

September 11, 2014

# It Takes a Library: Growing a Robust Institutional Repository in Two Years

Todd A. Bruns, *Eastern Illinois University*

Stacey Knight-Davis, *Eastern Illinois University*

Ellen Corrigan, *Eastern Illinois University*

Steve Brantley, *Eastern Illinois University*

## **It Takes a Library: Growing a Robust Institutional Repository in Two Years**

### **ABSTRACT**

In 2010, Booth Library began establishing an institutional repository, *The Keep*, an effort that involved multiple departments within the library. Potential content recruitment for the repository included large-scale digitization of archival materials and migration of previously created digital collections. Creation of the repository resulted in increased accessibility, better presentation of content that had existed on outmoded legacy web platforms, and the rescue of damaged content that had been disintegrating on other digital storage formats. By utilizing personnel across many departments and incorporating content from the Archives and Digital Collections areas, Booth Library has developed a robust institutional repository in only two years.

Keywords: institutional repositories, inter-departmental collaboration, digital preservation, legacy collection migration, organizational culture, new initiatives

### **INTRODUCTION**

Eastern Illinois University (EIU) is a mid-sized public comprehensive university established in 1895 and located in Charleston, Illinois. EIU is home to approximately 10,000 students and it offers bachelor's, master's, and specialist's degrees. Until the beginning of the 1990s, EIU did not have an official university

## It Takes a Library: Growing a Robust Repository

archives. Upon his arrival on campus, the new Dean of Library Services, Dr. Allen Lanham, sought to rectify this situation by establishing University Archives and Special Collections in Booth Library.

Nearly a decade later, the first digital collection was created at Booth Library with the establishment of a University Archives Photo Collection. This initial effort taught many lessons in project management and digital collection design that would inform subsequent ventures. The success of the photo collection also demonstrated a strong interest in and an eager audience for historical university documents.

Several years after the first digital collections were created, trends in scholarly communication, particularly the Open Access movement, led to the implementation of institutional repositories at several of EIU's sister institutions. Determined that EIU should not fall behind peer institutions, as had been the case with the absence of a university archive prior to his tenure, Dr. Lanham focused on two goals. The first was to establish an institutional repository for EIU, in keeping with similar efforts underway at peer institutions like Illinois Wesleyan University and Southern Illinois University, whose repositories were established to showcase faculty and student scholarship.

The second goal was to use the repository platform to support and contribute to the University Archives. In this way, EIU would position itself in the forefront of the trend in the digitization of historical documentation and also show that EIU's library understood that increasing numbers of documents-of-record and of historical interest are "born digital," blurring previously established lines that

## It Takes a Library: Growing a Robust Repository

separate a university archives collection from that of an institutional repository (Johnson 2002). *Please describe the lines that you have referenced.*

A common thread in early discussions about developing the institutional repository was to avoid playing catch up as had happened with the late establishment of University Archives. While many institutional repositories were established primarily to showcase an institution's scholarship (Shearer, 2002), EIU's repository would begin primarily as a platform for digitizing and hosting content from the University Archives. Although faculty and student scholarship would be an important element of the repository, the repository at EIU would focus on many different facets of university life, including archival records, historical events, student life, exhibits, and special collections.

### GETTING STARTED

In 2010, with the hiring of the Institutional Repository Librarian, Booth Library partnered with Berkeley Electronic Press (bepress) and began the design and implementation of the EIU institutional repository, *The Keep* (<http://thekeep.eiu.edu>). The selection of bepress' Digital Commons platform was based on several factors. By relying on Digital Commons for technology support, the library could focus on content. The stability of the Digital Commons platform meant that discoverability would remain high, as content can only be discovered when it resides on a platform that is available. The successful track record of the use of the Digital Commons platform by mid-sized academic libraries was a further factor (Lynch and Lippincott 2005).

## It Takes a Library: Growing a Robust Repository

The implementation of an institutional repository with a heavy emphasis on archival materials led to discussions about the commonalities and boundaries among the University Archives, the institutional repository, and the existing digital collections. It is essential and beneficial to have candid discussions about these overlapping roles in order to further the overall mission of the institution. At EIU it was determined that there was value in all three areas for discovery and preservation. For example, housing master's theses in the repository has led to higher levels of accessibility for Google and other search engine end users. Yet there is also evidence of users finding the theses via the different discovery stream of master's theses accessed via library catalogs (Bruns and Knight-Davis 2013). Maintaining the theses as print in University Archives and Special Collections represents another form of preservation. Indeed some collections exist in all three areas, providing multiple means of discovery for library patrons.

Large-scale digitization required laying a foundation of resources and personnel dedicated to supporting rapid scanning of massive amounts of documents. The first step in adding content to *The Keep* was to create the Booth Library Scanning Center.

### THE FIRST PROJECT – CREATING A SCANNING CENTER

The plan to include large amounts of materials from archives required that the material be digitized quickly, accurately, and with care. Previous digitization efforts had been small in scale, carried out by faculty and staff working independently at flatbed scanners in their individual work areas. The increased

## It Takes a Library: Growing a Robust Repository

scale of the digitization effort planned for *The Keep* required a large group of scanning operators working in shifts throughout the day. Additional equipment of higher scanning quality, capacity, and speed would also be required. This first project involved resources from Library Administration, personnel from the circulation department and Library Technology Services, and content from University Archives.

Library Administration allocated resources from a contingency fund for the purchase of three high-speed scanners and approximately seventy hours per week of student help. Additionally, personnel from Library Technology Services and the Circulation Department were reassigned to the scanning center staff. The result was the creation of the Booth Library Scanning Center, a new service point that would digitize all items for *The Keep*.

The establishment of the scanning center in [year] did not occur in a vacuum. Its creation was the result of several long-term developments in library practice, personnel, and space management, and it required the positive attitude and will to succeed from people across the library operation. A culture of shared responsibility and working for the mutual benefit of the organization is often an implicit assumption of library work. However, even if the tacit understanding of camaraderie exists, unit and task boundaries can become entrenched over time and the successful blending of roles cannot always be assumed. The history of successful multi- and cross-departmental projects at Booth Library demonstrates that this supportive and collaborative culture can succeed, and it is in no small part due to this organizational culture that *The Keep* owes its success. The proverb “It takes a

## It Takes a Library: Growing a Robust Repository

village to raise a child,” from which this paper adapts its title, provides an easy-to-grasp sense of group unity for the reader, but the prospect of duplicating a successful repository project in libraries with less unified or even fractious organizational cultures is not guaranteed.

At Booth Library, common purpose and collegiality allowed for reallocating space and staff. While years of stagnant materials budgets, increasing journal subscription prices, and the migration of journal content to electronic format created several years of cancellations of print journals, the resulting reduced need for serials processing and storage created the available space and personnel that could be put to excellent use in the Scanning Center. Supervision for the student scanning staff was provided from a .5 FTE civil service position from Circulation where the print journal cancellations had freed up staff time. A second staff member from Library Technology Services, already familiar with teaching students how to operate technical equipment, was also reassigned part-time to the scanning center. With facilities and adequate personnel in place, the Institutional Repository Librarian could oversee the rapid and efficient digitization of box after box of materials from archives. Early collections in *The Keep* included meeting minutes of faculty and staff governance committees, heavily used books and bulletins about the history of EIU, and historical press releases.

While the establishment of the Scanning Center ensured a rapid development of print-to-digital content, migrating many previously digitized special collections and “born digital” collections into the repository occurred simultaneously.

## EARLY DIGITAL COLLECTIONS – LESSONS LEARNED

To fully understand the success of *The Keep* and how a project of this scale could be completed so quickly at a modestly staffed library, it is important to step back and first consider the way earlier digital collections projects were processed and made available. These “baby steps” into digitization, taken prior to the establishment of the repository, taught lessons in creating procedures, establishing metadata standards, and implementing digital preservation processes.

The first digital collection at Booth Library was created in 2003 when photographs from the University Archives were scanned. In the first year over 700 photos were digitized. The images were hosted on the library website with metadata entered into a locally designed Microsoft Access database. The process of digitizing and adding photographs from the University Archives was developed organically in the spare time of an Archives faculty member. While an attempt was made to institute a controlled vocabulary for searching by subject, data were often entered by library staff with more knowledge of photography and institutional history rather than experience with metadata, cataloging, or digital collections. Public access to the collection was through a suite of locally designed ASP (Active Server Pages) Classic scripts. End users were able to scroll through menus listing names and subjects, but could not browse the entire collection. The project continued to add new content for several more years, thereby receiving high use both on and off the EIU campus.

The University Archives Photo Collection was digitized in order to provide a search system for Archives staff and a method for patrons to locate historical



## It Takes a Library: Growing a Robust Repository

images. No automated retrieval system or even a printed index existed prior to the Access/ASP system. Only paper finding aids and the institutional memory of the University Archivist were available to assist users in locating photographs relevant to EIU's history. Alumni, historians, and university departments were interested in the ability to search and retrieve these images.

The circumstances that enabled its creation depended on changes in technology and library facilities. A massive building renovation spanning the period 1999-2002 added workspace to the University Archives as well as network connectivity that had been lacking. These facilities improvements occurred at the same time that consumer grade flatbed scanners were undergoing technological advances and dropping in price. Digitizing key portions of the unique Archives photo collection became feasible. File name conventions based on the date of digitization were adopted, providing a unique identifier for each image. An MS Access database held image metadata. As the system was developed to assist Archives staff in locating photos, metadata included the names of buildings, places, and people, keywords, and the date the photo was taken if it could be determined. A web interface to the Access database was also developed to provide public access.

After the relatively successful launch of and positive response to the Photo Archive, a new digital project to digitize volumes of the EIU yearbook, *The Warbler*, was carried out in 2006. Funded by an Illinois State Library grant under the federal Library Services and Technology Act, the first eleven volumes of *The Warbler* (1919-1927), were digitized in a single year. As a condition of the grant, the completed project files were uploaded to the Illinois Digital Archive (IDA). Inclusion in the IDA

## It Takes a Library: Growing a Robust Repository

allows users to easily search for content and view the material in the larger context of Illinois history; consequently the material was for several years separate from the body of digitized EIU materials. *The Warbler* was the first digital project at the library to use a commercially produced platform (CONTENTdm) and also the first library project to use the Dublin Core metadata standard.

In 2008, the third phase of Booth Library digitization efforts began with the establishment of a digital collection hosted by the Consortium of Academic and Research Libraries in Illinois (CARLI). This collection, the theatre programs of the Little Theatre on the Square in Sullivan, Illinois, also utilized CONTENTdm and was supported at the consortium level.

It was during this phase of digital collection building that a more concerted effort was made to establish metadata protocols for improved harvestability of the content, as well as to document the process to guide future in-house projects.

In addition to these collections, Booth Library began scanning individual EIU master's theses in 2008 as a response to a recurring distance education class assignment. To address the needs of the class, a workflow was established whereby all new theses would be made available digitally. Upon depositing print copies of their theses with the library, authors submit a signed publication release form allowing the authors to retain copyright while granting Booth Library permission to make the theses publicly available in any format, including digitally. The document would be scanned and cataloged in the cataloging department, with the print format returned to the bindery for processing and shelving. The digital file would be stored

## It Takes a Library: Growing a Robust Repository

on Booth Library servers, and the catalog record made available through the EIU OPAC, I-Share (the CARLI shared catalog), and WorldCat.

The Booth Library Subcommittee on Digital Resources was formed in 2010 as a standing library committee to create and implement more consistent practices across ongoing and future projects, as well as to monitor and oversee related issues and procedures. During this time, production of Booth Library's digitized resources continued on an ad-hoc, cross-department basis.

### MIGRATING DIGITAL COLLECTIONS TO THE REPOSITORY

Early content recruitment for *The Keep* included massive digitization efforts of materials from University Archives and the migration of several of the previously described existing digital collections. Two of the digital collections, the master's thesis collection (<http://thekeep.eiu.edu/theses/>) and *The Warbler* collection (<http://thekeep.eiu.edu/Warbler/>), were essential collections to ingest into the repository given their importance as examples of student scholarship and creativity.

Migrating the master's theses into the repository served three purposes. First, moving theses to the repository provided *The Keep* with a jump start by adding born-digital content that increased repository numbers and visibility by hundreds of documents within its first month of production. This proved to be crucial for early efforts at faculty recruitment. It is a common complaint among institutional repositories that faculty are slow to contribute their work to a new, untested and frequently misunderstood initiative. Several studies have discussed faculty perceptions that contributing to an institutional repository requires work on

## It Takes a Library: Growing a Robust Repository

their part, such as researching the acceptable version to make available for open access and the creation metadata (Foster and Gibbons 2005; Kim 2007). In institutions where the repository infrastructure does not provide metadata or document ingest support for contributing scholars this is likely to be true. However, upon being presented with a well populated repository, EIU faculty responded positively, both to the repository as a viable platform – already having nearly 1,000 documents – and also to the significant download counts that were already being generated due to the popularity of the master’s theses.

The second purpose served by migrating the master’s theses into the repository was the greatly improved visibility and access to the collection. Although the theses had been digitized and made available beyond EIU via WorldCat, much greater exposure was now possible thanks to the added metadata fields, embedded social media sharing utilities and search engine optimization provided by the bepress platform. Additionally, bepress functionality in the EIU repository allows for the assertion of intellectual property rights by automatically assigning the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 license to each document in the repository.

It is essential to note that the original cataloging metadata proved invaluable in the ease of batch uploading these documents to *The Keep*, and the metadata provided additional access points for web crawlers. This is yet another instance where traditional librarian expertise was leveraged for the success of a leading edge technology project and were it not for the existing culture of collaboration, smooth

## It Takes a Library: Growing a Robust Repository

migration of the theses might not have occurred, potentially adversely affecting faculty buy-in.

An excellent example of the increase in accessibility can be seen in the download statistics of *Improvements to Vertical Axis Wind Turbine Blades to Aid in Self-starting*, a Master of Science thesis in the EIU School of Technology. In October 2011, under EIU and WorldCat catalog access, the thesis was downloaded thirty-five times, thirteen of those during a classroom demo. By comparison, in October 2012 the thesis was downloaded 729 times from the institutional repository. As of the date of this writing, the thesis has been downloaded well over 3,500 times. The success of the master's thesis collection in *The Keep* has encouraged deposits of undergraduate honors theses as well, the visibility of which helps demonstrate the centrality of undergraduate education, scholarship, and research at EIU.

The third purpose served by the master's thesis collection was the creation of second copies of the theses on servers beyond the EIU campus. The potential for catastrophic server losses at EIU now no longer threaten the entire digital collection of master's theses, as the same collection is cloned and kept on Berkeley Electronic Press servers.

Unlike other content in *The Keep*, most of the processing and metadata management for master's theses occurs in the library's Cataloging Services unit. MARC bibliographic records in OCLC WorldCat are created for the theses, with URLs linked to the digital file hosted on the library server. Minimal-level data is provided following the AACR2R (Anglo-American Cataloging Rules, Second Edition (Revised)) standard, plus Library of Congress Subject Headings are added with a variant

## It Takes a Library: Growing a Robust Repository

Library of Congress Classification call number assigned to the theses. Once produced, these records are imported into the local OPAC and then exported to the I-Share consortium catalog by means of a regular automated batch process. Once a month the data from the new thesis bibliographic records are transferred to *The Keep's* metadata spreadsheet and batch uploaded.

The effects of migration were similar for the first eleven volumes of the student yearbook, *The Warbler*. The files for these early yearbooks were copied to *The Keep*, allowing them to be searched for, discovered, and used in conjunction with the thousands of other documents related to the history of EIU. All 94 volumes of *The Warbler* have now been digitized, described, and added to the institutional repository.

The benefits of moving *The Warbler* collection to the repository were similar to those of moving the master's theses. In the repository, *The Warbler* collection benefited from increased visibility and access, social media utilities, the Creative Commons license, and preservation via multiple access copies. EIU has a strong alumni network and adding a high-profile collection like *The Warbler* to the repository serves to increase the gravitas of *The Keep* as a viable archive of historical documents of importance to the institution.

Cloning the master's thesis and *The Warbler* collections to the repository was relatively easy and represented early successes. Migrating other digital collection efforts into *The Keep* proved to be more difficult. Whereas *The Warbler* and master's thesis collections were already in stable web platforms, other collections existed on platforms that are prone to data loss and built on outdated coding or stored in

formats not intended for long term preservation. Some of these were older collections digitized by the library, while others were collections digitized by other university units and adopted by the library for preservation and access. Migrating these collections away from obsolete technology was essential to the continued existence of valuable local collections.

### **MIGRATING FROM UNSTABLE SOFTWARE PLATFORMS**

While not anticipated or planned for, a serendipitous benefit of the repository was that it provided a platform for migrating content that existed on unwieldy and difficult to update legacy web applications. After implementing the repository, it became obvious that two existing projects would receive a much needed upgrade by migrating to the new platform. Both the University Archives Photo Collection and the *Journal of Collective Bargaining in the Academy (JCBA)* required immediate migration.

These collections were built using Microsoft Access and ASP Classic. As early pilot projects, these applications were intended to provide a test bed (would platform be appropriate?) to gauge interest in library digital content. By hosting *JCBA*, Booth Library was taking the lead as an early adopter of library-as-publisher. While successful in some ways, the projects also exposed the limits of the underlying technology. While MS Access and ASP are sufficient for small projects of limited interest, they are not designed to support the large numbers of users that these projects were receiving. The eventual end of support for ASP Classic and the increasing difficulty of keeping old applications running on new servers was a

## It Takes a Library: Growing a Robust Repository

strong incentive to migrate the content of both collections to the new repository platform.

Migrating the Photo Archive ([http://thekeep.eiu.edu/archives\\_images/](http://thekeep.eiu.edu/archives_images/)) to the institutional repository required generating a content report from the current database and transforming that data into a spreadsheet for uploading into the repository. This process allowed staff to check and correct errors in the original data, a process which resulted in a stronger and more accurate collection. In addition to the increased visibility, access, and preservation via multiple access copies—functionality provided to all repository objects, document or image—the Photo Archive was far easier to browse in *The Keep*, than it had been in its original MS Access web interface. In *The Keep*, photos are arranged topically and can be displayed as a slide show or a thumbnail gallery. Another advantage of Image Gallery collections in *The Keep* are embedded social media sharing utilities making it easy for users to share the content, further increasing its visibility.

The second legacy collection existing on an unstable platform was the *Journal of Collective Bargaining in the Academy* (<http://thekeep.eiu.edu/jcba/>). Booth Library was an early adopter of the recent trend among libraries to serve as electronic journal publishers by hosting the journal and the annual conference proceedings of the National Center for the Study of Collective Bargaining in Higher Education and the Professions. Although the website provided some visibility for *JCBA*, overall use and discoverability were not optimal. Maintaining the site required a fair level of technical skill, and *JCBA* editors were dependent on library staff for



## It Takes a Library: Growing a Robust Repository

adding and editing content. Journal editors and reviewers used email to exchange manuscripts and accept submissions.

Migrating *JCBA* and the annual conference proceedings proved to be a painstaking process of collecting the full text of varied content (articles, presentations, handouts, manuscripts, and schedules) from server files and matching them up to the correct author(s). In the case of the proceedings, this included replicating the conference schedule while correctly matching materials to presentations.

Hosting *JCBA* in *The Keep* opened up an entirely new section of the repository. Under the current bepress agreement, *The Keep* is able to host five e-journals, with each journal able to have its own interface design and with built-in editor functions that make editing the journal a more streamlined process. These features provided functionality that wasn't possible with the original platform, adding value for the journal editors. Similarly, search engine optimization via bepress provided article authors with greater visibility and increased downloads, which in turn increases an author's scholarly impact. Based on reports from Google Analytics and *The Keep* statistical reports, downloads increased by a factor of ten, from 336 in 2011 (under the old MS Access/ASP platform) to 3,309 in 2012 when hosted in *The Keep*.

Utilizing the repository as a replacement for the ASP-coded website introduced new functionality for the editors and a much more stable platform for the journal. Accessibility and preservation were both improved due to the migration. By hosting the journal, *The Keep* provided EIU with marketing and

fundraising opportunities since it could showcase the journal and the repository as examples of EIU's partnering with the broader community to advance scholarship.

The materials collected for *JCBA* were all born digital, thus the main challenge was in collecting and correctly identifying the files. In other cases, collections migrated to the repository came from stored media where bit-rot had already caused damage.

### MIGRATION FROM DAMAGED, DISPARATE SOURCES

A serious challenge to digital library collections is the rapid loss of data and knowledge due to faulty formats and bit-rot. People from all walks of life have dealt with the fading of magnetized images on VCR tape, unreadable floppy discs, and the loss of data on CD-R, not to mention the obsolescence of formerly "hot" hardware. Laser discs are now nearly as quaint as 78-RPM vinyl records.

*The Keep* is intended to be the digital archives of the university,, as such, materials collections coming out of departments are just as important as the collected scholarly works of faculty and students. A major project of *The Keep's* first year was the creation of the EIU Department of Theatre Art's theatrical productions collection, a collection that unfortunately had suffered data loss due to format obsolescence.

The theatre arts department maintains a webpage called "Memory Lane" that lists theatrical productions going back to the 1940s. While the page includes some photos, the collection is largely incomplete, lacking programs, reviews, and additional photos. Photographs and other materials associated with the productions

## It Takes a Library: Growing a Robust Repository

that could enhance and complete this valuable collection existed in scattered locations and formats. Some were on servers in Booth Library, some on the hard drive of the EIU photographer's desktop computer, many were in print in file folders in the theatre arts department file cabinets and still more were stored as images in drawers full of CD-Rs.

### Creating the Theatre Arts Productions collection

([http://thekeep.eiu.edu/theatre\\_productions/](http://thekeep.eiu.edu/theatre_productions/)) in *The Keep* was an effective way to compile materials from these disparate sources. Print materials were digitized in the Scanning Center and uploaded to Booth Library servers for ingestion into the repository. Image files on other parts of Booth Library servers and image files on the photographer's workspace hard drive were easily migrated to *The Keep*. Most of this part of the operation was simply a matter of finding the files and getting access. The difficulty was in the files stored in CD-R format.

Although the theatre arts department had only been storing images on CD-R for a little over ten years, there was already significant data loss. Some of the discs could not be read at all. Some required older equipment in order for the images to be retrieved. Out of 150 CDs, half would not open on current computers and required use of an older CD reader. Three CDs were completely irretrievable.

The new Theatre Arts Productions collection in *The Keep* combines photographs, programs, newsletters, and posters, compiled from multiple sources, into collections of documents and images organized by theatrical production. As with other collections in the repository, the items are more accessible than before and are far more secure in terms of preservation. Images that previously resided only

## It Takes a Library: Growing a Robust Repository

on a desktop computer can now be discovered via any major search engine from anywhere in the world. Photographs that were slowly decaying in a file drawer on CD-R are now preserved via an access copy in *The Keep* as well as in master file TIFF format on library servers.

Preservation of the master files remains a challenge. While *The Keep* provides preservation of access copies (and introduces, in a very limited sense, the philosophy of Lots Of Copies Keeps Stuff Safe), the best practices for digital preservation of master files requires data-redundancy within a storage array, and off-site backups.

## MAINTAINING DIGITAL MASTERS

Up to this point, we have discussed only web-ready JPEG and PDF files associated with the digital collections that were incorporated into *The Keep*. In addition to updating and improving user interfaces and discoverability, the challenge of finding a way to store the TIFF master images generated by previous and ongoing library digitization projects had to be addressed. Master TIFF files were saved during the scanning process for each image in *The Warbler* collection, The University Archives Photo Collection, and all other archival materials scanned for *The Keep*.

Digitization of *The Warbler* was carried out following the Illinois Digital Archives Guidelines for Images, which specify resolution and file format. Collections created after *The Warbler* followed the standards of the Booth Library Digital Collections Manual produced by the Booth Library Subcommittee on Digital

## It Takes a Library: Growing a Robust Repository

Resources (2010). The Booth Library Manual is in compliance with the “Guidelines for the Creation of Digital Collections: Digitization Best Practices for Images” and “Guidelines for the Creation of Digital Collections: Digitization Best Practices for Text” developed by the Consortium of Academic and Research Libraries in Illinois (CARLI) Digital Collections Users’ Group (CARLI 2013a, 2013b).

The Booth Library Digital Collections Manual specifies file formats for use copies and digital master images. Following these guidelines from the outset of digitization projects results in data normalization practices being in place as the images are produced. In compliance with the CARLI Guidelines for the Creation of Digital Collections (2013a, 2013b), digital masters are saved in TIFF format. TIFF provides very good support for high-resolution images and the format continues to be one of the preferred formats for bitmapped images.

For the first few years of the photo archive project, as space on library servers was consumed, the large TIFF files were burned to CD-R in the hope that affordable and stable long-term storage options would become available in the future. Storing large master files on CD was not an uncommon practice, and major institutions such as the British Library also stored TIFF images on CD (NDIIPP 2009). However, the drawbacks of this storage system quickly became apparent. Retrieving the TIFF images from disc was cumbersome when these files were needed to fill patron requests for print-quality images. Re-burning the discs periodically to mitigate decay and manually checking the discs for integrity were time-consuming tasks that quickly proved burdensome to add to existing staff workloads. While the Optical Storage Technology Association estimated the life span

## It Takes a Library: Growing a Robust Repository

of recordable compact discs at fifty to one hundred years (SCAPE 2012), it was also noted that the life span of a recorded disc would be extremely difficult to estimate reliably (Bennett 2003). As was seen in the CD-Rs used by theatre arts, the life span for CD-Rs proved to be much shorter than fifty years. Even under controlled conditions in University Archives, damage has been discovered on some of the discs used for photo archive masters. Images display with discolored areas or cannot be opened at all. Fortunately, the photographs from the early years of the photo archive project are still available in the archives, and these images can be re-digitized when errors are discovered. However, this is an inefficient and potentially burdensome practice to adopt as policy, similar to using CDs for master files.

At the beginning of *The Warbler* project, digital masters were maintained on a server instead of a disc. Allocating space on existing file servers allowed for automatic error detection within files; however, digital projects require scalable space solutions. Additional drives require a system for distributing data across the disks. This management and distribution are accomplished with a RAID (Redundant Array of Independent Discs) array. RAID refers to a storage system composed of multiple disks with data copied and divided over several disks. By design, a RAID array holds at least two copies of each file.

Booth Library purchased an Apple Xserve RAID in 2007 expressly for storing digital master images. The RAID was in service until 2012 when it was replaced by an HP array system. Storage arrays do not eliminate the possibility of data loss, but they do greatly decrease the odds of non-recoverable damage. The manufacturer specified mean time before failure for individual drives in the system is 1,200,000

hours, or 136 years. This far exceeds the expected life span of other hardware components in the system. Taking these other variables into account, the calculated probability of non-recoverable data loss in seven years is 0.1342% (Whitehead 2009). Certainly this probability partnered with the security of continual error checking and multiple copies are acceptable odds and far superior to the CD-R storage.

For the Booth Library system, two copies of the master files are spread across twenty-two disks. The RAID controller continually checks the data for errors. On a weekly basis, empty space in the system is checked for errors and if bad sectors are discovered on the disk they are excluded from future use. For further security, files are backed up to a campus location outside the library.

Having multiple copies in multiple locations follows the now widely accepted digital preservation strategy of data redundancy. However this security comes with added costs. In his discussion of the practical limits of digital preservation, Kastellec states that “multiple copies necessarily increases expenses, therefore—given equal levels of funding— less information can be preserved redundantly than can be preserved without such measures” (2012, page). Thus, factors of cost, security, and the amount of data to be stored must be balanced. Cost has become the overriding factor limiting the growth of digital collections at Booth Library. Maintaining multiple copies and off-site backup requires not only equipment funds, but also skilled IT support. As the amount of data requiring long-term storage increases, the complexity of the hardware required to do so also increases.

One cost containment strategy is to limit the amount of material to be stored. The Booth Library Digital Collections Manual specifies rigorous evaluation procedures that are to be carried out before beginning a digitization project. The significance, audience, uniqueness, and the scope and size of the collection to be digitized must be established before work begins. All legal and technical issues, including storage costs, must also be considered. These steps, taken before any new digital collection is begun, allow us to select only those collections that will be of value to users and of sufficient worth to justify the resources allocated to them.

Another cost containment strategy is collaboration. Networks currently exist for the shared storage of use copies. Stanford University's LOCKSS (2013) network offers private networks for maintaining content in a distributed environment. While this system works very well for preserving web-ready content, the massive amount of data involved with digital masters comes with a very high cost in bandwidth usage to store in a network system. Persuading stakeholders to invest heavily in equipment that will be used by non-stakeholders, even if it is in the best interest of all parties, can also be problematic.

As new digital collections move into production, the demand for digital data only grows as faculty contributions increase and the use of existing collections expands. While advances in technology continually improve the options for data security, funding support tends to remain stagnant. Rosenthal (2010) aptly describes the problem: "Like almost all engineering problems, bit preservation is fundamentally a question of budgets." Working within this constraint, we strive to provide good security and room for growth.



## FUTURE PROJECTS

Like the Theatre Arts Productions collection, the current digital collection of the Little Theatre on the Square will also be expanded and migrated into the repository. Under the guidance of David Bell, Film and Theatre Librarian at Booth Library, programs, photographs, and newsletters important to the history of this community landmark are being collected, digitized, and organized for presentation in *The Keep*. Another important future project is the digitization of the EIU campus newspaper, the *Daily Eastern News*.

In 1998 Booth Library accepted hosting responsibilities for the campus newspaper, *The Daily Eastern News (DEN)*. This Journalism department publication consisted of PDF files beginning with 1993 editions. When the collection first launched, retrieval was accessible only by date and page number. In 2005, keyword searching of the content was achieved with the application of Google Mini Search. Keyword searching was a vast improvement in usability over date and page numbers. Yet, navigation issues remain due to the file structure. With each newspaper page a separate PDF file, articles spanning more than one page are difficult to use. Users searching historical issues of the *DEN* are interested in retrieving everything on an event or person. There is currently no article level or subject indexing for the student newspaper nor can it be searched in conjunction with any other sources.

The *DEN* archive, approximately fifteen GB of PDF files, receives about 300 hits per month. As the record of daily happenings at the University, the *DEN* is of great interest to students, faculty, alumni, and community residents. Due to its value

## It Takes a Library: Growing a Robust Repository

to the community, and the usability issues noted above, the *DEN* will be the next collection migrated to *The Keep*.

This will require some data massaging, as this important collection was handed off to the library staff in a fragmented condition. Some of the issues were in PDF, some in HTML, some were one file, and some issues were multiple files. Preserving the content in a consistent and stable form is essential since the newspaper is the document of record for historical events at the university.

The plan for the *DEN* collection is to collect disparate files and documents into a unified master file format, preserving masters in the RAID array, and present the collection via *The Keep*. Prior to ingestion into *The Keep*, the content will undergo Adobe full text recognition processing, allowing scholars to search at an article level without having to browse full issues. Google Analytics reports for the *DEN* repository page demonstrate that it already receives significant traffic, even though there is as yet no content. This is a clear indicator of strong community interest in the collection, making the *DEN* migration a priority.

## FROM PAST TO PRESENT

Launching *The Keep* with a strong foundation of archival documents and migrated digital collections ensured that from the beginning the EIU repository was growing and providing strong download counts. These qualities made *The Keep* attractive to faculty. High use collections like the master's theses and the *Journal of Collective Bargaining in the Academy* demonstrated to faculty that *The Keep* would offer their work greater visibility.

## It Takes a Library: Growing a Robust Repository

*The Keep* now showcases the work of over ninety-five current EIU faculty members representing every college in the university. Publications such as the history department's *Historia* have gained a larger audience with their inclusion, allowing the department's work to be communicated with scholars around the world. In addition to increased visibility to Internet search engines, content in *The Keep* is also discoverable through the Digital Commons Network. The Network provides the ability to search the repositories of over 300 institutions, including many high prestige universities, and hierarchical browsing by discipline. Through browsing, scholars can easily connect with other researchers in their field through the "popular institutions" and "popular authors" links, expanding their scholarly networks and increasing the potential for fruitful collaboration across geographical and disciplinary lines. EIU scholars are frequently featured as popular authors, providing further validation for *The Keep*, the Digital Commons Networks, and Open Access scholarly communication.

## CONCLUSION

For EIU, the implementation of an institutional repository had more to do with bringing a 20<sup>th</sup> century, paper-based University Archives into the 21<sup>st</sup> century and providing the library with a foundation for digital preservation than it had to do with collecting scholarly output. However, the archives content has served to stoke rather than stall faculty participation. In 2012, its first full year of operation, *The Keep* grew in content by over 10,000 documents, compared to the average first-year growth among repositories of 731 documents, as reported in a personal

## It Takes a Library: Growing a Robust Repository

communication from a bepress representative. The repository provided a significantly improved platform in terms of stability and preservation for a number of important collections at EIU, and, in some cases, provided a timely migration of data and content from failing digital storage environments.

The utilization of *The Keep* as a digital archive will continue into the foreseeable future, along with continued growth in faculty contribution and digital publishing initiatives. As noted by Mahyar Izadi, Dean of the College of Business, the establishment of *The Keep* represents an accomplishment no less significant than “building a second [digital] library for campus.” With dedication and attention to the continuous improvements in digital preservation, the EIU digital archives will grow and be preserved well into the future.

## REFERENCES

- Bennett, H. 2003. *Understanding CD-R & CD-RW*. Optical Storage Technology Association. Retrieved from [http://www.osta.org/technology/pdf/cdr\\_cdrw.pdf](http://www.osta.org/technology/pdf/cdr_cdrw.pdf).
- Booth Library Subcommittee on Digital Resources, Eastern Illinois University. (2010). *Digital collections manual*. Version 1.0. [http://works.bepress.com/ellen\\_corrigan/9/](http://works.bepress.com/ellen_corrigan/9/).
- Bruns, T., and Knight-Davis, S. 2013. Increasing the Visibility and Impact of Graduate Research with Electronic Theses and Dissertations. Webinar. Berkeley Electronic Press. <http://digitalcommons.bepress.com/webinars/46/>.
- CARLI (Consortium of Academic and Research Libraries in Illinois). Digital Collections Users' Group. (2013a). *Guidelines for the creation of digital collections: Digitization best practices for images*. Retrieved from [http://www.carli.illinois.edu/sites/files/digital\\_collections/documentation/guidelines\\_for\\_images.pdf](http://www.carli.illinois.edu/sites/files/digital_collections/documentation/guidelines_for_images.pdf)
- CARLI (Consortium of Academic and Research Libraries in Illinois). Digital Collections Users' Group. (2013b). *Guidelines for the creation of digital collections: Digitization best practices for text*. Retrieved from [http://www.carli.illinois.edu/sites/files/digital\\_collections/documentation/guidelines\\_for\\_text.pdf](http://www.carli.illinois.edu/sites/files/digital_collections/documentation/guidelines_for_text.pdf)
- Foster, N. F., and Gibbons, S. (2005). Understanding Faculty to Improve Content Recruitment for Institutional Repositories. *D-Lib Magazine*, 11, 1 <http://www.dlib.org/dlib/january05/foster/01foster.html>

## It Takes a Library: Growing a Robust Repository

- Johnson, R. K. 2002. Institutional repositories: Partnering with faculty to enhance scholarly communication. *D-Lib Magazine*, month  
<http://www.dlib.org/dlib/november02/johnson/11johnson.html>.
- Kastellec, M. 2012. Practical limits to the scope of digital preservation. *Information Technology and Libraries* 31: 63-71.
- Kim, J. (2007). Motivating and Impeding Factors Affecting Faculty Contribution to Institutional Repositories. *Journal Of Digital Information*, 8(2). Retrieved from <http://journals.tdl.org/jodi/index.php/jodi/article/view/193/177>
- Lynch, C. A., & Lippincott, J. K. (2005). Institutional repository deployment in the United States as of early 2005. *D-Lib Magazine*, 11(9). Retrieved from <http://www.dlib.org/dlib/september05/lynch/09lynch.html>
- NDIIPP (National Digital Information Infrastructure and Preservation Program). (2009). TIFF, revision 6.0. *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Retrieved from <http://www.digitalpreservation.gov/formats/fdd/fdd000022.shtml>
- Rosenthal, D. S. H. (2010). Bit preservation: A solved problem? *International Journal of Digital Curation*, 5(1), 134-148.
- SCAPE (Scalable Preservation Environments). (2012). *IS10 Potential bit rot in image files that were stored on CD*. Retrieved from <http://wiki.opf-labs.org>
- Shearer, M. K. (2002). Institutional repositories: Towards the identification of critical success factors. *Canadian Journal of Information and Library Science*, 27(3), 89-108.
- Stanford University. (2013). "Global & Private LOCKSS Networks." Accessed February 1. Retrieved from <http://www.lockss.org/community/networks/>
- Whitehead, J. (2009, June 10). Calculating mean time to data loss (and probability of silent data corruption). *Zetta Saclabytes Blog* [blog]. Retrieved from <http://www.zetta.net/blog/calculating-mean-time-to-data-loss-and-probability-of-silent-data-corruption>