Brittan Dunham
Digital Preservation
Howard Besser
December 15, 2010

Saving the Dissident Deep:
A Theory on Designing an Independent Strategy for Personal Blog Preservation

In summarizing what a blog is, Pamela Smith of the reBlog project also explains their importance: “The web log or "blog" is a specific form of web page that proliferates throughout the web, allowing users to publish electronic content second by second, day by day. The blog exists in a public space but there is a special first person immediacy and freedom in a blog post that feels more like a personal record made public.”[1] Indeed, blogs offer anyone the opportunity to add as much or as little as they like to the continual flow of online dialogue. They have taken various forms, from major sources of political news and opinion to outlets for family pictures and diary entries. The reBlog project is one of the first to explore curating blog content. This includes selecting blog posts and republishing them on the Eyebeam reBlog site, adding curatorial comments and images, introducing the blogger (often an individual, not affiliated with a major website or group), organizing posts and otherwise presenting personal blog posts as one might a selection of film or art. The project argues the importance of building a collection of blog posts and archiving them through Eyebeam. Building on Smith’s ideas, individual bloggers may find it pertinent to begin thinking of their blog posts as a collection that needs to be cared for and preserved.

[1] http://www.eai.org/resourceguide/preservation/computer/casestudy_reblog.html

The scope of preservation needs facing web archivists today is vast. The Library of Congress admits that its first challenge is to "define the scope of what to collect," and goes on to say that it is "impractical to collect everything on the web, so the Library focuses on significant themes or events."[2] Because of this, the main focus of the archiving community has to be on the broad, rather than the deep, on preserving a wide range of content on a subject to give that subject context, for example. However, we may find that it is also important to maintain the full history of an individual blog, even a personal blog, including all of the layers that make up its content and appearance. Personal blogs offer individual commentary and exhibition on a level that has never been possible before. Their typical layout offers a diary-like record and the opportunity to share links, embed media content and allow comments makes them an active part of the web. Thus, bloggers have the potential to become something like curators themselves, pulling together content from multiple sources and presenting it with commentary. These personal blogs should be preserved for both the role they are playing in society and for their unique content, but perhaps for now the responsibility should fall on the creators to preserve these records. Until the technology is such that personal blog content can be easily preserved and supported by archives, blog creators should at least have the tools available to them to preserve their own online content.

Crawl services like Archive-IT are attempting to preserve web pages with regular, scheduled capture and make them available for access on a third party host such as The Internet Archive's Wayback Machine. While these services appear to be

[2] http://www.digitalpreservation.gov/videos/docs/webarchiving_video_transcript.pdf

a useful way to capture the content on a website on any given day, there are many downsides that might keep the non-professional blogger from committing their personal content to them, or at least from relying on them entirely. Archive-IT, for example, requires a paid subscription. It also requires that users trust it to accurately capture and host their archived content. In our experience using Archive-IT in the Fall 2010 MIAP Digital Preservation class, we found that many of the "archived" sites had embedded media missing, broken links and did not maintain their layout. I am interested, instead, in exploring the possibility of a rogue system by which bloggers can capture their own information and preserve it in a secure way for future re-publishing or static access. I don't propose to present that tool, only to offer suggestions as to how we might arrive at one.

The first step in assessing the feasibility of independently preserving personal blogs is to set boundaries. First, limiting the process to blogs hosted on popular third party sites that offer preset layouts and plug-ins, such as Blogger, WordPress and Tumblr, keeps it simple and fairly widely accessible. Bloggers with knowledge beyond that should be able to adapt these steps to fit their needs. It is then important to look at what it would take to create a Digital Asset Management System that non-archivists can understand or be taught. There are three tiers to consider in outlining a DAM for blogs, and the solution to personal preservation may lie in further exploring them.

The first is target file archiving to preserve content, layout and style. For preservation purposes, the two main factors that make a blog different from most static websites are regular updates and the amalgamation of content from many

different places on the web. Regular updates, both in posts and comments, inform how frequently a blog needs to be preserved. (The designation of "preserve" in this context and throughout the rest of this paper refers to the all-compassing set of methods necessary in order to capture content in a complete sense). A blog's stylesheet, for example, is its most static element and may only need to be preserved once, or once each time it is changed or updated. CSS handles the appearance and visual behaviors of elements on a blog, and separates content from style. A blogger can simply alter the CSS file associated with the site to change its appearance, and the web browser downloads any style sheets linked to in an HTML header and applies them to the page. A well-archived blog would need to maintain all CSS changes for each blog post or other site update. This technique will ensure that any viewer can see blog posts exactly as they appeared on the site at the time of posting and will also document all visual transitions over the period of time that the stylesheets are archived. One thing to keep in mind is that any files referenced in the CSS need to be properly maintained to assure that they remain as they were at the time it was saved. This includes archiving all file-based elements used in the blog and coming up with a file naming convention that differentiates between each element. For example, a background image with the filename background.jpg, if replaced by another background later that is also saved as background.jpg, would result in the loss of the original background image in favor of the new one. Using timestamps on the filenames and updating the resource locations in the CSS will counter any issue of overlapping filenames.

Similarly, a blog's layout only needs to be preserved as often as it is updated. The layout may be a preset HTML template, a variation on a preset template (on sites like Blogger, Wordpress and Tumblr, bloggers with even limited HTML knowledge may make slight alterations to preset templates), a layout designed by a third party or a layout designed completely by the blog creator using more advanced HTML. In any of these cases the HTML template or code should include a space that is filled in with a post each time the blog is updated. It seems to make sense then that the HTML template and the content inside it could be saved as one file by viewing the blog's source code. To be safe, it would probably also be wise to save just the layout HTML (found, in most cases, on the main edit page). After the initial save, all future posts should be saved one at a time and a portion of the code in the HTML template should filter in new posts when the blog is republished. If nothing else, at least all components that make up the appearance of the blog will be saved. Other backup options for preserving the look of the layout at a given time might be to take screen grabs and save PDFs and Word documents of the HTML and its visual output, so that they may be reproduced if the saved code does not suffice.

While stylesheets and HTML templates may be saved on an as-needed basis, the blog content will have to be more routinely preserved. This includes all aspects of the physical content: embedded video and audio, photos, fonts, formatting and links. As discussed before, careful archiving of the file-based elements is key, including naming the files, organizing them in folders and sub folders on an accessible internal hard drive (with backup copies stored elsewhere) or, if they're stored online, making sure their correct location is referred to in the code. Most

major blog hosting sites offer and XML-based RSS feed that can be used to keep up with and save posts. An RSS aggregator may be used to trigger the frequency of these saves, and the blogger-turned-archivist must remain diligent in doing additional saves if changes are made to old posts that aren't captured by the RSS aggregator, or as comments are left on posts. Some blog sites offer the option to subscribe to comments, so this might also be an option. Several decisions need to be made here, beginning with an assessment of the tradeoffs between frequent and less frequent grabs. With each save, should you grab everything or just new content? What if old content has been changed, either by being edited, affected by layout changes or updated with comments? It is also important to take into consideration the intellectual aspects of the content: rights issues, relevance of posts, comments, etc. when deciding how often and how deep to go in saving posts.

The next tier in creating a Digital Asset Management System for preserving blogs is planning backup methodology. There is always the chance that a blog-hosting site will crash, be hacked or fade away in obsolescence to the newest host site or technological trend in blogging. For that reason, it is important that bloggers be proactive in preserving their own content, so that nothing is lost and they are able to migrate forward as the blogosphere changes. This begins with local storage, saving all of the files, as discussed above, to an internal hard drive, backup copies of each on an external hard drive and possibly copies on a flash drive. This is a good place to start, but to remove the concept of a single point of failure it is a good idea to also have off-site offline storage. This could mean simply having extra copies on an external drive kept at a separate geographic location. Though less important than

offline storage, one might consider the added benefit of data redundancy using a RAID array to preserve file integrity. This is not a backup solution, and may not be a viable choice for the non-archivist, but does give an added level of security when preserving file-based materials.

Finally, cloud storage offers multiple online sources for storage in case of DNS or Internet outages, and is accessible and easy to use. Cloud storage is an interoperable web-based service in which users lease storage space across a platform of servers that can be dedicated or shared based on the needs of the user and level of subscription. Cloud storage may be most appealing for blog preservation because it offers a place to store files and also the opportunity to backup versions of a work-in-progress. The Library of Congress partnered with DuraCloud in a pilot program to test the service as part of the National Digital Information Infrastructure and Preservation Program. The goal was to test the long-term sustainability and accessibility of digital files in cloud storage, with the DuraCloud service becoming widely available as a "cloud archive" in 2011.[3] Services like Amazon S3 and Windows Azure are already available to consumers for fairly low rates. Dropbox.com, though it offers a limited amount of space, has free web-based storage and sharing.

The final tier to consider in preserving a personal blog is that of access. One should strategize publishing tactics that will make it possible to access the blog in its complete form as well as its separate elements, and to migrate it forward to another host or domain as needed. This is where the question of trust comes in. If a blogger

[3] Information on DuraCloud taken from a PowerPoint presentation found on the LOC website. The project is mentioned here: http://www.loc.gov/today/pr/2009/09-140.html

is already hosting their un-preserved content on a third party site, chances are they are willing to give up keeping their content closely guarded for the opportunity to keep it published and accessible. It can therefore be assumed that they are willing to rely on accessibility through the URL by the DNS and domain solution company. In other words, they have to trust to some degree that it will continue to be hosted at the domain its already at for the foreseeable future.

From there, they may explore the possibility of hosting on another server or having a plan to migrate to another server in case of emergency. One lower cost option is web hosting on a shared server. This option remains debatable, as there is always the risk of being hacked, experiencing an outage or some other possible disruption or corruption when others are accessing the server your content is stored on. Choosing a shared server would have to be weighed against the resources at the blogger's disposal. There is also the option of a dedicated server, which poses a much lower risk, or a personal server. Using a personal server to host a blog is, of course, only feasible if the technological capabilities are already in place or there are reasonable means and motivation to have one. This is the most secure option, but not readily available or practical for most bloggers.

The most feasible option may be to rely on the current host, backup offline and/or in cloud storage, and migrate to the next third party host site that proves to be more user-friendly or offer more (like so many did in the great move from Live Journal and Xanga to Blogger and Wordpress) when the time comes. However, in any preservation plan it is always good to have a backup plan. While it may seem like an excessive measure to keep a personal blog accessible through a separate

server, it is important to remember that with blogs "the work is not just a static composite of text and images but depends on server functionality."[4] There is only one outlet for public access to a blog, so taking precautions at this level is a good way to guarantee staying online.

Beyond these three main components, there is the issue of metadata to explore. In Howard Besser's paper on Digital Longevity he states, "…the crucial step is to create extensive documentation for both the physical, structural components of the work along with the possible connections, interactions and choices driving these components."[5] This may be less relevant to the non-archivist, and one must decide how deep to take the aggregation of metadata. Rigorous file-naming conventions are a must for the reasons discussed previously. Folder naming and organization based on dates and timestamps is also important to keep track of the location of elements, and to differentiate between different codes and different iterations of saved content. From an archivist's point of view it might also be useful to include a text file with information about the provenance of the blog content, connections, choices made and possible rights issues to keep in mind when re-publishing. (Rights issues are another subject entirely and are not being explored in-depth for these purposes, as the goal here is to strategize preserving content. However, they are a major issue when it comes to publishing and must be considered at that point in the preservation process.) As far as individual posts are concerned, tags are a good way to develop a controlled vocabulary that is easily preserved along with the code for

[4] http://www.eai.org/resourceguide/preservation/computer/casestudy_reblog.html
[5] http://www.gseis.ucla.edu/~howard/Papers/sfs-longevity.html

your blog. They are picked up with each RSS trigger and create useful contextual metadata that connects posts and helps create structure.

It would be interesting, though difficult within the scope of this paper, to explore the possibility of using RSS feeds in a similar way for the purpose of capturing metadata about a blog. In 2005, a study was done on how the bookmarking site Del.icio.us collects metadata for its users using tags or keywords from the sites collected. Using a tool they developed, Del.icio.us programmers are able to download RSS feeds containing metadata on users and the sites they bookmark, parse those feeds using a Java-based XML parser and load that data into a database where it can be analyzed for statistics on the connectivity and similarity of the users and sites. It sounds complicated but the feeds look fairly straightforward:

- http://del.icio.us/rss/url?url=
(How users tagged a specific URL.)

- http://del.icio.us/rss/tag/tagname
(Which links were tagged with a specific tag.)

- http://del.icio.us/rss/username
(What links did a user tag, and with what tags.)

- http://del.icio.us/rss/popular
(What links are popular.)

- http://del.icio.us/rss/popular/tagname
(What links are popular pertaining to a certain tag.)

It would be interesting to explore the feasibility of setting up similar RSS feeds to collect metadata for a personal site.[6]

[6] http://www.metablake.com/webfolk/web-project.pdf

Blogger and Wordpress users have the advantage of an easy but thorough way to collect metadata from a link to the RSS file in their blog's source HTML, or even at the bottom of the main blog page. For example, at the bottom of a Blogger site there is a link labeled "Subscribe to: Posts (Atom)." Clicking on it takes you to the RSS file <http://brittanclaire.blogspot.com/feeds/posts/default> where the following data can be gathered:

Metadata describing RSS properties:

- version
  (This indicates the version of the RSS language in which the file is formatted.)

Metadata describing blog properties:

- atom:id
  (A unique ID for this blog, based on the domain name and when it was posted.)
- lastBuildDate
  (The latest update.)
- title
  (Title of the blog, different from the URL in many cases.)
- description
  (A search-engine friendly description.)
- link
  (Blog URL.)
- managingEditor
  (Email address and name of the main editor of the blog.)
- generator
  (The tool used to post the blog.)
- openSearch:totalResults
  (How many posts have been made.)
- openSearch:startIndex
  (Which post it should start with – 1 is the most recent.)
- openSearch:itemsPerPage
  (How many posts the RSS feed should show by default.)

Example of those values for my personal blog:

<atom:id>tag:blogger.com,1999:blog-6201689354935716185</atom:id>
<lastBuildDate>Tue, 14 Dec 2010 18:18:36 +0000</lastBuildDate>
<title>Brittan Claire</title>
<description></description>
<link>http://brittanclaire.blogspot.com/</link>
<managingEditor>noreply@blogger.com (Brittan)</managingEditor>
<generator>Blogger</generator>
<openSearch:totalResults>526</openSearch:totalResults>

<openSearch:startIndex>1</openSearch:startIndex>
<openSearch:itemsPerPage>25</openSearch:itemsPerPage>

Metadata describing individual post (item) properties:

- guid
  (Like the atom:id but a unique ID for the specific post.)
- pubDate
  (Date originally published.)
- atom:updated
  (The last time the post was updated.)
- title
  (Title of the individual post.)
- description
  (The body of the post. HTML is converted to codes. For instance a < is replaced with a < so the rss reader can choose how to display or ignore html code.)[7]
- link
  (URL to the specific post.)
- author
  (Email and name of the author.)
- media:thumbnail
  (A link to an image thumbnail to represent the post.)
- thr:total
  (The best I could find from the Blogger help forums is that this has something to do with comment threading. Regardless, it is useful to include any metadata that is given.)

Example of those values for my last post:

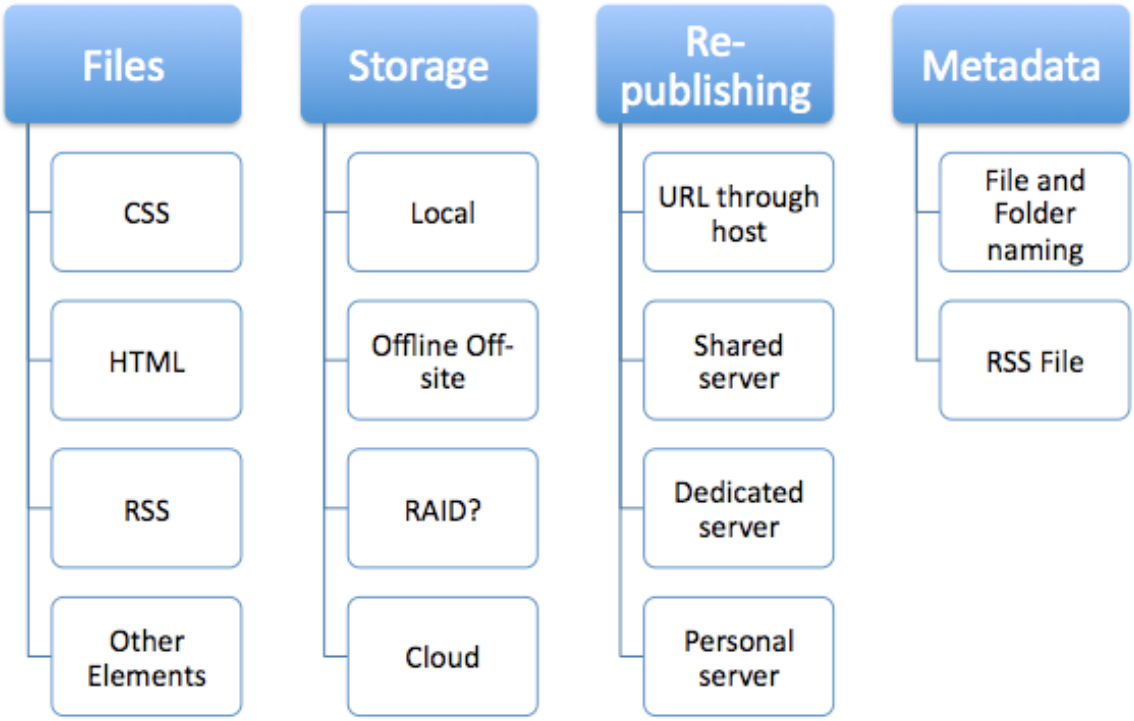
<item>
<guid isPermaLink='false'>tag:blogger.com,1999:blog-6201689354935716185.post-2993815083737138942</guid>
<pubDate>Mon, 13 Dec 2010 21:51:00 +0000</pubDate>
<atom:updated>2010-12-13T15:55:19.003-06:00</atom:updated>
<title>The time has come.</title>
<description><div style="text-align: center;">He has officially outgrown his small skateboard. </div><div style="text-align: center;">Time to [tell Santa to] go back to Index for the real deal.</div><div class="separator" style="clear: both; text-align: center;"></div><div

[7] I discovered this through much searching within the Blogger Help articles and forums. http://www.google.com/support/blogger/

class="blogger-post-footer"><img width='1' height='1'
src='https://blogger.googleusercontent.com/tracker/6201689354935716185-2993815083737138942?l=brittanclaire.blogspot.com' alt=''
/></div></description>
<link>http://brittanclaire.blogspot.com/2010/12/time-has-come.html</link>
<author>noreply@blogger.com (Brittan)</author>
<media:thumbnail xmlns:media='http://search.yahoo.com/mrss/'
url='http://1.bp.blogspot.com/_TkH26PJbRIs/TQaSoGo1UnI/AAAAAAAADno/Mp1vyUFqdcU/s72-c/IMG_1730.JPG' height='72' width='72'/>
<thr:total>0</thr:total>
</item>

Like with the HTML and CSS, the RSS file can be saved, copied and pasted into a text document, or parsed through for information and added to fields in a database or spreadsheet based on the guide given above.

All of these steps should be further tested for their functionality and longevity, but are a good place to start in understanding how to best preserve a blog. My next step in this study is to convey the following basic plan in simple, easy-to-follow instructions for the non-archivist:



These instructions will be posted on my blog at Blogger.com and made available to those who follow my blog. I also hope to reach out to popular personal bloggers (those with 500 – 1,000+ followers) to offer this as a tool for them to use and share with their followers. If surveys and posting trends can go viral among bloggers in a matter of days, I hope that a plan for the long-term care of the blog can do the same.

Sources

Archive-IT. About Archive-IT. <http://www.archive-it.org/public/about-us.html;jsessionid=FE07661AF438BA325E18D70CB7EB6376>.

Besser, Howard. "Digital Longevity" in Handbook for Digital Projects: A Management Tool for Preservation and Access. Ed. Maxine Sitts. National Document Conservation Center. 2000. <http://www.gseis.ucla.edu/~howard/Papers/sfs-longevity.html>.

Blogger Help. <http://www.google.com/support/blogger/>.

Eyebeam Art + Technology Center. reBlog. <http://www.eyebeam.org/reblog>.

IIPC. “Long-term Preservation of Web Archives – Experimenting with Emulation and Migration Methodologies.” 10 December 2009. <http://www.netpreserve.org/publications/NLA_2009_IIPC_Report.pdf>.

Internet Archive. Glossary of Web Archiving Terms. <https://webarchive.jira.com/wiki/display/ARIH/Glossary+of+Web+Archiving+Terms>.

Library of Congress. “Library of Congress and DuraCloud Launch Pilot Program Using Cloud Technologies to Test Perpetual Access to Digital Content.” News Release. 14 July 2009. <http://www.loc.gov/today/pr/2009/09-140.html>.

LOCKSS Datasheet. PDF. 2009. <http://www.lockss.org/lockysswiki/files/LOCKSS_Datasheet_2009.pdf>.

Shaw, Blake. “Utilizing Folksonomy: Similarity Metadata from the Del.icio.us System CS6125 Project.” 9 December 2005. <http://www.metablake.com/webfolk/web-project.pdf>.

Smith, Pamela. “Preserving the Dynamic.” Electronic Arts Intermix, IMAP. First published 2004. <http://www.eai.org/resourceguide/preservation/computer/casestudy_reblog.html>.