

## **Content Management – Web Sites**

Sub-group members: Kendall Hobbs and Mike Roy  
Consulted: Pat Leone, Jennifer Carlstrom, Ganesan Ravishanker

### **Domain/Scope**

"Web sites" includes all freely accessible web content. It does not include subscription sites (journals, databases, etc) or passworded sites (portfolios, blackboard, etc).

"Web sites" includes two main categories: Wesleyan sites and non-Wesleyan sites.

Non-Wesleyan sites are freely available web sites that are linked to, cataloged by, or otherwise compiled at Wesleyan. This includes the library (subject guides, external links in the online catalog), departments (department web pages that include links to subject resources), ITS, Project CuRL, and other official departments and campus web pages. It does not include personal pages that may link to external sites.

Wesleyan sites are all web pages hosted at Wesleyan that are freely accessible over the Internet. This includes personal sites as well as academic and administrative departments, student groups, projects such as the Social Psychology Network, etc. This is a much more extensive issue than non-Wesleyan sites, as it involves creating, maintaining, and archiving content rather than just pointing to someone else's work.

### **Existing practices**

For non-Wesleyan sites, the primary people involved are academic departments, individual faculty, and the library linking to and/or collecting information about sites. There are no standardized methods for departments or faculty to deal with external sites, but a typical way is for a department's web site to include a list of links somewhere on their site. Many faculty have a personal or course web page which includes links. While each department or faculty may have a system for compiling, managing, and updating links, there is no standardized methodology or set of guidelines on doing so.

In the library, most of the keeping track of outside web resources is done on the subject guide pages, which list some good indexes to sites in each subject/discipline area and also list a few good content sites. Rather than create extensive subject indexes to web sites, the focus is on finding a few other places that already do a good job of indexing web sites and point to them. These, however, are only listed on subject guides or course guides; there is no database where these are collected or cataloged, nor is there a standard method for indexing useful web sites that may not fit into a subject guide. To address this, the librarians are currently developing criteria for cataloging free web resources in the online catalog.

In addition to academic departments and the library, there are various other individuals or groups which index web sites, such as some of the web pages for student groups. Project cURL (curricular resource library), hosted at Wesleyan, is an attempt for librarian/faculty/IT cross-campus collaboration to index academic sites for specific subjects. There are also sites such as the Social Psychology Network that is hosted here but is funded, designed, and maintained by the project itself; ITS has very little involvement.

For Wesleyan sites, ITS provides the technical infrastructure and assistance and then leaves it up to departments, individuals, etc, to create and manage their own content. The library also offers some instructional assistance in using FrontPage, but most instruction is done by ITS, such as with group classes offered each semester, or individually by lab consultants for students or academic computing managers and desktop support specialists for faculty.

In addition to providing the hardware, networking, etc, ITS and University Communications have set up a templating structure for academic and administrative departments to manage their content more easily and in a more standardized format. The template system is not mandatory but it is strongly recommended and supported, and most departments use it. The template system allows for a standardized look to Wesleyan pages which helps users with navigation and orientation and helps content providers with consistent standards and methods for producing and managing content. It allows for updates in one area or one page to be automatically spread to connected content on other pages, and for pulling content from PeopleSoft and other databases and thus for being automatically updated when those databases are updated.

For all these sites, official and unofficial, ITS keeps backups of the most recent three changes to a page. The backup is run daily, so a page that is changed daily will have three day's worth of "archiving," one which is changed annually will have three year's worth.

All that, however, concerns only the hardware, software, and support to create web sites. Creating and managing the content that goes on those web sites is up to the departments, groups, and individuals. It is a very decentralized model for content. ITS responds to the content providers' needs, but allows them to determine their needs and to make their own decisions regarding content policies and procedures.

As for costs, it is difficult enough to calculate cost in terms of staff time and money for ITS infrastructure and support for the publicly available free web site since much of that same infrastructure and support overlaps with campus computing in general supporting all the other work now done using the Internet. It would be virtually impossible to calculate staff time and money for content production and management due to its decentralized nature. Other than storage space for putting publicly accessible files, most of what ITS provides for the infrastructure that web sites use would need to be there anyway for general computing needs. Departments, groups, and individuals would need to determine the cost efficiency of their web site related work for themselves.

## **Archiving**

Archiving is an unaddressed area of concern. Other than financial data (archived for a number of years by law) and WesMaps (systematically archived), any archiving beyond standard backups (which does not provide permanent archiving) is left to the content managers dispersed over campus in various departments, organizations, or on their own. There are no standard procedures or resources specifically for archiving, and providing such would be difficult.

Archiving web content is problematic for a number of reasons. Much of the web's content is ephemeral. Authorship or responsibility is often unclear, the reliability of the information is questionable, and a list of other problems it has in common with print sources which archivists must deal with to determine the value of spending time and effort to archive it. In addition to concerns common with print, though, much of the web's content is meant to be changed frequently or deleted soon, and all pages are subject to being revised at any time. The notion of an authoritative edition, or of clearly numbered revised editions, is problematic or nonexistent. Also, a web page's content can include links to other pages, images or sounds or other content pulled from databases or files, content generated by cgi scripts, or otherwise added to a web page on the screen and thus must rely on other sources for its content in a way traditional archival materials do not. Further, there is no guarantee that current content will be accessible by future technology. Perhaps the best way to archive a web page to guarantee you will be able to access all the content later is to print it out (provided, of course, the page does not include animations, sound, etc).

Another problem with archiving, at least at Wesleyan, is that as far as ITS has seen it is not thought of as a problem by anyone but librarians. As noted above, web site managers around campus set policies for content management, and ITS responds to their needs by providing the

hardware, software, and support for them to accomplish what they want. Archiving has not been a concern for web site managers. Any archiving they need to do is done on their own with their own computers or space on network drives. Archiving for more general scholarly or posterity purposes has not been a stated concern, and any such archiving offered or imposed by the library or ITS would need to avoid interfering with the content managers' purposes or procedures for their sites. It would have to avoid being a burden on the content providers. Also, the providers may for various reasons not want some things to be archived; they may want a say in, or a veto of, what gets archived and how, and how accessible the archived content should be.

### **Strengths of existing practices**

In general, the templating system has worked well. It is a home-grown system that was in ways ahead of the curve, and still works well to manage navigation and current information in larger and frequently updated sites. The system has continued to evolve with various tweaks and additions as they are requested by users.

Departments, groups, and individuals who have web sites outside the templating system are free to do pretty much what they want within technical and legal possibilities. ITS does not interfere with content or procedures unless it disrupts the network (e.g. using too many resources) or is brought to ITS's attention for doing something illegal (e.g. copyright violations). Technical and instructional support is available for general needs, and specialized or advanced needs are usually associated with a funded project which can acquire the extra resources it needs.

### **Weaknesses of existing practices – What can be improved**

Probably the primary problem with links to off-campus sites is maintaining the links. Pages change content and location, and often are deleted. There are link checkers which can assist and ITS can help set something up for that, but it is up to the various content people scattered around campus to maintain their links, nothing is done automatically for them. Another problem is that these indexes are scattered around the Wesleyan web site in various department, group, or individual sites, with no easy way to determine where they all are much less to search the contents of them all. Whether it would be worth the time and resources it would take to address this problem by, for example, creating a common index, is another question. The library's plan for cataloging web sites in the online catalog will address this for the library, but not for the rest of the campus. But there is not much evidence that those outside the library see it as a problem.

One problem with the template system is a lack of flexibility, especially with main/home pages for departments, but flexibility has been increased, and there is more flexibility on lower level pages. Another problem is with printing: the pages do not fit on the paper with a 'portrait' orientation, and must be printed as 'landscape'. The printing issue is not a major problem since there is the 'landscape' workaround, but it has been a common complaint since people tend to print with the default 'portrait' and then have to reprint (or complain and then be instructed on the workaround).

No archiving: there are no systematic means or general plans for identifying or permanently saving archive-worthy web content.

### **Opportunities – Emerging trends, benchmarks**

For archiving, the Internet Archive's "Wayback Machine" at [www.archive.org](http://www.archive.org) has been archiving web pages since 1996. It is not comprehensive, and archives primarily text, and to use it you must know the URL of an old page (or of another old page that linked to it) but it is probably currently the best general web archive. The LOCKSS project is working on providing another alternative, based on the idea that lots of copies keeps stuff safe. This is aimed at purchased or subscribed content, though, so it is not directly relevant for the task of archiving freely accessible web content created here or linked to elsewhere, but this or something like it may be adaptable to the task. Portico is another effort to archive academic journals, and may in the future be a model

for archiving web sites. Project Prism at Cornell is attempting to address archiving digital collections in general including web sites.

### **Threats – What could go wrong**

Storage space will likely be an issue. Although the cost of electronic storage media keeps going down, the demand is correspondingly increasing. The storage issue is particularly threatening in the context of archiving. With no archiving system and with an increasing demand to use storage space for current demand, older web content of archival value may be lost.

### **Short-term recommendations (within a year)**

Before the current template system, there used to be a campus web committee. This could be reformed, to have input from different areas of campus to suggest changes to, or respond to suggested changes by, Communications and ITS.

The printing problems with the templating system seems like it should be on a short-term list, but that is not on the plate of issues those running the templating system plan to address in the near term. This needs to be addressed whenever the current templating system is reviewed.

A short-term possibility for archiving Wesleyan web content is to take periodic snapshots of the entire Wesleyan web site and keep them in a permanent archive. This would address Wesleyan web content, but not content produced elsewhere.

### **Long-term recommendations (two to five years)**

For archiving content elsewhere, making our own archived copies would involve copyright and other issues. We probably have to wait for long term and wider answers to the issue of archiving non-Wesleyan web content. We should watch projects such as LOCKSS and Portico to see how they develop.

### **Readings**

An article with a good summary of the problem of preserving web content, and the efforts of Project Prism to address the issue, is:  
Kenney, A. & McGovern, N. (2002). Preservation risk management for Web resources: preserving Web content requires substantial resource commitments and flexible and innovative approaches to new technologies, organizational missions, and user expectations. *Information Management Journal*, 36 (5), 52-61.