
This study describes a content analysis of four departmental and administrative websites at the University of North Carolina at Chapel Hill. The analysis was conducted to determine the feasibility of describing website contents with the categories itemized on the record schedule for the relevant department. The majority of the material on three of the websites did not correspond to records listed on the relevant record schedules, though the fourth had material that corresponded better. While this study provides insight into the structure and content of the websites, it also provides significant evidence of the inapplicability of the record scheduling system as it currently exists for identifying record content on UNC’s websites for long-term preservation.

Headings:

Colleges and Universities-Internet Resources

Digital Asset Management

Electronic data archives
Contents

1. Abstract
2. Introduction
3. Intent of Study
4. Literature Review
5. Method
6. Results
7. Importance of Study
8. Summary and Conclusions
9. Future Study
10. Works Cited
11. Appendix A: Record schedules
12. Appendix B: Site Diagrams
Introduction

The more institutions depend on digital materials and publication, the more institutional archives will need to include preservation of those materials as part of their mission. Websites have moved far beyond simple groups of files to become portals for information and services. The structure and design of a site, the features and art used or avoided can create a positive or negative impression in visitors. This creates a much more accessible and detailed public face of an institution than was available through print publications. Redesign of departmental, institutional and even governmental sites such as "www.usa.gov", formerly "www.firstgov.gov" indicate a growing desire to design websites to serve particular purposes and audiences. These promotional decisions mean that websites are not only important to capture because of the information on them, but also to provide information about how an institution styles itself. In addition, records, important to both the running and the history of the institution, are being distributed through websites. A record is, “data or information in a fixed form that is created or received in the course of individual or institutional activity and set aside (preserved) as evidence of that activity for future reference.” While the main purpose of placing records on websites may be to make them more accessible, the online version is, in some cases, replacing the more costly printing and distribution process. This change is resulting in an increasing number of records available only online. These online records need to be saved along with their print predecessors.
Currently, a lack of literature suggests that many institutional websites are not being saved with any consistency or regularity. There are at least two reasons for this. First, a good case has not been made for the need to preserve these websites at the institution level. While there have been a number of web archiving conferences over the past 6 or more years, much of the focus has been on national projects or technological solutions. In the six International Web Archiving Workshops that have been held, only the 2004 workshop had a presentation focusing on domain level web archiving at an institutional level. Many organizations view websites as vehicles that provide access to information and create an online presence, and give little thought or resources to preserving that information over time. In a 2005 study of current practice at many institutions, Cloonan and Sanett noted that "the respondents focused on their institutional collections rather than on their internal administrative or financial records." The webmaster of a site may be the only one to know precisely what is on a particular website, and he/she generally does not advocate within organizations for preservation procedures. Second, websites fall between two existing definitions, that of records, and that of publications, which the SAA Glossary defines as, "A work that expresses some thought in language, signs, or symbols and that is reproduced for distribution." Despite this broad definition, preservation of publications is generally focused on well-documented, periodic output with identifiable metadata. Organizational newsletters are one example of printed publications that institutions have collected and preserved effectively. Falling between traditional records and traditional publications, many organizations have overlooked the preservation of websites either in whole or in part.
A system already exists to deal with print records. Records management is, “the systematic and administrative control of records throughout their life cycle to ensure efficiency and economy in their creation, use, handling, control, maintenance, and disposition.” Found within institutions with administrative, historical, or accountability needs for their records, records management programs are the intermediaries between the records creators and the storage or destruction of records. Records management programs are generally required for government institutions, including public universities, but many other public and private institutions have records management programs as well. Records managers make judgments as to the long term value of different documents. Those with such value are noted for transfer to an archive or other long term storage and those without are judged for their short-term value and given an interval after which they can be destroyed. These judgments are written in records schedules.

A records schedule, also called a retention schedule, is, “A document that identifies and describes an organization's records, usually at the series level, and provides instructions for the disposition of records throughout their life cycle.” UNC-Chapel Hill has an institution-wide schedule, and many administrative units have one more closely describing their particular records. By scheduling records, the information has a good chance of reaching the archive in predictable forms if the schedule is followed. Establishing that only subsets of the records have sufficient long-term value to be transferred to the archives discourages build up of bulky materials at the administrative unit, and reduces the amount of material the archive receives. Records schedules, if
followed, shield institutions from accusations of spoliation, or destroying records to conceal information.

Websites are both organizational publications, in the way that all materials posted to the Web are publications, and are records, or at least, they may contain records. Rather than developing an entirely new procedure for appraisal and preservation of websites, this paper explores whether or not records management approaches, and especially existing schedules, can be used to appraise websites and their content for harvesting and preservation. There are several questions involved in this exploration. First, how well do the websites correspond to the categories of the records schedules? Research into this relationship will demonstrate whether the same types of content are on the website and in the print records of an office. Second, is applying the record schedule to the website sufficient to identify all the materials with long-term value on the website? While following the record schedule provides a consistent value judgment for those materials listed on it, determining long-term value in a consistent manner for unscheduled materials can be challenging, as there are multiple sources to draw from. The digital curation project at UNC has also identified among digital assets for inclusion in the institutional repository, “research materials, learning materials, conference proceedings, and publications.” In addition, there are value judgments based out of archival theory, such as Schellenberg’s definitions of evidential and informational value. Records schedules are based at least partly on these judgments, and so it is consistent to apply these value judgments to website materials that potentially are not on the records schedule. The final part of the exploration is what can be learned from the characteristics of materials that do not correspond to the records schedule. Knowing what materials would and
would not be identified for collection as a result of the schedule means it is possible to judge the utility of identifying website materials via current records management programs, as well as how to adjust those programs to better fit collecting of institutional websites.
Importance of Study

The purpose of this study is to develop a method for appraising websites for preservation based on the record schedule for the administrative unit to which the website belongs. Websites, especially those of educational institutions or governments, are increasingly important as a public face of the institution, as well as giving access to materials that may or may not also be available in print. These websites, however, are vulnerable to being overwritten, having files go obsolete, or to otherwise being lost. Over time, digital materials are significantly less stable than the same information on paper. This fragility means that archivists need to take action earlier to preserve websites than they would for paper records. There is a need to develop systems not just for managing websites for current use but to support archiving websites for use in the future. One possibility might be an institution-wide content management system. While these systems work well at smaller institutions with significant top-down control, larger and more diverse organizations, such as universities, have difficulty implementing such systems. As a result, there is a need to develop a system for appraising and ingesting websites that can be administered by the archive or records management program of the institution without significant imposition on the creating departments. Such a system could fulfill both the historical needs of the archive and the risk management needs of the records management program.

The literature has revealed no systematic look at the applicability of the record scheduling process to website preservation. My study attempts to fill this lack, and support the need for a systematic, rather than piecemeal, approach to archiving websites.
Literature Review

The literature in digital preservation provides an overview of the field and of current practice. My research aims to build on this base of knowledge by devising a practical and usable preservation model that addresses areas of need and leverages good practices. Therefore, this section attempts not only to review the literature but to highlight areas of need, and concepts worth exploring.

In “National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and of Related International Activity,” Neil Beagrie reported on a survey of national digital preservation programs outside the US for the National Digital Information Infrastructure Preservation Program. He collected his information through site visits and phone interviews in 2002. While now rather out of date, this study does provide a starting point for identifying digital preservation programs. Not surprisingly, Beagrie found significant variation in both extent and goals of programs. He noted that, “in none of the countries surveyed is there a single national initiative for digital preservation. Rather, there are many institutional missions that are being extended into the digital domain, including those of national institutions such as the national archives and national libraries.” As of the time of the Beagrie study, the National Library of Australia had an active records management program and a growing digital archiving program, including a web domain harvesting study in the planning stages but no significant new funding, and no national mandate for the deposit of electronic records. The PANDORA project to archive digital publications, first started in 2001, has progressed to a third version. In addition, the related PADI Safekept program stores long term resources deemed to be valuable. The French
National Library split deposits by material, but deposit laws were being applied only to materials on removable storage physical media, such as CD-ROMs, which would exclude websites in 2002. A new law, passed in 2006, requires deposit of websites, and web harvesting tools and approaches are being developed. The French National Library is also experimenting with various digital preservation models, especially with OAIS compliance, while the French Archives is developing guidelines for electronic archives.\(^\text{13}\)

Beagrie also noted that the Koninklijke Bibliotheek (KB), the national library of the Netherlands, negotiates deposit of publications with publishers. The KB also has an extensive preservation program, and several projects to capture digital materials.\(^\text{14}\) The Dutch government is contributing extensive funding for the projects. The UK, like Australia, has a legal deposit system, one that did not include digital materials until 2003. As of the time of the study, the British Library was planning to start selective web archiving and regular snapshots of government websites. The British Library website now has their selection policy for websites as part of their collection development program.\(^\text{15}\) The National Archive of the UK has a web archive section for government pages. While Beagrie asserts that no comprehensive national programs exist, knowing the status and goals of projects currently underway in various countries provides valuable information for my study. Beagrie’s study shows that, in 2002, there is widespread agreement about the value of preserving websites. In addition, while there is still exploration of options by most of these programs, automatic harvesting, which I am also using, seems to be a common tool.

While the national programs generally focused on publications and on web domain harvesting in general, a more detailed focus is needed on electronic records.
Cloonan and Sanett, part of the InterPARES Preservation Task Force, conducted a study in 2002 of digital preservation programs at educational institutions. The first part was a survey of 13 institutions, collecting information on digital preservation, staffing and experience, technical decisions and policy. The second part is a series of interviews with 18 people at these institutions identified as familiar with the projects, and experienced in the field. The authors noted that, “The first challenge is identifying what is a record, the second is appraising it, and the third is accessioning and preserving it.” These challenges describe the main problems in both the general question of preserving electronic material, and the specific goal of preserving electronic records and websites for not just historical, but evidentiary purposes. While this article does not explicitly discuss preservation of websites, many of the goals, concerns and conclusions are still relevant. One key point was the difficulty of identifying a record among a mass of electronic information. My study can potentially address many of these concerns by creating a method for institutions to identify records and other key information within websites.

Terry Cook, in a commentary originally from the 1990 NAGARA Conference, addresses the question of how electronic records should be treated, and in so doing explains the debate between custodial and non-custodial digital preservation. Through a comparison of two other papers presented at the same conference, Alan Kowlowitz’ “Appraising in a Vacuum” and Michael Miller’s “Past is Prologue,” Cook addresses the debate on custody and policy regarding electronic records. Both articles address David Bearman’s theories regarding custody that electronic records should stay with the creating agency. Kowlowitz supports Bearman’s position that archival programs will need to develop policies for appraising electronic records and including handling
archiving at the creating agency. Cook compares that to Miller, who argues that the solution is on a higher level, making records systems that involve archiving and records management from the design phase on, so that appraisal and description are built into the program.

Both articles, as well as Cook’s response, are old enough that the electronic record systems they are addressing are very different from the websites I am studying. Despite this, however, these concepts have relevance for this study. Whether the electronic records are kept in the custody of the agency that created them, or transferred to an archive, there is a need to determine what should be saved and what can be deleted or ignored. Records schedules can either be built into the system or applied later, and so will be valuable whichever direction record management practice goes. In addition, material on websites may become obsolete quickly or slowly and identifying materials of value while they are still part of an active website will help preserve them in greater numbers down the line.

In Cathy Smith’s 2005 article, “Building an Internet Archive System for the British Broadcasting Corporation,” she explores how the British Broadcasting system has addressed the same challenges faced by other web archiving efforts. Unlike many of the web archiving projects addressed in other articles, the BBC’s main function is not saving their content for historical reasons. However, the BBC is required by law to keep all programming for a specified period of time. As part of this process, BBC staff needs to identify and safeguard personal data from user generated content, such as comments and message boards, and to have a method to deposit the electronic publications with the British Library if needed.
The BBC tasked the Information and Archives Department with “establishing a process for capturing, storing, and preserving the BBC’s online output,” while “considering how to make that collection accessible to the public.” A key decision in the process was whether to rely on the creators to capture the materials or to establish a centralized system. Based on interviews with producers of content, the Archives staff determined that “production departments did not want archiving to encroach upon content creation.” Content creators saw archiving as being I&A’s responsibility on behalf of the BBC and, while they would advise on its development, any archiving system would have to remain centralized and independent. The system they developed relied on automatic capture of the website as it appears to viewers. These decisions can also be applied to other web archiving projects, including that of UNC materials. While post-custodialists, like Bearman, claim that it makes more sense to leave electronic records in the custody of creators and let appraisal and preservation take place there, those duties are not ones that creating agencies take on willingly, at least at UNC-Chapel Hill. An invitation to send web materials to the archive resulted in only one submission over several years. An automatic crawling process, as conducted in this study, promises to yield much better results.

Julien Masanes, in his article “Web archiving methods and approaches: A comparative study,” compares web crawling and manual page collecting as methods for collecting websites for preservation. Websites pose challenges for collecting and preservation because of some characteristics common to web material. Both publishing and editing published materials have become easier and new pages often replace or update older ones. Also, the information, rather than being concentrated in a few pages,
is often dispersed among many with less information per page. Taken together, these characteristics mean that there are more websites where more pages need to be saved, and there is a shorter time to save them. In this article, Masanes compares the results of two web material collection methods, as applied to sites connected to the French elections of 2002. The first is automatic collection or crawling, where content is acquired by following links from a starting set of URLs. The second is a manual collecting of pages one at a time. The study evaluated the results based on how completely a website is archived. This was determined by how completely a website was captured and whether the archived version is navigationally functional. Masanes determined that automatic crawlers are more efficient at collecting the top levels of a large number of sites, or extensive archiving, than intensively archiving numerous levels of a small number of sites. Manual archiving produces higher quality results, but takes much more time and cost. Masanes noted specifically that in domain-based collecting, where sites are collected based on location and owner rather than subject, crawling does a poor job of capturing the most ephemeral pages. Without specific attention paid to them, they fall into the gap between crawls. This has major implications for preserving institutional websites, which are clearly domain collections. This article, along with Masanes’ previous works, highlights the benefits of combining automatic archiving with manual tailoring of crawls and limitations. By knowing generally how a website is arranged, and where there are valuable files, an automatic crawl can generate higher quality results. My research study aims to provide structural and appraisal information about the chosen websites.
In “Archiving and Accessing web pages: The Goddard Library web capture project,” the authors document the process used to capture selected portions of the Goddard Space Flight Center website, a process that may be applicable to other institutional websites. The Goddard Space Flight Center administers a large number of scientific and engineering research projects. In the last few years, the library of the center began exploring ways to preserve the information generated by these projects and stored on their websites. They quickly determined that the approaches used by national archives and libraries would not be suitable because they needed to capture a broader range of types of materials and formats than any of the national library or archive projects happening in 2001 or 2002. They also determined that the ‘crawl it all’ procedure of the Internet Archive was a poor fit because their goal was a selected subset of the website, namely the scientific and technical information within their own domain. The authors noted that, “The GSFC system needed to be more selective; yet that selection could not be based solely on an analysis of the domain components as represented in the URL.”

As a result, the library developed a system that was a hybrid, including both manual selection and automated crawling. The root and top two levels were manually selected, excluding any completely irrelevant sub-directories. The article states that, “This initial analysis provided important information about how to collect the sites. It identified anomalies that required human intervention, outlined requirements for the spidering software and provided statistics for the estimation of storage requirements for the captured sites.” This technique of combining crawling with partial selection is potentially very valuable for the harvesting of institutional websites. Since potentially not all the website has long term value, identifying parts of sites for exclusion or less
frequent crawling will improve the results of harvesting. Also, this technique will likely be necessary on sites that have content that is the intellectual property of individuals within the institution, rather than the institution itself. As will be addressed later, faculty directories on university websites pose special challenges. These functions are somewhat the reverse of how the GSFC plans to use the system, excluding some rather than selecting a little, but are applicable to both concepts.

In “An Arizona Model for Preservation and Access of Documents,” Richard Pearce-Moses and Joanne Kaczmarek lay out a model for handling websites that is different from either of those discussed by Masanes, while having elements of both, and shares characteristics with the Goddard project. Rather than the time- and quality-intensive selection of individual documents, a combination of manual and automatic scanning identifies websites as being collections of documents from the same creator. These sites are broken down into a hierarchy of subdirectories, much like an archival paper collection is broken down into series and sub-series. Pearce-Moses states that “Curating a collection of Web documents using archival principles is relatively straightforward. The archivist approaches the documents on a Web site as an organic whole, then, moving down the hierarchy, looks at each series in the collection as a whole. The archivist stops when further subdivision of the hierarchy is no longer useful.” I followed this approach with the websites I analyzed. This approach also allows for a powerful combination of full text searching and browsing. Lists of filenames can be easily generated by the computer and searched, and the results displayed in groups based on the collection and series they are in. Users can not only see all the results, but can browse them based on the categories, which helps exclude irrelevant results. There are
some challenges with this model, as websites are frequently not neat and tidy, and not all
important sites can be identified automatically. This process could be helped along by
having lists or charts of the governmental structure in which the sites reside. This paper
on the Arizona model, more than any of the others, reflects the direction my research
study will take. I will be treating each sample website as a collection, and then looking at
the series within. By taking the series level approach explored in this article, I hope to be
able to identify series within the websites that correspond to items in the record schedule.

Daniel Gomes, Sergio Freitas and Mario Silva, in “Design and Selection Criteria
for a National Web Archive,” conducted a study for the ECDL 2006 conference
exploring the criteria for harvesting national web sites with respect to Portugal.\textsuperscript{30} They
identified some key questions that need to be asked to determine the boundaries of a web
harvesting program, some of which have relevance for institutional websites as well.
First, what portion of the relevant websites does the harvesting entity have jurisdiction
over? In the case of this study, it is determining whether to include sites originating in or
containing content about Portugal, even if they are not on the country code domain (.pt).
The Art Department at UNC, at "http://www.webslingerz.com/depts/art/," is comparable
in this study, as well as websites on the university servers where the content belongs to
individuals or grant-funded projects. This article argues that web archive must select
types of content it will store. This runs counter to a certain amount of archival practice,
but may better reflect the realities of web preservation. In any case, decisions about
formats to be accepted impact preservation because, “Preservation strategies must be
implemented according to the formats of the documents.”\textsuperscript{31} Determining the boundaries
of a web harvesting project is more complex than simply identifying appropriate URLs.
While in their study of the Portuguese web space the authors of this article were dealing with a somewhat different situation, the main points are still relevant. First, the boundaries of a harvesting project must be identified, and the right to harvest must either be obtained or mandated. Second, the archive needs to identify what is feasible to harvest from a preservation standpoint, and identify preservation needs for the incoming materials.

In, “Accountability and accessibility: ensuring the evidence of e-governance in Australia,” Cunningham and Phillips discuss projects the National Library and National Archives of Australia are undertaking to address capture of government records and publications. Like other countries described by Beagrie, Australia has laws that address print records and publications, and more recently online publications and records, but the two systems remain separate. Publications are the responsibility of the National Library. In 1996, the Library established Preserving and Accessing Networked Documentary Resources of Australia (PANDORA). PANDORA accepts online publications and websites that fall into six categories: “Commonwealth and Australian Capital Territory government publications, publications of tertiary education institutions, conference proceedings, e-journals, titles referred by indexing and abstracting agencies, and topical websites in established subject areas or of current socials or political interest.” Based on these definitions, government or other institutional websites are not collected.

Government records are the domain of the National Archives, whose focus has been “helping agencies to design and implement record keeping systems that ensure the making and keeping of records.” This includes guidelines for describing and archiving web resources, most recently released in 2004. The Agency to Researcher project, which
I will discuss later, is the long-term preservation strategy for the National Archives. UNC-Chapel Hill has a similarly split system, with publications going to the North Carolina Collection and records going to the University Archives. Cunningham and Phillip’s article provides a basis for treating websites as records, or places where records are found, rather than as publications. My study also focuses on websites in relation to records, although UNC record schedules include publications.

Shelby Sanett discusses different models for the costs of digital preservation in her paper, “Towards developing a framework of cost elements for preserving authentic electronic records into perpetuity.” Models based on the life cycle of data include elements that fall broadly into: identification of materials, accession or ingest of materials, processing and description of materials, and preservation of materials. These elements are broken down differently in different models. A model developed by the Arts and Humanities Data Service divides cost elements into three broad stages: data design and creation, data accessioning into collections, and data use and administration. This model does not address, outside of acquisition a need for appraisal or selection. A different model, by Russell and Weinberger, starts their eight-element model with “selecting a particular digital object.” Sanett’s own model is based on a functional approach, which identifies types of costs at the preservation stage and at the use stage. It includes in the preservation stage’s direct operating costs: “identify potential records, evaluate/examine, acquire records, establish inventory record,” among others. These models, though in different ways, establish the identification, acquisition and description of records as vital parts of the digital preservation process. My methodology addresses all of these parts. By breaking down the website according to a records schedule and
identifying information about directories of files, some metadata elements can be filled in automatically. In addition, by knowing what is of value on the website, and excluding materials of no long term value, there will need to be less appraisal after ingest.

Catherine Nicholls looks specifically at high-level policy for web-page creation at the University of Melbourne in “Creating Road Signs and Encouraging Safe Driving on the Information Superhighway: Accountability and Compliance in the Web Archiving Environment”39 As part of the Web Archiving Project, Nichols identified the need to preserve web pages “that contain University records, that contribute to the business continuity of the organization, and that contain information that has significant cultural and historical value and falls under the collection strategy of University Archives and University Library.”40 Approaching it as a risk assessment program rather than as a preservation program, the Web Archiving Project encouraged departments to undertake risk assessments of their web presences. The aspects departments were told to look for: “whether the web pages contain information that is already recorded in another recordkeeping system: whether the information being presented on the web page is for an internal or external audience: whether the web page contains substantial unique information about a University project initiative, project, event or subject area,: whether the target audience of the information on the web page includes students: and the frequency of change and whether there are procedures in place to track changes and updates to the website.”41 The Web Archiving Project has gone on to explore university wide content management systems or other ways to have a top down approach to archiving, based on the categories of high-value materials identified in the risk assessment. This study makes a couple of important points. First, that everything with
long term value may not be identified in the record schedule. Second, appraisal is necessary to preserve websites. My research study includes data collection that addresses these points.

In “An Approach to Preservation of Digital Records,” Helen Heslop, Simon Davis and Andrew Wilson explain the Agency to Researcher project based at the National Archives of Australia as a new model for long term digital preservation. This project, which started in 2000, is based on a performance model where the digital object is made up of several parts. The data has to be mediated by hardware and software to create a performance, which is accessible to a researcher. Paper records, comparatively, do not require all of these steps. The Agency to Researcher project focuses on a goal to, “ensure that digital records of long term value will remain accessible for use over time.”

Rather than trying to emulate computing environments or repeatedly migrating files to newer systems, files will be converted to an open, common format for long term storage. It will be accessible to researchers in that format, or in formats common at the time it is requested. This approach depends on the ability of archivists to develop processes to convert a file from its original format to the format used by the archive without losing the characteristics that make up the ‘essence’ of the record. While the National Archives of Australia begins its set of principles for the project with the statement, “The digital preservation program must be able to preserve any digital record that is brought into National Archives’ custody regardless of the application or system it is from or data format it is stored in,” the authors have only vaguely addressed the challenges implicit in this goal.

This paper addresses the essence part of the process, noting that, “archivists in the National Archives digital preservation program will need to spend time analyzing
genres of records in our custody to identify and document their essence.” While the essence of a record might be the same in multiple formats, such as images stored as GIFs or JPGs, the exact process to normalize, as the paper calls it, the record will need to be tailored to each individual format, if not each version of each format. This will require, at minimum, knowledge of what formats are going to be used by creating agencies. It may also require guidelines or mandates to reduce the number of formats. The first step, however, is to identify what formats are being used by creating agencies. The Agency to Researcher project, while it did not discuss websites, is one where preservation of web materials is possible. Using the model addressed in this paper for web materials, as well as most others will require more data being gathered, either at the time of acquisition, or by an analysis of the records being created. My survey of websites at UNC fulfills the latter option of the requirement.

In 2004, the Metrics and Testbed Working Group of the International Internet Preservation Consortium conducted an analysis of various content types likely to be found on websites to be harvested, and looked at issues that might arise in the acquisition, parsing and presentation of the content, as reported in “Web Harvesting Survey.” They then rated the ability of current web harvesting tools to accomplish the different phases of harvesting on a three point scale: Easy, where current tools can harvest content now; Difficult-Tools may or may not be able to harvest content now; Future – Current tools cannot harvest content now. While it is not surprising that different file formats rated differently, the results identify major challenges in preserving many common file formats. Open, common file formats such as html, gif, and jpg were all rated as easy. While acquisition of proprietary formats such as Microsoft office documents is fairly
straightforward, parsing is difficult and presentation not guaranteed into the future, and may require plug-ins. Non-streaming media is similar, while streaming media was rated as ‘future’ for all three phases. Interestingly, server-side scripts were rated as ‘easy’ for each of the phases. While this article is slightly out of date, this sort of analysis is exceedingly valuable for web archiving projects to determine what formats can be accepted, and to set standards for creation of web materials.

The articles discussed above, taken together, provided not just an overview of the field, but guided my methodological decisions in this study. Beagrie established that websites should be saved, while Cloonan and Sanett identify a need for a way to save websites at the institutional level. Smith, looking at the BBC, identified problems that would make automated crawling more likely to succeed than relying on submitted materials. The study by Masanes not only supports Smith’s assertion but refines it with the conclusion that tailoring the crawling improves the results. The Goddard Space Flight Center library project introduces the idea of a partially manual selection process. Gomes’ study of Portugal’s web presence identifies some of the issues in defining the boundaries of a web harvesting project. My decision to look at websites as records rather than publications reflects the actions in Australia discussed by Cunningham and Phillips. While slightly less relevant, Sanett’s article on cost models for digital preservation highlights the cost of setting up new systems, whereas using the records management patterns would require less new work. Nicholls article on the Australian Web Archiving Project, while focused on content management systems, identified two key things to consider: what metadata can be generated automatically during the appraisal process and the need to look for important information that is not identified by the record schedule. Most relevant of all,
Richard Pearce-Moses’s article on the Arizona model not only puts web archiving solidly within the sphere of records rather than in publications, it supported the idea of looking not at the item level, but at the series level. Finally, while the focus of this paper is on appraisal for selection, appraisal for preservation is a parallel function. Helen Heslop’s paper on the Agency to Researcher system highlights some of the appraisal that will need to be done in order to preserve digital materials, while the IIRC study highlights the difficulty in preserving some types of digital materials. These articles, along with elements from others, influenced my decision to include documentation of formats in my analysis. While a certain portion of my methodological decisions depended on availability and feasibility, behind those was a solid grounding in the literature reflecting other projects being carried out.
Methodology

Based on the literature, and available resources, I developed a procedure for selecting and analyzing departmental and office websites at University of North Carolina at Chapel Hill.

1. Selection of websites
2. Descriptions of selected websites
3. Goals and harvesting of website content
4. Website analysis
5. Comparison to Record Schedules

**Website Selection.** In choosing websites to explore for this study, I had to balance three variables. First, I was looking for websites that corresponded directly to schools, departments and other administrative units within the University of North Carolina at Chapel Hill, as opposed to websites such as Carolina Community Resources, at "http://www.unc.edu/community/." Second, I wanted websites that were not subsets of larger sites. As a result, I excluded the Nutrition program because their site, "http://www.sph.unc.edu/nutr," is an offshoot of the website of the School of Public Health. Finally, I needed websites for which the corresponding department had reasonably current and comprehensive records schedules. The greatest limiting factor, as it turned out, was the relative scarcity of recent records schedules. In theory, every department, school and other unit has an up-to-date records schedule that reflects the functions and requirements around their work. Currently, however, UNC has an extremely small Records Management program, thus many departments have schedules that are significantly out of date. Only 15-20% of the schedules are from 2000 and later and not all offices are scheduled. Updating of schedules is dependent on interest from the
department in question, and does not reflect any systematic pattern on the part of the Records Manager, though there is an attempt to keep schedules for very important offices more up to date. These requirements limited my options and made it impossible to do any sort of probabilistic sampling among all the websites on campus.

Instead, I drew websites from broad categories of university functions. I wanted a graduate and professional school within UNC-Chapel Hill, an undergraduate academic department, an administrative services program such as Human Resources or Faculty Governance, and a central administration website containing materials of high value to the University. While more subjective than the others, my final preference was for websites that had features that would prove “interesting,” but were manageable in both size and complexity. For this reason, I avoided the Medical School, at "http://www.med.unc.edu," as too large, and the PID Office, at "http://www.pid.unc.edu/default.htm," which assigns student ID numbers, as too simple.

I selected four websites: School of Information and Library Science (http://sils.unc.edu), Department of Mathematics (http://www.math.unc.edu), Faculty Governance (http://www.unc.edu/faculty/faccoun/) and Office of the Chancellor (http://www.unc.edu/chan).

**Website descriptions.** The School of Library and Information Science is a graduate school at UNC-Chapel Hill, offering Doctoral and Master’s degrees, as well as an undergraduate level major and minor. The school has its own web server, which has multiple virtual hosts. While “www.ils.unc.edu” was the initial website, “sils.unc.edu” is now the name of the main site. There are parts of the website under both names. Generally, the main pages of the site are on sils.unc.edu, while faculty student, project
and organizational web space is on ils.unc.edu. Between the two sections, the site is very large, containing more than 17,000 files. This website is fairly hierarchical, with a number of distinct sections. It is fairly well organized, though shows signs of having grown rather than being designed as this size.

The Department of Mathematics is part of the College of Arts and Sciences. Degrees are given by the College, not the department. While the department has graduate students, most of the focus is on undergraduates. Not surprisingly, this website is smaller than the SILS site, only about 3,200 files. It is well organized, with predictably named folders, although there are a lot of files in the top directory.

The third website is the Faculty Governance site. This site serves as the point of public access for materials connected to the faculty council and associated committees. This site is smaller still, at about 2000 files, and has a narrower set of functions than either the Math Department or SILS sites. It is somewhat organized, but in a way that reflects growth rather than design, much like the SILS site.

The final site of the study is the Chancellor’s website. This is a much smaller site, only around 250 files. The fact that these are the Chancellor’s materials means that it is important to capture them. Most notably, there are speeches given by the Chancellor, stored in a combination of text and video. This site is the least organized of the four sample websites.

**Goals and Website Harvesting.** The goal of this study is to use these websites to answer the questions addressed in the paper. First, how well does the content and structure of the websites correspond to the categories of materials on the records schedules? Second, does the set of materials generated by applying the record schedule
to the website contain the materials with long term value? Third, where it doesn’t, what is
significant about the materials not corresponding to the record schedule? There are,
however, some limitations on how this can be done. The size of even the smallest
website makes analysis at the item level impossible. The approach of analysis using
upper level directories as the unit of analysis, much as in Richard Pearce-Moses’s article
is not just more feasible for this study, it also provides the potential that university
archivists will use this methodology outside of this study as a method for appraisal of
websites prior to ingest.

In order to get any sort of organized look at the websites, it made sense to use a web
crawling program. The arrangement of the files on a website is not readily apparent just
from looking at the website, as only rarely will you see lists of what is in a directory. For
this purpose, I chose HTTrack. This program is a free open source website copying
software created by Xavier Roche. It accepts one or more URLs as input, as well as
parameters about file types, links and crawling depth. It copies to a designated hard drive
location a mirror image of the website, with directories and subdirectories set up as
folders and subfolders. This allowed for mass analysis of the sites using Windows tools
such as file search and properties. For the mirroring of the four test websites, HTTrack
was programmed to collect all textual formats, image formats, video formats and
compression formats, but excluding advertising pop-ups. Each site was first captured
limiting crawling to the domain initially specified. The crawls for these sites were fairly
straightforward but with a few complications. First, the sites that are quite large took an
expectedly long time to crawl, over 8 hours. Second, the crawler ran into significant
difficulty with capturing wikis, of which there are at least two on the SILS site. When
the process finally ended, it was unclear if the wikis were captured in their entirety. Because of documenting at the directory level rather than the item level this did not pose significant problems, as a decision on whether to preserve the wiki would be made based on the entire wiki, and not each page.

Two of the four chosen departments, however, had their website under more than one domain name. The SILS site is split between "sils.unc.edu" and "ils.unc.edu." The Chancellor’s website is mostly in "www.unc.edu/chan/," but also in "www.unc.edu/realchan," "stateofuniversity.unc.edu" and "streams.sph.unc.edu." After discovering this, these two sites were re-crawled to capture these additional materials. In order to study the web presence of each department as an intellectual whole, it was necessary to look at all of the different parts of the website.

**Website Analysis.** The first data I collected was about the organization of the websites. I started by identifying the top level directories. The Chancellor’s site, for instance, has only five first level directories: “China”, “fyi”, “new_images”, “Singapore_visit”, and “speech_archive”. In the SILS site, on the other hand, because of its size and complexity, I ended up identifying second and third level directories in order to document the information variety. The first level directory “news” contains the calendar, information from a symposium, fliers about the degrees, and the newsletter. The other first level directories I documented in more detail were “events,” “itrc,” “people,” and “research,” each also because of the variety of content within. The high-level directories resulting from this first step served as my units of analysis for the rest of the data collection. This follows the method discussed by Richard Pearce-Moses, where he states that, “The archivist stops [moving down the hierarchy] when further subdivision
of the hierarchy is no longer useful.” I stopped at a level where I judged the contents of each sub-series were sufficiently similar to be dealt with as a unit.

Once I established my units of analysis, I collected some technical and content information about each one. I documented how many files in how many folders were in each unit. Next, I identified what file formats were present in each unit. These pieces of data were collected by applying the properties and search functions of the Windows operating system to the material collected by HTTrack. I then briefly described the contents of each unit for reference purposes. I also identified units of analysis mainly containing student-created materials and those directories that are university employees’ personal directories. These materials are significant because the copyright for them may rest with the creator and not with the University. The UNC-Chapel Hill copyright policy states that, “While, as a general rule, all rights to copyrightable material are the property of the creator, and the distribution of royalties, if any, is a matter of arrangement between the creator and his or her publishers or licensees, different treatment may be accorded by the University in case of specific contracts providing for an exception, in cases where the University is a joint author with the creator.” Based on this policy, a university-wide website archiving project will need to, at minimum, obtain clear copyright permissions before archiving faculty or student directories. Depending on law and policy changes, they may need to exclude those directories from the crawl, making documenting their location valuable. This descriptive information served to remind me of the contents of the units of analysis during the comparison with the record schedules.

**Record Schedule Comparison.** I next compared the data collected on the websites to the relevant record schedules. The records schedules produced in the last few years by
the University Archives and Records Service follow a consistent format. Each category of material, labeled as an ‘item,’ has a description of materials, followed by a disposition instruction (see Appendix A). While in rare cases an “item” on the schedules could be an individual document, most often the term is used to indicate a type of material such as a correspondence file. The disposition instructions, while varying slightly, fall into two broad categories. First, items without long term value have disposition instructions such as, “Erase/destroy in office records in paper and electronic formats when administrative value ends.” Some have a length of time as part of the instruction, such as, “Erase/destroy in office records in paper and electronic formats 1 year after graduation or date of last attendance.” I decided to consolidate all of the variations of this into a single ‘discard’ determination. The second group of instructions, which I am calling ‘keep’, involves transfer to the University Archive after some period of time. I have also included publications in this group, which call for transfer to the State Library as well as the University Archives.

The Chancellor’s Office and the Department of Mathematics had a single schedule each. SILS has five schedules, directed at different functions of the school, including the Dean’s Office and Director of Communications. Faculty Governance has one schedule for each committee, plus several more for general functions. In order to do the same sort of analysis with each of my samples, I created a list of items that included those from each of the schedules from SILS. If there was overlap with different disposition instructions, I chose ‘keep’ rather than ‘discard.’ There is a risk that this could skew my data, but since the question of ‘scheduled versus not scheduled’ is more important than ‘scheduled to keep versus scheduled to discard’, the effect should be minimal. I
combined the Faculty Governance schedules the same way. The different committees connected to the Faculty Council had very similar schedules.

I added the titles of the items, or categories, of the schedule to my data collection sheet as column labels. I then identified web directories that contained content described by the record schedule items, and in the column for each particular schedule item, stated whether the directory contained materials from that schedule item. Both browsing of the websites and text searches through the harvested material were necessary to find relevant material. Because, as noted above, I treated each unit of analysis as a unit, I categorized them as a unit as well. If a directory had material that corresponded to an item in the record schedule, then I labeled the directory based on the disposition of that scheduled item. Each unit of analysis fell into one of three categories: Scheduled to be kept, scheduled to be discarded after a period of time, or not scheduled. These were labeled “keep”, “discard”, and “unscheduled,” respectively. While with paper records, provision needs to be made for records that remain necessary to the running of the department, website crawling duplicates the material, making observing a period of waiting unnecessary. On the SILS site, for instance, the "news/calendar” directory contains materials that correspond to the item on the Student Services record schedule. The schedule item contains the disposition instruction: “Erase/destroy in office records in paper and electronic formats when superseded or obsolete.” In the column for this schedule item, I indicated in the news/calendar row the word “discard.” Once I finished comparing every schedule item, as defined above, against the list of directories, I labeled each directory that did not fall under any scheduled item as ‘unscheduled.’ I then consolidated the results into a single column that labeled each directory ‘kept’, ‘discard’
or ‘unscheduled’. As with the schedule consolidation, I resolved conflicts toward the ‘keep’ side. For each website, I went through the methodology:

1. merge schedule headings if needed
2. create a row of column headings from the schedule
3. compare each column to the units of analysis through folder names and Windows searching
4. Label units of analysis that contain material matching schedule items with the appropriate headings.

The different pieces of information gathered through this methodology provide the ability to answer the questions addressed by this study. Establishing information about the structure and content of the websites makes it possible to do comparisons with the record schedules. It also provides the technical information necessary for a more tailored crawl for archiving purposes. Since I know how many files are in each directory, and how many directories fall into each of the three categories, it is possible to answer the first of three study questions with straight numbers. For each unit of analysis of each website, I have documented the number and format of files, the number of folders, basically what the contents are, whether the contents are scheduled and the disposition.
Results

The websites for the Chancellor’s Office, Faculty Governance, SILS and the Math Department each have significantly different characteristics, and vary in size, organization, and how easily they can be categorized with records scheduling. Following my methodology, I analyzed each website in turn, and documented both information about the structure of the website, and how well the website corresponded to the record schedule for the department.

Office of the Chancellor. The website of the Office of the Chancellor has 246 files arranged in 5 first level folders with 17 files loose in the main folder. While the directories are fairly self-explanatory and organized, the speeches section of the site is rather disorganized in design, with different page designs, and the speeches stored in other web spaces. A number of speeches that the Chancellor has made are available through the website as streaming audio. These files are made available though directories on "streams.sph.unc.edu_8080" and "mediaserv.unc.edu_7070." Several other real media files are in "www.unc.edu/realchan." It is unclear whether these separate directories reflect the need to make the files stream-able as well. One final connected site is an access point for the State of the University Speech of 2002. Located at "stateofuniversity.unc.edu," it links to the State of the University speech for that year, which is located in "www.unc.edu/chan/speech_archive." This would simply be an external site, except from "http://www.unc.edu/chan/speech_archive/" it is necessary to access it to reach the speech. (See Fig. 1) While the State of the University site may have been designed to provide easy access to the speech for that year, now it causes problems with accessing the materials.
As is also apparent from Fig 1, there is no consistency in naming files in "speech_archive", as four files describing the same speech in different years are named differently. This may reflect either different page creators over time, or multiple people working on the website. For example, there are accounts of trips to China and to Singapore, general information about the office, and a collection of recordings and transcripts from speeches given by the current Chancellor. These speeches are very valuable for historical purposes, and materials of the Chancellor are generally kept in larger numbers than materials from lower level administrators. The record schedule, however, reflects more closely the duties of the Chancellor than the content of the website. As a result, the only schedule item that appears on the website is speeches by the Chancellor. These constitute only twenty-six percent of the website based on numbers of files. The bulk of the pages on the website are in the "China" and "Singapore_visit" folders. These folders have significant numbers of images, which are historically valuable, and the documentation of the trips is also important. Based on what has been saved from past chancellors, I would judge that these folders, along with several pages in the main folder about the history of the office, have significant long term value. This site does not contain any material with apparent privacy or intellectual property concerns. Because of the poor correspondence with the record schedule, this methodology is not suitable for this website.

**Faculty Governance.** The Faculty Governance site contains 1,960 files, arranged in 23 first level subfolders, though 22 files, many of them main site pages, are in the top level folder. There are agendas, minutes and/or transcripts, resolutions, codes and other output from the Faculty Council and its subcommittees. There are also a small number of
reference materials originating from other authorities, such as the Board of Trustees
determination on tenure rules. While the site is broken down by committee, with the
links to the relevant documents, the actual files are arranged by document type. Faculty
Council materials are mixed in with committee materials in folders with labels such as
“reports.”

This website contains a significant portion of materials that are electronic versions of
types of records that occur offline. Some older Faculty Council materials are in a folder
called “archives,” but other materials of that age remain in the main set of folders.
Eighty-one percent of the files on the website are in directories with files that appear on
the record schedule. The only directory with scheduled material to be discarded is
“elections,” which contains results from elections in 2004. The unscheduled materials
consist mainly of information and history about the Faculty Council, news and resources
for faculty, and the directories of the committees. None of these are materials that cannot
be kept for privacy or copyright reasons. Unlike the website of the office of the
chancellor, a records-management harvesting program could be much more effective for
this due to its strong correlation with the record schedule.

School of Information and Library Science. Both of the previous sites contain
materials that are very administrative in nature. There is either a single person or a small
number of people who contribute to the website and all the material on the websites could
be archived without any privacy or intellectual property concerns for content. The
Department of Mathematics and SILS websites are more complex. The SILS site has
17,146 files divided between the two parts of the website, and 111 first level folders.
There are a variety of materials, ranging from high to low value. Not only is there degree
and program information, and history and descriptive information about the department, there is also significant numbers of pages connected to research projects, plus the personal directories of faculty and doctoral students. The website materials are split between two virtual servers. “Sils.unc.edu” has the vast majority of the main portions of the site; “ils.unc.edu” has mainly single purpose directories, including the majority of the student and faculty directories. As a result, ils.unc.edu is more consistently organized, while sils.unc.edu has broader first level folders encompassing more areas.

Both halves are generally well organized, but there is no consistent identification of personal directories. Most of them have “_” before some contraction of the name, but there are two student organizations also labeled such, and there are personal directories that do not have such an identifier ("bmh," "Griffiths," "daniels"). There are also student materials whose copyright issues have been cleared, such as master’s papers, and some Digital Project Repository materials. Each organizational directory, research project, and personal directory, is arranged and administered by someone different, creating a much more distributed website. By numbers of files, fifty five percent of the SILS website is in personal directories of faculty or doctoral students. Unlike the previously discussed websites, a significant portion of the SILS website (approximately twenty two percent) is devoted to research projects. Neither personal nor research materials are included in the SILS record schedules. In addition, personal and research materials potentially pose copyright or other intellectual property challenges. These materials would need to be assessed prior to ingest into an archive. Due at least partly to personal and research materials, and to a fair number of pages describing and promoting the library school, only about 13 percent of the SILS website could be identified in the record schedule. Among
the unscheduled materials, there were a number with potential historical value, such as the websites of the student organizations, general information about research at the school, and syllabi, lectures, research data and other output from the aforementioned personal and research directories.

Most of the categories in the record schedule were completely inapplicable to the website, such as Accounting and Finance, Purchase Orders and Receipts. In addition, a number of schedule items were for materials that are confidential in nature, and as a result wouldn’t be on the website. These include Search Committees, Tenure considerations, and Student Test Scores. There were a few general documents in these areas, but for the most part these categories are not reflected on the SILS website. While the record schedule did correspond to some directories, especially ones such as publications and course and degree descriptions, the amount of material unscheduled means that the method studied here does not apply well to this website.

**Department of Mathematics.** The Department of Mathematics website has some of the same characteristics of the SILS site, though it is far simpler. It has 3,183 files, in 13 first level folders. There is course and faculty information; and a lot of exam, study, and math assistance for lower level math students; as well as general descriptive and historical information about the department. As with the SILS site, there are also personal faculty directories. This site is less complex than the other three sites, due to being within a single domain. There is some information of value, but also a lot of current resources without long term value, such as practice tests and other resources for students. The most complex part of the site seems to be the Help Center, which requires a login by the student. All of the faculty directories fall under the single first level
directory “faculty,” though there is a first level directory for a post-doctoral research fellow. These two directories account for 81% of the website. Due to this, and to the Help Center and other student assistance pages, only ten percent of the website is identifiable in the record schedule. As occurred with the SILS schedules, most of the record schedule is not at all relevant to the website. This is somewhat mitigated by the fact that if the personal directories were excluded from the crawl, the percentage of the site scheduled rises to almost sixty percent. Even when the comparison is altered, however, the correspondence between the website and the record schedule is not strong enough to make identifying website content for harvest based on the record schedule a valid methodology.

**File Formats.** Analysis of the four websites provides information about the file formats used by the web site creators that is essential to any long-term preservation strategy. Due to the need for ongoing migration, much of the challenge of preservation is dependent on the complexity and the available information about a format. Not surprisingly, the most common file formats found on these four websites are html, image types GIF and JPG, and text formats PDF and XML (Fig 6). The SILS and the Dept of Mathematics sites have more file formats, 39 and 21, respectively, than the Chancellor’s and Faculty Governance sites (8 and 11 formats). This likely reflects both the larger size of these sites and the greater use of different software programs for research or teaching. Of the 46 total formats (49 if the different html types, addressed below, are counted separately), 25 of the formats are found only within faculty and student directories on the Math Department and SILS websites. Twelve of these formats, in fact, are found only in the directory of one SILS professor.
The use of a relatively low number of file formats for the main pages of the websites will make preservation easier, as each file format may require development of preservation tools or methods. One point of concern is that each of the sites contains html files identified by Windows as being related or created with Microsoft Word, PowerPoint and Excel. These files could potentially pose more preservation challenges than files originally created for web design, because they call multiple files to work properly. Figure 4 shows the breakdown of types of html pages on each of the websites.
Figure 1: How the Chancellor's website is located on multiple domains

http://www.unc.edu/chan/speech_archive/

State of the University Speech, 2001

State of the University Speech, 2003

State of the University Speech, 2004
http://www.unc.edu/chan/speech_archive/04stateofuniv.html

stateofuniversity.unc.edu/

State of the University Speech, 2002

Fig 2: Percentages of files on websites.

<table>
<thead>
<tr>
<th></th>
<th>Scheduled to be kept</th>
<th>Scheduled to be discarded</th>
<th>Scheduled total</th>
<th>Unscheduled total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chancellor’s Site</td>
<td>28%</td>
<td>0%</td>
<td>22%</td>
<td>72%</td>
</tr>
<tr>
<td>Faculty Governance</td>
<td>80.8%</td>
<td>.3%</td>
<td>81%</td>
<td>18.9%</td>
</tr>
<tr>
<td>Dept of Math</td>
<td>9.7%</td>
<td>1.2%</td>
<td>10%</td>
<td>89.0%</td>
</tr>
<tr>
<td>SILS</td>
<td>12.7%</td>
<td>.6%</td>
<td>13%</td>
<td>86.7%</td>
</tr>
</tbody>
</table>
Fig 3: Distribution of file formats within SILS website

*Other includes one each of .bin, .ex_, .gz, .ico, .inf, .ini, .ins, .lid, .php, .ps, .qt, .spec, .tag
Fig 4: Distribution of file formats within Mathematics Department website
Fig 5: Distribution of Formats in Chancellor's Office and Faculty Governance websites

**File formats for Chancellor's Office**

- .jpg: 145
- .html: 63
- folders: 25
- .gif: 15
- .rm: 9
- .pdf: 5
- .css: 2
- .txt: 2

**File formats on Faculty Governance**

- .html: 1732
- .gif: 119
- .pdf: 67
- .pot: 12
- .xml: 9
- .doc: 8
- .jpg: 6
- .xls: 5
- .css: 1
- .png: 1
Figure 6: Comparison of the frequency of the different types of HTML files
Conclusion

In this study, I have addressed three questions:

- how well does the content and structure of the websites correspond to the categories of materials on the records schedules?
- does the set of materials generated by applying the record schedule to the website contain all the materials with long term value? and
- is there anything significant about the materials that do not correspond to the record schedule?

The answer to the first question was that it varied significantly between the different sample sites. The Faculty Governance website had significantly better correspondence to their record schedules than either of the academic websites or the Chancellor’s website. This was partly due to the presence of faculty and student materials on the academic websites, which are not included in record schedules. On the SILS site, it also reflected a larger number of pages devoted to the online presence of the school, while the Department of Mathematics devoted a larger portion of their website to academic assistance for math students, which is also not covered by the record schedule. These differences are also reflected in the answer to the second question. Because the goal of record scheduling is to identify both materials of long term value, and materials that can be discarded, simply being scheduled is not a sufficient testament to value. It does mean that these materials have been appraised for their long term value, and that materials with disposition instructions of ‘keep’ do have long term value as determined by the records manager. However, in the websites studied, only a fraction of the material that is scheduled is destined for discard. The answer to the second question, therefore, is that
while the materials identified by applying the record schedule to the website have long
term value, it is not even a majority of material with potential long term value on the
websites. The first thing that is clear given these results is that the current record
schedules do not include the range of content that is on the websites. Three of the
websites had only a small portion of high level directories that corresponded to the record
schedule. In addition, the lack of updated record schedules at UNC-Chapel Hill means
that significant work would need to be done just to be able to use this comparison method
on a larger scale, regardless of the results.

While not being able to prove or disprove the usability of a records-management-
based web harvest project, this study has identified characteristics of institutional
websites that would need to be taken into account either in updating record schedules to
include web content or in developing a separate web harvesting program. Besides the
aforementioned personal and research materials, each of the websites devoted a
percentage of the pages to creating an online presence of the office or department. This
presence is a combination of information about the department and promotion of the
department, answering the question “What does this department do?” The Faculty
Governance site had the least online presence, which may have contributed to the high
correlation with the record schedule. Finally, the effort needed to put information on
websites is much smaller than is needed to create print publications. Information such as
procedures or requirements for a summer program
(http://sils.unc.edu/programs/international/prague.html) may not have justified a print
pamphlet or flyer, but do rate a web page or three. The increase in use of websites to
transmit information means that more information is available to be captured than was taken into account by the record schedule.

The structures of the websites also proved enlightening. With only a few exceptions including the top folders of a site, directories within a site contained content that was thematically linked and further subdirectories were narrower subjects within those topics. I was able to analyze these websites without making separate decisions about each individual page. This corresponds well to the treatment of a website as an archival collection as discussed by Richard Pearce-Moses in the literature review. While the arrangement of the files on the Faculty Governance website was slightly less logical, the folders were accurate for the contents within.

Another important discovery is the diversity of file formats. While each of the sites chosen were based around interlinked html pages with images attached, rather than being database driven, there was use of a significant number of file formats, which pose different preservation challenges. Despite many of the less common formats existing only in faculty directories, there are a sufficient number of uses of proprietary formats to create preservation concerns with important content. Most notable in this is the speeches of the Chancellor stored in Real Media format. As the IIPC web survey demonstrated, the tools to preserve a number of common web technologies do not exist yet. As a result, there may need to be appraisal decisions that take into account the complexity of the preservation challenge along with the value of the content. While the records schedules are not sufficiently applicable, it may be possible to develop a schedule that does not indicate preservation for technically challenging, low value content. A rough example of this is the calendar and room reservation pages on the SILS site, both run by JavaScript.
Among the few scheduled items whose disposition is to be discarded, they present an intriguing example of the possibilities of taking preservation into account in the appraisal decisions.

Finally, I was setting out in this study to develop a methodology that could potentially be used on a wider scale at UNC-Chapel Hill or elsewhere. While comparing the websites to the record schedules, in their current condition, did not prove to be useful, the steps taken to analyze the websites prior to that comparison can still provide a guide to appraising a website for content and for preservation challenges. Despite doing much of this analysis manually for this paper, there is potential for this methodology to be automated, at least to the extent of determining the structure of the website and the file formats it contains.

**Future Study**

Because the record schedules used in this study varied so much from the content of the website, the first question for future would be: How would the content of a website be documented on a record schedule, and would that provide a good automatic appraisal tool for web harvesting? How would it work on more complex, database driven websites? Very recent record scheduling includes websites as a schedule item, with the instruction to save it all periodically and send to the archive. Given the size and complexity of the websites, this may not be a feasible strategy for long term preservation and access. In the past records managers have literally gone to an office’s file cabinets and built records schedules, at least in part, from the categories of materials found in those drawers. If websites become the file cabinets of the future, housing both records of long-term value
and more ephemeral materials, what methodology will archivists use to appraise the content of the websites and not the websites in totality? Finally, how will institutions balance the need to preserve research with the intellectual property and privacy issues of personal files? These questions may have answers that require legal, technological and procedural changes.

3 Cloonan and Sanett, 219.
10 Beagrie, 12.
14 Beagrie, 25.
17 Cloonan and Sanett, 213.


Smith, 20.

Smith, 22.


Alessandro, section 2.

Alessandro, section 3.1


Pearce-Moses, 6


Gomes, section 2.4.


Cunningham and Phillips, 309.

Cunningham and Phillips, 311.


Sanett, 391.


Sanett, 396


Nicholls, 9-10.

Nicholls, 10


Heslop, 6

Heslop, 14


Pearce-Moses, 6


See Appendix A, SILS student services schedule

same

same
Works Cited


<http://www.research.unc.edu/otd/policies/uncpat.html>


<http://www.archivists.org/glossary/term_details.asp?DefinitionKey=52>

<http://www.archivists.org/glossary/term_details.asp?DefinitionKey=54>

<http://www.archivists.org/glossary/term_details.asp?DefinitionKey=200>

<http://www.archivists.org/glossary/term_details.asp?DefinitionKey=125>


APPENDIX A

RECORDS RETENTION AND DISPOSITION SCHEDULE

The University of North Carolina at Chapel Hill
School of Information and Library Science

STUDENT SERVICES
(Electronic file maintenance is provided by and security copies of electronic data are backed up by the School of Information and Library Science, Information Technology.)

Item 1. Alumni Lists File. AutoText 5 (7/15/01)
Record copies of lists created and/or maintained in paper and electronic formats concerning university alumni from department and/or school. List includes names, addresses, phone numbers, and other related information.

Disposition Instructions: Erase/destroy in office records in paper and electronic formats when superseded or obsolete.

Item 2. Annual Reports File. AutoText 6 (7/19/01)
Record copies of annual reports created and/or maintained in paper and electronic formats concerning departmental activities.

Disposition Instructions: Transfer paper records to the custody of the University Archives after 5 years for appraisal and final disposition. Erase in office electronic records when administrative value ends and after records have been printed and filed into the office’s filing system for eventual transfer to the University Archives.

Item 3. Applications for Graduation File. AutoText 10 (7/15/01)
Applications created and/or maintained in paper and electronic formats for graduation.

Disposition Instructions: Erase/destroy in office records in paper and electronic formats 1 year after graduation or date of last attendance.

Item 4. Calendar of Events File. AutoText 11 (7/15/01)
Calendars created and/or maintained in paper and electronic formats of university events.

Disposition Instructions: Erase/destroy in office records in paper and electronic formats when superseded or obsolete.

Item 5. Career Planning/Placement Records File. (Restricted Access) AutoText 12 (7/15/01)
Record copies of records created and/or maintained in paper and electronic formats
concerning materials used to assist students in locating employment after graduation. File includes job interview forms, resumes, student profile data sheets, and other related records. (Comply with 20 U.S.C. 1232g (Family Educational Rights and Privacy Act of 1974) regarding confidentiality of student records.)

Disposition Instructions: Erase/destroy in office records in paper and electronic formats when administrative value ends.

Item 6. Class Schedules File. AutoText 211 (7/15/01)
Reference copies of school/department class schedules created and/or maintained in paper and electronic formats. (Record copies are maintained by Office of University Registrar except for the School of Medicine MD program.)

Disposition Instructions: Erase/destroy in office records in paper and electronic formats after 5 years.

Item 7. Comprehensive Examinations and Results File. (RESTRICTED ACCESS) AutoText 15 (7/15/01)
Records created and/or maintained in paper and electronic formats concerning completed student comprehensive examinations for degrees and results. (Comply with 20 U.S.C. 1232g (Family Educational Rights and Privacy Act of 1974) regarding confidentiality of student records.)

Disposition Instructions: Transfer to Student (Active-Graduate/Professional) File 1 year after completion.

Item 8. Correspondence File. AutoText 19 (7/15/01)
Record copies of records created and/or maintained in paper and electronic formats concerning the administration of the office. File includes directives, memorandums, official office correspondence, reports, and other related records.

Disposition Instructions: Transfer paper records to the custody of the University Archives after 5 years for appraisal and final disposition. Erase in office electronic records when administrative value ends and after records have been printed and filed into the office’s filing system for eventual transfer to the University Archives.

Item 9. Course Listings File. AutoText 217 (7/15/01)
Reference copies of course listings created and/or maintained in paper and electronic formats concerning courses currently offered by school/department. Listings include course schedules, enrollment numbers, and other related records. (Office of University Registrar, except for the School of Medicine MD Program maintains record copies.)

Disposition Instructions Erase/destroy in office records in paper and electronic formats when administrative value ends.

Item 10. Drop/Add Forms File. (RESTRICTED ACCESS) AutoText 219 (7/15/01)
Reference copies of completed forms used by students to drop/add courses. (Comply
with 20 U.S.C. 1232g (Family Educational Rights and Privacy Act of 1974) regarding confidentiality of student records.)

Disposition Instructions: Destroy 1 year after date submitted.

Item 11. Enrollment File. (RESTRICTED ACCESS) AutoText 220 (7/15/01)
Reference copies of enrollment reports created and/or maintained in paper and electronic formats concerning students enrolled for each course in school/department. (Record copies maintained by the Office of University Registrar.) (Comply with 20 U.S.C. 1232g (Family Educational Rights and Privacy Act of 1974) regarding confidentiality of student records.)

Disposition Instructions: Erase/destroy in office records in paper and electronic formats after 5 years.

Item 12. Examinations, Tests, Term Papers, and Homework Records File. (RESTRICTED ACCESS) AutoText 27 (7/15/01)
Records created and/or maintained in paper and electronic formats concerning a student’s academic performance. File includes final examination booklets, examination tests and scores, graded term papers, and other related records. (Comply with 20 U.S.C. 1232g (Family Educational Rights and Privacy Act of 1974) regarding confidentiality of student records.)

Disposition Instructions: Erase/destroy in office records in paper and electronic formats 1 term after completion, if grade results are unchallenged. If challenged, destroy in office after resolution of challenge.

Item 13. Grades File. (Restricted Access) AutoText 225 (7/15/01)
Reference copies of records created and/or maintained in paper and electronic formats concerning grades earned by students. File includes Distribution of Grades by Instructors Reports and final grade rolls, except for MD students. (Comply with 20 U.S.C. 1232g (Family Educational Rights and Privacy Act of 1974) regarding confidentiality of student records.) (Record copies are maintained by the Office of University Registrar for eventual transfer to University Archives.)

Disposition Instructions: Erase/destroy in office records in paper and electronic formats after 5 years.

Item 14. Internships/Fellowships File. (RESTRICTED ACCESS) AutoText 37 (7/15/01)
Record copies of records created and/or maintained in paper and electronic formats concerning scholarships, internships, and fellowships within the school/department. File includes accounting statements, applications, correspondence, recommendations, reference copies of award notifications, descriptions of awards, eligibility criteria, guidelines, procedures, regulations, and other related records. (Names of recipients and financial aid award letters are maintained in the Office of Scholarships and Student Aid, if need-based funding is awarded to the student.) (Comply with applicable provisions of 20 U.S.C. 1232g (Family Educational Rights and Privacy Act of 1974) regarding
confidentiality of records.)

Disposition Instructions: Erase/destroy in office accounting statements and applications in paper and electronic formats 3 years from award year and after released from all audits. Erase/destroy in office remaining records when superseded or obsolete.

Item 15. Permanent Record of Courses File. AutoText 44 (7/15/01)
Record copies of lists created and/or maintained in paper and electronic formats providing a permanent record of courses offered by the school/department.

Disposition Instructions: Transfer paper records to the custody of the University Archives when administrative value ends for appraisal and final disposition. Erase in office electronic records when administrative value ends and after records have been printed and filed into the office’s filing system for eventual transfer to the University Archives.

Item 16. Placement Tests Scores File. (RESTRICTED ACCESS) AutoText 241 (7/15/01)
Reference copies of placement test scores maintained in paper and electronic formats concerning results of administered tests that are used to determine a student’s aptitude. (Record copies for undergraduate students are maintained by the Office of University Registrar.)

Disposition Instructions: Erase/destroy in office records in paper and electronic formats 5 years after student's last attendance or date of graduation.

Item 17. Policies and Procedures File. AutoText 49 (7/15/01)
Record copies of policies and procedures created and/or maintained in paper and electronic formats concerning office administration.

Disposition Instructions: Transfer paper records to the custody of the University Archives after superseded or obsolete for appraisal and final disposition. Erase in office electronic records when administrative value ends and after records have been printed and filed into the office’s filing system for eventual transfer to the University Archives.

Item 18. Programs File. AutoText 50 (7/15/01)
Record copies of records created and/or maintained in paper and electronic formats concerning school/department academic programs. File includes correspondence, program proposals and descriptions, and other related records.

Disposition Instructions: Transfer paper records to the custody of the University Archives after 5 years for appraisal and final disposition. Erase in office electronic records when administrative value ends and after records have been printed and filed into the office’s filing system for eventual transfer to the University Archives.

Item 19. Prospective Graduates File. AutoText 51 (7/15/01)
Record lists created and/or maintained in paper and electronic formats concerning prospective graduates at end of each semester.
Disposition Instructions: Erase/destroy in office records in paper and electronic formats after 1 year.

Item 20. Reports File. AutoText 57 (7/15/01)
Record copies of reports created and/or maintained in paper and electronic formats concerning office programs.

Disposition Instructions: Transfer paper records to the custody of the University Archives after 5 years for appraisal and final disposition. Erase in office electronic records when administrative value ends and after records have been printed and filed into the office’s filing system for eventual transfer to the University Archives.

Item 21. Student Awards and Honors File. (RESTRICTED ACCESS) AutoText 66 (7/15/01)
Record copies of records created and/or maintained in paper and electronic formats concerning departmental awards and honors to students. File includes approval forms, recommendations, and other related records. (Comply with applicable provisions of 20 U.S.C. 1232g (Family Educational Rights and Privacy Act of 1974) regarding confidentiality of student records)

Disposition Instructions: Erase/destroy in office records in paper and electronic formats after 5 years.

Item 22. Student Financial Aid (SFA) File. AutoText 67 (7/15/01)
Record copies of records created and/or maintained in paper and electronic formats concerning students receiving financial assistance and/or scholarships. File includes program participation agreements, work-study payroll forms, financial aid disbursement histories, recommendations, financial aid award notification letters, descriptions of awards, eligibility criteria, guidelines/procedures and regulations, applications submitted to Department of Education or lender, and other related records. (Comply with applicable provisions of 20 U.S.C. 1232g (Family Educational Rights and Privacy Act of 1974) regarding confidentiality of student records.)

Disposition Instructions: Erase/destroy in office records in paper and electronic formats 3 years after date financial aid package was awarded to student and after released from all audits.

Item 23. Summer School File.
Records created and/or maintained in paper and electronic formats concerning courses taught during each summer school session. File includes correspondence, reference copies of listings of instructors, reference copies of course schedules, and other related records. (Original records maintained by the Office of University Registrar for eventual transfer to University Archives.)

Disposition Instructions: Erase/destroy in office records in paper and electronic formats after 5 years.
APPENDIX B
Chancellor’s Office

http://www.unc.edu/

realchan

chan

17 files in main folder

China

fyi

new_images

singapore_visit

speech archive

Stateofuniversity.unc.edu

streams.sph.unc.edu_8080