Diverse approaches to blog preservation: a comparative study

Richard M. Davis
University of London
Computer Centre
Senate House, Malet Street
London WC1E 7HU
+44 20 7692 1350
richard.davis@london.ac.uk

Edward Pinsent
University of London
Computer Centre
Senate House, Malet Street
London WC1E 7HU
+44 20 7692 1345
edward.pinsent@london.ac.uk

Silvia Arango-Docio
University of London
Computer Centre
Senate House, Malet Street
London WC1E 7HU
+44 20 7692 1343
silvia.arango-docio@london.ac.uk

ABSTRACT

This poster presents highlights of a comparative study of three distinct approaches to preserving the content of blogs, to consider the relative benefits of each approach in meeting the requirements for blog preservation, in different contexts. Assessment criteria are drawn from key publications and frameworks on digital preservation as well as practical considerations derived from the authors' experience as users and designers of digital archiving tools and systems.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous; D.2.8 [Software Engineering]: Metrics / complexity measures, performance measures

General Terms

Management, Performance, Design, Reliability, Human Factors, Standardization, Theory.

Keywords

digital preservation, digital curation, designated community, authenticity, intellectual entity, archive, web archive, blog, weblog

1. INTRODUCTION

The importance of blogs as a distinct class of Web resource has received considerable attention in recent years, notably at iPRES (Pennock and Davis, 2009 [1]; Kim and Ross, 2011 [2]; Stepanyan et al, 2012 [3]). The need to capture this dynamic, cumulative content for future access has been recognised by several institutions and projects and a variety of tools and approaches have emerged.

This poster will present, in graphic form, a summary of key results of interest from a comparative analysis of three distinct approaches to blog-archiving, each of which differs significantly in its methodology, strategy, and delivered outcomes.

The study is based on key criteria derived from study of a wide range of established frameworks in digital preservation, including:

- Reference Model for an Open Archival Information System (OAIS) [4]
- Preservation Metadata Implementation Standard (PREMIS) [5]
- Metadata Encoding and Transmission Standard (METS) [6]
- Trustworthy Repositories Audit & Certification (TRAC) [7]
- Digital Repository Audit Method Based On Risk Assessment (DRAMBORA) [8]

The study compares the relative strengths of three types of digital archive/repository in the context of blog preservation: one created specifically for blogs; another designed for institutional publications; and a third designed for Web content.

The study identifies a number of indicators for success in web-archiving, and a select range of metrics to the effectiveness of each approach against established criteria, derived from the authors' experience and review of literature on best practice for web archiving projects. The study will be completed during June 2013 and the highlights are presented in the accompanying poster.

2. THREE APPROACHES TO BLOG ARCHIVING

1. The BlogForever project, funded by the European Union, has developed an integrated platform, comprising a harvesting methodology and associated content management system, for creation, management and preservation of blog collections.
2. The London School of Economics (LSE) preserves its academic blogs by creating and depositing PDF renditions of blog posts into an existing Institutional Repository.
3. The UK Web Archive, operated by the British Library, which collects and preserves blog content from the UK Blogosphere. This collection represents a cross section of UK Web logs containing a wealth of material which will be of value to researchers now and in the future.

3. ASSESSMENT AND SELECTION CRITERIA

The assessment criteria are derived from definitions and understanding of digital preservation as expressed in the following standards, projects and reports.

1. Long Term Preservation (OAIS): does the repository offer sufficient control of the content to ensure long-term preservation?
2. Designated Community (OAIS): does the repository identify a Designated Community who should be able to understand the information provided; and is the content independently understandable and available to the Designated Community?
3. Preservation metadata (PREMIS): does the repository support the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context?
4. Metadata encoding and transmission (METS): is there metadata necessary for both the management of digital objects within a repository and exchange of such objects between repositories (or between repositories and users)?
5. Long-term Access (TDR and TRAC): can the repository provide reliable, long-term access to managed digital resources to its designated community?
6. Digital curation risks (DRAMBORA): does the repository demonstrate it effectively and efficiently manages the risks associated with the process of curating digital materials?
7. Completeness: is the collection underpinned by a sound selection policy to ensure comprehensive coverage. (IIPC Selection for Web Archives)?
8. Preservation of the blogosphere: does the repository succeed in capturing and rendering something of the whole extent, nature and context of the blogosphere?
9. Sharing and Interaction: can users instantly disseminate archived content using major social web platforms; and can they easily recommend new blogs for inclusion/archiving?
10. Meeting immediate user needs: do the archived blogs participate in the overall “scholarly record” [9], and how best to preserve this?

Out of scope are considerations of the different methods of harvesting / content creation between the three methods, which will not be explored in this study.

To ensure consistency of comparison across the platforms, a defined set of interesting and exemplary blogs has been selected, each of which is available for comparison in at least two of the platforms being studied.

4. PRELIMINARY CONCLUSIONS

Preliminary conclusions of the comparative study, are:

- That preserving parsed blog content (BlogForever) offers greater benefits in terms of discovery and fine-grained retrieval than preserving entire crawled websites (as per UK Web Archive)
- That websites stored in the WARC format (UK Web Archive) are more robust and better supported as coherent, preservable digital entities
- That PDF renditions of blogs (LSE) are easier and quicker to produce than using traditional web-archiving methods, but may in turn introduce additional preservation challenges
- That renditions of blog content viewed through the Wayback Machine (UK Web Archive) are perceived as more complete with regards to look and feel, attachments and layout than pre-processed renditions stored in XML (BlogForever)
- That a user-centric platform with tags, shopping baskets and other social media features (BlogForever) addresses the needs of user communities and curators more effectively than an inflexible and non-customisable view of the data
- That research value to scholars is enhanced by maintaining and indexing an aggregated collection of micro-detail from the blogosphere (authors, tags, comments)
- Aggregated collection of textual blog content will potentially be extremely useful to text-mining projects that are concerned with finding particular types of patterns, e.g. the evolution of language used on the internet, that cannot be easily discerned through the more usual title-based approach
- That XML-based blog content, capable of being exported into numerous library and metadata formats such as MARC XML, Dublin Core and METS, offers more flexibility for interoperability and sharing than WARC
- The three methods vary considerably in their searching facilities (speed of search, intuitiveness, interpretability of results)

5. ACKNOWLEDGEMENTS

Our thanks to the BlogForever project, the London School of Economics and the British Library for their assistance in conducting this survey using their materials.

6. REFERENCES

[1] Pennock, M. and Davis, R. 2009. ArchivePress: A Really Simple Solution to Archiving Blog Content. In: *Sixth International Conference on Preservation of Digital Objects* (iPRES 2009), 5-6 October 2009, California Digital Library, San Francisco, USA.

[2] Kim, Y., and Ross, S. 2011. Preserving Change: Observations on Weblog Preservation. In Proceedings of the 8th International Conference on the Preservation of Digital Objects (iPRES 2011)

[3] Stepanyan, K., Gkotsis, G., Kalb, H., Kim, Y., Cristea, A. I., Joy, M., Trier, M., Ross, S. 2012. Blogs as Objects of Preservation : Advancing the Discussion on Significant Properties. In: iPres 2012: Proceedings of the 9th International Conference on Preservation of Digital Objects.

[4] Consultative Committee for Space Data Systems, June 2012. Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2.

[5] Library of Congress PREMIS Editorial Committee, July 2012. PREMIS Data Dictionary for Preservation Metadata version 2.2.

[6] Library of Congress Network Development and MARC Standards Office, ND. Metadata Encoding & Transmission Standard (METS).

[7] Center for Research Libraries (CRL) and Online Computer Library Center, Inc. (OCLC), 2007. Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist.

[8] Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE), 2009. Digital Repository Audit Method Based On Risk Assessment (DRAMBORA).

[9] Hank, C. 2011. Scholars and their blogs: Characteristics, preferences and perceptions impacting digital preservation (Doctoral dissertation).Available from ProQuest Dissertations & Theses database (UMI No. 3456270)