Practical Options for Archiving Social Media

Content Summary for ALGIM Web-Symposium Presentation 03/05/11

Euan Cochrane,
Senior Advisor, Digital Continuity,
Archives New Zealand,
The Department of Internal Affairs.

Te Rua Mahara o te Kāwanatanga
ARCHIVES NEW ZEALAND

INTERNAL AFFAIRS
Te Tari Taiwhenua

Table of Contents

1. What is Social Media?

"Social media is the use of web-based and mobile technologies to turn communication into interactive dialogue."
http://en.wikipedia.org/wiki/Social_media

Social media is difficult to define and can cover a wide range of content. The topics covered in this document are outlined in Section 2. below.

2. Contents of This Document

This contains a summary of the presentation content for the ALGIM Web symposium held on the 3/5/11. The document identifies practical tools and services for use in archiving content produced through/in the following formats/services:

- Facebook
- Twitter
- Blogs
- Flickr
- General Website or Cross-Service Preservation Tools

It also briefly summarises the pros and cons of each tool or service.

In order effectively conduct social media archiving it is best to start with a plan that includes such things as consideration of stakeholders needs and outcomes to be achieved. This document does not cover these issues or anything other than that which is stated above.

This document does not constitute guidance from Archives New Zealand but is a brief research summary for presentation to ALGIM members.

3. What are Central Organisations Doing?

3.1. National Library of New Zealand Web Archive

http://www.natlib.govt.nz/collections/a-z-of-all-collections/nz-web-archive

The Web Archive forms part of the Alexander Turnbull Library collections.

The Library began selecting websites in 1999 and the collection has continued to grow with active development since 2005, reflecting New Zealand's growing online cultural and historical presence. The selected websites in the collection cover a diverse range of subjects including significant events, and is strong in the following areas:

- government
- politics, including blogs, and general and local body elections
- Māori, including iwi and Treaty of Waitangi
- community and ethnic groups
- music, including labels, organisations, artists and directories
- sport
- the arts
- the environment.

The collection also provides a visual history of how websites change over time and most of the websites in the Archive are collected at regular intervals to ensure new content is captured. This content includes web pages, images, multi-media, and publications, such as journals, that are made publicly available.

Selective Harvests:

Around 150 websites archived each month
Limiting Factors:
People: Limits frequency of harvesting each site
Time: Length of time it takes to harvest a website
Bandwidth: The number of concurrent harvests
Prioritisation necessary!

Whole-of Domain Harvests:

There were conducted in 2008 and 2010 and covered the whole .nz domain (where possible).

You can submit websites for regular harvesting,
http://www.natlib.govt.nz/collections/a-z-of-all-collections/nz-web-archive/natlib-forms/nominate-a-site

3.2. Archives New Zealand

Archives New Zealand has signed an MOU with the National Library in which the Library has agreed to preserve public sector websites.

For more guidance on managing web records see:

The Decommissioning Websites Fact Sheet
http://archives.govt.nz/decommissioning-websites-factsheet

The Guide to Managing Web Records
http://archives.govt.nz/g20-guide-managing-web-records

3.3. Library of Congress Twitter Archive

The Library of Congress in the United States has been preserving open-access tweets since 2010.

What is in the Archive?

Twitter has been a public and open communications platform since its beginning. Twitter is donating an archive of what it determines to be public. Private account information and deleted tweets will not be part of the archive. Linked information such as pictures and websites is not part of the archive, and the Library has no plans to collect the linked sites. There will be at least a six-month window between the original date of a tweet and its date of availability for research use.

Pros: Most comprehensive in the general sense (covers the most number of tweets), free.
Cons: Not currently accessible, does not cover personal information, controlled by another (trustworthy) institution.

3.4. The Internet Archive (www.archive.org)

The Internet Archive Wayback Machine is a service that allows people to visit archived versions of Web sites. Visitors to the Wayback Machine can type in a URL, select a date range, and then begin surfing on an archived version of the Web. The internet archive has copies of many websites from 1996 onwards. They have a backup of the archive in Alexandria, Egypt.

Pros: Most comprehensive web-record available, free.
Cons: Irregular crawls, doesn't preserve all content, only accessible via their site, content can be removed at any time.

4. Facebook

4.1. Facebook export

http://facebookexport.com/

Facebook Export uses the Facebook Open Graph protocol to export your Facebook data to an xml file. Facebook Export does not store any data about you. You can then use this xml file to import your data to other services and websites that support the Facebook Export.

Pros: simple service
Cons: Need login access, produces only an xml file with textual information, giving away private information.

4.2. Facebook's export

http://www.facebook.com
(in Account → Account settings → Download your information)

Included information:

- Your profile information (e.g., your contact information, interests, groups)
- Wall posts and content that you and your friends have posted to your profile
- Photos and videos that you have uploaded to your account
- Your friend list
- Notes you have created
- Events to which you have RSVP'd
- Your sent and received messages
- Any comments that you and your friends have made on your Wall posts, photos, and other profile content

Pros: Most comprehensive export available, simple to use.
Cons: You need to have account access to do this, may not preserve all content.

5. Twitter:

5.1. Twapperkeeper

http://twapperkeeper.com/index.php

Online

Twapperkeeper Online is a web-based tweet archiving service:
"Do you want to archive tweets from your conference? Maybe archive trending hashtags or keywords for historical or analysis purposes? Maybe save your own personal tweets? Twapper Keeper is here to help!
How does it work?

1. Create a new Twapper Keeper archive based upon hashtag, keyword, or person
2. Tell your friends about the archive
3. Read, track, export, and analyze as much as you want!"

Pros: Simple comprehensive service, free for up to 2 'archives'.
Cons: cannot export tweets at all, need to login to get personal tweets

Offline server-side Twapperkeeper

Twapper keeper also provides a downloadable installable version of their service from which you can export tweets for preservation.

Pros: Comprehensive, simple to use once set-up, unlimited 'archives'.
Cons: Technically challenging to install/set-up, requires a web-server

5.2. TweetDoc

http://www.tweetdoc.org/
Tweetdoc is a service that produces document that brings together all the tweets from a particular event or search term. A tweetdoc allows you to keep a record of an event through twitter.

Pros: Simple, easy to preserve outputs.
Cons: Not comprehensive, will not gather all tweets from a particular user for example, limited in scope due to Twitter terms of Service.

5.3. Archivist Desktop

http://visitmix.com/work/archivist-desktop/

The Archivist is a Windows application that helps you archive tweets for later data-mining and analysis. Start a search with The Archivist and get as many results as it can.

Pros: Simple, powerful, will work well if just beginning (see cons)
Cons: Relies on twitter search which limits it to 1500 results per search, the search will only go back in time for a set amount, usually around 3-4 weeks.

6. Blogs

6.1. ArchivePress

http://code.google.com/p/archivepress/

What parts of a blog does ArchivePress harvest?

For each blog (or feed) URL added to the control panel, ArchivePress harvests all the information available from the RSS or Atom feeds:

- The (hyper) text of blog posts
- The (hyper) text of comments on posts
- Image media embedded in the post as long as they are on the same host as the blog itself

What parts of a blog doesn't ArchivePress harvest?

ArchivePress ignores any features rendered solely in the HTML of a blog, such as colours, logos and widgets. If this is important to you in your blog archiving efforts, ArchivePress is probably not for you (but feel free to use any part of the ArchivePress code to make your own blog harvester, and be sure to let us know about it!).
We feel that, particularly in the academic and research communities, and small institutions there is a strong case for accepting the compromise involved in creating essentially text-based collections of blog content at low cost (the cost of installing and maintaining WordPress + ArchivePress). By comparison, there are substantial costs involved in setting up and running web harvesting applications like PANDAS and Web Curator tool to capture entire blog websites. Again, many of these use cases and issues are described and discussed on the ArchivePress project blog and several articles written by the ArchivePress team.

Pros: Simple to use once set-up, manageable
Cons: Does not preserve all components of blogs, relies on wordpress, requires a wordpress installation, technically challenging to set-up and install. Does not preserve all blog content.

6.2. See also: RSS feed preservation/website preservation below

7. Flickr

7.1. Downloadr

http://janten.com/downloadr/

Downloadr is a photo downloader for Microsoft Windows. It provides a simple interface to download large sized images from Flickr to your computer. You do not need a Flickr account to use Downloadr, though you do have even more functionality (like the ability to do a complete backup of all your photos) if you register at flickr.com.

Pros: Free, comprehensive, easy to use
Cons: Download and installation required.

7.2. Bulk Image Downloader

http://bulkimagedownloader.com/
Downloads images and videos from multiple different websites.
Pros: Extensive functionality that covers video as well as images
Cons: Requires purchase.

7.3. PhotoGrabber – for Mac OSX

http://www.malarkeysoftware.com/projects_PhotoGrabbr.html

Downloads copies of your flickr albums
Pros: Effective, Mac compatible, free.
Cons: Limited functionality.

7.4. FlickrEdit

http://sunkencity.org/flickredit
FlickrEdit is a Java Desktop application that allows you to display and edit your photos in a variety of ways. It also allows you to download/backup or upload your photos to and from Flickr. FlickrEdit is written in Java and it uses flickrj framework to access Flickr.
Pros: Nothing significantly better than other options but good. Free
Cons: Not as comprehensive as others.

7.5. FlickrTouch

https://github.com/dan/hivelogic-flickrtouchr#readme
A Python script to grab all your photos from flickr and dump them into a directory, organized into folders by set name.

Pros: Free, effective
Cons: Limited Functionality.

8. Others:

8.1. Trunk.ly

http://trunk.ly/
Web based tool that organises and saves links your facebook, twitter, delicious, pinboard and RSS feeds. It will preserve all the links and make them exportable.

Pros: Single tool to cover many services, simple to use
Cons: Only preserves links in the services and not other content, needs your login information for each service.

8.2. Archive-it

http://www.archive-it.org/

Archive-It, a web archiving service from the Internet Archive, allows institutions to harvest and preserve collections of digital content and create Digital Archives. Through a user-friendly web interface, Archive-It partners can catalogue, manage, and browse their archived collections. Collections are hosted at the Internet Archive data centre and are accessible to the public with full-text search.

Pros: Powerful solution for website archiving.
Cons: Slightly technical, cost component, may over-lap with National Library Web-harvesting work.

8.3. Heritrix

http://crawler.archive.org/

Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler. This is the software that the National Library of New Zealand's archiving service is based on.
Pros: Best web-archiving tool available.
Cons: Technically sophisticated to use and set-up, produces files that can only be accessed easily through the wayback machine tool.

8.4. HT-Track

http://www.httrack.com/page/1/en/index.html
HTTrack is a free (GPL, libre/free software) and easy-to-use offline browser utility. It allows you to download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer. HTTrack arranges the original site's relative link-structure. Simply open a page of the "mirrored" website in your browser, and you can browse the site from link to link, as if you were viewing it online. HTTrack can also update an existing mirrored site, and resume interrupted downloads.

Pros: Easy to use, extensive functionality captures websites in easily accessible format. Free.
Cons: Non proven scalability does not capture everything.

8.5. The Web Curator Tool

http://webcurator.sourceforge.net/

The Web Curator Tool (WCT) is an open-source workflow management application for selective web archiving. It is designed for use in libraries and other collecting organisations, and supports collection by non-technical users while still allowing complete control of the web harvesting process. It is integrated with the Heritrix web crawler and supports key processes such as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata.

WCT was developed in 2006 as a collaborative effort by the National Library of New Zealand and the British Library, initiated by the International Internet Preservation Consortium. From version 1.3 WCT software is maintained by Oakleigh Consulting Ltd, under contract to the British Library. WCT is available under the terms of the Apache Public License.

Pros: See Heritrix above, easy to use
Cons: See Heritrix above.

9. Other Resources:

Report: Agencies unsure how to archive social media

Report: Agencies unsure how to archive social media - FierceGovernmentIT http://www.fiercegovernmentit.com/story/report-agencies-

unsure-how-archive-social-media/2011-04-03#ixzz1LFB5rvpp

Best Practices Study of Social Media Records Policies ACT-IAC Collaboration & Transformation (C&T) Shared Interest Group (SIG). March 2011
http://www.actgov.org/knowledgebank/whitepapers/Documents/Shared%20Interest%20Groups/Collaboration%20and%20Transformation%20SIG/Best%20Practices%20of%20Social%20Media%20Records%20Policies%20-%20CT%20SIG%20-%2003-31-11%20%283%29.pdf

Chris Prom's Practical e-records Blog
http://e-records.chrisprom.com/

Chris's post on web-archiving tools
http://e-records.chrisprom.com/?p=1873