Preservation Trends in the Blogosphere

Banos V.                      Aristotle University
Baltas N.                     Imperial College
Manolopoulos Y.      Aristotle University
# Table of Contents

1. Context & definitions
   - Blog, Blogosphere, Preservation, Blog Preservation, Issues, Relevant projects

2. Issues in blog preservation
   - Aggregation, Preservation, Management

3. Approaches to blog preservation
   - Study weblog structures and semantics
   - Define blog digital preservation strategy
   - Software and case studies
Part 1

Context and Definitions
Definitions

• A blog is a web service that encompasses an accessible and widely accepted mechanism for creating, maintaining and automatically distributing chronologically published material on the Web, along with the feedback and user domain associated with it.

• The Blogosphere is the collection of all blogs in the internet and their inter-connections.
State of the Blogosphere

- Blogs have become fairly established as an online communication and web publishing tool.

Number of Blogs Tracked by NM Incite

Read as: In October 2011 NM Incite tracked 173 million blogs as sources of online buzz.

Source: NM Incite
State of the Blogosphere

• Hundreds of millions of blogs are published about every conceivable subject

• There are many different types of blogs
  – Personal blogs
  – Corporate and organizational blogs
  – By genre (political, health, technology, etc)
  – By media type (vlog, linklog, photoblog, text blog, etc)
  – By device (composed by mobile phone or other computers)
  – Reverse blogs (aggregates of other blogs)
The problem of Blog Preservation

• But despite the fast growth of blogosphere, there is still no effective solution for ubiquitous semantic weblog archiving, digital preservation, management and dissemination.

• No current Web Archiving effort has ever developed a strategy for effective preservation and meaningful usage of Social Media.
Examples for the necessity of blog preservation

• A study (2010) evaluated blog posts for the Iraq war in 2003
  – “Blogs of War: Weblogs as News” (2005) documented 29 blogs on the Iraq war with their names
  – of those 29 blogs,
    - 9 (31%) no longer exist on the Internet,
    - 6 (20.1%) either were abandoned or now deal with a different subject
    - only 13 blogs (45%) were still active

• blogs on major events have already been lost
• Many single owner blogs lose their author through death
  – i.e. Mohammed Nabbous, the most active blogger during the Libyan revolution
  – he was killed in combat and his videoblog still remains today due to his wife who maintains and keeps it alive for the public

• Blogs like these can be lost simply due to the fact that hosting bills are not paid or the blog is deleted due to inactivity.
Digital Preservation

- **Digital preservation** is the set of processes, activities and management of digital information over time to ensure its **long term accessibility**.

- The goal of digital preservation is to preserve materials resulting from digital reformatting, and particularly information that is born-digital with no analog counterpart.

- Because of the relatively short lifecycle of digital information, **preservation is an ongoing process**
Blog Preservation

• The capture, management and preservation of blogs and their resources.

• Aims:
  – availability, viability, fixity, identity, authenticity, understandability, and renderability

• Issues:
  – Frequency of change, quantity and range of resources, integrity of web resources, database driven websites, ownership and DRM, +++
Google may, at its sole discretion, at any time and for any reason, terminate the Service, terminate this Agreement or suspend or terminate your account. In the event of termination, your account will be disabled and you may not be granted access to your account or any files or other content contained in your account although residual copies of information may remain in our system for some time for back-up purposes.

<table>
<thead>
<tr>
<th>Provider</th>
<th>Based</th>
<th>Global Reach (Unique Users, in Millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blogger</td>
<td>US</td>
<td>61</td>
</tr>
<tr>
<td>fc2</td>
<td>JP</td>
<td>120</td>
</tr>
<tr>
<td>LiveJournal</td>
<td>US</td>
<td>37</td>
</tr>
<tr>
<td>Tumblr</td>
<td>US</td>
<td>110</td>
</tr>
<tr>
<td>TypePad</td>
<td>US</td>
<td>28</td>
</tr>
<tr>
<td>Wordpress</td>
<td>US</td>
<td>310</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>blogger</th>
<th>wordpress</th>
<th>tumblr</th>
<th>typepad</th>
<th>mokono</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic archives</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Manual backup</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>XML format</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Export files</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
</tr>
</tbody>
</table>
Web Preservation Projects and Initiatives

1995 1997 1999 2001 2003 2005 2007 2009 2011 2013

- The Internet archive
- Library of Congress Web Archive
- Bibliothèque nationale de France - Archives de l'Internet
- PLANETS
- The UK Government Web Archive
- International Internet Preservation Consortium (IIPC)
- CASPAR
- UK Web Archive
- LiWA
- SCAPE
- LAW
- VaultPress.com
- PageFreezer.com
- ARCOMEM

ICEIS'2012 Conference, Wroclaw, Poland
# Current Web Archive Services Examples

<table>
<thead>
<tr>
<th>Organization</th>
<th>Year</th>
<th>Access Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliotheca Alexandrina's Internet Archive, Egypt</td>
<td>1996</td>
<td>URL Search</td>
</tr>
<tr>
<td>Bibliothèque nationale de France - Archives de l'Internet</td>
<td>2002</td>
<td>URL Search, Keyword Search, Full-Text Search, Topical Collections</td>
</tr>
<tr>
<td>Government of Canada Web Archive</td>
<td>2005</td>
<td>URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search</td>
</tr>
<tr>
<td>The Internet Archive (International)</td>
<td>1996</td>
<td>URL Search, Topical Collections</td>
</tr>
<tr>
<td>Library of Congress Web Archive, USA</td>
<td>2000</td>
<td>URL Search, Alphabetic Browsing, Subject Browsing, Topical Collections</td>
</tr>
<tr>
<td>National Library of Korea</td>
<td>2005</td>
<td>URL Search, Keyword Search, Subject Browsing</td>
</tr>
<tr>
<td>PANDORA Australia's Web Archive</td>
<td>1996</td>
<td>URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search, Subject Browsing</td>
</tr>
<tr>
<td>UK Web Archive</td>
<td>2005</td>
<td>URL Search, Alphabetic Browsing, Full-Text Search, Subject Browsing, Topical Collections</td>
</tr>
<tr>
<td>Web Archive Switzerland</td>
<td>2008</td>
<td>URL Search, Keyword Search, Full-Text Search, Subject Browsing, Topical Collections</td>
</tr>
</tbody>
</table>
Relevant EU Projects

- **LiWA** – Living Web Archives
- **ARCOMEM** – collect-all ARchives to COmmunity MEMories
- **SCAPE** – SCAlable Preservation Environments
- **LAWA** – Longitudinal Analytics of Web Archive Data
- **BlogForever** - create a software platform capable of aggregating, preserving, managing and disseminating blogs
LiWA - Living Web Archives

• 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**
• developing methods which improve archive fidelity
• developing methods for improved archive coherence and interpretability
• focusing on **audiovisual streams** and **social web** content
• **http://liwa-project.eu/**
ARCOMEM - collect-all ARchives to COmmunity MEMories

- innovative models and tools for Social Web driven content appraisal and selection, and intelligent content acquisition
- novel methods for Social Web analysis, Web crawling and mining, event and topic detection and consolidation, and multimedia content mining
- reusable components for archive enrichment and contextualization
- two complementary example applications, the first for media-related Web archives and the second for political archives
- a standards-oriented ARCOMEM demonstration system
- http://www.arcomem.eu
**SCAPE – SCAlable Preservation Environments**

- develop **scalable services** for planning and execution of preservation strategies on a **web scale**
- develop infrastructure and tools for scalable preservation actions;
- Provide a framework for automated, quality-assured preservation workflows
- integrate these components with a policy-based preservation planning and watch system.
- validated within three large-scale testbeds from diverse application areas.
- [http://www.scape-project.eu/](http://www.scape-project.eu/)
LAWA – Longitudinal Analytics of Web Archive Data

• The LAWA project on Longitudinal Analytics of Web Archive data will build an Internet-based experimental testbed for **large-scale data analytics**.

• Develop 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.

• [http://www.lawa-project.eu](http://www.lawa-project.eu)
# BlogForever Partners

<table>
<thead>
<tr>
<th>Name</th>
<th>Short name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aristotle University of Thessaloniki</td>
<td>AUTH</td>
<td>Greece</td>
</tr>
<tr>
<td>CERN, The European Organization for Nuclear Research</td>
<td>CERN</td>
<td>Switzerland</td>
</tr>
<tr>
<td>University of London</td>
<td>UL</td>
<td>UK</td>
</tr>
<tr>
<td>University of Glasgow</td>
<td>UG</td>
<td>UK</td>
</tr>
<tr>
<td>University of Warwick</td>
<td>UW</td>
<td>UK</td>
</tr>
<tr>
<td>Technical University Berlin</td>
<td>TUB</td>
<td>Germany</td>
</tr>
<tr>
<td>Software Research and Development Consultancy</td>
<td>SRDC</td>
<td>Turkey</td>
</tr>
<tr>
<td>Tero Research &amp; Technology Consulting</td>
<td>Tero</td>
<td>Greece</td>
</tr>
<tr>
<td>CyberWatcher</td>
<td>CW</td>
<td>Norway</td>
</tr>
<tr>
<td>ALTEC Software Development S.A.</td>
<td>ALTEC</td>
<td>Greece</td>
</tr>
<tr>
<td>Phaistos Networks S.A.</td>
<td>Phaistos</td>
<td>Greece</td>
</tr>
<tr>
<td>Mokono GmbH</td>
<td>Mokono</td>
<td>Germany</td>
</tr>
</tbody>
</table>

ICEIS'2012 Conference, Wroclaw, Poland
The BlogForever Platform

BlogForever will create a software platform capable of aggregating, preserving, managing and disseminating blogs.

Key Objectives

BlogForever will develop robust digital preservation, management and dissemination facilities for weblogs.

These facilities will be able to capture the dynamic and continuously evolving nature of weblogs, their network and social structure, and the exchange of concepts and ideas that they foster; pieces of information omitted by current Web Archiving methods and solutions.

Scientific Objectives

- Study weblog structure and semantics
- Define a robust digital preservation policy for weblogs
- Implement a weblog digital repository
- Implement specific case studies
The BlogForever Platform

- **Unstructured information**
  - Blogs
  - Web services, Blog APIs

- **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,

- **Web services**
  - Digital preservation
  - Data migration
  - Quality assurance checking
  - Collections curation
  - Public access APIs

- **Web interface**
  - Personalized services
  - Data mining, memetracking
  - Information retrieval
  - Public Web Interface, Browse, search, export

- **Weblog Digital Repository**
  - Microformats
  - Associations between different blogs, posts, users, etc.
The BlogForever Platform

**Impact**

The final output of BlogForever will be a simple weblog digital archiving solution that any user, user group or institution could use to preserve their weblog(s) and ensure their authenticity, integrity, completeness, usability, and long term accessibility as a valuable cultural, social, and intellectual resource.

A multitude of parties will benefit from the project (Bloggers, Universities, Libraries & Information Centers, Museums, Education, Research, Business)

**Advances to the state of the art**

- Definition of a generic data model for weblog metadata and semantics
- Weblog digital preservation strategies
- Weblog spider
- Weblog digital repository web application
Part 2

Issues in blog preservation
Blog Content Aggregation Issues

• Web content aggregation scheduling
  – Weblogs are extremely volatile
  – Not considering web site updates, only perform periodic aggregation

• Web content aggregation performance
  – Brute force techniques, waste computing resources
  – Important components of blogs not identified

• Quality assurance checking
  – No automated archive checking
  – Not identify important information (metadata, microformats, specific blog elements of importance)
Blog Content Preservation Issues

• Current web preservation initiatives are geared towards aggregating and preserving **HTML pages** and not **information entities** (posts, comments, authors, metadata, dates, pingbacks, etc).

• **Current web archiving efforts** disregard the preservation of **Social Networks** and **interrelations** between the archived content (meme-effect).

• **Current web archives** cannot identify **topics**, **subjects** or **events** (monolithic). There is no generic web archiving solution capable to implement arbitrary subjects and topic hierarchies.
Blog Archive Management Issues

• Regardless of the way a weblog is archived, current solutions do not provide users with meaningful management features of the stored information.
• Current solutions completely disregard the social aspect and interrelations of weblogs or other social media.
• Even browsing the preserved Blogosphere through current Web Archiving solutions, like Internet Archive remains a desired if not impossible task.
• No meaningful information - need of ontologies
Part 3

Directions towards robust blog preservation
BlogForever

• Study weblog structure and semantics
• Define a robust digital preservation strategy for weblogs
• Implement a weblog digital repository
• Implement specific case studies
• http://www.blogforever.eu
BlogForever Survey
Main aims include the study of:

- **Common blog authoring practices**
- **Patterns in blog structure**
- **Blog lifecycles and ranking**
- Blog author **intention** to contribute to a blog archive
- **Factors for search** strategies in a blog archive
- **Common blog preservation aspects**
BlogForever Survey Responders

Responders:

✓ Readers: 517
✓ Authors: 428
Importance of Preservation of Blog Elements
BlogForever Survey – General Insights

✓ Majority of authors (90%) never use external services to preserve their blog, relying on their blog provider.
✓ Using a blog archive is welcome but authors think blog providers are the leading bodies for this role at the moment.
✓ Motivations for maintaining blogs were primarily personal (80%) for sharing information and promoting discussion topics.
✓ Perceived importance of rich media, links and citations
✓ Wide variation in the use of media objects
✓ Primary use: textual content (98%) in blogs
✓ Frequent use of photographs and moving images (40%-83%)
BlogForever Technology Survey

Main goal was to study large number of blogs and evaluate the use of:

✓ Third-Party Libraries
✓ External Services
✓ Semantic Mark-Up
✓ Metadata
✓ Web Feeds
✓ Various Media Formats
Dataset by weblog.com

Number of included URLs:

✓ Overall total of accessed resources: 259,930

✓ Overall total of valid records: 209,830
Blog Technology Survey: Blog platforms

[Bar chart showing the frequency of different blog platforms:
- Unidentified: 41%
- Wordpress: 36%
- Blogger: 19%
- Other: 2%
- Typepad: 1%
- Joomla: 0%]
**Feed type**

Frequency of Feeds per total URI (blog):

- ✓ 83% - RSS feeds
- ✓ 43% - Atom feeds \(\subseteq\) RSS
- ✓ 22% - RSS Comments
Most popular Libraries:

- Dojo
- Ext Core
- JQuery
- JQuery UI
- MooTools
- Prototype
- YUI Library
<table>
<thead>
<tr>
<th>File Ext.</th>
<th>Application</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc</td>
<td>Word Processing</td>
<td>1097</td>
</tr>
<tr>
<td>docx</td>
<td>Word Processing</td>
<td>147</td>
</tr>
<tr>
<td>odt</td>
<td>Word Processing</td>
<td>51</td>
</tr>
<tr>
<td>pdf</td>
<td>Word Processing</td>
<td>13731</td>
</tr>
<tr>
<td>txt</td>
<td>Word Processing</td>
<td>641</td>
</tr>
<tr>
<td>mp4</td>
<td>Video/Audio</td>
<td>3265</td>
</tr>
<tr>
<td>mpeg</td>
<td>Video</td>
<td>36</td>
</tr>
<tr>
<td>mpg</td>
<td>Video</td>
<td>613</td>
</tr>
<tr>
<td>avi</td>
<td>Video</td>
<td>3265</td>
</tr>
<tr>
<td>mov</td>
<td>Video</td>
<td>71</td>
</tr>
<tr>
<td>3gpp</td>
<td>Video</td>
<td>1429</td>
</tr>
<tr>
<td>xls</td>
<td>Spread Sheet</td>
<td>138</td>
</tr>
<tr>
<td>xlsx</td>
<td>Spread Sheet</td>
<td>24</td>
</tr>
<tr>
<td>ods</td>
<td>Spread Sheet</td>
<td>722</td>
</tr>
<tr>
<td>ppt</td>
<td>Presentation</td>
<td>67</td>
</tr>
<tr>
<td>pptx</td>
<td>Presentation</td>
<td>20</td>
</tr>
<tr>
<td>odd</td>
<td>Presentation</td>
<td>618</td>
</tr>
<tr>
<td>odf</td>
<td>Math Formulas</td>
<td>63</td>
</tr>
<tr>
<td>odg</td>
<td>Graphics</td>
<td>4</td>
</tr>
<tr>
<td>mdb</td>
<td>Database</td>
<td>0</td>
</tr>
<tr>
<td>ccdb</td>
<td>Database</td>
<td>0</td>
</tr>
<tr>
<td>odb</td>
<td>Database</td>
<td>153</td>
</tr>
<tr>
<td>vCard</td>
<td>Card</td>
<td>14</td>
</tr>
<tr>
<td>mp3</td>
<td>Audio</td>
<td>10231</td>
</tr>
<tr>
<td>wav</td>
<td>Audio</td>
<td>13</td>
</tr>
<tr>
<td>vrml</td>
<td>3D</td>
<td>0</td>
</tr>
</tbody>
</table>
Comparing blogs and web in general

Errors

The Web

HTTP Response Codes
- Errors
- No Errors

25%
75%

The Blogosphere

HTTP Response Codes
- Errors
- No Errors

5%
95%
Comparing blogs and web in general
Use of Flash

The Web

The Blogosphere
Comparing blogs and web in general
Common JS Libraries

The Web

<table>
<thead>
<tr>
<th>Library</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>jQuery</td>
<td>48%</td>
</tr>
<tr>
<td>YUI</td>
<td>2%</td>
</tr>
<tr>
<td>Dojo</td>
<td>1%</td>
</tr>
<tr>
<td>Google Analytics</td>
<td>63%</td>
</tr>
<tr>
<td>Quantcast</td>
<td>13%</td>
</tr>
<tr>
<td>AddThis</td>
<td>5%</td>
</tr>
<tr>
<td>Facebook</td>
<td>15%</td>
</tr>
<tr>
<td>Google +1</td>
<td>7%</td>
</tr>
<tr>
<td>Twitter</td>
<td>6%</td>
</tr>
<tr>
<td>ShareThis</td>
<td>1%</td>
</tr>
</tbody>
</table>

The Blogosphere

<table>
<thead>
<tr>
<th>Library</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>JQuery</td>
<td>10%</td>
</tr>
<tr>
<td>YUI</td>
<td>3%</td>
</tr>
<tr>
<td>Dojo</td>
<td>0%</td>
</tr>
<tr>
<td>Facebook</td>
<td>4%</td>
</tr>
<tr>
<td>Google+</td>
<td>17%</td>
</tr>
<tr>
<td>Twitter</td>
<td>0%</td>
</tr>
</tbody>
</table>
Comparing blogs and web in general

Image Formats

The Web

The Blogosphere

ICEIS'2012 Conference, Wroclaw, Poland
Blog Technology Survey Conclusions

- 469 platforms (with the dominant WordPress & Blogger)
- Wide range of encoding standards
- RSS/Atom feeds, CSS and JavaScript are among most widely used technologies
- Use of images and their formats are similar to their general use on the Web
- The use of Dublin Core, Open Graph, FOAF and SIOC is not widespread
- Quick adoption of Google+ and evidence of some integration of Facebook and Twitter
Blog modeling and semantics
BlogForever: Conceptual Data Model
Version 0.6
Future Work: Analyse Archived Blogs

• Analyse blog archives to gain a better understanding of the content and provide new services:
  – Use **Linked Open Data** to link archived blog content with other web content
  – Apply **Semantic Extension of Tags** to understand them better and reuse them for multiple purposes.
• In any case, use **Ontologies** to interpret and reason with information.
Linked Open Data

• Linked Open Data can help us link archived blog content with other web content.

• Improve interoperability by using common vocabularies:
  – Friend of a Friend (FOAF)
  – Dublin Core (DC)
  – Semantically Interlinked Online Communities (SIOC)

• Example: Link archived blog content to the description of scientific publications in DBLP
Semantic Extension of Tags

• User generated tags and resulting folksonomies are widespread in social media applications.

• Aims:
  – Identify and expose the meaning of tags using MOAT (Meaning Of A Tag) to resolve various problems
  – Deduce tag relationships to utilize the social aspect of blogs
Semantic Extension of Tags: identify and expose their meaning
Semantic Extension of Tags: Deduce tag relationships
BlogForever Platform Testing

- 2 academic blogs cases
  - University of Warwick (58 blogs)
  - University of London (70 blogs)
- Multilingual blogs cases
  - 356 blogs in 4 languages with diverse topics
- Multimedia blogs case
  - 1000 Greek blogs with diverse multimedia content
- Large case
  - 500,000 blogs from blog.de
Summary

• Introduction to blog preservation
• Identified a number of open issues on blog preservation
• Presented an outline of current research work on the BlogForever project:
  – Survey
  – Modelling
  – Interoperability
  – Preservation
  - Spam filtering
  - Crawling - Ingestion
  - Software platform
  - Testing using real data
Thank you!

Any Questions?

Visit: http://blogforever.eu to learn more.

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