Planning the Archiving of Project Websites and Digital Materials

Larry S. Jackson, Ph. D.
Graduate School of Library and Information Science (GSLIS)
University of Illinois at Urbana-Champaign

Technical Report ISRN UIUCLIS/CIRSS/DCEPH--2012/1/VER01+DCEPH
May 18, 2012

Introduction

Universities have long retained writings and project materials of faculty members in some combination of archives and/or libraries. Gathering and organizing digital materials from multiple web servers or computation-host computers, though, adds new challenges to archival ingestion processes. Further, constructing suitable facilities to provide ongoing access to this material can be both challenging and expensive. With some forethought and deliberateness in the construction and location of project material, advocated herein, archival processing might be significantly simplified. This document can serve as a guide to points deserving of consideration in the planning of how a project is to be archived. If such planning occurs early on the life of a project, both the completeness and the affordability of the archive might benefit.

This paper discusses considerations in assembling an archive of the digital materials of a project. It draws extensively on the author's experience in archiving hundreds of websites for the State Libraries and/or State Archives of seven U.S. states, plus the archiving of several faculty projects at the Graduate School of Library and Information Science (GSLIS) of the University of Illinois, Urbana-Champaign. GSLIS project archiving was motivated by termination of the projects involved, due to reaching the end of the sponsoring grant and/or faculty retirement or relocation.

Principal Problems

Although at one level, directives such as "archive the website" or "archive the project" are sensible, the implementation of such directives raises a great many difficulties. Before embarking on a project archiving effort, principal investigators and archivists should come to a mutual understanding of these areas;

1. What exactly is the scope implied by the phrases "the website" or "the project?"
2. Can the material be legally and ethically retained in an archive?
3. How can copies of all the material be obtained?
4. Will the archived material remain available and useable?

Boundary Delineation

What constitutes a project website or the set of artifacts of a project? Stated another way, what rules delineate the content to be included within the archive which is to be built? As websites are implemented with hyperlinks, and as content authoring of academic projects may involve multiple research assistants or classroom students, and varying degrees of editorial oversight, project hyperlinks might point to information content in a great many locations. Project datasets may additionally reside in storage systems located at, or used with the computation-host computers used in the project. Offline materials, too, may be, or may have been, held by any of multiple project participants.

The process of delineating the artifacts a project comprises probably involves less total effort, and certainly involves less speculation if done at project inception, and/or in an ongoing way throughout the life of the project, rather than primarily being done after the fact. Administrative policies permitting the inclusion of facilities within a project only as an intentional act would reduce the total number of information storage locations, and may also reduce redundant storage of specific items of information content.

Perhaps a faculty member might have multiple projects to be archived, or there may be material which transcends a single project, or material which applies to no specific project. But, a project-oriented archival grouping or finding aid provides at least some collection-like context information of the individual artifacts of a project, and might be expected to enhance the potential for reuse of data or computer programs of that project. Project-oriented access mechanism need not be the only access mechanism the archive maintains.

The web presence of a project might be spread across multiple servers. The content within these servers may or may not hyperlink to information on other servers. Example content-host websites for projects include;

1. the department's website,
2. the university's website,
3. the sponsoring agency's website,
4. faculty members' personal and professional homepages, websites/webpages, blogs, and social media pages,
5. students' homepages, websites/webpages, blogs and social media pages,
6. the project's public website,
7. websites (e.g., wikis) for internal communication within the project staff or department,
8. websites created for the convenience of a larger community of practice (e.g., websites with *.org names),
9. web servers where the parent organization allows scripts to be run,
10. shared disk storage provided, for both faculty and students, by (a) the department, (b) the university, and (c) several web-focused commercial corporations, and
11. websites where content of a certain specific nature are widely shared (e.g., Flickr, YouTube, Facebook) or where website implementation is facilitated (e.g., Macromedia Flash, stylesheet and schema libraries, stock image/photography libraries).

It may be that some of these hosts may take archival responsibility for projects. For example, the relevant department, university, or sponsoring agency may centrally archive materials for both completeness of their own historical record, or for other purposes. However, prudence suggests that depending on archiving by others be confirmed rather than presumed. If recurring costs are involved, a correspondingly ongoing means to pay those costs will be required to sustain the archive.

Requirements, Prohibitions and Limitations on Retaining Material

Consideration of ownership rights not only applies to the postulated faculty member chairing the project being archived, but also to contributing colleagues, research assistants, (classroom) students, and website/journal/textbook publishers. Resolving ownership of the intellectual property is usually a prerequisite to obtaining the necessary permission for inclusion of the material in an archive. Ownership rights must be considered for both the time of construction of the archive and on into the future.

In rather the reverse direction, project records may be required to be retained for a certain period of time (e.g., when research involves identified individuals, or when a sponsoring agency requires record retention). Mandated retention must be compared against other factors arguing for selective (e.g., sampling-based) or term-limited retention.

An "expectation of privacy" may apply to included materials (e.g., especially concerning identity-disclosing information), possibly making Institutional Review Board (IRB) review of the proposed archive appropriate.

Obtaining Copies of Digital Materials

Automated acquisition of materials from facilities provided by web-oriented businesses may run afoul of the usage policies of those sites. It may not be permitted to acquire content posted on some websites, making the obtaining of those project materials for archival uses problematic, even for the creator of the material. Perhaps author-private pre-publication draft copies may be the only copies which can be legally retained in the archive.

The mechanisms which need to be used to obtain the various types of digital information vary (see following). Unless a project's documentation and data locations were controlled, project materials might reside in many different locations. Obtaining copies of project materials can be expected to be a significant undertaking, growing more so with the number of years the project information facilities have operated without centralized administrative rules governing its construction.

If physical media are obtained for use in the archive, then media obsolescence considerations (following) also apply. In the cases of both the original material and the digital copies made by the archive, multiple media types and file formats must be supported. At the outset of gathering the material, media items might be held by any of several members, or former members, of a project's staff. To the extent to which existing arrangements for physical custody are known, they should be provided to the Archivist by the Principal Investigator.

Content-hosting firms may encounter fiscal difficulties or undergo organizational change, and either cease operation or change access policies suddenly, suggesting project archives should be constructed promptly upon project termination, or continuously throughout the life of a project.

Media, File Formats and Obsolescence

The set of those physical recording media types commercially available undergoes continual change as new products enter the market. File formats also change, however the driving forces behind those changes are changes in software (i.e., either operating system or application software) and not hardware. Even if a file is recorded on a still-viable media type, continuing changes in the applicable vendor's operating system may abandon support for that media type, or the original file/directory structure as used within a certain media type. Revised application software may not be capable of reading all its earlier formats with complete fidelity. Disappearance from the marketplace of software products is a frequent occurrence, being a particularly severe form of software "change." As discussed more extensively following, software and computer system changes frequently occur, often with minimal notice.

Assembling Archive Contents

Maintaining an archive has its costs. The Principal Investigator and the Archivist should plan for these costs at the outset of the project, ensuring there are suitable and sustainable sources of funds. It seems prudent to assemble copies of all digital material to be archived under the physical and administrative control of the archive, although the archive may contract for certain services with third-parties. Once so assembled, the material can be copied onto different media types, reformatted for different application software or for retention under different operating systems, distributed to offsite storage, or loaded into online digital libraries/archives. These curation activities will have associated costs, but at least the work will be as implementable as possible, given that the material to be processed will be ready to hand.

Archiving an ongoing project might best be done by making a "snapshot" copy of all the essential files at one point it time, and then archiving that copy. Conversely, some portions of a project undergo very rapid change, so perhaps administering that section of the project with the aid of a version control system, and then archiving the version control system, complete with all its records of changes, would permit a more fine-grained ability to recover a certain specific version of the archived material. Both approaches may be simultaneously applied, if considered appropriate. An issue is whether or not those assembling the archive content have the full cooperation of those operating the host computers currently housing/serving the information content.

Database Contents

Many web servers provide some form of gateway access to "back-end" systems (e.g., formatting rather terse database contents into a more human-readable webpage). An online telephone directory would be an example of such a system, where the name and phone number of the searched-for individual are obtained from a database and formatted nicely for web presentation. Blogs and Wikis generally store the text of their pages as database entries too, composing the webpage only upon browser-user demand. Gaining access to the back-end, though, quite universally requires the full cooperation of the operators of the current content-hosting computer.

For database systems, deciding what is to be archived and the formats in which it is to be archived are policy decisions of the archive. As has long been the practice in archives of physical material, retention of all the included records might give way to the retention of representative samples (e.g., typically for cost-limiting reasons) [Berner, 1983].

To best support data reuse, archiving the database contents, the database schema definition, and the scripts which generate the sought webpages is the more complete answer. However, a future user of the information would have to reconstruct a suitable run-support environment in order to obtain what one specific resultant webpage would have looked like. Archiving all possible resultant webpages might be desirable, but in cases involving retrieval based on multiple search parameters, this might be combinatorially impracticable. Perhaps archiving only a few example renderings, selected because of differing display/genre type, would suffice (e.g., the phone number display of an individual, an academic department, and the display for any public "help" numbers). Archiving both the machine-readable form plus at least some of the more human-friendly renderings could also be done.

Data can widely be archived in delineated text files, though narrative fields can introduce difficulties to the process, and alternative character sets must either be supported or disallowed. Encoding in markup languages with a schema is a more generalized alternative to delineated field/record storage. Database schemas, though, need some form of "stand along-side" retention, or embedding within an archived markup language file.

Computer programs manipulating the database are another separate item for retention. These computer programs might have remote access facilities which utilize the web server, so those web scripts, too, should be archived. Database copies in vendor-specific formats are more easily reified onto extant computer systems (e.g., in "recover from backup" scenarios), but can be considerably more problematic to reify in the more distant future. Retention in open formats would be expected to be more survivable, though proprietary formats might also be used, as a convenience-oriented supplement for the near-term.

Website Crawling and Scraping

If the cooperation of the database operators cannot be obtained, often the only practicable means of obtaining copies of project information from a website is by copying all the pages of the website using a web "spider," or "crawler."[1] The many technical, practical, and procedural limitations of a spider-based approach to content gathering are discussed in [Chakrabarti, 2003], [Jackson, 2003 & 2005], [Masanès, 2005], [Pardo, Burke & Kwon, 2006], [Thelwall, 2004], and [Wells & Pearce-Moses, 2006]. If the implementation technologies and/or the design of a website do not support crawling by a spider, there may be no means outside of the cooperation of the website's operators to obtain copies of the digital

[1] The definition of these terms varies considerably. Here, they refer to some computer program which recursively transits some subset of the web, retaining copies of files on a local disk. Other possible software functionality is not germane to this discussion.

material sought for the archive. Again, crawling may be combinatorially impracticable, especially in situations where multiple parameters are simultaneously used in database queries. Use of spiders on certain websites may violate terms of use agreements. And, spider operations on certain websites may be blocked through technical means.

Another considerable disadvantage of crawling as the only content acquisition approach is that the relatively small amount of data, formatted into an attractive display, will need to be parsed ("scraped") out from the surrounding invariant ("boilerplate") markup text in order to support future reconstruction of the database. For example, if one person's name and phone number are presented to users in a webpage, then only those few dozen characters of information are unique to that webpage, as opposed to all other webpages of formatted results of searches for telephone numbers in the database. To reconstruct a database table of names and phone numbers would involve obtaining all the possible search result pages, and then discarding all the invariant content, populating the records of the new table with only the characters of each name and phone number.

Spiders encounter fundamental limitations in content retrieval when they encounter active technologies embedded in webpages [Thelwall, 2004] [Weideman & Schwenke, 2006]. In-line displays of non-markup formats may not communicate to the spider the identity of the necessary files to be downloaded and retained. Scripts retrieving files at run-time may or may not contain spider-recognizable web address, though spiders differ in the extent to which they can discover a usable address from an otherwise unrecognizable character string fragment. Wholesale duplication of downloaded content may occur. A spider-constructed archive should generally be considered incomplete and/or not compliant with archive organizational policies until manually proofread.

Spider-generated copies of websites also raise the issue of whether or not embedded hyperlinks should/may be spider-revised so that the copied website is itself navigable by web browser users via clicking the displayed links. If the links are not modified, they often point back to the original website host computer, which may well not exist in the future. If the links are modified, then the archived copy of the website is not a verbatim copy, but an automatically generated "work-alike." Depending on the motivation behind creating the archive, either approach, or both approaches, might be chosen.

Further, much website content incorporates active technologies and scripts. These aspects of webpage displays may simply not work, when delivered from the archive computer, or after a referenced online facility is no longer available at the original web address. There are some webpage implementation choices which will not be supportable from archives, or at least, not supportable once a necessary other server ceases to be available at its former web address.

Sustainable Retention

In archiving a website, a facility is needed where digital materials may reside with minimal risk of content corruption. It may also be desired to have the content remain web accessible, in some fashion, at least for the foreseeable future. Further, an archived copy probably should reside in some readily discovered place, and should be under the continuing cognizance of some official of the archiving agency.

Third-Party Archives and Trust

As examples of archives aligned with the missions of the research sponsors or hosts, the University of Illinois has recently augmented its systems for the long-term retention of digital papers and data from faculty projects, and some federal research sponsors have begun to operate their own retention systems for results from research they have funded. If the mission foci of the content creators and the archival facilities correspond, the content itself would seem to be more "core" to the archive mission, with the expectation that the content might survive longer or in a more complete form.

Copying of great numbers of websites via spiders is currently being done by official, quasi-official, and self-appointed groups. Some of these groups may intend, for the moment, to retain the copies as a long-lived archive. As these groups do not have the same commitment to the information content of a website as that of the content creators, it seems imprudent to assume long-term viability of an extra-organizational archive copy. If some agency in the administrative or funding hierarchy of the content-creating agency is unwilling to make the curatorial commitment to the archived content, an outside group may also stop short of making a binding commitment.

Third-party organizations or groups might be hired to provide retention facilities as effective as in-house facilities. However, as any archive is subject to organizational change or equipment damage, the viability of the archive itself needs to be continually reassessed by those stakeholders needing access to the information content. Fiscal realities might suddenly change the availability of any archival facility, or of the content stored therein.

Media Obsolescence

New digital media are continually invented, and less capable media types fall into disuse. If archive content is recorded on media that becomes completely obsolete, access to that content is lost. In a situation largely different from archives of paper copies, digital archive curators must act promptly and continually to copy digital materials onto new media whenever the recording media initially used begins to be less commercially supported. An undisturbed paper copy of a document may remain viable for a long period without much additional curatorial expense, but an undisturbed digital copy may become useless as a means of access to the file contents in ten or twenty years, due to the progress of commercially supported technologies in the world outside the archive.

Media also have vulnerabilities arising from their various physical natures, such as the susceptibility of magnetic recordings to heat or magnetic fields, or the propensity of early CD-ROM/DVD plastics to discolor to the point the substrate could not be reliably read by laser. Another requirement of digital curation is the continuing appraisal of media viability, and the possible need for timely re-copying of information content to another media type, or to another item of the same media type.

File Format Obsolescence

If archived content is formatted for display using a certain software application, running under a certain operating system, and that application and/or operating system becomes completely obsolete in the computer marketplace, all access to that information content might be lost. Archive curators must act promptly to copy digital materials into new file formats whenever the application or operating system initially used begins to be less-widely commercially supported. Again, as in the case of changes in media popularity and support, while an undisturbed paper copy of a document may provide viable access for a long period without much additional curatorial expense, an undisturbed digital copy may become useless as a means of access to the file contents in ten or twenty years, due to the progress of commercially supported technologies in the world outside the archive.

When a particular type of application software is withdrawn from the market, conversion of the document information content to some other application software is generally the only practicable way to go about preserving the content. Unfortunately, such conversions often have less than complete fidelity to the original, and some may be quite nearly illegible (e.g., compare HTML table layouts where a "save as PDF" has been done). In general, the more exotic the original document formatting directives, the less likely those directives will happen to be identically supported by some other software application. Simplified, or open content-formatting directives seem the most likely to survive.

Offsite Backups

An archive needs to be replicated at more than one physical location in order to survive events which might destroy any one instance. Hard disk crashes are an example of a damage event which might be defended against by simply providing a second, redundant disk within the same computer enclosure. But, a lightning strike might simultaneously destroy the entire contents of that enclosure. Considering a larger radius of damage, building fires or weather events are common examples of damage events sufficient to cause the complete loss of all the contents of a building, including both a project's computer and a department's in-house archive.

With digital information, content replication offsite is considerably more affordable than for physical documents, and especially more affordable than for unpublished manuscript documents. Of interest is then the probability that at least one copy of the information survives. If so much as one copy survives, more copies can quickly be regenerated. The probability that at least one copy survives is greater than the probability that exactly one copy survives, as the survival of two or more copies will also support recovery operations. Rather than sum the probabilities that exactly some *n* copies survive, it is simpler to instead calculate the probability that no copies survive, and then to take the complement of this probability to determine the probability that one or more copies survive [Hoel, Port & Stone, 1971].

$$\text{p(at least one survives)} = 1 - \text{p(all copies destroyed)} = 1 - (\textstyle\sum_{i=1}^{n} p_i).$$

If the probability of a copy being destroyed is p_i, then the probability all copies are simultaneously destroyed is $\sum_{i=1}^{n} p_i$ If p_i is very small, and (say) the same values for all values of i, say a disastrous building fire occurs on campuses once per 1000 building years, then, for 3 non-co-located copies, the probability that all three are simultaneously destroyed some year is $(1/1000)^3$ or .000000001 or

.0000001%. The probability that at least one of the three copies survive is one minus this value or 1-.000000001 or .999999999.

Adding another offsite copy increases that probability to $(1-(1/1000)^4) = 0.999999999999$. While the cost of the duplicated computer hardware for offsite backup facilities increases linearly with the number of such offsite facilities, the likelihood an incremental facility will actually be utilized in recovering the only viable copy of the archived content decreases exponentially, as the exponent increases around the very small value p_i. In deciding the number of independent backup copies to maintain, costs of additional backup hardware must be weighed against the cost, however calculated, if all copies of that information were to be lost.

While automatic backup systems might assist in moving these copies to their off-site locations, the mathematics of the survival probability only requires that the copies somehow reached their off-site retention location. Mailing a DVD would be functionally equivalent, providing the archiving staff has that as a formal step in the ingestion process. If the archive staff is forgetful or undisciplined, automatic copying of files might provide increased assurance. However, automatic systems can themselves fail (e.g., if the cryptography or DNS infrastructure linking two computer systems is changed, presumably the automatic transmittal offsite will fail). Whether or not the archive operator is notified of such failures is dependent on the design of the systems involved.

It is important for archive staff to act promptly in the event a copy of the archive is lost as the survivability of the information is compromised until such time as a replacement copy is brought to full capability. For example, if the archive resides on only two on-line disks and one is lost, then during the time interval until the failed disk is replaced and the archive content re-copied, there exists only the one surviving copy. A second casualty damaging that surviving disk within the recovery time period is not defended against, and will result in the complete loss of the archive content. The number of copies established by archive policy, less one, is momentarily not in compliance with the policy on the degree of redundancy in backups.

Summarized Heuristics

At a high level, then, planning and implementing a digital archive for a research project reflects multiple steps. Note that each step listed generates its own archive-contextual metadata documentation, which should also be saved with the archived material.

1. Identify all the digital/digitized materials which should be retained in the archive of this project. For each, identify where they reside, and how they are currently accessed.
2. Make survivable arrangements to support all existing access mechanisms into the content (e.g., finding aids, indexes, inventories), for both the current day and into the future. That is, treat the access mechanisms as information content in their own right, and ensure that they too are fully, functionally archived.

3. Review the entire planned content of the archive to ensure all those materials may be legally and responsibly be retained in the archive, considering both the purpose of the archive and the institution's policies on archives.
4. Review the entire planned content of the archive to identify supplemental file formats appropriate to the long-term survivability and probable reuses of each item.
5. Obtain copies of all these materials, retaining one or more access mechanisms to each information item, as appropriate.
6. Convert original materials into the other formats and media types identified for better/best survivability and/or reuse, supplementing the original materials but not replacing them. (Content conversion may engender considerable cost.)
7. Find a digital archiving and/or on-line access facility (a) with motivations and sustainable funding appropriate to the importance of this archive, and (b) which possesses suitable reliable infrastructure to provide viable copies in the event of plausible damage events, and (c) which is staffed and equipped so as to be able to monitor and respond to shifts in commercially viable technologies before complete obsolescence occurs.
8. Post a copy of the all the digital materials of the archive into the archiving facility, including the generation of all appropriate metadata and access mechanisms to support both inventory and discovery of the material.
9. Keep abreast of developments with the archiving facility, especially with regard to their funding and staffing.
10. Keep abreast of developments in file formats and media types which may necessitate large-scale porting of project materials within the archive.

References

Berner, R.C., 1983. *Archival Theory and Practice in the United States -- a Historical Analysis*. Seattle: University of Washington Press.

Chakrabarti, S. (2003). *Mining the web - discovering knowledge from hypertext data.* San Francisco, CA: Morgan Kaufman Publishers.

Hoel, P.G.; Port, S.C.; Stone, C.J (1971). *Introduction to Probability Theory*. Boston, MA: Houghton, Mifflin Co.

Jackson, L.S. (2003). Preserving state government web publications – First-year experiences. In *National Science Foundation National Conference on Digital Government Research (DGO2003), Boston, MA, 18-21 May 2003*, 109-114. Available at http://www.ideals.illinois.edu/handle/2142/16400.

Jackson, L.S. (2005). Difficulties in electronic publication archival processing for state governments. In *Proceedings of the 1st International Conference on Universal Digital Library, ICUDL2005, Hangzhou, China, October 2005*, 175-185. Pan, Y., Reddy, R., Gao, W., Zhuang, Y., Balakrishnan, N., Chen, C-C., et al (Eds.). Hangzhou, China: Zhejiang University Press. Available at http://www.ideals.illinois.edu/handle/2142/16401.

Masanès, J. (2005). Web archiving methods and approaches: A comparative study. *Library Trends 54(1)*, Summer 2005, 72-90.

Pardo, T.; Burke, G.; Kwon, H. (2006). *Preserving state government digital information: A baseline report.* Albany, NY: Center for Technology in Government.

Thelwall, M. (2004). *Link analysis: An information science approach.* Amsterdam, The Netherlands: Elsevier Academic Press.

Weideman, M.; Schwenke, F. (2006). The influence that JavaScript™ has on the visibility of a website to search engines - a pilot study. *Information Research, 11(4)*, July 2006.

Wells, G.; Pearce-Moses, R. (2006). From bibliographer to curator: archival strategies for capturing web publications. *IFLA Journal* 32(1), 41-47. Available at http://ifl.sagepub.com/content/32/1/41.full.pdf.