

**The Eyes of Texas:
What can archivists learn from working with a digital institutional repository?**

Patricia Galloway
Assistant Professor
School of Information
University of Texas at Austin

Abstract:

Establishing even a small-scale archival digital institutional repository calls on every skill archivists can muster, both inside and out of the (technological tool-) box. In this case study we discuss one category of collections, faculty papers, as acquired and added to the departmental institutional repository created in 2005 for the School of Information, University of Texas at Austin. The nature of such collections requires of the digital archivist technical skills ranging from retrospective digital archaeology (presaging migrations and emulations to come) to prospective records creation management (asking archivists to manage an ongoing interaction among creator, records, and repository).

Origin of the project:

We wanted to establish a departmental institutional repository to serve University of Texas School of Information faculty, staff, and students by providing a secure and persistent environment to preserve and expose faculty scholarship, to archive scheduled departmental administrative records including faculty committee work, to preserve departmental digital productions (websites, tutorials) for historical reasons and as teaching resources, and to provide a repository for the preservation of the preprofessional work of students. We felt that the repository, if managed to instantiate archival standards on various levels, would exercise many of the skills needed for student archivists to learn to work with digital records, while additionally becoming an asset to other interests in the School.

How does this project address the issue of “new skills”?

Skills in the analysis and facilitation of the social ecology of record creation and use are vital to the management of all records, but they are especially important in the case of digital records because new environments have led to new uses and communication practices. Like all electronic records projects, however, this one also calls on an amazingly broad panoply of technological skills, because it demands not only that we understand the objects themselves and their formats, but also that we develop a sophisticated understanding of the system environment in which the objects must be processed and the repository environment in which they must be placed and managed. At the present stage where most of us find ourselves, the focus is on pre-processing of records to be ingested into repositories and metadata capture to support future preservation activities. I anticipate that our students will be struggling in the future to

solve the problems raised by the materials we ingest today, but if we wait to ingest until we know everything, there will be nothing for them to struggle with.

Background of the case:

Since 2003 I have been working with graduate students at the University of Texas School of Information (iSchool) on problems of archival permanence for digital objects using the open-source DSpace repository software as a testbed for our thinking and practice. We have worked on a design and policies for a centralized email repository for Texas state government (2002),¹ a website repository for the School (2003, 2005-2006),² preservation of papers and publications of iSchool faculty (2005), administratively created materials belonging to the iSchool (2005 and 2006), and special projects not directly involving the iSchool except as temporary digital repository, such as policies and practices for the permanent preservation of online scholarly journals and anthropological fieldnotes for the American Anthropological Association (2004 and 2006)³ and the Michael Joyce Collection hosted for the Harry Ransom Humanities Research Center (see Stollar and Kiehne project).

In 2005 we undertook the formal establishment of a digital institutional repository for the iSchool, beginning with faculty records and the recovery and preservation of the School's website in its historical incarnations. The faculty records subset of this task we called the "Eyes of Texas," in reference to the UNC-SLIS "Minds of Carolina" project for collecting the digital records of retiring faculty at the University of North Carolina at Chapel Hill. We used version 1.2.1 of the DSpace repository software. Theoretical principles and grounds for practice for all the procedures not instantiated in DSpace were adopted from the Open Archives Information System reference model and broad readings in the work of digital projects worldwide, with special reference to big projects like InterPARES and NARA research and special thanks to the industrious Dutch and Australians.

For the purposes of this case, I will concentrate on the "Eyes of Texas" project from 2005, although developments this year, with projects just completed, have greatly expanded some findings and will be mentioned as appropriate. I received a small grant sponsored by Ingenta and granted by the ALA Research Round Table to focus on the role of these collections in the institutionalization of a departmental repository, and as a result we have followed up on these collections and the institutionalization process this year. The creator-participants in the "Eyes of Texas" project were four faculty members, three at the end of their faculty careers and one (our dean) who is at active mid-career. We asked them to tell us what they wanted to preserve as emblematic of their work and potentially make available in this way to both our School and to broader university and public audiences, and their perspective was vital to our understanding of the contexts of record creation and use. We built on work done in the same seminar in 2003 to develop

¹ Marlan Green, Sue Soy, Stan Gunn, and Patricia Galloway, "Coming to TERM: Designing the Texas Email Repository Model," D-Lib 8(9), September 2002.

² Anne Marie Donovan, Maria Esteva, Addy Sonder, and Sue Trombley, Proposal for Establishment of a DSpace™ Digital Repository at The School of Information, University of Texas at Austin, May 10, 2003, available at <http://hdl.handle.net/1721.1/1273>.

³ See documentation filed in Portal Archiving Study Project, <https://pacer.ischool.utexas.edu/handle/1721.1/45> and in the ASSC Special Project, <https://pacer.ischool.utexas.edu/handle/123456789/1333>

repository policies on levels of service and support for different formats and types of materials and intellectual property, privacy, and confidentiality concerns, but we added to that work as additional issues emerged in the process of actually committing to support real materials.

What did we do with electronic records?

- Understand the creators' relationship with their records. Our first concern was to elaborate on typical inventory practice and emergent interests in self-representational aspects of private records to understand how the collection of digital materials each faculty member offered also mirrored the history of the computing environments each has experienced.
- Establish submission agreements with creators incorporating intellectual property and preservation commitments. As we all by now well appreciate, a typical SIP agreement goes far beyond the usual donor agreement (even one that anticipates accretions), especially in that it must remain a dynamic document.
- Capture the features and structure of the environment in which the records were created and used. In most cases we were able at least to control this information partially through historical institutional knowledge of the IT environment of the iSchool.
- Capture and catalog the technical characteristics of the records themselves, capturing and providing object-level metadata where feasible.
- Design a repository structure for the records to inhabit, incorporating collection-level metadata. Here our archives students had the opportunity to bring their skills to bear on digital collections.
- Preprocess the records prior to ingest. This preprocessing is not built into DSpace, so we followed best practices to set up workflows that would preserve intrinsic metadata intact while carrying out tasks like copying and testing for viruses.
- Ingest/accession the records, adopting and adapting the DSpace workflow.
- Prepare access versions of some formats.

What did we learn about working with electronic records? What skills did we need?

- Complex problems of format/operating system interactions (like the Apple/Macintosh resource fork) meant that we could not even capture records and move them into a preservation environment without a sophisticated knowledge of both the creation and preservation digital environments. Further, because even in the case of the active faculty member, files to be dealt with had been created with several different versions of software with some time depth, we had to proceed carefully so as to preserve the object intact for archiving before we moved to effect access. These complex issues meant that it was vital for us to consider from the outset which significant properties of each collection we needed to be able to preserve (beyond foundational bitstream preservation) and how we planned to render them on access.
- In one case, a faculty member wanted to include postprints but did not have any digital versions of older publications, so the student team set up a mini digitization project to satisfy this requirement. Here we instantiated digitization best practices as carried out the UT General Libraries' Digital Library Services department.
- In supplying keywords to assist in retrieval, we found that LCSH were both too formal and not detailed enough to be of much help with highly technical articles,

especially when compared with new access methods like tagging and full-text search. Accordingly, one student undertook an independent study to experiment with using data mining techniques on the texts of submitted publications for extracting keywords from the deposited texts themselves in order to establish a controlled vocabulary for subject description in that collection. This year other projects have made use of this procedure.

- Diagnosing obscure file formats and capturing environmental metadata required the use of tools as various as hex editors, directory listers, metadata extractors, legacy viewers, and web crawlers. Where these tools were open source and open access, we collected and archived them to form the nucleus of a working digital toolshed.
- Conventional archival description is aimed at the aggregate levels of archival collections rather than the granular individual object level. This means that for paper archives there is in fact—apart from ill-codified practice in the making of calendars and the like—very little “best practice” for metadata allocation at the individual object level. DSpace itself is designed to capture a certain amount of technical metadata on ingest, but the depositor is required to provide a range of Dublin Core metadata chiefly aimed at resource discovery. We are still in the process of wrestling with appropriate descriptive practice at this level and plan to delve into the consistent structuring of collection-specific metadata this fall in a class on metadata, now that we have worked through several solutions.
- Decisions favoring access were taken to produce more accessible “use copies” through conversion from the original objects (for example, JPEGs derived from TIFFs). It is clear that for larger-scale projects than we have yet tackled, a migration on demand tactic will be more appropriate, which will mean the provision of a digital workspace and appropriate tools for users to wield.
- Intellectual property issues proved to be quite complex. We made use of the online SHERPA registry (although initially in early 2005 it was still very much in formation and did not cover all the periodicals/publishers we needed) to determine self-archiving policies of publishers for published materials that we would instantiate in providing access copies.
- We referred to archival ethical practices regarding privacy to deal with a large quantity of emails, which we decided to preserve as an encapsulated object to remain closed for the present and to be used for research on email use in the medium term only with permission.

How did working with these materials differ from working with analog materials?

- Obviously the materials themselves are mediated: it *mattered* whether we looked through a Windows, Mac, or Linux lens. Students learned to *see* their mediating tools and not just *use* them. This made it clear that there is a need to establish and document a consistent preprocessing environment for handling digital objects.
- We were much more dependent on creators for learning about the records themselves and the environment in which they functioned than archivists of analog records usually feel they need to be. In addition, simply because most electronic records are being obtained much closer in time to creation than are

- paper records, our creators were still very much alive and concerned with their materials. This meant that our practice, like that of collecting archives, had to be much more participatory and to take much more account of the creators' wishes.
- We had to be concerned from the beginning with intellectual property issues; the fact that digital records have the potential for nearly instantaneous availability, together with the enhanced awareness of these issues in a university environment, meant that IP and access management had to be fundamental from the start to a degree that is quite foreign to paper environments where processing alone can take years and soften the problems attached to IP issues.

Postscript: Repository as Social Object

I was and remain interested in the problem of institutionalization as crucial to permanence in digital archiving, and I believe that institutionalization depends vitally on local commitments and participatory development. Campus-wide repositories are receiving all the promotion that university PR departments can offer, but in many cases they have found it difficult to attract materials. As a result of our work on all the projects mentioned above, we have developed the idea of a campus federation of departmental and special library digital repositories with a central repository located in the general library institution. This layered concept, which we have discussed with digital library principals in the UT General Libraries, answers both to the need for secure dark-archive storage of original bitstreams to guarantee authenticity and to the now-accepted standard requirement that any trusted digital repository must provide a succession plan in the case of its dissolution. This concept, however, needs to be proofed in practice, and that is what we are engaged in now.

We feel that it is particularly important to capitalize on the strong sense of local ownership that develops within a community of practice such as an academic department, and to recognize that such an institution consists of several important stakeholder groups—faculty, administrative staff, IT staff, and students—each of which needs to be involved in the project as it progresses. The iSchool repository is being anchored in iSchool practice in several ways:

- We are beginning to archive materials that must be kept to a specific standard because they are administrative documents under an approved records schedule. This year we began with website postings that constitute the record copy of policy statements, and we anticipate working next year with the office manager to archive digital meeting minutes going back several years.
- Faculty collections are being monitored via ISI and Google Scholar for their effect on faculty scholarship impact, and preliminary results have attracted the interest of a second potential cohort of faculty member-donors, including Professor Lorlene Roy, president-elect of ALA. Sharing scholarship has long been a goal of early adopters of repository technology, and even faculty who may want to post writings on their own websites are interested in archiving them in the repository.
- Among the faculty collections are learning objects and other materials which, because their creators are our own faculty, are relevant to the ongoing

instructional activities of the School. These are the kinds of materials that are rarely collected by university archives, yet the digital environment permits them not only to provide expanded provenance for more conventional faculty output, but also to be deployed for their functionality both as relevant now and as examples of information objects for future studies.

- Close collaboration with IT staff and webmaster have moved us closer to embedding the repository and its support in the ongoing IT service of the School. Because of the special security that the repository offers, IT staff are beginning to see benefits in what I would call “archiving in the strong sense” to protect their investment in historical versions of the website, for instance.
- Student interest in depositing work in the repository to support future professional careers has manifested itself this year in the form of projects laying the groundwork for the ongoing capture of an online student publication in preservation and conservation and for the self-archiving of student portfolios.

Publicity, as always, has helped in addition to collaborative efforts and as such has become an organic part of the process. In September of 2005 students exhibited five posters about repository-based projects at the annual meeting of the Society of American Archivists in New Orleans, and one of them was honored with one of two prizes. Three of the projects produced papers that were submitted to *Provenance*, a journal about archival practice, and the work of the Michael Joyce project, which prepared and deposited the first digital collection accepted by the Harry Ransom Humanities Research Center, is now being featured in an exhibit, “Technologies of Writing,” at the HRC. In addition, a brief report on the SAA posters was featured in the University of Texas-Austin alumni magazine, *The Alcalde*, and I was invited to give a presentation at the 2006 ALISE meeting on the use of an institutional repository for LIS teaching. Finally, I was granted tenure toward the end of 2005, and that fact, since it makes it likely that I will be involved with the repository for some years to come, has led to additional support emerging for the repository, totally apart from the recognition of its use to faculty.

Other specific events, not uncommon or unavailable to others, had an impact as well in providing opportunities for the repository to demonstrate its usefulness. Because a redesign of the School’s website was undertaken in fall 2005, the value of the repository as a secure archive for past designs came to the forefront. In addition, a new sequence of digitization courses, designed to serve digital library requirements, has made the necessity for a secure archival repository for such materials obvious, especially since I am involved in team-teaching the introductory course that stresses archival care for these derivative digital objects. During 2005 also, an online publication of the School’s Kilgarlin Center for the Preservation of the Cultural Record, *The Cochineal*, was targeted as the first candidate for deposit of student work, and this project has gone forward this spring. Finally, student interest in learning about archival institutional digital repositories has grown and students are beginning to undertake individual research projects using the repository. The overall result of this increased interest is that we obtained a new and more powerful server this year and are now in discussion with the director of information technology services in the School about the possibility of a .25 FTE student assistant to support repository activities.

This part of the story possibly does not suggest new skills, but I hope that it provides a case study arguing for the notion of a goal of systemic ubiquity for digital preservation. In that context, there is no avoiding the necessity for sophisticated digital skills requiring ongoing maintenance, but our work and its history has shown the importance of getting the maximal number of people in on the act, not by knowing what is good for everyone, but by asking them. There is so much too much work to go around in this task that we need all the help we can get.

Applied technical skills:

- Understanding of networking, operating systems, and related file systems, including tools to work within such spaces and to visualize virtual “original orders”
- Beyond this, understanding of standard system management practices, including security, backup, and software installation issues
- Excellent skills as a user of the desktop systems and web-authoring tools used to generate most of the file types at issue
- In many cases, at least modest digitization skills (we had to ingest some paper documents for efficiency of use)
- Good database skills
- Understanding of e.g. message digest tools for establishing fixity data
- Familiarity with text analysis tools for deriving document-level access points
- Familiarity with the metadata mining process and tools, including forensic tools and various kinds of metadata and file format registries
- For work with websites, familiarity with web crawlers and how they work

Questions for discussion:

- How can we get the profession past the idea that it is sufficient for a practicing archivist to take a weekend workshop in order to be able to function as a digital archivist? In other words, should archival educators be demanding that their students learn technical skills? It is a lot more likely that most archives will prefer to hire another archivist than that they will work to justify hiring a technologist, so archivists with good technology skills will be required in large numbers.
- While we wait for technologically-qualified archivists to emerge, are there ways that we can coopt computer scientists and information managers into the field?
- If archives are to support digital collections, many of them are sorely in need of better computer equipment. The new repository certification document mentions this but is not specific enough about technological requirements. Almost nobody talks about the necessity for dark mirrors and federated peer-to-peer collection sharing (i.e., multiple servers and storage sites) except at the national level. Should we come up with more detail along these lines?