. Technical report, 2011.

[RDM+11] Thomas Risse, Stefan Dietze, Diana Maynard, Nina Tahmasebi, and Wim Peters. Using Events for Content Appraisal and Selection in Web Archives. In *Proc. of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011)*, Bonn, Germany, 2011.

[RP12] Thomas Risse and Wim Peters. ARCOMEM: from collect-all ARchives to COmmunity MEMories. In *21st international conference companion on World Wide Web (WWW '12 Companion*, pages 275–278, New York, New York, USA, 2012. ACM Press.

[SBVW12] Marc Spaniol, András A Benczúr, Zsolt Viharos, and Gerhard Weikum. Big Web Analytics: Toward a Virtual Web Observatory. *ERCIM NEWS*, 89:23–24, 2012.

[Sch12] Rainer Schmidt. SCAPE — An Architectural Overview of the SCAPE Preservation Platform. In *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES 2012)*, pages 85–88, Toronto, 2012.

[SGK+12] Karen Stepanyan, George Gkotsis, Hendrik Kalb, Yunhyong Kim, Alexandra I. Cristea, Mike Joy, Matthias Trier, and S amus Ross. Blogs as Objects of Preservation: Advancing the Discussion on Significant Properties. In *iPres 2012*, pages 218–224, Toronto, Canada, 2012.

[SJC+11] Karen Stepanyan, Mike Joy, Alexandra Cristea, Yunhyong Kim, Ed Pinsent, and Stella Kopidaki. Blogforever: Weblog Data Model. Technical report, 2011.

[SLY+12] Arif Shaon, Simon Lambert, Erica Yang, Catherine Jones, Brian Matthews, and Tom Griffin. Towards a Scalable Long-term Preservation Repository for Scientific Research Datasets. In *The 7th International Conference on Open Repositories (OR2012)*, Edinburgh, UK, 2012.

[SPNT13] George Sfakianakis, Ioannis Patlakas, Nikos Ntarmos, and Peter Triantafillou. Interval Indexing and Querying on Key-Value Cloud Stores. In *29th IEEE International Conference on Data Engineering*, Brisbane, Australia, 2013.

[SW12] Marc Spaniol and Gerhard Weikum. Tracking entities in web archives. In *21st international conference companion on World Wide Web (WWW '12 Companion*, pages 287–290, New York, New York, USA, 2012. ACM Press.

[Tia13] Q Tian. Social Anxiety, Motivation, Self-Disclosure, and Computer-Mediated Friendship: A Path Analysis of the Social Interaction in the Blogosphere. *Communication Research*, 40(2):237–260, February 2013.

[TNTR10] Nina Tahmasebi, Kai Niklas, Thomas Theuerkauf, and Thomas Risse. Using word sense discrimination on historic document collections. In *Proceedings of the 10th annual joint conference on Digital libraries (JCDL '10)*, pages 89–98, New York, New York, USA, 2010. ACM Press.

[TRD11] Nina Tahmasebi, Thomas Risse, and Stefan Dietze. Towards automatic language evolution tracking A study on word sense tracking. In *Proc. of the Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn 2011)*, Bonn, Germany, 2011.

[TZ09] Vicente Torres-Zúñiga. Blogs as an effective tool to teach and popularize physics: a case study. *Latin-American Journal of Physics Education*, 3(2):4, 2009.

[TZIR10] Nina Tahmasebi, Gideon Zenz, Tereza Iofciu, and Thomas Risse. Terminology Evolution Module for Web Archives in the LiWA Context. In *10th International Web Archiving Workshop*, pages 55–62, Vienna, Austria, 2010.

[VdSNS+09] Herbert Van de Sompel, Michael L Nelson, Robert Sanderson, Lyudmila L Balakireva, Scott Ainsworth, and Harihar Shankar. Memento: Time Travel for the Web. Technical report, November 2009.

[VdSNS13] Herbert Van de Sompel, Michael L Nelson, and Robert Sanderson. HTTP framework for time-based access to resource states, 2013.

[WDSW12] Yafang Wang, Maximilian Dylla, Marc Spaniol, and Gerhard Weikum. Coupling label propagation and constraints for temporal fact extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers (ACL '12)*, pages 233–237. Association for Computational Linguistics, July 2012.

[WJM10] Xiaoguang Wang, Tingting Jiang, and Feicheng Ma. Blog-supported scientific communication: An exploratory analysis based on social hyperlinks in a Chinese blog community. *Journal of Information Science*, 36(6):690–704, December 2010.

[WNS+11] Gerhard Weikum, Nikos Ntarmos, Marc Spaniol, Peter Triantafillou, András A Benczúr, Scott Kirkpatrick, Philippe Rigaux, and Mark Williamson. Longitudinal Analytics on Web Archive Data: It's About Time! In *5th Biennial Conference on Innovative Data Systems Research (CIDR '11)*, pages 199–202, Asilomar, California, USA, 2011.

[YBE+12] Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Maya Ramanath, Volker Tresp, and Gerhard Weikum. Natural language questions for the web of data. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*, pages 379–390, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[YWX+13] Qiang Yan, Xingxing Wang, Qiang Xu, Dongying Kong, Danny Bickson, Quan Yuan, and Qing Yang. Predicting Search Engine Switching in WSCD 2013 Challenge. In *Workshop on Web Search Click Data 2013 (WSCD2013)*, 2013.