MODULE 15

COLLECTING DIGITAL MANUSCRIPTS AND ARCHIVES

Megan Barnard and Gabriela Redwine

SAA

SOCIETY OF
American
Archivists

Case Study 3: Stanford University Archives

By Josh Schneider, Assistant University Archivist, and Daniel Hartwig, University Archivist, Stanford University

Introduction

Stanford University Archives has adopted many strategies to support the appraisal and acquisition of contemporary university records bearing enduring historical value. These include approaches to facilitate review of born-digital files for potentially sensitive, restricted, or legally protected information, both before and after acquisition. Several of these strategies and approaches are discussed in this case study.

Background

Established in 1965, and today forming part of Stanford University Libraries, the Stanford University Archives collects materials in any format that document the Stanford University campus, founders, administration, faculty, teaching, learning, and research, as well as campus events and student life.

To ensure the availability of a broad spectrum of materials for research and scholarship, and in keeping with a robust institutional collection development policy, the archives has adopted an aggressive outreach strategy encompassing a wide variety of campus record creators. The archives provides guidance on its website for campus units regarding the scope of our collecting, as well as the various means available to transfer materials to us.[13] We also meet regularly with campus units on both a programmatic and ad hoc basis to educate and train staff on the role of the archives. These meetings often afford us the opportunity to appraise campus records on an iterative basis.

The diversity of content generated by the university includes materials that may fall into the category of sensitive, restricted, or legally protected materials. The archives must satisfy donor agreements surrounding these issues and can face legal exposure for materials inadvertently or inappropriately acquired, maintained, or shared. In addition, the university has undertaken a comprehensive security overhaul of its entire computing infrastructure, including formulation of strict

---

13 Stanford University Libraries, Special Collections and University Archives, "For Staff," http://library.stanford.edu/spc/university-archives/transferring-your-records/staff, captured at https://perma.cc/E929-QFVJ.

protocols for dealing with high-risk content.[14] For these reasons, we are especially cognizant of ensuring that we meet our custodial obligations with regard to this content.

The following three examples underscore the archives' need for a flexible and scalable approach to accommodating sensitive, restricted, or protected information in appraisal and acquisition.

Example 1: Student Group Policy Documents, Office of Student Activities and Leadership

Stanford University's Office of Student Activities and Leadership (SAL) sets policies for over 700 active student organizations, overseeing their operations and management. In fulfilling this function, SAL maintains these organizations' original and updated policy documents, which they began to accept electronically in 2012. The transfer of SAL's electronic records, including the records of student organizations, to the University Archives commenced in 2014. The process by which the records were appraised, captured, and accessioned provides a strong illustration of how preacquisition appraisal and ongoing relationship building can benefit a collecting program.

While SAL transfers materials to the University Archives on a regular basis, the 2014 transfer represented the first time these policy documents were transferred using CardinalSync, a Web-based platform for student organizations that provides a membership directory, schedules, and messaging service. Prior to transfer, we worked closely with SAL to identify student information restricted under campus guidelines and the Family Educational Rights and Privacy Act (FERPA) to ensure that this information was not exported from CardinalSync for transfer to the archives. Working with SAL, we were able to export updated student group policy documents to a shared network folder. We then accessioned these into the Stanford Digital Repository. Because a student's membership in a particular student organization can be a sensitive matter, access to the digital and paper records requires the explicit permission of SAL.

---

14 Stanford's Information Security Office (ISO) defines high-risk content as health information, including protected health information (PHI); health insurance policy numbers; Social Security numbers; credit card numbers; financial account numbers; export controlled information under U.S. laws; driver's license numbers; passport and visa numbers; and donor contact information and nonpublic gift information.

Building this ongoing relationship offered additional benefits beyond ensuring these important materials were provided to us in a timely fashion. After collaborating with SAL on this initial transfer, they agreed to grant the University Archives administrator privileges with CardinalSync, allowing us to export new and updated policy documents on our own, as well as generate and download reports, including CSV files containing descriptive information on all of the groups and schedules of SAL-sanctioned events. CardinalSync's indexing of student contact information has also simplified our ability to contact student group leadership to solicit other types of materials and identify URLs for student group websites for inclusion in Stanford's instance of Archive-It, as well as request permission to archive Web-based student group materials not hosted on the Stanford domain.

Example 2: Campus Maps and Architectural Drawings, Office of Land, Buildings, and Real Estate

The upcoming acquisition of 140,000 campus maps and architectural records from Stanford's Office of Land, Buildings, and Real Estate (LBRE) provides an additional example of how preacquisition appraisal and relationship building has benefited the University Archives. The collection comprises both paper holdings and related scans, as well as extensive quantities of born-digital files, including those in specialized architecture and design formats such as CAD and BIM files, as well as Lidar-generated 3-D campus scans and historical campus base maps.

In the past, LBRE has resisted transferring born-digital materials to the University Archives due to security concerns around providing access to certain classes of drawings, specifically mechanical and electrical plans. We have been able to overcome this hesitation through several means, building on an already strong relationship with a University Archives champion within LBRE. This individual has a specific interest in and responsibility for the preservation of historic campus buildings. In addition, we have proven our value by undertaking a high-profile internal digitization project with LBRE funding, a project that is currently making its way through the University Library. These factors have enabled us to position the University Archives as a trustworthy and digitally savvy partner. LBRE now views us as having the digital infrastructure in place to protect Stanford University's

security interests while still ensuring the digital preservation of the transferred materials and enabling public access to approved drawings.

We have performed extensive preacquisition appraisal of this collection to assist us in determining how this project will fit with current and upcoming programmatic needs, such as physical and digital storage and accessioning workflow priorities. LBRE currently provides mediated access to its digital files to authenticated users through its own Web-based database. We are currently in the process of reviewing its internal metadata to begin planning for the accessioning of the files and related metadata into the Stanford Digital Repository.

Example 3: Email Correspondence, Knowledge Systems Laboratory

The email correspondence of campus administration presents great potential to support scholarly research into the history of the university. The University Archives recently appraised the email of Richard Fikes, professor emeritus in the Department of Computer Science and director of the Knowledge Systems Lab (KSL) from 1999 to 2006. The correspondence includes discussion of administrative matters falling under Fikes's purview as laboratory director.

Our appraisal of this email was greatly facilitated through use of ePADD, freely downloadable open source software developed by Stanford University Libraries and partners. ePADD harnesses machine learning, natural language processing, automated metadata extraction, and other batch processes to address challenges of appraising email archives. For example, it allows archivists to screen email for messages containing potentially sensitive, confidential, and legally protected information that might preclude transfer or the provision of access.[15]

The ePADD software package includes an appraisal module a creator can use to take advantage of many of ePADD's appraisal functions prior to transferring custody of the email to a repository. Although in this case the creator did not request use of this module, he did provide the archives with a list of several hundred email addresses to restrict, prompting the ePADD development team to build new functionality to support taking bulk actions on a list of correspondents imported as a CSV file.

---

15   For more information, see https://library.stanford.edu/projects/epadd, captured at https://perma.cc/4DEB-L6HB.

Once the email was transferred, we were able to make extensive use of three core features included in ePADD's processing module in reviewing the email for sensitive materials: (a) a regular expression search, which helped flag credit card and Social Security numbers; (b) a lexicon, which supports tiered thematic searching of the email archive for terms and phrases associated with sensitive materials such as "tenure" and "performance review"; and (c) a custom-named entity recognizer (NER) that also supports extraction and browsing of disease names bootstrapped from *Wikipedia*, which improved our ability to screen the email for protected health information.

Conclusion

These three examples illustrate the need for adopting a broad range of approaches when appraising and acquiring born-digital files at a contemporary research university. They underscore the value of building strong relationships with campus records creators, as well as the need for flexibility and patience to accommodate their concerns. They also highlight the potential of preacquisition appraisal, both to promote planning for the scale of projects, as well as to help determine the best means of screening for potentially sensitive, restricted, or legally protected information.

In the future, we expect to take further steps to integrate the archives into the university's digital ecology, including continuing to take advantage of network and cloud-based file-transfer services that build upon campus infrastructure to minimize the learning curve and effort required for campus units to transfer materials. We also expect to take further advantage of software and services that support regular expression search, natural language processing, and related methods to aid in the batch appraisal of born-digital files.