Victoria McCargar
Consulting in digital archives

Digital archives at MU: the J-School and beyond

**Report to the director of University Libraries,
University of Missouri-Columbia**

Table of Contents

Executive summary

Preserving the legacy of the School of Journalism

Introduction

As the J-School prepares to celebrate its centenary in 2008, it is also heading toward a digital crossroads.

The new Donald W. Reynolds Journalism Institute, with its mandate to explore new frontiers of newsgathering in the Digital Age, will train the next generation of journalists in a host of technologies that are already wreaking fundamental changes in the profession. Even as the very existence of traditional newspapers is increasingly called into question, the University of Missouri is poised to produce visionary leaders and practitioners who will guide journalism and publishing through the current technological upheaval into the next hundred years. Web pages, video, RSS feeds, pod- and videocasts, and delivery devices yet to be invented will be there to challenge and inspire students and faculty in their state-of-the-art new facility.

And where will all this multimedia journalism end up? News archives, famously the in-box for the first draft of history, must also rise to the requirements of this flood of digital output. Unfortunately, the fragility of digital information in any form is a threatening paradigm in its own right. There are no assurances that *any* digital content produced *tonight* by *any* newspaper will survive in its database or on its CD-ROM disks for ten or fifteen years, let alone the next hundred. The complexity of current media (revisit the list above: web pages, video, RSS feeds, pod- and videocasts and those media still to be invented) only work to shorten that time frame. For all the technological wonders the J-School will be producing, the legacy of that material is at best unknown, and at worst, vanished.

I hope this short statement of the seriousness of digital preservation issues will set the stage for the rest of this report.

Missourian archives

My first charge in being brought in to consult for the University of Missouri was to determine whether there was any way to rescue the defunct digital archives of the *Columbia Missourian,* some fifteen years of text and seven of images. Losing the archives in a series of unanticipated, even tragic, events is not unusual for a smaller paper where resources are always stretched. It remains to be determined whether the archives can be salvaged: I learned that backup data may still exist and that efforts to extract it from a vendor's proprietary format may even have been launched, but the outcome of either was unknown at the time if my visit. If the material is found, J-School and university stakeholders will have to determine whether to proceed with a rescue operation. Thousands of dollars and a commitment in staff time are likely to be required, so the costs and benefits will have to be carefully examined.

Archives policy

There are currently a number of *Missourian* and J-School products that are not archived at all, or are captured piecemeal. This confuses archives users, and the process for capturing archival content is labor-intensive and not necessarily the best use of the professional librarian's time. A written policy should determine what is ingested and what is omitted from the *Missourian* collection, and should guide a review of existing workflows. In that way, efficient processing, strategic building of the collection, and user expectations will all be in sync.

Reynolds Institute technology planning, POYi and other important collections

The situation with the *Missourian* archives and the loss of the first fifteen years of the collection provided a template for looking at the plans for the digital collection in the new Reynolds Institute. I was pleased to hear that planners considered an information professional to be an integral part of institute personnel, but not particularly surprised to find that only a part-time archivist position had been budgeted. Given the complexity of the digital collection and the necessity of building a system from the ground up, a part-time position is possible only after the system is in place and functioning, tended by a host of automated workflows to ensure quality and durability of formats. Getting there will be a substantial challenge, especially in light of the current cross-domain confusion surrounding technology support for the *Missourian* archives.

Fortunately, that confusion is a byproduct of one of the university's great assets — the cross-domain knowledge and expertise of a big institution. Properly channelled, there are probably more than enough resources in Columbia to sustain a great journalism archives.

That archives already has a tremendous foundation in Pictures of the Year International and Freedom of Information Center. These and other collections, which are on paper, have value on a national and international scale and should enjoy the same consideration in technology planning for the Reynolds Institute as the new tide of born-digital content.

Toward an institutional repository

I had the privilege of reviewing a copy of the University Libraries' task force report on developing an institutional repository for digital content. It was encouraging to see the high level of awareness of the serious issues of digital preservation among senior librarians. The report thoughtfully sets out the role of such a collection not just for the University of Missouri in Columbia, but on a truly global scale.

It is hard to avoid the convergence here between the commitment of the University Libraries to begin developing an institutional repository and the evident, immediate need for the School of Journalism to have a landing spot for its new Reynolds Institute content.

I know from my own newspaper experience in the no man's land between newsroom technology and the corporate IT department that journalists prefer to manage things themselves. In this situation, given the difficulty of digital preservation, I believe that would be a mistake. Digital content, very democratically, does not pay attention to which discipline has produced it; a Picture of the Year in TIFF format is indistinguishable from a TIFF scan of a Gutenberg leaf. The objects' bytestreams can occupy adjacent sectors on a storage disk; the difference will be made in the metadata. A University Libraries setting offers the internal cohesiveness of form and description that will allow the J-School content to be viewed and accessed all over the world. The expertise that can be leveraged across the university would be exciting to investigate and tap into.

The J-School has experienced the sad, but instructive, loss of its archives. It is happening on a smaller, slower but inexorable scale to newspapers all over the world. A major cross-discipline effort at building a 21st Century system would not only assure future MU scholars (and journalists) some semblance of useable content, but it might well serve as a beacon for other historic collections as well.

. Interviews and input

Schedule of contacts

I had excellent access to representatives from several constituencies within the University Libraries and Journalism School, although the schedule was quite compressed. These meetings varied in context, from formal presentations to informal conversations, but all were useful in providing a picture of the digital landscape in Columbia. Most of the schedule was arranged and facilitated by Nina Johnson, *Missourian* librarian.

Below is a recap of my schedule and brief synopsis of the subject(s) discussed.

Contact/affiliation	Meeting	Topics
Tom Warhover, executive managing editor, *Columbia Missourian*	9-10:30 a.m., **Tuesday**, Aug. 8	Basic history of events leading to the loss of *Missourian* archives, problems presented by that loss
Nina Johnson, *Missourian* librarian	10:30 a.m.-12 p.m. Tuesday	More detail on lost archives, basic description of the *Missourian* backfiles in other formats
Kathleen Edwards, coordinator, National Freedom of Information Coalition	12-12:45 p.m. Tuesday	Tour and overview of Freedom of Information Center archives and finding aids
Marty Steffens, professor of business journalism	1-2 p.m. Tuesday	Former colleague from *Los Angeles Times*; topic was outreach to China, Russia, etc.
Digital Collections task force, University Libraries	2-4 p.m.	Presentation on digital preservation and PREMIS project, "Preservation Metadata for Long-Term Digital Asset Management: a Beginner's Guide," followed by Q&A period
Jim Cogswell, director of University Libraries	4-4:30 p.m.	First-day wrap-up
Nina Johnson	9-10:30 a.m. **Wednesday**, Aug. 9	Discussion of archives, JNET, IT, access issues
Sue Schuermann, School of Journalism library	10:30-11:30 a.m. Wednesday	Tour of temporary J-School library, plans for new location

Contact/affiliation	Meeting	Topics
Dean Mills, dean of the School of Journalism	2-3 p.m. Wednesday	Discussion of potential issues of access to and retention of multimedia journalism produced by the new Donald W. Reynolds Journalism Institute, its context within larger university, and related collections (e.g., FOIC)
Pamela J. Johnson, executive director, Reynolds Institute	3-4 p.m. Wednesday	Follow-on to meeting with Dean Mills; discussion of digital collection development at the institute
Nina Johnson	4-5 p.m. Wednesday	Planning next day
Nina Johnson	8:30-9:30 a.m., **Thursday**, Aug. 10	*Missourian* archives
Brian S. Brooks, associate dean of School of Journalism	9:30-11 a.m.	Discussion of IT and JNET operations in School and within the university, planning for Reynolds Institute, efforts to retrieve lost *Missourian* archives
John S. Meyer, systems administrator, School of Journalism	11 a.m.-12 p.m. Thursday	Discussion of efforts to retrieve lost archives, possible disposition of retrieved data, general archives issues
Journalism School faculty, library staff, other librarians	12-2 p.m. Thursday	Presentation "Avoiding Digital Doom" — basic primer in digital preservation, Q&A period
Representatives of *Digmo*, *eMprint* and *Vox* products	2-3:30 p.m. Thursday	Discussion of material produced by J-School that is not routinely archived
Ashlee Erwin, J-School special events and communications coordinator	3:30-5 p.m. Thursday	Interview for "J-School Alumni: Profiles in Success" web site
Jim Cogswell	8:30-9:30 a.m. **Friday**, Aug. 11	Wrap-up of visit, likely next steps
Rick Shaw, director of Pictures of the Year International (POYi)	10-11 a.m. Friday	Discussion of POYi's existing archives, digital issues, copyright concerns and planning for Reynolds Institute
Nina Johnson	11-11:30 a.m. Friday	Brief wrap-up of visit
Daryl Moen, professor of journalism studies	11:30 a.m.-1 p.m.	Planning for Reynolds Institute, changes at J-School (former professor of mine)

Findings

Digital Archives: Issues for consideration and further discussion

My initial charge was to identify prospects for rescuing and retrieving the lost *Missourian* archives. It grew during the course of my three and a half days in Columbia into a somewhat broader mission: My time was divided between *providing* information on preservation issues and *gathering* information on the current state of digital archives at Missouri, particularly in the School of Journalism. The following table presents a brief rundown of topics that will be discussed in more detail below.

Topic	Synopsis
Missourian archives, missing years 1986-2002 (text) and ~1995-2002 (images)	The data may still be extant and accessible in one form or another.
Current *Missourian* archiving practice	The newspaper archives exist in multiple formats. Some products are not archived because data flow to archives are cumbersome or nonexistent.
Planning for Donald W. Reynolds Journalism Institute	Besides a line item for a half-time archivist, planning for access and retention technologies is not fully under way. The institute will house a number of existing archives which are currently stand-alone collections.
Pictures of the Year International, Freedom of Information Center, Committee of Concerned Journalists	These are three among possibly more archives that are highly valuable, if not priceless, and headed for the new institute. Description ranges from good to spotty, and physical preservation varies. There may be inhibiting copyright issues in any digital system envisioned for the Reynolds Institute.
Collaboration	There seem to be abundant opportunities for collaboration university-wide for digital archives strategies, especially in the context of a digital repository, among the J-School, University Libraries, School of Information Science and Learning Technologies and the Department of Computer Science.
Digital repository	The university's Task Force on Developing Digital Collections for the MU Libraries has begun the task of establishing priorities for a digital repository, which might now be expanded to include the developing needs of the new Reynolds Institute.

Discussion of findings

Possible rescue of the missing *Missourian* archives

The loss of the *Missourian* text and photo archives may be the most spectacular I have yet encountered, but the situation that led to the data loss is certainly typical of a smaller paper, where resources are stretched, individuals are sole proprietors of critical systems knowledge, and the need to "just get the paper out" overrides — often justifiably — all other considerations, including the archives.

Talking with *Missourian* Librarian Nina Johnson and Managing Editor Tom Warhover, and looking at some of the documentation of the data loss,[1] I came to understand the series of events that rendered the 1986-2002 archives inaccessible. Greatly condensed, here is what happened:

1. The newspaper's first and largely home-grown archives database is migrated to a newer system. Some years later, an important software upgrade is skipped and the format becomes obsolete, although the database is still functional.

2. The archives server experiences a major crash. Although it seems to have been backed up (this is uncertain), the archives become physically inaccessible. The vendor no longer supports the obsolete version.

3. In response to the *Missourian* librarian's query about attempts to rescue the data, newspaper technical support declares that the project is not worth pursuing until more pressing production issues in the newsroom are resolved.[2]

4. A new archives system is installed ~2002, but it contains only content from that time forward. Users needing to find a story from the missing years frequently turn to the competitor *Columbia Daily Tribune,* whose archives are free on the paper's website. Extracting data from the production system for entry into the archives is a highly manual process, and the librarian reports spending about half her time archiving *Missourian* material.

Obviously, no data rescue project can proceed without the backup data in hand in some form. In the course of our meetings, two interesting possibilities were raised. (1) Brian Brooks recalled that the *Missourian* had paid to have the data retrieved by the successor company that holds the license to the obsolete software. He said he did not know the outcome and would follow up. (2) John Meyer said he thought the backed-up data had been placed on a set of CD-ROM disks. He said they might still be around and he would look for them.

If the backup data exists in an accessible format (i.e., if it's on CD-ROMs, we assume they're not physically degraded, and if on tape that there is a compatible drive) there are two more significant hurdles: (1) establishing that the data is not

[1] Some of the existing documentation on the data loss was made available to me by Johnson, comprising a short series of e-mails. Others involved in the situation said there might be additional documentation but they were not sure, nor did they know where it might be located if it were still extant. Documentation is important to data recovery because it should contain pointers to past vendor relationships, support agreements, file formats, etc., that are useful in determining whether a rescue operation is feasible. A former support technician, now believed to be in New Mexico, may have additional information, according to Brooks.

[2] Nina Johnson, e-mail to John S. Meyer and others, Feb. 2, 2004, 4:11 p.m.; same-day reply from Meyer to Johnson and others, 4:27 p.m.

locked in a proprietary format that is impossible to parse out, and (2) that it can be successfully mapped into the data structure of the new archive. It is a positive sign that, according to Meyer, the obsolete data was written in some flavor of Unix. Mitigating that prospect is the "newer" archive, Merlin, which is already quite old. The vendor may not be willing to attempt to map the rescued data into another obsolete format. All this will require further exploration once it is established that the backed-up data actually exists.

Key points

- There does not appear to be much institutional commitment to the *Missourian* archives other than a basic recognition that they're probably important. Without that commitment, there is no clear responsibility for archival outcomes, and decision-making tends to be ad hoc. That has more or less acted to ensure that the lost data has stayed lost.
- The archives server was apparently backed up and the backup data even extracted by the vendor, but what happened to those files since then is not clear.
- The requirements of creating and sustaining the archives are consistently subordinated to more important technology projects having to do with producing the paper.
- When newsroom production systems were changed and improved, archival workflows do not appear to have been integral to the planning process. Opportunities for automating archival feeds may have been overlooked.

Suggested next steps

- Establish that the lost data still exists. Follow up with Brian Brooks and John Meyer on their speculation regarding the backup data: Did Meyer locate the CD-ROMs, and did Brooks determine what the *Missourian* paid the vendor to do. Also, any documentation on the system and its crash (e-mail threads, specification paperwork) might prove helpful in a data rescue operation; this should be gathered if possible.
- Determine whether the surviving data is in Unix, as Meyer suggested.
- If it is established that the data exists and is accessible, the *Missourian's* stakeholders will need formally to make the decision of whether to attempt a data rescue, weighing the benefits of electronic access to the files (some of which are otherwise available in hard copy form and on microfilm) against the potential costs (in time and money) of reverse engineering, metadata mapping and other processes that may be required to restore viability.

Current *Missourian* archives practice

The *Missourian's* peculiar position in relation to the School of Journalism and university at large bears some responsibility for the ad-hoc and unsupported nature of the archives. As a 501(c)(3) entity it is to some extent a freestanding organization, but personnel and technical support are divided among the *Missourian*, the J-School, MU's information technology operations and University Libraries.

Technology decision-making in the newsroom, which has a direct impact on the archives, may be made without much substantial input from the person with direct oversight of the archives, the *Missourian* librarian, whose position comes under the auspices of the University Libraries. The archives system, while administered by the J-School, is physically hosted by the university IT department, which determines critical access issues. The actual flow of content into the archives originates with the newsroom, but not according to any established policy.

These sometimes conflicting interests combine to create a system where standardized workflows, automation of basic processes, consistent metadata and data provenance are difficult if not impossible to establish and maintain. Instead of supporting the newsroom with professional research services and fact-checking — an increasingly important journalistic role in the era of Google — the newspaper's librarian spends as much as half of her workday tracking down missing content and manually coding it for ingest into the archives. Moreover, a software incompatibility problem between the librarian's PC and the newsroom's Macintoshes prevents her from acquiring *Missourian* metadata intact, so she has to enter it manually or correct it record by record.

Further complicating the picture, from an archival standpoint, is the *Missourian's* charter as a research and educational organization. Special projects, products and publications are generated by faculty and graduate students, many of which result in a body of published work. The expectation of these producers is that the content somehow ends up in the archives, but this is not always the case. The *Missourian* librarian fields many queries from disappointed searchers.

I had an opportunity to meet briefly with representatives of three news products that typify this situation: the *Digmo* (shorthand for *Digital Missourian*) website, the weekly magazine/tabloid section *Vox* and *eMprint*, an experimental product aimed at delivering a print look and feel in electronic form.[3]

None of this content is routinely archived. In certain instances, unique web content will be captured and ingested in the archives database, but these actions depend on an editor remembering to alert the librarian that there is archival material ready for processing. It must then be manually tracked down, cut and pasted into a Merlin record. This human-to-human system works fairly well (and is quite typical among newspapers) but it falls apart often enough to leave gaps in the archives. And, of course, it results in a lot of repetitive, clerical tasks.

Digmo content has at least a chance of being captured; neither *Vox* nor *eMprint* is saved at all. The reason appears to be that the librarian is locked out of even hunting-and-gathering activities by platform incompatibility: the software used to create each product does not run on the librarian's PC. Because the products are captured outside of the newsroom's regular production system, established workflows are bypassed and the content rendered inaccessible to the librarian.

Still another issue has to do with university policy regarding access to content on its servers. A security policy handed down through IT suddenly placed the *Missourian* archives — until then a popular resource with the J-School, subscribers and the public — behind the university firewall. I am under the impression that there was little or no prior consultation with the *Missourian* library, whose staff of two now answers calls from (sometimes irate) members of the public to look up articles they used to find for themselves. Another unintended consequence has been to drive those public users to the online archives of the competition, the *Columbia Daily Tribune.*

Together, these situations suggest a need to try to bring some coherence to a technology infrastructure that has grown from a series of cobbled-together solutions to one that is trying to be an integrated system.

It's clear that because resources are so tight, pulling this off will be a complicated challenge. But before any J-School technology project is undertaken, stakeholders need to make some basic decisions about the archives themselves.

There is no policy in place to determine what is saved and what is not. Certainly not everything produced under the J-School's roof is worth keeping, but what is? What are expectations by current students, faculty and other users? What would they hope to have available in the future? As a research institution, what are the obligations to the university and public? Who is allowed access?

It would be perfectly appropriate for the School to determine, for example, that retaining *eMprint* content in the *Missourian* archives is not cost effective (particularly when issues of its long-term viability in a format other than flat ASCII are raised — a much more complicated issue). Between a print version and longer web version of a story, is it

[3] *eMprint* is the latest iteration of media technologist Roger Fidler's work with reformatted and original print content packaged to offer newspaper readers an electronic alternative to browser-enabled websites.

appropriate to keep both? These are tough questions that archivists must answer all the time, but without them the archives will remain ad hoc, inefficient and incomplete.

These questions and answers potentially form the basis of a written policy for the *Missourian* archives. Such a policy would represent a commitment to that content. Existing workflows and systems must then be reexamined to see how well they fulfill policy. Those processes that do not accomplish this policy would need be refined until they do, or, if that proves prohibitively expensive or impossible, the policy needs to be revisited and user expectations managed accordingly. If the *Missourian* cannot or prefers not to retain its published content, users must be aware of what is reliable and available, perhaps in the form of a notice on the archives' home page..

Key points

- The *Missourian* lacks a clear policy for managing its archives. In its absence, the newspaper library struggles to meet a range of user expectations and lacks the systems to meet those needs.
- The academic agenda of the J-School mandates an ever-evolving set of projects, publications and products, but whether this content is suitable for long-term retention is not examined. The growing sophistication of electronic publishing products will vastly complicate the archives process over time.
- The technology infrastructure of the archives crosses several blurry boundaries, so there are no clear lines of responsibility.

Suggested next steps

- Post a notice on the archives' web site stating clearly that *Vox, eMprint* and most *Digmo* content is not archived. Monitor response and reaction from users.
- Work with the *Missourian's* systems vendors (Falcon, MerlinOne) to automate workflows under existing contracts. Upgrade the library's equipment if platform compatibility is an issue. If automation is not an option, weigh the cost and benefits of continued manual processing.
- Begin work now on a long-term archival policy for the J-School, including the *Missourian.*
- Revisit the university firewall issue and determine what steps are necessary to restore the *Missourian* archives to public access.
- Involve the *Missourian* Library and Journalism School Library in all newsroom technology discussions.

The Donald W. Reynolds Journalism Institute

"The Donald W. Reynolds Journalism Institute engages media professionals, scholars and other citizens in programs aimed at improving the practice and understanding of journalism in democratic societies."
— Statement of Mission

"The institute will focus on innovation and problem-solving across the diverse specialties within journalism: media convergence, news content and methods, new approaches to advertising, innovation in management, the impact of new technologies and new developments in law and regulation."
— Statement of Focus

The issues surrounding *Missourian* archives readily carried over into discussions of the digital material that will be generated by the new Journalism Institute. Its research mandate and cross-discipline contributors suggest data retention and discoverability issues that go well beyond those at the *Missourian* — by a factor of magnitude. In frank conversations with Dean Mills, dean of the School of Journalism, and Pamela Johnson, the institute's executive director, I tried to raise various considerations in dealing with the sort of material the new breed of converged "iJournalist" is likely to produce.

News in multiple manifestations (print, web, still images, video, RSS feeds, podcasts, video podcasts, blogs, blog feedback, wikis, and media yet to be developed) presents an overwhelming tide of data and data formats that need to be captured, processed, managed, retained, and possibly preserved in perpetuity. It is becoming axiomatic in digital preservation that the cost of sustaining files rises in direct proportion to the number of formats to be preserved.

Moreover, the institute's archives will become the umbrella organization for at least three already significant collections: the Freedom of Information Center, the acquisition of the files of the Committee of Concerned Journalists, and the archives of the renowned Pictures of the Year International photojournalism competition. (I offer a brief discussion of the POYi issues below.) These represent cross-platform, cross-media materials that will need to be brought together under one conceptual roof, with all the issues of technology, access, format and description — not to mention copyright — yet to be resolved.

It became apparent that these issues had not been altogether ignored, however: A line item in the institute current budget calls for a part-time archivist. I expressed the view that while a part-time position might eventually be feasible, that would only happen *after* the comprehensive development of such a data framework was complete. A part-time archivist at the outset would probably drown. Indeed, a professional archivist would think twice before taking a position where a definitive program for the digital content was not spelled out.[4]

I hope that I provided Mills and Pamela Johnson with a different view of how to think about the Reynolds Institute's future archives. (In fact, we joked about the need for another, less eye-glazing term.) The considerations of digital repositories — institutional, intellectual, and technological — all come into play and need to be part of ongoing planning, not merely within the School of Journalism, but university-wide.

There is, I believe, a real opportunity here for collaboration throughout the university. There are not only technical reasons for identifying and understanding all the cross-domain considerations that came up in discussions of the *Missourian* archives, but there are abundant intellectual benefits as well. Large-scale repositories have the potential to positively engage resources across the university: not merely the J-School, IT and University Libraries, but the schools of engineering, library science and (thinking about copyright and intellectual property) law.

Key points

- As an outcome of its research and publication activities, the Reynolds Institute will produce digital records in unprecedented numbers of media formats and platforms.
- It is a safe bet that students, faculty, other research institutions and the outside news media will expect and seek some form of access to the institute's output and findings.
- Planning for the future of the content, published and unpublished, is only in its earliest stages and will require extensive development.

Suggested next steps

- Form a cross-discipline, university-wide task force to begin forming policy for the institute's day-to-day digital asset management and archives.

[4] The Society of American Archivists' Code of Ethics states that archivists "have a fundamental obligation to preserve the intellectual and physical integrity of [the] records." (http://www.archivists.org/governance/handbook/app_ethics.asp)

- Begin investigating possible funding sources for the repository.[5]
- Come up with a more compelling, more palatable (and fund-able) and less boring word for *archives*.

Pictures of the Year International

The newly named director of POYi, Rick Shaw (a former Times Mirror and Tribune colleague of mine), and I met to discuss his questions and concerns regarding the 62 years' worth of photographs generated by the prestigious annual contest.

There is a consensus that this virtually priceless collection of photojournalism, which documents some of the greatest historic events since World War II, has been neglected from an archival standpoint, and that a few steps have been taken to put some controls in place. We talked briefly about the physical condition of the prints, storage (some of which is offsite), and intellectual access. I learned that a recent doctoral graduate undertook to catalogue the collection, but he apparently did so without input from information science professionals, raising such issues as metadata compatibility (it's unlikely he used MARC or DACS) and consistency in subject classification.[6]

Also of interest in the current discussion of digital archives and preservation is the last five years' worth of winners, which are all "born digital." While everyone agrees that this is a permanent collection, plans for archiving the material digitally are unformed at this point, except to say that the Reynolds Institute will have some role. These recent files, it should be noted here, are currently on CD-ROM — not a mid- or long-term archival medium.

The copyright issues with this collection, unfortunately, are complex and serious. Ownership of the winning photos is by the photographer's newspaper or magazine, with limited republication rights granted to POYi via a contract signed by the photographer. Lacking direct rights, POYi as a nonprofit organization (like the *Missourian*, a free-standing entity) is, according to Shaw, prohibited from taking certain measures that might be construed as being simply for preservation, such as placing the digital images in a piece of university-owned hardware like a file server.[7] While the contract does permit POYi to use the images for "research and education," Shaw said, he was not sure how this might be interpreted in practical terms — i.e., in a Reynolds Institute repository.

[5] In various conversations I mentioned the Library of Congress' National Digital Information Infrastructure Preservation Program (NDIIPP) as a possibility; the implications of that resource are beyond the scope of this report but are certainly worth serious discussion, especially in light of the School of Journalism's ties with journalism foundations and private industry.

[6] There has been work done in the past few years by Associated Press to develop a hierarchical authority list for describing news events, but it does not reference LCSH or other schemas, such as the Thesaurus of Geographic Names, in use at other institutions.

[7] This aspect of copyright law is currently in flux. The so-called Section 108 exception to the Digital Millennium Copyright Act is aimed at permitting certain otherwise-proscribed actions to prolong a digital object's viability. It is currently under study by the Library of Congress; see http://www.loc.gov/section108/.

The rights terms raise important issues about university access, funding and technology support that must be resolved as a key part of archives planning.[8]

Key points

- Missouri has physical control of one of the world's preeminent photojournalism collections but evidently lacks rights to the material, including the ability to store digital manifestations on university hardware.
- The collection's intellectual controls ("library science") have been neglected for decades, but recent efforts have tried to impose some order.
- The physical collection is stored in an archival facility some distance from Columbia.
- The collection is a component of the Reynolds Institute, but how that will be managed is unclear.

Suggested next steps

- Involve a university intellectual property attorney in establishing what the POYi contract permits and proscribes. Determine whether the university may have a role in archiving the collection digitally.
- Determine what portion of the collection involves orphan works.
- Have an information professional evaluate the cataloguing project done by the journalism doctoral student and determine what might be useable — and even useful — in a standardized repository.
- Involve POYi in asset management and archives planning for the Reynolds Institute.
- Have an information professional/trained picture conservator evaluate and benchmark the physical collection, both onsite and offsite.

Digital repository

While my initial task in visiting MU was to assess the possibilities for rescuing the *Missourian* archives, I was also asked to spend some time educating library staff on preservation issues and especially preservation metadata in a digital asset management environment. The University Libraries formed a task force to investigate implementation of an institutional repository, and, en route to mid-Missouri, I had a chance to read the group's *Report of the Task Force on Developing Digital Collections for the MU Libraries*.[9]

This thorough, insightful document raises many of the considerations I have attempted to bring up in analyzing and discussing asset management and archives issues at the Journalism School, *Missourian* and future Reynolds Journalism Institute in this initial brief trip. Such a large topic bears much more careful planning and discussion than is feasible here, but this is a good opportunity to review some of the key questions, most of which are covered in the report:[10]

1. Is the larger institution adequately prepared to undertake a large-scale digital project with all the contingencies identified, such as repository policy, metadata development, long-term (*very* long-term) funding, object life cycle and preservation management?

[8] An excellent 2003 white paper on the interplay between rights ownership and preservation outcomes is Brian Lavoie's "The Incentives to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making," available on the OCLC website at http://www.oclc.org/research/projects/digipres/incentives-dp.pdf.

[9] My commented copy is being submitted separately.

[10] I am drawing here on research I have been doing for the Center for Research Libraries in developing audit and certification criteria for a trusted repository for news content and the Audit Checklist for the Certification of Trusted Digital Repositories by the Research Libraries Group (http://www.rlg.org/en/page.php?Page_ID=20769).

2. Is staff sufficiently versed in current and ongoing research, established and emerging standards, and the nuances of digital copyright law?

3. Can MU's reputed strengths in cross-discipline cooperation be leveraged here?

4. With what other important projects will a repository program compete? For example, will digital preservation draw resources from traditional preservation and conservation?

5. How will the repository interface with established analog collections?

6. In what contexts will authenticity be an issue? Is it important to establish provenance, chain of custody, document authenticity?

7. How important will it be to preserve original software and functionality indefinitely? For what portion or portions of the collection?

8. What are future users (say, a hundred years hence) going to require in order to understand the legacy material?[11]

9. Is the technical infrastructure cohesive enough across disciples to support this kind of endeavor (cf. the situation that led to the lost *Missourian* archives)?

If I were to encapsulate the preceding dozen pages in one statement, it would be to say that benign neglect (of the sort that has been the rule with POYi, for example) does not carry over into the digital environment. If that is the default practice, you end up with the *Missourian* situation.

Digital archives require regular attention to shifting technology, formats, metadata and user expectations. The metaphor I found myself using constantly in conversation and presentation was "having a patient on life support." There can be no letting go of attention, no failure of mechanical equipment, no complacency with obsolete equipment and — crucially — no reduction in best practices in order to save money. If something is vitally important, it should be on paper or film.

Much of the material that will be created in the new Reynolds Institute, as has been the case for years university-wide, has no paper original, no paper equivalent, and does not survive reformatting to a simpler medium. Committing groups within the university, and the university as a whole, to long-range digital archiving is to sign up for a long-term, multi-decade experiment, but it is one for which there are currently no other options. The important thing is to understand the limitations, ask the right questions, and plan for how resources will be spent.

Key points

- The University Libraries are preparing for a digital repository, asking the right questions, and have a good grasp of the issue. This effort can be expanded to investigate the developing plans for the School of Journalism
- Terms like "institutional repository," "digital repository" and "trusted repository" have subtle philosophical distinctions. How the J-School plans will fit into one or the other should be part of the planning process.

[11] The concept of understandability has to do with the ability of a future user to interpret the preserved data. Not only is its intellectual context important (e.g. "where did these numbers come from?"), but how the data is rendered may be critical. For example, if a 3-D modeling program is necessary to interpret data triples, the information must be stored in transparent a way that allows it to be interpreted in whatever rendering program is available in the distant future. The technical metadata must also declare that 3-D rendering is required for understanding.

Suggested next steps

- Investigate concretely how the Reynolds Institute's program for handling digital content might intersect with the University Libraries' plans.
- Begin developing cross-discipline cooperatives that can help tend and grow the repository; an example might be a joint J-School/Library School M.A.-MLIS degree. Other opportunities may not be obvious. [12]
- Familiarize staff with emerging "audit criteria" for trusted digital repositories and explore how those criteria will be met over time. Even without establishing a formal "trusted" repository, the criteria ask a series of important questions about institutional suitability, technical environment, and the needs of future users. Begin with the article describing the RLG-CRL project and go from there: http://www.rlg.org/en/page.php?Page_ID=20793&Printable=1&Article_ID=1780.

[12] The University of Rochester enlisted the help of an anthropologist to determine why contributions to the repository by various departments were so uneven. For a feature on the project, see http://www.rochester.edu/pr/Review/V67N4/feature3.html

Conclusions

This was a great deal of ground to cover in three and a half days, but I think we accomplished a number of things:

- Raised awareness of digital preservation issues at a fundamental level (presentation to J-School faculty) and in a more sophisticated context (university librarians).
- Discovered that the Missourian archives (1) may have been extracted from their obsolete format by the vendor and (2) may still exist on a set of CD-ROM disks. Follow-up by the stakeholders involved will point to further possible actions to restore the archives — assuming it is something the J-school wants to commit to.
- Discussed short-term solutions and longer-range possibilities for handling non-archived material produced by the J-School: *Vox, Digmo* and *eMprint*.
- Helped *Missourian* librarian to develop a short current policy statement on these three products for immediate posting to the archives website.
- Initiated discussion of the need for archives planning at the new Donald W. Reynolds Journalism Institute and raised basic issues for consideration.
- Discussed analog and digital preservation issues, collection value and copyright concerns with the POYi director.
- Reviewed institutional repository report and discussed issues informally with library staff.
- Pointed staff to some online resources for further study.

Near-term next steps

Each section of this report has concluded with a short list of possible next steps. The next step for me is to hear back from MU about progress on any of these fronts (especially the specific, short-term items like the missing *Missourian* disks). We can look at how to further refine, advance or omit some of the proposals and, finally, to zero in on the most important issues to tackle when I return in the fall.

A short list of initial areas to look at:

- Rescuing the Missourian archives (or not) — willingness of J-School to pursue
- Refining existing newsroom workflows, vendor involvement
- Asset management and archives planning for the Reynolds Institute — organizing and agenda-setting
- Refining repository planning by University Libraries