# Diamonds in the Rough: capturing and preserving serving online content from blogs

**Primary author** (First name, Surname): Maureen Pennock  
Affiliation: The British Library  
Primary author email address: Maureen.pennock@bl.uk  
Primary author mobile phone*: 07983426574

**Additional Authors:** Richard Davis (presenter)  
Affiliation: University of London Computing centre (ULCC)

## Abstract (300 words)

As the public and private sectors increasingly embrace the interactive power of Web 2.0 technologies, a growing number of organisations permit or encourage staff to blog or microblog about their work activities, using a diverse range of software platforms and applications. The value of these activities is commonly placed firmly in the 'here and now', with heightened user engagement and increased communications typically identified as primary benefits. Yet the value of these outputs is often more enduring: some are official organisational outputs and ought to be preserved as such; others offer valuable insights into the inner workings of an organisation, potentially critical cultural records for future historians or researchers. In academia, scholarly blogs created by researchers can provide important context and background to the research process, no less than traditional note-books and lab-books, and increasingly form an essential part of the institutional record of a university, faculty or project.

Preservation of personally published online blog content raises a number of interesting questions that have as yet been little discussed in the web and archiving communities. What are the requirements for archiving and preserving blog and microblogged material? What are the technical options? Can they replace or complement 'traditional' website archiving and preservation approaches? Can collections be effectively created and maintained at the institutional level?

This paper will address these and specifically consider the practical issues in capturing and preserving blog posts and comments. It will introduce "ArchivePress", a new approach for institutions large or small that wish to archive blog content. It leverages the native data structures and formats of blogs and the dynamic capabilities of associated newsfeeds and APIs, to create flexible, manageable, scoped archives of multiple blogs. A pilot case study will be presented to show how the concept can work in practice and audience feedback invited.

**Keywords**

Web 2.0, preserving blogs, web archiving, ArchivePress

**Explicit fit with conference theme (not more than 50 words)**

This proposal fits with the call’s ‘collection & capture’ and ‘preservation’ sections as an illustrative case study of preservation challenges and/or tools for collecting web archives.

Please send to info@dpconline.org by Friday the 8th May 2009 (deadline for submission of paper abstracts)