**Brigham Young University**
**Harold B. Lee Library**
**L. Tom Perry Special Collections**
**Web Archive Guideline**

## Purpose:

The mission of the L. Tom Perry Special Collections Web Archive at Brigham Young University is to enhance scholarship and learning by documenting, providing access to, and preserving the state of Mormonism in all of its variations, as they exist online. To accomplish this purpose the Web Archives will harvest websites and other Internet content through the use of a web archiving service.

Since its inception, the L. Tom Perry Special Collections has sought to document Mormonism to the fullest extent possible through the collection of personal and corporate documents, images, moving images, personal narratives, and primary and secondary source publications. The web archive seeks to build upon and complement these traditional collecting pathways. More and more creators of Mormon-themed works are turning to the Internet as the sole source in disseminating their efforts. Documenting this work through the capturing of Internet content is a natural and inevitable extension of the functions curators of Special Collections are already performing, and will ensure the relevance and authority the L. Tom Perry Special Collections presently enjoys in Mormon primary research.

It is estimated that in 2011, there were over 500 million websites[1] on the Internet; of that over 300 million[2] were created in 2011 alone. These large numbers begin to pale in comparison to Internet-generated content such as email (estimated over 3 billion email accounts registered by 2011[3]) and social media (Facebook alone reached 800 million users by the end of the year[4]). Media such as images and video, which have been traditionally collected by archival repositories around the nation, have also seen a virtual swell. In 2011, there was an average of forty-eight hours of video uploaded to YouTube every minute[5] and 4.5 million images uploaded to

---

[1] Pingdom, "Websites", pingdom.com. http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/ (accessed November 23, 2012)

[2] Ibid

[3] Pingdom, "Email", pingdom.com. http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/ (accessed November 23, 2012)

[4] Pingdom, "Social media", pingdom.com. http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/ (accessed November 23, 2012)

[5] Pingdom, "Videos", pingdom.com. http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/ (accessed November 23, 2012)

Flicker each day[6]. With new material becoming available on such a frequent basis it can cause content that is even just one-day old to become culturally irrelevant, and potentially lost in the wide array of superfluous material being published online. The fluidity of the Internet causes it to be ephemeral in nature; consequently the risk of losing material of enduring value is prodigious unless such material is purposely sought after, captured, and retained in perpetuity. Archiving selections of the Internet presents many challenges, yet equal to the challenge is the opportunity to play a role in documenting and preserving what has become an essential part of human life and experience. Without the deliberate and focused efforts of the L. Tom Perry Special Collections Web Archive, the present depth and knowledge of Mormonism being circulated online will be lost to history.

## Terms used in Web Archiving:

**Archive:** A repository containing records, documents, or other materials of enduring, evidential, legal, or historical value that are preserved so as to provide continual access in accordance with user access policies.

**Capture:** The process of copying digital information from the web to a repository for collection or archival purposes.

**Collection:** A group of resources related by common ownership or a common theme or subject matter. A web collection consists of one or more crawls that harvest a group of related websites (e.g., candidate websites for state election campaigns). Collections are owned and/or maintained by an organization or institution.

**Crawl:** The content associated with a web capture operation that is conducted by a crawler.

**Crawler:** A software agent that captures information from the web. Our crawler starts with a list of URL's to visit. As it visits these URL's it captures the documents on these web pages.

**Curators:** persons responsible for building collections of web-based resources, and specify seed lists for specific crawls.

**Digital Archive:** A digital collection for which an institution has agreed to accept long-term responsibility for preserving the resources in the collection and for providing continual access to those resources in keeping with an archive's user access policies.

---

[6] Pingdom, "Images", pingdom.com. http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/ (accessed November 23, 2012)

**Digital Collection:** A collection consisting entirely of born-digital or digitized materials.

**Document:** A document, or web document, is a resource on the World Wide Web that has a distinct web address. It could be an embedded image, whole web page, PDF, or any other component of a web page. A document can be any kind of MIME type.

**Domain:** A resource on the web that has a distinct web address. It could be an embedded image, whole web page, PDF file or any other component of a web page.

**Harvest:** Another name for the act of capturing web content as a part of crawling.

**Host:** A single networked machine, as usually designated by its Internet hostname (i.e. byu.edu). The hostname can be identical to an URL's domain name, but not always.

**HTML:** A markup language used to structure text and multimedia documents and to set up hypertext links between documents, used extensively on the web. It can be created and processed with a wide range of tools from simple text editors to sophisticated authoring software.

**MIME:** Stands for multipurpose Internet Mail Extensions. This is a specification for formatting non-text content to be sent over the Internet. A MIME file can be just about any kind of non-text file, i.e. gif, jpeg, html, etc.

**Repository:** The physical storage location and medium for one or more digital archives. A repository may contain an active copy of an archive (i.e. one that is accessed by end users) or a mirror copy of an archive for disaster recovery.

**Robots.txt:** A text file placed in the root directory of a website that prohibits crawlers from indexing all of specific pages of the site. The Robots Exclusion protocol provides a format for designating which directories and files are off limits to the crawler.

**Robots.txt Query Exclusion Error**: A message indicating that the requested information was not captured by the crawler because the site owner asked that the information be excluded from capture using a robots.txt file.

**Seed:** Any URL that tells the crawler you specifically want to capture. Seed can be an entire website, a specific part of a website or a specific URL of a document. Also called a targeted URL.

**URL (Uniform Resource Locator) :** A web address (for example: http://lib.byu.edu/sites/sc/), usually consisting of the access protocol

(http), the domain name (lib.byu.edu), and optionally the path to a file or resource residing on that server (/sites/sc/).

**Web Archive:** A collection of web-published materials that an institution has either made arrangements for or has accepted long-term responsibility for preservation and access in keeping with an archive's user access policies. Some of these materials may also exist in other forms but the web archive captures the web versions for posterity.

**Web Archive Service:** Enables curators to build collections of web-published materials that are stored in either local and/or remote repositories. The service includes a set of tools for selection, curation, and preservation of the archives. It also includes repositories for storage, preservation services (e.g., replication, emulation, and persistent naming), and administrative services (e.g., templates for collection strategies, content provider agreements, repository provider agreements.)

**Web Page:** A resource on the web, usually in HTML/XML format and with hypertext links to enable navigation from one page or section to another, displayed with a web browser. A web page can contain any of the following:
*Text*
*Graphics* (.gif, .jpeg, or .png)
*Audio* (.mid or .wav)
*Interactive multimedia content* that requires a plug-in such as Flash Shockwave or VML
*Applets* (subprograms that run inside the page) which often provide motion graphics, interaction, and sound

**Website:** A website is a collection of related web resources, usually as grouped by some common addressing – as when all resources on a single host, or group of related hosts, are considered a 'website'.

## Collection Development:

The Archive will focus mainly on Mormonism as it relates to culture, expression, history, philosophy, ideology, society, and theology. Secondary focus will be given to areas of interest such areas as business, education, politics, activism, and philanthropy. In order for content to be collected in any of these areas there must be an established connection to Mormonism either through the content, the creator, or the subject matter.

The primary emphasis of the Web Archive will be to capture the previously emphasized priorities of Mormonism, as they exist online in a publicly accessible and freely available format. In general, for Pay Sites will not be collected; examples of this include databases or news outlets such as the Salt Lake Tribune or the Daily

Herald. Material types to be collected will consist of, websites, blogs, wiki's, social media, news articles, videos channels (such as YouTube and Vimeo), PDF's, image files, audio files, and any other relevant Internet format that has either not been stated or has yet to be created.

The Web Archive will seek to capture, if possible, all sites under the http://BYU.edu/ domain and other sites associated with Brigham Young University-Provo, such as https://byucougars.com/. If the harvesting of http://BYU.edu/ in its entirety is not feasible due to volume or resource limitations then specific sites under the domain will be identified and a crawling rotation established for only those selected sites. The Web Archive will not seek to capture the BYU affiliated institution http://BYUI.edu/ in its entirety. Specific sites or pages within that domains may be considered for capture depending on content related to the main areas of interest as stated in the collection development guideline.

Additionally, the Web Archive will not seek to capture any domain that is officially owned and operated by the Church of Jesus Christ of Latter-day Saints as Intellectual Reserve, Inc, Deseret Digital Media, or Deseret Book.

The Archive will potentially seek to capture websites of other religious organizations that have a direct or in-direct association with Mormonism such as the site http://www.churchofchrist-tl.org, owned and operated by the Church of Christ (Temple Lot).

## Acquisition:

The Web Archive will rely heavily on a curated approach to building the collection; meaning that it will be the responsibility of the curator(s) to actively seek material for inclusion into the archive. Web content will be selected based on the guidance of the collection development guideline stated above.

Unlike the traditional acquisition practices of archives in acquiring the physical material along with the ownership rights and copyright, The L. Tom Perry Special Collection Web Archive will not actively seek to acquire ownership or copyright of any collected web content. It is therefore expected that the creator or owner of web content captured by the Archive will retain full copyright and intellectual property rights of their content at all times, unless otherwise agreed upon. Furthermore, by only capturing web content for transformative proposes the Web Archive assumes no liability for direct, indirect, special, incidental, or consequential damages, that are in any way related to the content. Rather, the Web Archive will acquire web content based on the Best Practice Guide issued by the Association of Research Libraries

(ARL)[7] that the Harold B. Lee Library (of which the L. Tom Perry Special Collection is a division in) is a supporting member of. The standard, as put forth by the library community represented in the ARL Best Practice Guide states "It is fair use to create topically based collections of websites and other material from the Internet and to make them available for scholarly use."[8] ARL provided further explanation for archives and libraries to exercise fair use in building web collections:

> Gathering impressions of ephemeral Internet material such as web pages, online video, and the like is a growth area in academic and research library collection-building, with activities typically focusing on areas in which the institution has an established specialty, or on sites specific to its local area. Such collections represent a unique contribution to knowledge and pose no significant risks for owners of either the sites in question or third-party material to which those sites refer. In the absence of such collections, important information is likely to be lost to scholarship.
>
> Selecting and collecting material from the Internet in this way is highly transformative. The collecting library takes a historical snapshot of a dynamic and ephemeral object and places the collected impression of the site into a new context: a curated historical archive. Material posted to the Internet typically serves a time-limited purpose and targets a distinct network of users, while its library-held counterpart will document the site for a wide variety of patrons over time. A scholar perusing a collection of archived web pages on the Free Tibet movement, or examining the evolution of educational information on a communicable disease, seeks and encounters that material for a very different purpose than the creators originally intended. Preserving such work can also be considered strongly transformative in itself, separate from any way that future patrons may access it. Authors of online materials often have a specific objective and a particular audience in mind; libraries that collect this material serve a different and broader purpose and a different and broader network of users. Libraries collect not only for a wide range of purposes today, but also for unanticipated uses by future researchers.[9]

As demonstrated throughout this guideline, the intent of the L. Tom Perry Special Collection Web Archive is to align its practices to be congruent with ARL's Code of Best Practices in Fair Use for Academic and Research Libraries (http://www.arl.org/pp/ppcopyright/codefairuse/index.shtml), specifically principal Eight (http://www.arl.org/pp/ppcopyright/codefairuse/code/eight-collecting.shtml).

It is the nature of most websites to be constantly adding or augmenting content. It is therefore expected that acquisition of this content will be a continuous process. As a site or page is identified for capture it will also be assigned a capturing schedule. The schedule will be based on a fixed rotation of crawls to be performed:
Daily: (depending on size and speed of the harvesting tool)

---

[7] See Appendix One

[8] EIGHT: Collecting Material Posted on the World Wide Web and Making It Available, "Principle", arl.org. http://www.arl.org/pp/ppcopyright/codefairuse/code/eight-collecting.shtml (accessed on November 20, 2012)

[9] EIGHT: Collecting Material Posted on the World Wide Web and Making It Available, "Description", arl.org. http://www.arl.org/pp/ppcopyright/codefairuse/code/eight-collecting.shtml (accessed on November 20, 2012)

Weekly
Semi-weekly
Monthly
Semi-Annually: (based on a year rotation)
Annually: (based on a year rotation)
In addition to the schedule, a stop date will also be assigned at the time of the initial acquisition. For example, a web page may be assigned a daily crawl schedule but also be issued a stop date of one week, thus the web page will be crawled everyday for only one week. Reviews will be continually conducted to ensure that each URL intended for crawling will be matched with an appropriate schedule and stop date.

For Pay sites, or websites that require a fee in order to access the content of, will not be captured by the Web Archive. If content on the site is deemed significant, based on the standards of the collection guideline, the Web Archive will contact the site owner and seek formal permission through the Web Capturing and Archiving Agreement to archive such content.

The Internet content that the L. Tom Perry Special Collections Web Archive does capture will be archived as is without any changes purposely made to augment or diminish content. In essence, the Archive will strive to take a "snap-shot" of each page or item harvested.

## Metadata:

Websites or web pages will be associated with corresponding metadata records that will be created by the Web Archive. At present, the standards used for creating metadata will be based on Dublin Core (dublincore.org/) or on MARC (loc.gov/marc/bibliographic/) with local elements used as well. Unless specifically identified as a sole resource, individual web pages will not be given corresponding metadata. At minimum, the metadata that will be generated and displayed is:

**URL:** The current Uniform Resource Locator given to the resource.
**Title:** The name given to the resource.
**Type:** The nature or genre of the content of the resource, blog, wiki, webpage, website, etc.
**Creator:** An entity primarily responsible for making the content of the resource. Examples of a Creator include a person, an organization, or a service. Typically the name of the Creator should be used to indicate the entity.
**Owner:** the primary entity holding ownership of the content or resource, if different from the Creator.
**Creation Date:** Date associated with the creation or availability of the resource.
**Harvest Date:** Date associated with each web capture of the resource.
**Harvest Method:** The tool or system by which the resource was captured.

**RightsHolder:** A person or organization owning or managing rights over the resource.

Metadata is subject to change over time as new standards are developed and implemented.

## Access:

The Web Archive of the L. Tom Perry Special Collections has been founded in part to serve the needs of current and future researches. It is expected that the majority of researchers utilizing the Web Archive will be doing so virtually however, it is intended that access still be limited only to those who seek the content for research proposes, as is consistent with the current practices and policies of the L. Tom Perry Special Collections. The following statement will be branded on all of the website and pages harvested by the Archive: "This website has been captured by the L. Tom Perry Special Collections Web Archive and is intended for research purposes only. By viewing this page you are agreeing to the term of use. Any deviation in usage will jeopardize your potential access to this material in the future."

Access may also be limited based on content. If deemed inappropriate for the average user, a site may still be captured for preservation but will be made publically unavailable and thus restricted. Restriction of web content will be made based on the same consideration given to the restricting of printed and manuscript material. Access to restricted web content will only be granted on an individual basis and in most cases will require the researcher to be physically present at the L. Tom Perry Special Collections and undergo the same process for accessing restricted material as currently practiced with the printed and manuscript collections.

Please direct any concerns or complaints to BYU's Copyright Licensing Office by contacting them through phone: 801-422-9339, email: copyright@byu.edu, or visiting their webpage at: http://lib.byu.edu/sites/copyright/. If you are an owner of content that has been harvested by the Web Archive and wish your material not be included in the Web Archive please use the above information to contact the Copyright Licensing Office with your concerns.

# Appendix One:



## Association of Research Libraries (ARL)

http://www.arl.org/pp/ppcopyright/codefairuse/code/eight-collecting.shtml

**Code of Best Practices in Fair Use for Academic and Research Libraries**

**Code of Best Practices**

**EIGHT: Collecting Material Posted on the World Wide Web and Making It Available**

**Description**

Gathering impressions of ephemeral Internet material such as web pages, online video, and the like is a growth area in academic and research library collection-building, with activities typically focusing on areas in which the institution has an established specialty, or on sites specific to its local area. Such collections represent a unique contribution to knowledge and pose no significant risks for owners of either the sites in question or third-party material to which those sites refer. In the absence of such collections, important information is likely to be lost to scholarship.

Selecting and collecting material from the Internet in this way is highly transformative. The collecting library takes a historical snapshot of a dynamic and ephemeral object and places the collected impression of the site into a new context: a curated historical archive. Material posted to the Internet typically serves a time-limited purpose and targets a distinct network of users, while its library-held counterpart will document the site for a wide variety of patrons over time. A scholar perusing a collection of archived web pages on the Free Tibet movement, or examining the evolution of educational information on a communicable disease, seeks and encounters that material for a very different purpose than the creators originally intended. Preserving such work can also be considered strongly transformative in itself, separate from any way that future patrons may access it. Authors of online materials often have a specific objective and a particular audience in mind; libraries that collect this material serve a different and broader purpose and a different and broader network of users. Libraries collect not only for a wide range of purposes today, but also for unanticipated uses by future researchers.

**Principle**

It is fair use to create topically based collections of websites and other material from the Internet and to make them available for scholarly use.

**Limitations**

- Captured material should be represented as it was captured, with appropriate information on mode of harvesting and date.
- To the extent reasonably possible, the legal proprietors of the sites in question should be identified according to the prevailing conventions of attribution.
- Libraries should provide copyright owners with a simple tool for registering objections to making items from such a collection available online, and respond to such objections promptly.

**Enhancements**

- Claims of fair use relating to material posted with "bot exclusion" headers to ward off automatic harvesting may be stronger when the institution has adopted and follows a consistent policy on this issue, taking into account the possible rationales for collecting Internet material and the nature of the material in question.
- The more comprehensive a collection of web impressions in a given topic area is, the more persuasively the inclusion of any given item can be characterized as fair use.

---

—> Next: Credits

<— Previous: SEVEN: Creating Databases to Facilitate Non Consumptive Research Uses (Including Search)

Appendix Two:

# Web Capturing and Archiving[1] Agreement (Draft)

| | | |
|---|---|---|
| | **Brigham Young University (BYU)**<br>**Harold B. Lee Library**<br>**L. Tom Perry Special Collections**<br>1130 Harold B. Lee Library<br>Provo, Utah 84602 | Agreement Date: |

| | | |
|---|---|---|
| This Web Archiving Agreement ("Agreement") is entered into on the date set forth above between BYU, and: | | |
| Web Content Owner's name(s) (collectively "Owner"): | | |
| Telephone number(s): | | |
| Address(es): | | |
| E-mail address(es): | | |
| Birth date(s) (dd/mm/yyyy): | | |
| Website Name: | | |
| Website URL: | | |
| **Detailed Description of Web Content Being Captured and Archived (collectively "Web Content"):** Please include a description of the textual, visual, or aural content that is encountered as part of the user experience on the website to be Archived. The Web Content may include, among other things: text, images, animations, hyperlinks to third-party websites, and audio and video content. | | |

1 Web Archiving is the capturing and preservation of selected websites that are of interest to the L. Tom Perry Special Collections Web Archive and that is of benefit and use to current and future researchers.

**TERMS.** Owner agrees to allow BYU to capture the Web Content as described above for archiving and distribution by the L. Tom Perry Special Collections Web Archive for education, private study, and research ("educational purposes"). The Owner agrees to allow the Web Archive to periodically capture Web Content to document changes.

This Agreement provides the legal permissions and warranties needed to allow BYU to capture, archive, preserve, and make accessible, in a variety of formats and media now known or later developed, Web Content maintained by the Owner. This agreement permits use of the captured and archived Web Content only for non-commercial purposes, research, private study, and educational purposes.

This is a non-exclusive agreement, which ensures that Owner rights are not transferred by this agreement. The Owner retains copyright and intellectual property rights and is free to use or publish the Web Content elsewhere.

Cataloging information and documentation can be publicly available but access to the captured and archived Web Content will only be available to authorized users who have agreed to abide by access conditions set by the L. Tom Perry Special Collections unless the Owner has stated that the Web Content can be available to any user.

It is understood that the Web Content may contain third-party copyrighted material and/or links.

**MODIFICATIONS.** BYU or its authorized users may not modify or change or create a derivative work of the captured and archived Web Content in any manner.

**WARRANTIES.** Web Content is captured and archived "as is" without warranty of any kind, either expressed or implied. In no event will BYU be liable for direct, indirect, special, incidental, or consequential damages that are in any way related to Web Content.

**NO COMPENSATION.** Owner acknowledges and agrees that Owner shall not receive any monetary compensation in exchange for the captured and archived Web Content.

In consideration of the mutual promises and covenants herein contained, and for other good and valuable consideration, Owner and BYU indicate their agreement with the descriptions and terms above, and with the Terms, by signing below:

Brigham Young University L. Tom Perry Special Collections Web Archive

Print Name: _____ Date: _____

Signature: _____

Web Content Owner

Print Name: _____ Date: _____

Signature: _____

**For Internal Use Only**

| Item received by: | Date: |
|---|---|
| | |
| Acknowledged by: | Disposition: |
| | |