From the College Democrats to the Falling Illini: Identifying, Appraising, and Capturing Student Organization Web Sites

Christopher J. Prom Ellen D. Swain

ABSTRACT

For nearly a decade, student organizations have used the Internet to promote events, record activities, issue statements and maintain officer rosters. Student Web sites are potentially an important resource for scholarship on twenty-first century student experience, providing both evidence about the activities of the organizations and information about the student body. Based on findings from a multi-component case study at the University of Illinois at Urbana-Champaign, this article analyzes the documentary value of student organization Web sites and discusses ways in which archivists can use offline browser technology to capture extracurricular activity more fully.

Over the past thirty-five years, scholars, students, alumni, and campus administrators have examined student life to explore the origins and history of such topics as the civil rights movement, anti-war protests, the sexual revolution, the growth of consumerism and the resurgence of conservatism. Scholarship on student life has informed not only our understanding of broader social phenomenon, but, more practically, curriculum development and higher education administrative practices. In fact, some scholars have insisted that the out-of-classroom experience is “one of the most potent educational forces affecting the
Prior to the late 1960s, many archivists believed extra-curricular documentation was unimportant or outside their collection scope. But by the mid-seventies, archivists wrote of the important research value of student records, and a Society of American Archivists subcommittee proclaimed that “institutions of higher learning have a special obligation to preserve the records of individual students, student organizations and campus life.”

Although many archivists value student documentation in the abstract, recent studies suggest that few are actively collecting student material. Why? Most likely, it is because of the transient nature of student population—student organizations form and dissolve frequently and their leaders are in office only briefly. Documenting students requires great amounts of time, effort, and determination. New media, such as Web sites and email, would seem to further complicate the documentary process. Or do they?

In 2003-2004, the authors conducted a National Historical Publications and Records Commission (NHPRC) funded project to investigate the ways in which archivists can document the lives of students in the electronic age. More specifically, our study focused on

evaluating the potential research value of student organization Web sites, identifying students’ documentation patterns and Web-site use, and determining a “best practice” approach for capturing Web site content. We divided our work into five stages, each represented by a question:

1) What range of documentation do students currently produce?
2) What materials do student organizations—the focus of much student extra-curricular activity—generate and collect?
3) What information is contained in student organization Web sites?
4) How do universities administer and provide access to such records?
5) How might archivists identify and capture such student organization records, particularly Web sites concerning student organizations?

By examining these issues, we hoped to develop a set of criteria that would help archivists identify, appraise, and capture documentation related to students’ extra-curricular lives. Based on the results of our project, we argue that the advent of electronic media, specifically the proliferation of student organization Web sites, actually improves archivists’ ability to document student activity and experience. In addition, our study sheds some light on practices that can be used to organize, preserve, describe, and provide access to student organization Web sites

**Student Documentation Patterns**

Before initiating a documentary program focused on student organizations, archivists must ask a fundamental question: What types of documentation are student organizations producing and how well do those documentary forms represent student activities?
Developing and implementing Web-capture guidelines first required a thorough investigation of how student organizations used their Web sites, how well the sites represented organizational activities, and what information the sites contained (or did not) contain, relative to other documentation the organizations produce.

In our NHPRC application, we proposed to answer these questions by conducting a survey of registered student organizations (or RSOs) at the University of Illinois at Urbana-Champaign (UIUC). UIUC provides a good test case for several reasons. As a public land-grant institution, with a large enrollment (26,995 undergraduate students and 9,717 graduate/professional students in spring 2004), UIUC’s student population includes members of many racial, ethnic, religious, social, and cultural groups. Students have organized themselves into a wide range of organizations (1,075 in fall 2004), and the University Archives includes a dedicated Student Life and Culture Archival Program, which provides a point of contact on campus and with many groups.

Two earlier smaller studies informed our research. First, a survey of 96 students who visited the UIUC Archives in fall 2003 indicated that participants saved photo albums or scrapbooks and course notes in high numbers (see Figure 1). Further, only 21% of the 96 participants indicated that they had a Web site and only 6.3% authored a blog. Figure 2 shows the typical topics discussed in Web sites and blogs by the small number of students who authored them.5

---

Second, to gain additional insight into student record generation and record keeping, Ellen Swain consulted the Student Life and Culture Archives’ newly-established Student Advisory Committee. The Committee, comprised of 12 student leaders from a wide range of
student organizations, offered insight into the types of materials student groups kept, how these materials were maintained, and what problems and issues student organization leaders faced in keeping records. From the discussion, it became clear that leaders had no systematic method for saving records long term. Each organization used electronic communication heavily, but few knew whether hard copies of their communiqués were being saved. In addition, leaders indicated that they were frustrated with the amount of information lost by frequent leadership turnover.  

**Student Organization Record Keeping: RSO Survey**

Both the class survey results concerning individual students’ recordkeeping practices and the student advisory committee discussions jump-started our thinking about how and why student organizations use Web sites. How important are Web sites to student groups? What content do they contain and why? Further, are Web sites the primary recordkeeping means for student groups? Would the capture of a student organization’s Web site sufficiently document the organization’s activities? Could better management practices regarding the group’s Web site facilitate organizational memory? In March 2004, we administered a survey targeting RSO (Registered Student Organization) leaders on campus and addressing four subject areas: demographics, record keeping practices, Web site content, and Web site use and purpose. We posted the 22-question survey online and invited leaders of all RSOs to participate. Over a three week period, 375 of the 1,075 registered student organization leaders on campus completed the survey, a response rate of 34.8%.

---

6 For an analysis of the class survey research and use of the student advisory committee, see Swain, Ellen, “College Student as Archives’ Consultant?: A New Approach to Outreach Programming on Campus” *Archival Issues* 29:2 (2005).
The first section of the survey provided basic demographic information about the participating groups. Upon registering with the RSO Office, each student organization at Illinois places itself into one of 15 subject categories (Figure 3).

![Figure 3: Organization Types (N=375)](image)

The majority of respondents represented academic, athletic, cultural or fraternal organizations. Tracking these categories allowed us to compare responses to questions by group type. In addition, we asked participants to indicate the size of their organization (Figure 4), and found that the largest number of organizations in the survey had a membership of one to thirty persons.
The second section of the survey focused on the RSO’s recordkeeping practices. We asked participants to indicate each category of documents they produced, saved in hard copy, and saved electronically. Results, shown in Figure 5, indicated that student organizations most often document their activities by e-mail, photographs, membership records and fliers. Not surprisingly, the members of student organizations produce a large amount of e-mail regarding their group’s activities, yet they save little of it. Findings also indicated that student organizations save documents in hard copy more frequently than in electronic form. Photographs are the most likely document type to be saved electronically. Students saved e-mail and videos in hard copy.
*Due to an error in the survey instrument, data that would allow us to differentiate electronic vs. analog recordkeeping practices was not gathered for e-mail and video.

The president of the group was most often responsible for maintaining the group’s records, followed by the secretary, as shown in Figure 6. The “other” category included positions not listed among the survey choices, such as “the director of information,” “interns,” and “the membership chair.” Turnover in the record keeper position occurred annually in the vast majority of student organizations. Responses concerning the physical location of records netted a variety of responses—including the cubicle in RSO complex, on the president’s computer, in closets in the chapter house or in e-mail accounts. One leader indicated that the only records kept were on the Web site. This range of responses illustrates the unsystematic nature of student record-keeping. Although this finding may not be surprising, it suggests that considerable effort is needed to preserve student records before
they disappear.

![Figure 6: Person Responsible for Organization's Records (N=375)](image-url)

Questions about content of Web sites comprised the third section of the survey. Seventy-eight percent of the participants indicated that their organization maintained a Web site. Figure 7 provides a breakdown of the Web content for those organizations that maintain a site, showing that contact and event information, photos, member lists, mission statements, and links to national parent organizations are included most often. Chat and blog options are not widespread. When asked if organizations kept paper copies of the Web site content, 56 percent of the 179 students who answered the question said that they did not. While some of the information on Web sites—for example officer lists and press releases—is likely kept in paper format, it is extremely unlikely that any of the student groups are systematically printing their entire Web sites for purposes of recordkeeping.
The final section of the survey—Web site use and purpose—offers insight into the importance of a Web presence for student organizations. When asked why the organization maintained a Web site, 91 percent checked publicity. Large numbers also indicated that membership recruitment (69 percent) and information sharing among members (65 percent) were important reasons.

When asked to identify the primary users of the organization’s Web site, almost half of the participants listed “members of the organization,” while 23 percent cited prospective members and 18 percent listed the campus community (see Figure 8). The most heavily used features or type of information on the Web site were contact lists, event/publicity information, and photographs. As seen earlier, these features also were cited as the most common features on sites, and publicity was the most common purpose for all sites.
Finally, when asked about the importance of Web sites and e-mail to the operation of the organization, 11.6 percent of participants indicated that the Web site was critical, and 24.8 percent felt it was very important to their group, while 53.4 percent indicated that their site was moderately or somewhat important. In contrast, participants overwhelmingly responded that e-mail was critical to the operation of their organization. Nearly 89 percent indicated that someone in their organization regularly updated their Web site.

The results from the RSO survey offer several insights into the relatively informal ways that ways students document their lives. First, Web sites are important to the operations of student organizations and Web sites have clear evidential value. Results indicated that Web site use is widespread among student organizations (78 percent). Web sites provide evidence regarding existence of the organizations (and the social trends they illustrate) as well as the basic activities and functions the groups serve within broader campus community. Web sites are used for publicity, membership recruitment, and information sharing. The most utilized features of the sites include contact information, event publicity items, photos, and member lists. Importantly, 90 percent of participants responded that their
organization’s site was *at least* moderately or somewhat important to the group’s operation.

Second, the survey provides an understanding of how well Web sites represent student activities. The Web sites have informational value that should be of future research interest, but documents central to the operation of these groups—such as by-laws, minutes, reports, newsletters, and correspondence—are less likely to appear on the Web sites than are officer lists, biographies, mission statements, event calendars, and photographs. Archivists who wish to implement a full collecting program for student life must continue to document organizations’ activities in more traditional ways. Archivists might also have an opportunity to encourage students to post these documents on their Web sites, perhaps through the development of Web site guides produced in conjunction with registered student organization offices.

Finally, the survey results illuminate student organizations’ recordkeeping practices. Although students are not saving Web content in hard copy, they continue to save hard copies of other documents created by their groups. In addition, students responding to our class survey stated that they save scrapbooks, photo albums, ticket stubs, and course materials in paper copy. Not surprisingly, e-mail, a critical communication form central to group operations and to students personal lives is not being saved long term, although those groups that produce minutes, by-laws, reports, histories, and scrapbooks tend to save these documents with great frequency.

Web sites offer an important window into student activities, but they are not *the* definitive source regarding student culture. These sites provide information about the organizations to an outside audience, but they rarely include more “meaty” sources such as minutes, reports, or correspondence that document the activities central to their operation.
They are, however, an important source that should be captured with the caveat that substantial records may be missing. To capture the full range of documentation produced by a student organization, archivists need to collaborate with that organization on a personal level to acquire records that the sites do not contain.

Documenting student life in the electronic age will not eliminate the necessity for personal contact with groups. However, Web sites serve functions beyond the original reasons for which they were created, which gives them more archival value in the aggregate than each site may hold individually. An individual organization creates a Web site to promote its cause and facilitate communication among members. But the universe of such Web sites at any institution documents the entire range of student activities and extracurricular activities taking place at it. By preserving student organization Web sites, an archivist without the time or resources to implement a full program to collect student records can nevertheless capture core documentation about extracurricular activities on campus.

Description of Typical Student Organization Web Sites

An examination of student organization Web sites at UIUC confirmed our initial impression that the entire body of student organizations Web sites on one campus has archival value greater than the sum of their individual parts. While we were unable to complete a full content analysis of the Web sites for all 1,075 student organizations active during the project period, we reviewed a sample of UIUC student organization Web sites, selecting a few well known campus organizations as well as several random organizations from the categories maintained by the Registered Student Organizations Office. Several are cited here as typical examples of the document types and content available.
The Web site of the University of Illinois Graduate Employee’s Organization is one of the most extensive on campus. The Web site includes bylaws, meeting minutes, press releases, photographs, and other materials documenting the activities of the organization. The materials offer insight into labor organization on campus, student attitudes about financial aid, and teaching conditions.

The Web sites of other major organizations on campus, such as some of the major religious organizations, include similar items, as well as a wealth of photos and materials documenting student activities. For example, the Web site of graduate Intervarsity Christian Fellowship contains a doctrinal statement, events calendar, photo gallery, reports, and other documentation. The Web site of the campus Hillel Foundation contains similar materials documenting student activities on campus, social action, and religious belief.

Political organizations, such as the College Democrats, typically include officer lists, press releases, events calendars and photos. The Web site of our local conservative journal, the Orange and Blue Observer, contains outspoken political statements, news items, e-mail, and photos.

Social or recreational organizations typically include announcements, photographs, officer lists, and promotional materials. Even the Web sites of smaller and more transient organizations, such as the “Seinfeld” fan club, are of potential value, if only because they

7 http://www.shout.net/~geo/ (Accessed March 4, 2004). The descriptions of all student organizations websites are taken from the copy of accessioned into the University of Illinois Archives on March 4, 2004. Copies of the site as of that date are available by contacting the University of Illinois Archives.


contain a large number of identified photographs of students—thus providing the UIUC archives a potential way to supplement its photographic holdings.

For the archivist, the existence of these Web sites and other forms of electronic student communication offers many possibilities. At UIUC, we are lucky to have an endowed Student Life and Culture Archival Program that can also solicit student information, in both digital and analog format, directly from the organizations. But when so many other issues beg for attention, it is unlikely that archivists who do not command such resources are able to undertake a full-fledged program to collect paper or other analog materials directly from the groups—and we doubt many students are contacting archives on their own to deposit such records. Because, however, the process of capturing student organization Web sites can be automated in whole or part, institutions without a dedicated student life program can capture some student records on a systematic basis,

Administration of Student Organization Web sites

The environment in which the records created by student organizations are managed is essential for understanding how archivists might also be able to find and capture student organization records on-line. How can student organization Web sites, be located, appraised, and captured by archivists, if we do not understand the context in which they are managed?

In order to provide some recommendations that would apply beyond the University of Illinois, we studied administrative practices for student organizations at a sample of U. S. colleges and universities. We examined practices at 369 institutions of higher education, including doctoral and masters Universities as well as baccalaureate and tribal colleges.12
The data gathered during the examination of these 369 Web sites was subsequently loaded into a database, sorted and analyzed.\textsuperscript{13} We found enough consistency in practice across the United States to provide some general recommendations, although each archivist will need to study local practices carefully before undertaking a capture project.

Our survey showed that most colleges and universities have developed an administrative structure to manage student organizations and to provide them an on-line working space. As Figure 9 shows, most institutions have a “registered organization office” that allows the college or university to exercise some supervision over the groups. For archivists, these offices provide a contact point. Currently, many of these offices provide on-line lists of student organizations or policies regarding the management of the organizations.

\textbf{Figure 9: Institutions with Student Organization Office}

\begin{figure}[h]
\centering
\includegraphics{figure9.png}
\caption{Institutions with Student Organization Office}
\end{figure}

\begin{verbatim}
Doctoral  Masters  Baccalaureate  Tribal
66  56  149  115  147  86  7  2
\end{verbatim}

\textsuperscript{12} The sample was drawn from \textit{The Carnegie Classification of Institutions of Higher Education, 2000 Edition}. Electronic data file, third revision (2003) from a population of 1506 Doctoral, Masters, Baccalaureate, and Tribal institutions listed, by selecting every fourth institution listed. This provided an over-sample of 63 institutions at a 95% confidence level. Data for seven institutions was not included because the institutions were located in Puerto Rico and information regarding student organizations could not be easily located on the Spanish-language Web sites.

\textsuperscript{13} Our graduate assistant located and entered data for each of the institutions using electronic and print resources. Information was entered into an Access database, queried, and analyzed using statistical analysis tools in Microsoft Excel.
In the past, the University of Illinois Archives has had some success in soliciting paper-based documentation from our RSO Office, including constitutions and publications produced by organizations, but it is unlikely that similar paper-based documentation will be forthcoming in the future, either at UIUC or elsewhere, because the paper-based filing systems are no longer maintained. However, the on-line list of student organizations serves as a key access point for identifying organizations and capturing their electronic documentation. The list typically includes the organization’s name, some contact information, and a brief description of the group. In most cases, a link to the organization’s Web site is also provided.

In our survey, we correlated the existence of such an on-line list with the existence of an archival program. Of the 174 institutions with a university archives 162 (or 93 percent) also provided a list of registered organizations. For these institutions, our analysis identified 21,254 distinct student organizations. In other words, a great deal of on-line documentation regarding student groups exists at most institutions with established archival programs. But how might an institution identify the records of an organization, and, more importantly, decide which ones best fit into the archives’ scope? How can such a task be accomplished, given the other priorities that beg for the archival staff’s time and attention?

**Basic Appraisal Decisions**

Past literature on the appraisal of electronic records offers only partial guidance in dealing with student organization Web sites. As Adrian Cunningham has noted, writing about electronic records have been dominated by a “corporate myopia” in focusing only on the records of large organizations and in particular on those created by large administrative...
However, most writers on the appraisal of electronic records have agreed that “the preservation of evidence documenting the functions, processes, activities and transactions undertaken by the institution or individual” must be the primary aim of appraising electronic records. If one assumes that evidential value is the only thing that matters in dealing with electronic records, the practical choice in dealing with student organization Web sites is relatively straightforward: should one take a selective approach, capturing information only for selected sites, or cast the net as broadly as possible?

Assessing the function and context under which the records were created can help answer this conundrum. Terry Cook has argued that appraisal is the “articulation of the most important societal structures, functions, records creators, and records creating processes, and their interaction, which together form a comprehensive reflection of human experience.” If one is seeking to document the life of a heterogeneous, disparate, and decentralized student body, it makes sense to capture as many student organizations websites as possible.

The alternative, selecting to document only particular student organization websites for capture, would appear to find little support in the literature about the appraisal of electronic records, if only because so little attention has been paid to the notion of appraising the informational value of electronic records. Linda Henry, while defending the applicability of traditional archival concepts to electronic records, notes that the contents of electronic records...
records, not just their function, must be measured. Preliminary principles for appraising the
archival value of Web content emphasize that decisions should take place at the series, not
the document or item level, and that appraising materials at this level can facilitate
harvesting, aggregation, description, and retrieval by preserving context and provenance.

At the University of Illinois, the content of the student organization websites is most
analogous to two existing paper-based record series, Student and Faculty Organizations
Constitutions and Registration Cards and Student Organizations Publications. Both of these
record series were generated by the Registered Organizations Office, which gathered the
information for a number of organizations. The Web sites also include a limited number of
records that are typically found in separate records series that the UIUC Archives has
established for individual organizations (such as minutes, correspondence and photos).

Since the entire set of Web sites is managed under the purview of the Registered
Student Organization office, we appraised it as a discrete record series which contained
records from many creators, but which shared a functional provenance in their management
and arrangement by the RSO office. Viewed from this perspective, to entire set of Web sites
appeared worthy of accession, preservation and access under a common reference point in
our descriptive and access system. Further, the Registered Organizations Office provides an

---


18 Richard Pearce Moses and Joanne Kaczmarek, “An Arizona Model for Preservation and Access of Web

19 In the UIUC Archives, a “Record Series” is functionally equivalent to way many repositories term a “Record
Group” or a “Collection.” These record series are more fully described at
http://web.library.uiuc.edu/ahx/archon/controlcard.php?id=936 and

20 At the University of Illinois Archives, a record series is established for each student organization, containing
all of the materials produced by that organization and related by provenance. Generally, these series then
include subseries for minutes, scrapbooks, correspondence, photographs or other types of documentation
produced by the organization.
Identification and Capture

The final part of our project tested methods for identifying and capturing student organization Web sites at the University of Illinois at Urbana-Champaign. Capturing the Web sites presented some challenges from both an administrative and technical point of view. We eventually found a method that allowed us to capture many student organization Web sites in a semi-automated fashion, and we believe this method could be used by other archives.

For testing, we decided to try to systematically capture as much information as possible about all student organizations rather than focus on a few selected groups. We thought that it would be very difficult, if not impossible, to decide which Web sites might hold future research interest, and capturing as many as possible fit well with our prior appraisal decision to capture the entire range of Web sites as a single record series. Furthermore, we anticipated that users may be interested in finding information about a specific organization, and saving something for all organizations meant that we would be able to provide some information about any student organization that existed. In addition, the entire set of Web sites would be valuable because it would provide insight into student culture as a whole. Finally, we wanted to put Web capture methods through a real test. In the past, we had used Web capture software, such as offline browsers, to some good effect, but on much smaller and less dispersed resources.

In the end, we decided it made more sense to attempt the capture the Web sites
ourselves rather than work with each of the individual records creators in order to accession copies of their sites into the Archives. Some might object to this approach because emerging appraisal theory for electronic records places much emphasis on working with records creators during the records creation stage in order to identify records of potential interest and to ensure that adequate documentation is being produced.

But aside from the sheer impracticality of working with 1,000 or more creators, it must be noted that student records are created and organized with low levels of formality and funding. Since the groups are born (or reborn) each academic year, the sites they create seem to reflect the typical life cycle mode of records rather well, and it makes much sense for archivists to capture them, take custody over the copies, and provide access to them, much as advocates of applying traditional archival models to electronic records recommend.21

Active involvement of the organization during the capture process, while desirable, is not strictly necessary, since the websites are publicly available to anyone in the world sitting at a computer connected to the Internet.22 Although a copyright lawyer might object that someone copying a Web site in toto is engaging in unauthorized third party copying, we believe it is possible to mitigate such objections by taking several simple steps. Mass e-mail makes it possible to inform leaders of the organizations that the sites will be captured for inclusion in the Archives. At UIUC the director of the RSO office has been willing to send such e-mail on the Archives’ behalf. In addition, Web site administrators can block capture

---

21 Luciana Duranti and Heather MacNeil, “The Protection of the Integrity of Electronic Records: An Overview of the UCB-MAS Research Project,” Archivaria 42 (Fall 1996), 46-67. For records created by informal groups, applying notions common to those advocating a “new paradigm” for managing electronic records (such as non-custodianship, extensive consultation with records creators, and appraisal by function) would seem particularly inappropriate.

22 Our study did not attempt to harvest website on intranets or behind firewalls.
by using a robots.txt, .htaccess, or similar file. In conducting our download, we obeyed the
instructions in these files, so that organizations that restricted copying were self-selected out
of our sample. Finally, if representatives of the organization object to inclusion of the
material in the archives at a later date, the archivist can negotiate terms of restriction or, if
necessary, withdraw the Web site from the collection.

This is not to say that the archivists should totally disengage from contact with the
creators. Before capture, we might take some basic steps to help organizations create an
adequate record, such as issuing general recommendations via e-mail on the type of content
most likely needed to be of archival value, as Adrian Cunningham recommends for electronic
records created by individuals.

However, problems of a practical nature made it impossible to pursue detailed work
with the creators to actually capture the Web sites. Since these issues were rooted in the
nature of the records we were trying to capture, we believe similar problems would be
inevitable at other institutions. For example, we initially contacted the student organization
officer for the UIUC campus, and he offered to provide a mirror of Web sites stored on his
server. However, we had no luck actually getting a copy of his files from the system
administrator. We later found out they were stored in a non-standard system and would have
been unusable without the database that stored the files and converted them into standard
HTML output.

This problem may or may not apply at other institutions. But even more troubling,
we discovered that the files stored on the central RSO server represented only a small

percentage of the total number of student organization Web sites. Many organizations store their files on departmental or non-university servers. The distributed nature of this content makes it difficult to capture. Since it was not feasible to solicit the sites from hundreds of students, we had to find an automated method to harvest these sites. Here the on-line list of organizations, mentioned previously, proved extremely helpful. We used the list as an entry point for the automated capture of information regarding all student organizations linked.

Capturing distributed web content for archival purposes is an extremely tricky business. Adobe Acrobat has a web capture feature that allows one to create copies of Web sites. We rejected Acrobat because, it changes the Web site to PDF format and is not flexible enough to allow automated download of distributed content or large Web sites.

Other institutions, particularly the Library of Congress and Cornell, have partnered with the Internet Archive\textsuperscript{25} to receive content from its webcrawls and to develop standards for indexing the pages. This has resulted in some very successful projects, particularly the September 11 Archives.\textsuperscript{26} However, the Internet Archive appears to be no longer capturing student organization Web sites at the University of Illinois, and our attempt to use the resource from the last capture they had undertaken in July 2004 showed many broken links.\textsuperscript{27}

We realized that we might have better success capturing the files ourselves, so we investigated two basic options. Heavy-duty webcrawlers, such as the open source Heritrix


\textsuperscript{26} http://www.loc.gov/minerva/ (Accessed November 5, 2006).

\textsuperscript{27} The Internet Archive now provides a subscription service, Archive-It, that institutions can use to capture information by harvesting information from seed URL’s. Content is stored on the Archive-It servers. The service was not available at the time this study was completed, but it may offer a good option for archivists seeking a relatively simple capture mechanism and off-campus storage options. http://www.archive-it.org (Accessed November 5, 2006).
can be used to download large amounts of data. However, Heritrix would be very difficult for all but the most tech-savvy archivists to install and use. A convenient alternative is off-line browser software. Off-line browsers are intended allow computer users to download Web sites for browsing while not connected to the Internet.

We tested three: the trial version Offline Explorer Enterprise ($400); a trial version of the mid-range Black Widow ($40), and the free HTTrack (open source). Offline Explorer has some nice features, such as the ability to export Web sites and to schedule crawls in advance, but it repeatedly crashed when we it followed links away from the student organizations server. It would, however, be very useful for small scale crawls of individual sites, and archivists could use it to set up an automated crawl of selected sites, at monthly or yearly intervals. Black Widow did not include filtering options sophisticated enough to be of much use. We quickly abandoned it.

The free tool HTTrack proved to be flexible, stable enough, and easy to use. We judged it suitable not only for our project but can recommend it to other archivists who wish to capture and preserve web content. It allowed us to capture the distributed content of student organization Web sites with a moderate degree of success.

The filters in HTTrack were powerful enough so that we could direct the crawler to use the RSO list to harvest student organization sites from distributed servers. For the 715 organization Web sites active on the day we attempted the test, 366 (51 percent) were completely captured and an additional 123 (17 percent) were partially captured. Although this may not seem satisfactory at first glance, it represents an enormous amount of information regarding the organizations—26,778 files and 2.3 gigabytes. The paper based

\[ \text{\footnotesize 28 Heritrix is the webcrawler underlying the Internet Archive. http://crawler.archive.org/ (Accessed November 5, 2006).} \]
files currently held by the University of Illinois Archives that are most analogous to the Web sites are even less complete in their coverage (see footnote 20). And it should be noted that the files of many of the largest organizations on campus, such as political organizations and social clubs, were captured in near-perfect detail, including not only basic information such as officer lists and mission statements, but newsletters, photographs, and other materials of likely research interest.

Nevertheless, the failures need explanation. They can be attributed to dead links, inconsistent coding practices on the pages, the use of technologies like Flash, and references to third party sites, such as photo galleries. An examination of the sites where failure occurred indicated that it would be difficult, if not impossible to tweak the web capture software to achieve total success. They include too many external links to allow the construction of an effective algorithm. In addition, many organizations nested detailed content, such as newsletters or articles, several links deep into the structure of the website, making them difficult to capture in a general crawl designed to capture information about all organizations, rather than complete information for a selected number of groups. For organizations for which we wanted to achieve a more complete capture of information, we were forced to undertake more targeted crawls of their Web sites by including the site’s specific URL directly in the capture software. For example, the Web site of Illini Life, a religious organization, contained a large number of links to video and audio files. The links to articles in a local feminist magazine, *Pandora’s Rag*, were so deeply buried in the structure of the website that they were not automatically captured on the first pass. For both Illini Life and *Pandora’s Rag*, we captured additional content after manually adding the URL’s of the website into the list of sites being searched by HTTrack.
The entire process of capturing, describing, and placing the websites into storage took approximately 5 hours of staff time (although the crawl ran for over 24 hours once started.) Since we will reuse the same search algorithm, the archivist’s time to complete the process will be considerably lower in future webcrawls.

**Description, Preservation, and Access**

One advantage to capturing student organizations on the level of an entire record series is that our post-capture routine to arrange, describe, preserve, and provide access to the files was greatly simplified. The off-line browser saved the lead-in-list of organizations that the RSO office provides. The list functions as a de facto finding aid for the record series, and it can be easily browsed or searched to provide access to the records of any individual group. In other words, the collection partially described itself. We created a collection-level description for it in our database but did not need to create a detailed finding aid, as we would have for a similarly complex paper-based record series. The web capture software saved the files into a series of nested subdirectories and converted some (though not all) interactive pages to HTML. The total amount of data captured was 2.3 gigabytes, and we store the files on a networked server that is backed up daily. In addition, copies of the websites were saved to Archival Gold CD’s and accessioned into the Archives like any other record series.

The record series representing the Web sites is described in our on-line database,\(^\text{29}\) and the files have been accessed by staff and users several times since being accessioned. Because of copyright considerations, we do not provide public access to the files through our website, but since they are stored on a networked server, remote access can be provided on a [link](http://web.library.uiuc.edu/ahx/archon/controlcard.php?id=2527) (Accessed November 5, 2006).

case-by-case basis. We may provide access through our website to the UIUC community, if it proves to be technically feasible.

While this description, preservation, and access regime could be enhanced, we implemented it with a minimum amount of time and fuss. We believe it provides an effective short term model for managing electronic records stored in standard formats such as html, jpg, and mp3, since the files in question are open standards and can be accessed using a web browser. Significantly more research is needed to discuss long-term preservation, and we hope to address this issue at a future time, as part of the University of Illinois Library’s digital preservation planning activities.

Conclusion

As a profession, we now have the ability to capture a wide range of documentation from distributed and diverse groups of people as that documentation is being created. Capturing it is a critical weapon in our professional arsenal because the lifecycle of many records is becoming shorter and shorter, both for individuals and for organizations. It gives us the power to capture the records of many organizations and individuals that in the past would have disappeared from the historical record with nary a trace. But because most student documentation remains off-line, web capture is only one of many strategies to document student life.

Individual students and student organizations produce a variety of paper-based and electronic information that may be of archival value, including scrapbooks, hard copy, and electronic correspondence. Student organization Web sites provide only one element in documenting student life. Taken as a whole, they provide a snapshot of all student life on
campus at any given time, but they are a bit like the kiosks where students post fliers, providing an interesting yet somewhat superficial view of student life. Further, they are often technically difficult to capture in their entirety.

Based on our experience, we recommend that archivists use web capture software to attempt a once-yearly capture of all student organization Web sites for their campus, by seeding the software with the URL for the list of student organizations on their campus. Ideally, the capture should take place late in the spring semester, after the organizations have posted most of the information they are likely to post for the year, but before a new crop of students has revamped the website. While the approach we are recommending is likely to miss some pages and accidentally capture unrelated information, it does preserve a representative sample of student life as a whole at the moment when the files are captured. We believe it is suited to institutions that may not have the resources to implement a full collecting area in student life.

Alternatively, or a supplement, archivists might consider several other strategies to assist them in compiling a record of student organizations on their campus. Those interested in capturing student on-line information for a select number of student groups could use web capture software to undertake a highly-targeted approach by seeding the software with the URLs of those websites they wish to capture. Similarly, archivists might undertake targeted crawls in response to particular campus events, such as controversies or disasters. Other archives might focus on a limited number of organizations and attempt to solicit and capture both electronic and non-electronic documentation. Appropriate criteria to be used when contacting groups would include factors such as the size, age of organization, the amount of influence or prestige it holds, and the nature of the activities it documents. Archivists
pursuing a targeted approach should seek balance among the different types of organizations whose records they solicit. Since working with individual records creators will be more resource intensive than simply harvesting all student organizations Web sites at a surface level, such a strategy should only be undertaken after full assessing personnel and technical capabilities in the light of the repository’s overall mission.

Whichever method is choosen to document student culture, the UIUC project has shown that the Internet provides our profession with a new opportunity. Few archives are lucky enough to have an endowment to support a full documentary program in the area of student life, but the techniques outlined above can be applied without such resources. A good relationship with library or campus IT offices can help in the process. All that is needed is a little time, a good sense of what is worth capturing, and a little technical expertise.