Practical Blog Preservation

Richard M. Davis
University of London
Computer Centre

A Workshop for IWMW 2009
28th July 2009
Practical Blog Preservation

1. Introductions

2. Blogging and digital preservation in HE:
   - What have we got?
   - What do we want?

3. Breakout & discussion

4. A possible approach

- Identify issues and scenarios
- Identify possible solutions ...
- ... and which to use when
- An effective plan to preserve blogs at your institution
What is digital preservation?

“a series of managed activities necessary to ensure continued access
[...] for as long as necessary”

Digital Preservation Coalition, 2002
What are blogs?

• “Web logs” - online diaries

• Chronological format

• Single- or multi-author

• Internal or public

• One-way or two-way (comments)
What are blogs used for?

• Sharing ideas
• Disseminating information
• Eliciting feedback
• Personal reflection
• ...
Who uses blogs?

- Students
- Teachers
- Researchers
- Project teams
- Admin departments
- Who doesn’t?!
Animal rights activist placed on terrorist list
April 24, 2009 at 12:40 pm - Filed under News Items

A vegan computer specialist from America has been placed on the terrorist list, which up until today has only included alleged Islamist terrorists.

The FBI have been quoted as saying “Animal rights and environmental extremism pose a significant domestic terror threat.”

This is a perfect example of when animal rights activism can be taken too far and to the extreme. The man in question, Daniel Andreas San Diego, has been accused previously of causing property damage with an explosive device and has been on the run since 2003.

Read the rest of this entry »

Learning blogs
Researcher blogs
Institutional blogs
Why are blogs important?

- Communication
- Discussion
- Communities, connections, contacts
- Personal publishing platform
- Record of activities
- *Different from “ordinary” websites?*
Why should institutions preserve blogs?

- Record of institutional activities:
  - Social
  - Professional
  - Intellectual
- Study, research, re-use
- Citation, quotation, reference
- Alternative to journal articles(?)

Who should preserve blogs?

- Institution
- Individual

National archives/libraries

Internet Archive etc.
What if we didn’t preserve blogs?
Blog preservation: some stakeholders
What are the issues in preserving blogs?

- Policy
  - Selection
  - Retention

- Technology
  - Tools for capture
  - Tools for management
  - Tools for access
  - Formats, obsolescence?

- Resources
  - Whose responsibility?
  - How much will it cost?

- Copyright, IPR
  - Who owns the content?
  - Are we allowed to copy it?
How do we preserve blogs?

Selection

Selection: Bloggers were asked to identify characteristics for appraisal.

<table>
<thead>
<tr>
<th>Very Important or Important</th>
<th>Characteristic</th>
<th>Very unimportant or unimportant</th>
</tr>
</thead>
<tbody>
<tr>
<td>88.6%</td>
<td>Unique/innovative content</td>
<td>3.3%</td>
</tr>
<tr>
<td>84.4%</td>
<td>Cultural significance (blogger)</td>
<td>3.8%</td>
</tr>
<tr>
<td>75.3%</td>
<td>Popularity</td>
<td>8.7%</td>
</tr>
<tr>
<td>69.7%</td>
<td>Number of links to blog</td>
<td>8.6%</td>
</tr>
<tr>
<td>66.8%</td>
<td>Number of regular readers</td>
<td>9.8%</td>
</tr>
<tr>
<td>66.3%</td>
<td>Number of hits</td>
<td>14.2%</td>
</tr>
<tr>
<td>65.4%</td>
<td>Preservation preference</td>
<td>10.4%</td>
</tr>
<tr>
<td>63.4%</td>
<td>Unique/innovative functionality</td>
<td>12.6%</td>
</tr>
<tr>
<td>54.3%</td>
<td>Unique/innovative graphic attributes</td>
<td>16.3%</td>
</tr>
<tr>
<td>52.5%</td>
<td>Popularity of (other) blogs linking to blog</td>
<td>17.5%</td>
</tr>
<tr>
<td>33.9%</td>
<td>Number of occasional readers</td>
<td>20.7%</td>
</tr>
<tr>
<td>32.6%</td>
<td>Age of blog</td>
<td>26.5%</td>
</tr>
<tr>
<td>23.5%</td>
<td>Size of blog</td>
<td>30.8%</td>
</tr>
</tbody>
</table>

Hank et al. 2007.
How do we preserve blogs?

*Content*

Bloggers identified the following elements as most important (ranging from very important to important) to capture:

- Text: 98.3%
- Images: 95%
- Video: 88.9%
- Audio: 88.4%
- Other multimedia: 81.9%
- Comments: 78.9%
- Keywords – e.g., tags: 71.1%
- Linked content: 62%
- Imported images – e.g., flickr: 51.7%
- Graphic layout/design: 50.6%
- Imported tags – e.g., del.icio.us: 41.3%
- Bookmarks – e.g., blogroll: 40.8%
- RSS feeds: 39.1%

Hank et al. 2007.
Capturing blog content

• Possible approaches:
  • Web crawling and harvesting
  • Database backup
  • XML export
  • RSS feed harvesting
Providing access to archived blog content

• What sort of system would we like to provide?

• Access via library database portal?

• Full-text, keyword searching and indexing?

• Unified view, e.g. of single author contributions across multiple blogs?

• Respect original access restrictions?

• Conflict between ‘live’ and archived content?
Who owns the content?

- Copyright issues: does your institution have a policy on...
  - Content created as part of employee duties? In which case copyright may reside with employer?
  - Content created as part of a course by students? In which case copyright probably is with the student?
Who owns the content?

Brian Kelly talking about UK Web Focus blog:

“A rich copy of the contents of the blog will be made available to UKOLN (my host organisation) if I leave. Note that this may not include the full content if there are complications concerning third party content (e.g. guest blog posts, embedded objects, etc.), technical difficulties, etc.

“Since the blog reflects personal views I reserve the rights to continue providing the blog if I leave UKOLN. If this happens I will remove any UKOLN branding from the blog.”

Kelly, 2009
Who owns the content?

- Suggested approaches:
  - Explicit statement of policy with regard to rights, ownership and long-term access (including if author leaves)
  - Explicit Creative Commons licensing statement on blogs, regarding
    - For blogger content
    - For comments
  - For embedded content: only use material that is appropriately and explicitly licensed
Breakout Exercise: Preserving Institutional Blog Content

Your University has established a Blog Server. All current students and staff will be able to create private or public blogs.

You have been asked to research the policy implications and technical requirements for mid-to-long term access to the blog content, and make recommendations.

1. Try to identify at least 3 key issues and suggest ways to address them. Describe the advantages and any disadvantages of your solutions.

2. Suggestions you have already received include:
   - Deleting all student blogs on graduation
   - Printing staff and project blogs to PDF and storing them in the IR
   - Making all staff and students sign a copyright agreement with the University

Are any of these suggestions acceptable? If not, why not?
Resources and further reading


• JISC: ArchivePress project. http://archivypress.ulcc.ac.uk/


Credits


• Screenshots:
  • http://learninglab.lincoln.ac.uk/blogs/oj1/
  • http://wwmm.ch.cam.ac.uk/blogs/murrayrust/
  • http://www.arts.ac.uk/index-blogs.htm
  • http://www.bloggled.com/

• SouthPark cartoons created using SP Studio. http://www.sp-studio.de/