Currents Spring 2014 Workshop

Digital Archiving for Collective and Personal Use

Archiving for Collective and Personal Use

Web resources:

- Websites can be saved as complete websites with internal hyperlinks inserted to other complete web pages stored in a directory.
- Internet Archive accepts requests to save pages not yet archived in the Wayback Machine
- Browser-based page saving utilities that function across platforms like Instapaper, Laterloop, and the Firefox extension Scrapbook
- Zip or export wiki/blog content

Additional Strategies:

Cloud Storage

- Constantly updating cloud drives like Google Drive, Dropbox or Microsoft OneDrive – Apple iCloud for iWork files

Hard Backups

- Hard drives or external drives regularly backed up
- Storage media – DVD-R(W), CD-R(W)

Currents Restoration

- Fixing broken links on the site itself
- Missing images sourced through Internet Archive's "Wayback Machine" < http://archive.org/web/>
- Internet Archive is a non-profit founded to build an internet library by offering permanent access to content published on the web on sites viewable by its webcrawlers (provided by Alexa Internet)

Internet Archive has two practical considerations in dealing with digital collections:

How to store massive amounts of data
How to preserve the data for posterity

Storage

Storing the Archive's collections involves parsing, indexing, and physically encoding the data. With the Internet collections growing at exponential rates, this task poses an ongoing challenge.

Our hardware consists of PCs with clusters of IDE hard drives. Data is stored on DLT tape and hard drives in various appropriate formats, depending on the collection. Web data is received and stored in archive format of 100-megabyte ARC files made up of many individual files. Alexa Internet (currently the source of all crawls in our collections) is proposing ARC as a standard for archiving Internet objects. See Alexa for the format specification.

Preservation

Preservation is the ongoing task of permanently protecting stored resources from damage or destruction. The main issues are guarding against the consequences of accidents and data degradation and maintaining the accessibility of data as formats become obsolete.

Accidents: Any medium or site used to store data is potentially vulnerable to accidents and natural disasters. Maintaining copies of the Archive's collections at multiple sites can help alleviate this risk. Part of the collection is already handled this way, and we are proceeding as quickly as possible to do the same with the rest.

Migration: Over time, storage media can degrade to a point where the data becomes permanently irretrievable. Although DLT tape is rated to last 30 years, the industry rule of thumb is to migrate data every 10 years. We no longer use tapes for storage, however. Please take a look at our page on our Petabox system for more information on our storage systems.

Data formats: As advances are made in software applications, many data formats become obsolete. We will be collecting software and emulators that will aid future researchers, historians, and scholars

- Wayback Machine provides website captures over time viewable through either calendar or sliding timeline modes.
- Internet Archive does not capture protected or secure content on sites that restrict data crawlers (e.g. social media profiles) using 'robots.txt' or similar scripts.
- Internet Archive's Archive-It offers specialized services to Colleges and Universities and other organizations < https://archive-it.org>