Title:
ArchivePress: A Really Simple Solution to Archiving Blog Content

Author:
Pennock, Maureen, British Library
Davis, Richard, University of London

Publication Date:
10-05-2009

Series:
iPRES 2009: the Sixth International Conference on Preservation of Digital Objects

Publication Info:
iPRES 2009: the Sixth International Conference on Preservation of Digital Objects, California Digital Library, UC Office of the President

Permalink:
http://www.escholarship.org/uc/item/7zs156mb

Multimedia URL:
http://www.cdlib.org/services/uc3/iPres/video.html?file=ipres/Pennock&title=Maureen%20Pennock%3A%20ArchivePress%3A%20A%20Really%20Simple%20Solution%20to%20Archiving%20Blog%20Content

Abstract:
Blog archiving and preservation is not a new challenge. Current solutions are commonly based on typical web archiving activities, whereby a crawler is configured to harvest a copy of the blog and return the copy to a web archive. Yet this is not the only solution, nor is it always the most appropriate. We propose that in some cases, an approach building on the functionality provided by web feeds offers more potential. This paper describes research to develop such an approach, suitable for organisations of varying size and which can be implemented with relatively little resource and technical know-how: the ArchivePress project.

Supporting material:
Presentation

iPRES 2009

THE SIXTH INTERNATIONAL CONFERENCE ON THE PRESERVATION OF DIGITAL OBJECTS

Proceedings

October 5-6, 2009
Mission Bay Conference Center
San Francisco, California

California Digital Library

ArchivePress:
A Really Simple Solution to Archiving Blog Content

Maureen Pennock[1], Richard Davis[2]

The British Library, Boston Spa, Wetherby, West Yorkshire. LS23 7BQ[1]
University of London Computing Centre, 20 Guildford Street, London, WC1N 1DZ[2]
maureen.pennock@bl.uk; r.davis@ulcc.ac.uk

Introduction

Blog archiving and preservation is not a new challenge. Current solutions are commonly based on typical web archiving activities, whereby a crawler is configured to harvest a copy of the blog and return the copy to a web archive. Yet this is not the only solution, nor is it always the most appropriate. We propose that in some cases, an approach building on the functionality provided by web feeds offers more potential. This paper describes research to develop such an approach, suitable for organisations of varying size and which can be implemented with relatively little resource and technical know-how: the ArchivePress project.

Blogs: new medium, old genre

People have been keeping diaries and journals for centuries, and early forms of diaries have been found that date back as far as the 2nd century AD. Their value to Archives is indisputable, and many Archives hold the diaries of historically notable figures within their collections. Today's modern equivalent is the blog – or online web log.

The term 'weblog' was coined in 1997, though early online diaries pre-date this. These early blogs were simply regularly updated websites, and it was not until late 1998 – '99 that modern blogging platforms such as LiveJournal and Blogger began to appear. Regardless of their supporting software, blogs now typically almost always follow a dated post/comment structure and provide web feeds (Atom or RSS) to alert users of new posts or content. It is upon these main features that the ArchivePress project will draw.

The need for a new tool

The common and familiar scenario is that an organisation runs a web crawler to capture copies of content – in this case, blogs – and provides subsequent access to the web site (blog) as an integral whole. This is perfectly acceptable if the requirement is that the site is presented as an integral whole. It allows institutions to capture a copy of a website as an historical record, even specifying the depth of crawl and omitting certain parts of a site if necessary. Yet whilst this approach successfully renders blogs as a set of hyperlinked web pages, it overlooks the utility of the blog as a rich data object and does not capitalise on the kind of functionality and flexibility inherent in structured data stored and manipulated in either a relational database or with XML. ArchivePress is based upon the premise that not only is this a desirable utility, but also that organisations can have different reasons for wishing to capture copies of blog content, and different intentions for using the content once they have it.

For example, an increasing number of institutions allow staff or users to contribute to blogs in a professional or institutional capacity, on internal or externally hosted platforms. In particular, academic and scientific sector organisations increasingly use blogs, internally and externally, for many activities that merit preservation as part of the institutional and scientific record. As a matter of recordkeeping, the institution may wish to capture and consolidate the raw content on such blogs into one easily manageable and searchable resource – and typical 'crawler based' web archiving approaches do not easily facilitate this. In other cases, cultural heritage organisations may wish to generate a collection of blog content around a given theme and to present it as a single and searchable resource, rather than a set of individual blogs.

For these organisations, it is the raw and aggregated content, the structural relationships between each content component, and the accompanying metadata (eg date headers, author attributions) that may be most important. Typical 'crawler-based' web archiving approaches do not easily facilitate an approach that distinguishes between content components in this manner, particularly since for them the 'atomic' unit is the HTML web page structure, not individual blog posts or comments as discrete entities and data objects. Another complicating aspect of current approaches to automated web preservation relates to calculating the optimal frequency of harvesting and providing the computing and network facilities to continually re-crawl and re-collect entire websites. Explicit signals of significant updates to websites in general are many and varied, where they exist at all. Crawler schedules are not therefore typically established in response to such signals, and thus are not usually responsive to the direct

publication of new content. In contrast, the use of newsfeeds and APIs as triggers for keeping 'up-to-date' with cumulative web content is well established.

All of this suggests that typical 'crawler' based approaches, whilst satisfactory in many cases, are not always as flexible as might perhaps be required, and that a feed-based approach could better serve institutional owners and future owners, particularly where aggregated and real-time re-use of content is a requirement. It is for this purpose that the ArchivePress project was established.

The ArchivePress project

The ArchivePress project addresses what we perceive as a gap in the market and has been developed in response to the changing nature of content published on the web. It is a collaborative initiative between the University of London Computing Centre (ULCC) and the British Library (BL). Funded by the UK Joint Information Systems Committee (JISC), it brings together the combined experience of experts in digital preservation and web archiving, particularly in such initiatives as the JISC PoWR project and the UK Web Archiving Consortium (UKWAC), to address the challenges and potential of blogs as a class of dynamic information resource, distinct from other types of websites.[1]

ArchivePress is intended to be a lightweight, off-the-shelf, blog content archiving tool that can easily be installed by organisations and projects, in order to facilitate preservation of blog output by multiple individuals or groups potentially using diverse blogging tools, platforms, or hosts. It works by drawing on the full power of the WordPress plug-in system and the rich functionality offered by associated RSS feeds. The project's target group is, in the first instance, academic institutions, though we appreciate that the tool has far broader reaching applicability, particularly in the cultural heritage and scientific sectors.

The project has three key aims:

- To create plug-ins for WordPress that will enable it to work as a blog archiving tool;
- To create demonstrator instances of the ArchivePress system in use, using groups of blogs nominated by institutions with ongoing, active blogging outputs, in order to demonstrate, compare, and analyse the results of the project approach;
- To assess the effectiveness of this approach and promote discussion and debate among the community including web managers, bloggers, and web archiving specialists.

[1] JISC PoWR (Preservation of Web Resources) - http://jiscpowr.jiscinvolve.org/; UKWAC - http://www.webarchive.org.uk/ukwa/ . (retrieved: Sept 17th 2009).

Legal issues, including permissions and copyright are out of scope of the project, though we recognise that users of the tool need to ensure they have appropriate permissions before collecting and repurposing content.

Participating institutions are:

- The UK Digital Curation Centre (DCC). DCC runs at least three blogs on Blogger.com. Most posts are text only.
- UKOLN at the University of Bath. UKOLN staff run approximately twelve blogs on a variety of platforms, some on their own self-hosted installations of WordPress, some at WordPress.com, and others on Blogger.com. Posts typically include a range of content, including text, images, and embedded slideshare files.
- Lincoln University. Lincoln provides a multi-user WordPress instance for staff use. Approximately nine are currently available to the project. As with UKOLN, posts typically include a range of content types.
- The British Library. The BL provides a Typepad platform to support corporate blogging activities. Approximately twenty blogs are currently hosted, incorporating a range of content from alphanumeric text (including unusual characters) to embedded multimedia files.

The main outputs of the project include:

- An open source WordPress plug-in based on the Wordpress plug-in API. This is being developed using Google Code, and is published and publicised using the WordPress themes website;
- Supporting research into academic attitudes towards blog archiving and the significant properties of blogs;
- Documentation and guidelines on installing and configuring the plug-in(s), and on managing the archive.

The project has so far explored two key aspects: academic attitudes to blogging and blog archiving, and development of the demonstrator tool.

Academic attitudes to blog archiving

An information gathering exercise was launched to identify academic participants' attitudes to blogging and blog archiving. The exercise aimed to elicit information on:

- How staff typically digest blog content;
- Why they blog and the values of blogs;
- Current blog archiving practices;
- Which elements of blogs were most valuable;
- Which elements of blogs ought to be preserved.

Twelve participants were sampled from our three participating academic institutions. They represented a variety of users including technical staff, managers, subject librarians, researchers, and learning & support coordinators. Responses overall indicated a high level of support for the project and its underlying premise that not all blog content requires preservation. Almost all participants either had their own blog or contributed content to an organisational blog, with only one exception. All reported using content from blogs from academic research, though they were equally divided between occasional use and regular use, and all users reported leaving comments on blogs managed by other people.

Participants were asked how they typically digest blog content. Three quarters of participants utilised RSS feeds, though most of these followed this up with a direct visit to the website. Only two participants relied on the RSS feeds alone, without recourse to the host website.

When asked to identify the most valuable functions served by their blogs, the majority of participants said it was to discuss ideas (Average Rating: 4.55), closely followed by dissemination of ideas (AR: 4.36). This was followed by providing a record of personal professional impact (AR: 4.09) and a record of personal activities (AR: 3.55). Citation was marked relatively lowly in comparison (AR: 3.05).

Participants were asked which measures they relied upon (if any) to currently protect/archive their blogs:

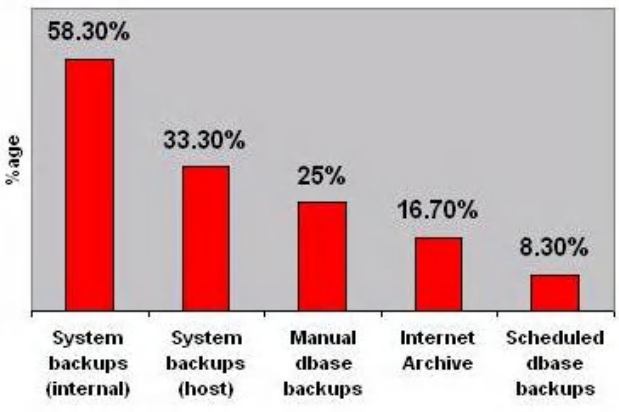


Table 1: What measures do you rely on, if any, to protect and archive your blog content?

Most bloggers rely upon system backups performed by the institution to protect and archive their blogs. None of the participants used the UK Web Archive or any other eb archiving service, other than the Internet Archive (IA). Upon checking, we discovered that the Internet Archive does not currently contain archives of any blogs identified by participants. One institution even blocks the IA crawlers through use of the robots.txt file, so none of its blogs are captured – a fact that our participant was unaware of. Another participant submitted her blog URLs to the IA around the time of the survey so they are not yet accessible (there is a six-month delay between submission of a URL/crawl and access to the archived site via the IA's WayBack machine).

Participants were asked 'what do you think are the most valuable elements of a blog?' and to rate options on a scale of 1 – 5 (5 being most valuable):

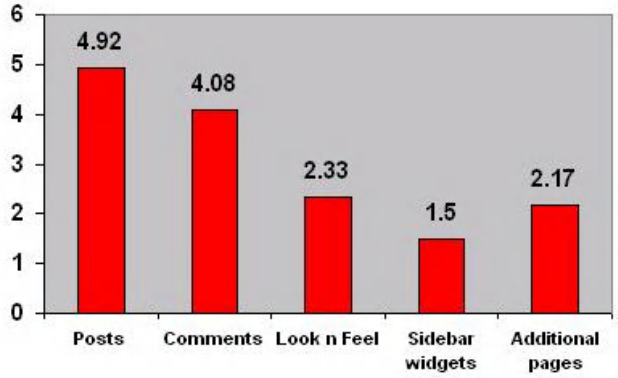
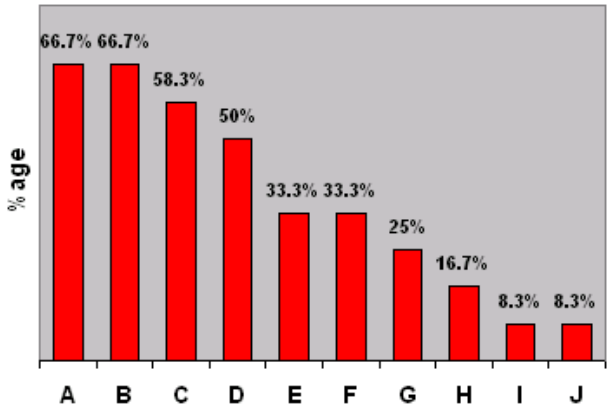


Table 2: What do you think are the most valuable elements of a blog?

Figures in the table are the average ratings (from a total possible score of 5). A forced rating system was in operation, with respondents placing features in order of value. Eleven out of twelve participants stated the posts were most valuable, followed by comments. One respondent selected comments over posts. Participants were almost equally divided over the third most important feature, with six selecting 'look and feel' and five selecting additional pages; most relegated sidebar widgets to last place. This supports the ArchivePress premise that posts and comments are more important than 'peripheral' features, though one participant noted that additional pages can have contextual value if they provide personal information about the blogger – 'it helps credibility to know the environment the blogger is coming from'.

Participants were also asked which types of data they thought blog archiving applications ought to capture. This question went into more detail than the question on valuable elements, and was intended to identify additional data on, for example, supporting metadata. Participants were able to select as many options as they wished.



KEY:

A. Posts
B. Comments
C. Blog name & URL capture
D. Tag/category names
E. Embedded objects (eg images)
F. All of it!
G. Additional pages
H. Author profiles
I. Commenter profiles
J. Associated feeds (e.g. from microblogging services such as Twitter)

Table 3: Which types of data do you think blog archiving applications should capture?

All participants thought that posts and comments should be captured, though a third suggested that this should take place as part of capture of the whole site, i.e. that all of it should be captured. Breaking down the demographics, these responses typically came from non-technical staff, mainly librarians, all of whom currently access content first via RSS feed then via direct visits to the website.

In terms of supporting information, almost all participants believed that blog name and URL should be captured, followed by tag and category names. Surprisingly few people placed value on capture of author and commenter profiles. This could be for a variety of reasons, for example, thinking only of short term re-use by existing users, not new users who have no familiarity with the original context of the post or posters.

The relatively small scale of the survey means that we cannot interpret the findings as overwhelmingly conclusive. Nonetheless, they represent of a range of user-types within our target audience, and are indicative of the responses a larger survey may elicit. They are, by and large, in keeping with the much larger survey carried out in 2007 by Carolyn Hank et al into 'Blogger perceptions on digital preservation', and we intend to follow up our initial work with a larger scale survey over the next few months.[2] In the meantime, we draw the following tentative conclusions:

- Content posted to blogs has value to bloggers and readers for a range of reasons. Whilst real-time interaction in the form of discussing and disseminating ideas is the most valuable function of a blog, their value as records of personal professional impact and personal activities is also appreciated.

- A blog archiving service that targets and stores only certain types of content from blogs is an acceptable approach

- Such a service should focus primarily on archiving blog posts and comments. Tag & category names, and blog names & URLs should also be captured. Capture of embedded objects may also be desirable, though look and feel is not so important.

- 'Peripheral' features, such as associated feeds and author profiles are not a priority for most users (though it is this author's belief that profiles may have longer term value than identified within the confines of our survey results).

Significant Properties

The INSPECT framework defines significant properties of digital objects around the high-level categories of Content, Context, Structure, Rendering, and Behaviour.[3] We are drawing upon input from project participants (above) and the expert knowledge of the project team to develop a set of significant properties for blog content based around the same framework. We perceive significant properties as a set of features specific to a content type and a given preservation objective. In this case, the objective of preservation is to preserve content from blogs for re-use, rather than to preserve the entire blog website. Version 1 of the model includes the following features:

---

[2] Hank, Carolyn et al *Blogger perceptions on digital preservation*. Poster for JCDL, 2007 (retrieved 18 September 2009). http://www.ils.unc.edu/%7Ehcarolyn/blogsurvey/poster.pdf; Hank, C., Choemprayong, S., and Sheble, L. 'Blogger perceptions on digital preservation' in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (Vancouver, BC, Canada, June 18 – 23, 2007). JCDL '07. ACM, New York, NY, 477-477.

[3] Knight, Gareth *Framework for the Definition of Significant Properties,* 2008 (retrieved 18 September 2009) http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf.

- **Content:** Blog posts; Blog comments; Embedded objects; Tags/Categories.
- **Context:** Blog title & URL; Primary contributors; primary contributor profiles; Post authors; Content dates.
- **Structure:** Post IDs, Comment IDs; Component relationships
- **Rendering:** Text formatting
- **Behaviour:** Hyperlinks

The set of properties will be revised during the course of the project upon further input from participants and feedback to the demo installations using content from participating institutions.

Technicalities

WordPress is arguably the most widely used blogging application for self-hosted blogging. It is also the engine behind many prominent free and commercial blog providers, including WordPress.com, Edublogs and JISCInvolve. It is Open Source, GPL licensed, based on PHP and MySQL, and uses well-defined open data schemas (using SQL natively and XML for data exchange). These features make it eminently suitable for digital preservation applications. It also has rich inbuilt functionality, including support for importing content from a range of diverse blog platforms (Blogger, Typepad, MT, LiveJournal) using a variety of approaches (such as RSS, export XML, APIs).

Many third-party plug-ins have been developed: one in particular enables WordPress to be used as an aggregating blog, automatically importing posts from other blogs via RSS feeds. This is the FeedWordPress plug-in.[4] The ArchivePress project builds on that approach and has developed a plug-in to enable an appropriately configured WordPress instance to harvest not only posts, but also comments, embedded content and rich metadata from other blogs, regardless of platform, as long as an RSS/Atom feed is available.

The main ArchivePress plug-in is currently being developed, using PHP. On completion of the project it will be available for download from the project website and Google Code. It will require a compatible WordPress installation of WP 1.5 or higher with the FeedWordPress plug-in installed. Configuration depends to some extent on the requirements of the user institution, though basic configuration is intended to be as simple as possible, so that the archive manager can set a blog to be archived by providing the following information:

- Enter the main feed URL for blog;
- Select whether comments are required;
- Select whether embedded content is also to be imported.

Further actions will also be supported, such as providing descriptive information about the post authors and the archived blogs, and any additional metadata that may be necessary. The extent to which different configuration options may be required is still being ascertained, based on ongoing feedback to the demo systems populated with sample content from participating institutions.

Other third-party plug-ins have also been identified that provide additional support and add value to the archive as a tool for research. For example, the Academic Citations (AC) plug-in adds text citations in various common formats (Harvard, Chicago) to the foot of a post, and a modified version of the same plug-in displays a metadata summary for each post. ArchivePress is therefore emerging as a suite of complementary tools, rather than a single addition.

Data is thus harvested from blogs and archived in a single WordPress instance under the control of the institution running ArchivePress. We recognise that the creation of archives in this manner does not of itself fulfil all the goals and requirements of digital preservation, any more than setting up a repository or crawler. It will, however benefit many key preservation activities, such as selection, classification, access and migration, within any broader digital preservation system. Furthermore, we believe the open source nature of the installation, along with our use of standards and the resulting highly structured datasets, will facilitate digital preservation with a minimum of subsequent interventions to the format of the data.

Limitations of the approach

ArchivePress relies on RSS feeds for delivery of content. Capture is therefore limited to that content which can be delivered by RSS feed. Required RSS 2.0 elements are:

- Title (the name of the blog)
- Link (URL)
- Description (in brief)

However there are far more optional elements, including:

- Language
- LastBuildDate (date on which content last changed)
- Generator (source software)
- ManagingEditor (person responsible for editorial content)
- Copyright (copyright notice)
- Item elements:
    - Titles (post title)
    - Link (URL)
    - Comments (URL)
    - PubDate (publication date & time)

[4] Johnson, Charles *FeedWordPress plug-in* http://wordpress.org/extend/plugins/feedwordpress/ (retrieved 18 September 2009)

- Author (name of author)
- Category (categories allocated; repeatable)
- Enclosure (describes media object attached to item)
- Guid (a string that uniquely identifies the item)
- Description (introduction)
    - Content (post content)

Since both Atom AP and RSS are XML applications, any number of other embedded, namespace-qualified elements can also be included by the application generating the feed. One external namespace of particular significance is the Well-Formed Web (WFW) project: this provides extensions for RSS which specifically support better handling of comments.[5] WordPress blogs include the WFW tag "commentRss" for each post, which is a link to the feed of comments for that post. It is essential for the ArchivePress approach that the harvester is able to automatically locate not only the blog's feed of posts, but also each post's feed of comments.

In addition to comments, and embedded content hosted at the target blog, it remains to be established if ArchivePress can effectively harvest content from external sites embedded in blog posts (for example, YouTube videos). This may be an issue in some cases, for example with academic blogs in which the clarity of a post is dependent on embedded resources from external sites; when such external content is not captured, it can become more difficult to understand the argument being made in the post and the subsequent comments. In such instances, apart from a link to the content in its original location, ArchivePress may not be able to capture all significant properties of blog content and an extended solution sought.

AP1: The DCC collection

The first ArchivePress Demonstrator (AP1) has been set up to harvest and re-present content from three DCC blogs: The DCC Digital Curation Blog; the DCC BLawG; and the Research Data Management Forum blog, all of which run on the Blogger platform. Posts are contributed by a total of eight different staff from across the DCC project.

AP1 uses the existing FeedWordPress plug-in, which harvests only posts and associated metadata. Comments functionality is being developed under v2.0 of ArchivePress. Blogger, the host software for all blogs in AP1 delivers posts via RSS and/or Atom. AP1 uses the RSS option, though we plan to research closely the differences between the RSS and Atom content to establish if either offers richer content. The AP1 demonstrator was a first opportunity to explore how the installation needed to be configured in order to capture and re-present all required information in the required order. It represents our initial efforts and has been a valuable learning experience.

Different access interfaces are under development and are posted on the ArchivePress blog for feedback. The current access interface provides a homepage to explain the context of the Archive, its source blogs, authors, and the main features of the archives. Author profiles are not yet enabled but will feature in a future version of ArchivePress. Users can immediately access archive content by author via the homepage, by tag, or by date.

Users can tab through different pages of the site. For simplicity, these are currently restricted to the 'Home' page, and the 'Archive' page which lists all the harvested content. Future iterations may present further options, for example with further information about the source blogs. Other views are also possible – one option we wish to explore is a more 'repository-like' interface, with the homepage featuring options to 'view latest additions', 'search repository', 'browse repository' etc.

The 'Archive' tab on the current interface offers access to all posts from the collection, listed by default in chronological order (most recent first). A by-line under the post title identifies the source blog, author, and date of posting. Embedded hyperlinks are active but most link to external resources: the longevity or completeness of such content, as noted earlier in the paper, cannot therefore be guaranteed. Blog post headings are hyperlinks to separate pages containing content from that post only: this is not obvious in the current version and will be amended in future versions to accord with the accepted convention for denoting hyperlinks. Individual pages provide further details on the content, including the time of posting.

Most of the content in the DCC collection is text based: none of the posts we have captured to date have contained images or other embedded objects. A validation process has been initiated to check that all required content has been captured and is accessible. This has already revealed problems in three areas:

- Tags and categories have been captured in the RSS feed but are not displayed with posts: this is due to a configuration issue and is being remedied in ArchivePress 2.0.
- Presentation of metadata in associated plug-ins, namely the Academic Citation plug-in and the modified AC plug-in for metadata. Neither plug-in is consistent in showing author details: we believe this to be another configuration issue and aim to remedy it in V2.0 of ArchivePress.
- Links to the original blog are redirected to point to all content from that blog on an ArchivePress page: it must be clear whether the links are intended to point to the Archive or not.

AP2: The UKOLN collection

AP2 Demonstrator has been set up to harvest and re-present content from a collection of twelve UKOLN blogs. Most run on a UKOLN-hosted multi-user instance of

[5] Well Formed Web project: http://wellformedweb.org/.

WordPress; some run on self-hosted installations of WordPress; others at WordPress.com, and others on Blogger.com. Most blogs feature posts from a single author, though a small number are multi-authored.

AP2 is intended as a testbed for the development of ArchivePress 2.0. We are experimenting with harvesting comments and have discovered that this requires a different approach depending on the hosting platform of the original blog. WordPress is currently the only platform we have identified that explicitly declares the location of each post's comments feed: the ability to know, or at least accurately infer, the address for the comments feeds is essential to achieving the best results for a blog archive.

Three blogs are currently represented in AP2, all from different platforms, though we intend to harvest content from all twelve sources when the demonstrator is fully established. We expect that the number of blogs in this collection will produce a lengthy tag/category list and that usability of the list may be compounded by the diverse nature of the blogs and different users' usage of the same tags. Both issues will be explored during the remainder of the project. Other activities for AP2 and the remainder of the project will include:

- Resolving outstanding issues from ArchivePress 1.0 (mostly rendering issues)
- Clarifying the differences between content harvested from Atom newsfeeds and those from RSS feeds
- Exploring how TypePad, Blogger and WordPress differ
- Identifying a core set of metadata requirements and mapping newsfeed elements to the core set
- Involving participants in evaluation and validation of the demonstrator collections
- Assessing options for collecting copies of embedded content, including for example SlideShare files and YouTube videos.
- Writing installation and configuration guides to support future ArchivePress administrators and coders who wish to extend the plug-in further.

Conclusions

ArchivePress is intended to offer a way for institutions, regardless of their size, to collectively harvest blog content and repurpose according to their needs. The tool draws on features of the WordPress software and RSS technology that are conducive towards preservation and will, in our opinions, generate sustainable corpora of blog content that are relatively easily accessible, for as long as required.

Our research so far has indicated that the underlying premise to ArchivePress is sound: for many users, it is acceptable to archive only certain elements of blogs, i.e. those with primary importance and re-use value. Our technical work is proving that this can be achieved through relatively (compared to 'typical' web crawling activities) small-scale development work. Though we have encountered unexpected intricacies in configuration issues, particularly relating to variations in feed content from different platforms and subsequent configuration requirements, we expect to resolve these and document them in such a manner as to make use of the plug-in a straightforward and low-resource activity.

ArchivePress is an enabling tool that will not only facilitate archiving and preservation of blog content, but will also produce new corpora of aggregated blog content that have as yet unknown research potential. From this perspective, it is not only an archiving tool, but also a curatorial and re-purposing tool. Ongoing evolution of the web demands that we continue to adapt and develop new approaches for archiving and managing web-based, or 'born online' content', in order to take advantage of and respond to all the opportunities and complexities it entails not only now but also for the future. ArchivePress is one such solution.

References

ArchivePress project: http://archivepress.ulcc.ac.uk/

Hank, Carolyn et al Blogger perceptions on digital preservation. Poster for JCDL, 2007 (retrieved 18 September 2009). http://www.ils.unc.edu/%7Ehcarolyn/blogsurvey/poster.pdf

Hank, C., Choemprayong, S., and Sheble, L. 'Blogger perceptions on digital preservation' in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (Vancouver, BC, Canada, June 18 – 23, 2007). JCDL '07. ACM, New York, NY, 477-477.

JISC PoWR (Preservation of Web Resources) project: http://jiscpowr.jiscinvolve.org/

Johnson, Charles *FeedWordPress plug-in* http://wordpress.org/extend/plugins/feedwordpress/ (retrieved 18 September 2009)

Knight, Gareth *Framework for the Definition of Significant Properties,* 2008 (retrieved 18 September 2009) http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf

UKWAC: http://www.webarchive.org.uk/ukwa/

Well Formed Web project: http://wellformedweb.org/