Ctrl+Alt+Archive: Navigating Born-Digital University Records and Publications

Laurel McPhee, Tori Maches, and Marlayna Christensen

ABSTRACT

Over the last few years, a new approach to collecting born-digital university archives content in Special Collections and Archives at UC San Diego Library has evolved as archivists incorporate flexible and forward-thinking accessioning practices. In the early 2000s, breaks in professional library staffing, the rise of digital records output, and a dramatic increase in the size and scope of the campus all contributed to significant gaps in the university archives. Archivists knew they had to rectify workflows and fill years-long gaps in the collections. The archivist and digital archivist now proactively collect born-digital records and publications by engaging directly with creators, using innovative techniques to capture files. This includes comprehensive planning, troubleshooting, and processing while accessioning. The authors present a scoped literature review and some examples of accessioning digital university records and publications that led to public access.

KEY WORDS

Accessioning, Born-digital records, Collaborations, Digital collections, Iterative processing, Podcasts, University archives, Website content, Workflows

In 1963, the University of California president, Clark Kerr, established a system-wide records management program, recommending a university archives (UA) and an archivist for each of the seven campuses to collect and preserve historically significant documents. A records schedule identified records for archival retention, defining "files, records, or documents, including but not limited to correspondence, reports, writings, and other papers, records, maps, tapes, photographic films and prints, magnetic and punched cards, discs and drums," as the property of the university.[1] Since the 1980s, archivists in the Special Collections and Archives program (SC&A) at UC San Diego traditionally focused on collecting paper records from a small core group of campus leadership offices. However, by the early 2000s, breaks in professional library staffing, the rise of digital records output, and a dramatic increase in the size and scope of the campus all contributed to the development of significant gaps in the university archives. For example, when a key shared governance committee switched from paper to digital agendas and file management in the late 1990s, accessions from that committee simply came to a halt. The archives lacked staff trained to manage born-digital materials, which hampered archival appraisal and acquisitions efforts across campus. Consequently, over the years, relationships between UA and campus constituencies weakened, and large coverage gaps emerged in the archives. While library staff began capturing campus websites in 2007, other digital records and publications were sorely neglected.

The library hired its current university archivist in 2016 and digital archivist in 2018, and they have developed a new approach to acquiring and accessioning born-digital university records. In this article, we present examples of how accessioning born-digital content while actively planning for researchers' use helps archivists address acquisition challenges and the life cycle of digital collections care and informs decisions for tasks such as file migration and metadata extraction. In this introduction, we briefly describe the landscape of digital records management infrastructure in place at the library. We present a scoped literature review that highlights some of the conversations over the past dozen years surrounding born-digital records that frame our approach. Then, we describe two themed categories of projects in which UA has pivoted from paper to born-digital, met creators where they are, and designed new workflows to capture records and make them accessible in a timely fashion.

We are fortunate to have a strong foundation of technical infrastructure for digital collections. Over the last fifteen years, the UC San Diego Library has digitized primary source materials and ingested born-digital content into our local Digital Asset Management System (DAMS),[2] building high-quality, accessible online collections. The DAMS content is freely available on the library's Digital Collections[3] website, a public search and discovery interface that currently features over 400,000 digital objects. Content from the site is regularly harvested and made discoverable on other platforms, including Calisphere (a California Digital Library

product) and the Digital Public Library of America. Digital content includes images, text, audiovisual recordings, and data. For researchers and our community, online access to University Archives collections improves discovery and creates potential for new analysis and use of the materials. Sensitive or restricted records are displayed in a metadata-only view and can be made available upon request. Currently, SC&A has approximately two terabytes of digital content available online, and annual intake of born-digital records varies greatly in the number of files and total volume.

Digital collections and their preservation require a highly collaborative, cross-programmatic approach at the UC San Diego Library. Content is proposed by both collection curators, such as the university archivist, and format specialists. Metadata Services provides consultation, mapping, and enhancement of metadata for more efficient discovery and management. Technology and Digital Experience (TDX) develops and manages the systems for digital preservation and access. Collectively, these programs ensure the creation, management, delivery, and preservation of digital assets in support of the university's mission and goals. Preparing materials for long-term preservation and discovery in our DAMS requires extensive work from a designated archives project manager in SC&A. The project manager coordinates the digitization (or, as required, normalization or migration of born-digital files), description in a metadata spreadsheet, and ingest of objects. Project documentation and progress are tracked using Trello and Jira in Confluence, the library's internal wiki. Finally, after the digital objects are saved in the DAMS, they are backed up in UC San Diego's preservation repository, Chronopolis (as of the writing of this article, this tool is in transition as we plan migration to a new platform).

Literature Review

To contextualize our work among evolving conversations in the field, this literature review highlights resources that connect digital records accessioning to a greater vision of discovery and access. As a foundation, we would like to define our intended use of the terms "acquire" and "accession." The Society of American Archivists dictionary defines them both as "materials physically and officially transferred to a repository as a unit at a single time," but with slightly nuanced second meanings.[4] Acquisition may include the "process of seeking and receiving materials," whereas accessioning implies a change to the intellectual and physical custody of materials from a source to the repository. A subtly enhanced comparison of the terms, which will be used throughout this paper, was put forward by Heather Ryan and Walker Sampson:

> *Acquisition* refers to the physical retrieval of digital content. This could describe acquiring files from a floppy drive, selecting files from a donor's hard drive or

> receiving files as an e-mail attachment from a donor. . . . *Accessioning* refers to integrating the content into your archives or collections: assigning an identifier to the accession . . . and adding this administrative information into your . . . collection management system.[5]

This distinction is important. With born-digital content, the concept of physical retrieval (to the extent that bits are subject to the notion of physical control) interfaces with tasks associated with accessioning to the point where the terms can often be muddied and used interchangeably. It is relatively simple to consider this with pieces of individual media, such as floppy disks; they are acquired when passed to the archivist, but not accessioned until a number has been assigned, the existence of the files is represented in donor or transfer agreements, and a database record is created. But with files that exist on the Web and need to be harvested, or digital publications locked in proprietary tools, the act of acquiring requires forward-looking judgment and intervention from the archivist and has immediate implications for accessioning workflows. To further confuse the downstream work—and we argue this is part of the ongoing challenge of clearly defining and documenting born-digital workflows—accessioning checklists can bleed into processing, to the point where so many tasks are front-loaded into accessioning that the time investment is significant.

Literature from the last dozen years relating to born-digital accessioning touches on both the friction and slipperiness found along the spectrum of accessioning terms, tasks, and values. In 2012 and 2013, several pieces of new guidance helped archivists frame out born-digital acquisitions and accessioning work. Gabriela Redwine et al. wrote the report *Born Digital: Guidance for Donors, Dealers, and Archival Repositories* to address the fact that the increasing volume of digital content was still often treated as an afterthought in workflows and policy. The report, which almost exclusively uses the word "acquisition" to encompass both acquiring and accessioning born-digital materials, recommends digital materials at archival repositories across four areas: initial collection review (information gathering, surveying); privacy and intellectual property; acquisition of digital materials; and post-acquisition review. As a reflection of how quickly digital materials can incur legal responsibilities, the very first recommendation in chapter 4, "Key Stages in Acquiring Digital Materials," is for an archives to draft an agreement or contract with the records creator *prior* to survey or acquisition.[6] To form agreements with donors before survey activity—surveying being an initial step to ascertain the scope of the collection—reflects an extraordinarily cautious and conscientious approach to working with digital records. Recommended practices include conducting detailed surveys; communicating fully with donors and dealers; reviewing inventories prior to transfer; reaching consensus on rights and intellectual property considerations; identifying sensitive or private files and how they will be screened, restricted, or redacted; screening emails; instituting gift or purchase agreements; following

handling protocols; creating disk images; etc.[7] When interpreted together as a firm list of requirements, the process arguably strays from what many archivists feel is practically achievable at the point of transfer. Particularly at academic libraries and manuscript repositories, the option to converse extensively with records creators often doesn't exist: we sometimes acquire computers and hard drives after the creator's death.

Another framework was published in 2012 by members of the AIMS Project, an interinstitutional effort to investigate and document recommendations for good methodologies and practices for born-digital records. Their whitepaper[8] includes a lengthy summary of resources and references available to archivists. The report is forward-thinking in its ready acceptance and clear articulation of the need for born-digital records stewardship to be seen, from the start, as an iterative and holistic ecosystem of work:

> Some tasks must be carried out at a specific place or order . . . while others are relevant to all or can be done at different points. In some cases the deciding factor was archival, sometimes practical or technical, sometimes ethical. . . . With born-digital material there is a greater need to understand, analyze, and assess the implications of decisions made at a particular stage of the workflow to avoid problems or conflicts later.[9]

While the AIMS authors do not use the term *flexible* here, it is implied that when tasks occur at differing points and levels, and early decisions impart downstream effects on future work, flexibility is key. The AIMS approach stresses that work can occur in concentric, overlapping circles while pushing forward in key areas. The four major circles of activity are identified as collection development, accessioning, arrangement and description, and discovery and access. Waiting for a perfect tool or solution for each is not an acceptable choice. Unlike some other early guidance that emphasizes technical tools and approaches, such as forensics and reference models,[10] the AIMS report is purposefully high level and sets main objectives for each function. Each objective is then presented with greater granularity, including outcomes, decision points, tasks, and "keys to success."[11] The guidance on collection development equates to acquisition; the accessioning section, meanwhile, is defined as the phase where active management of the content begins, including establishment of legal and administrative custody, and where baseline processing can begin.

While a linear outline of these steps may sound familiar to most archivists, an important aspect of this framework is the acknowledgment that these functions overlap, may not occur sequentially, and directly impact each other. The final function, enabling discovery and access, directly relates to the first module on collection development.

The outcomes of those steps have a significant impact on what is either appropriate or achievable in terms of discovery and access. It is therefore crucial

to consider issues relating to discovery and access as early as possible—beginning with the collection development phase—and continuing to update and revise plans as work on the collection progresses.[12] In other words, the report emphasizes that accessioning may not function best as a strictly siloed task or a matter of punching checklists. All work must be considered iterative and interrelated, and open to new techniques and strategies.

Case studies published around the same time on acquiring and accessioning born-digital materials emphasize this feeling of being overwhelmed, but showing a path through the forest, nonetheless. In 2014, Cyndi Shein wrote about the J. Paul Getty Trust Institutional Archives' initial effort to process an "incoming born-digital collection from the time of transfer to the time of public dissemination."[13] Near the end of their robust literature review on born-digital stewardship, Shein points out the gap between recommendations on acquisitions and accessioning, and models for providing access and discovery for records. The story of the transfer of *Pacific Standard Time: Art in L.A. 1945–1980* records from the Getty Foundation to the Institutional Archives illustrates how limits and roadblocks that occur quite naturally during acquisition have an immediate impact on accessioning (in this case, nineteen separate accessions forced certain workflows) and access (challenges of scale).

Other case studies identify similar stresses. In 2016, Laura Uglean Jackson and Matthew McKinley wrote about their experience archiving a 2.5 terabyte collection from the Office of Strategic Communications that documented planning and celebrations for the University of California, Irvine's (UCI) fiftieth anniversary.[14] At the time of the acquisition in 2014, UCI's archivists had an established workflow for born-digital records. However, they quickly realized that the size of this single accession, the file formats (such as raw video), and the sprawling, organic disorganization were far beyond what normal procedures could accommodate. Jackson and McKinley walk readers through their major pain points for this collection, identifying appraisal, preservation repository ingest, and developing an access model (still unclear at the time they published their study) as the major steps. Creativity, communication, and a frank admission of their assumptions make this case study honest and relatable. Prior to this acquisition, Irvine archivists had always refrained from appraising born-digital materials prior to ingesting them into the preservation system.[15] But, due to several issues with this collection, especially size, they had to change the order of events. Performing tasks out of their usual order became necessary for forward movement in stewarding this material, showing the necessary overlap of functional spheres.

The Society of American Archivists presented several short case studies from practitioners on born-digital accessioning in 2016 as part of its series *Trends in Archival Practice* that highlight the diversity of digital records (in resource type, format, and volume) and repositories.[16] The authors all show careful, ethical

handling of the records following established guidelines, while acknowledging that each example presents complex new challenges that require boutique solutions. In their conclusion, Stanford University archivists Josh Schneider and Daniel Hartwig state that their case study shows the need for "adopting a broad range of approaches when appraising and acquiring born-digital files at a contemporary research university. [The approaches] underscore the value of building strong relationships with campus records creators, as well as the need for flexibility and patience to accommodate their concerns . . . "[17] Trust and communication between archivists and creators, in addition to flexibility, is foundational to the whole process. Those principles are just as important to their stories as technological solutions.

Further emphasizing the importance of mutual trust, collaboration, and empathy between archivists and creators, Itza A. Carbajal calls for "mechanisms and opportunities to engage directly and indirectly in archival decision-making" in her analysis of interviews with musicians donating materials to archives.[18] She highlights the importance of "the implementation of more transparency around [archival] practices and decisions," even when there is not an established model for donor agreements and other documents that reflects this transparency.[19] This approach mitigates power imbalances between archivists and creators, which can certainly exist between an office with a mandate to acquire materials, like university archives, and the campus community. She recommends building trust by "attending to and caring for a donor's well-being" related to their materials and undertaking ongoing conversations about materials and their use, instead of employing strategies "focused almost exclusively on avoiding litigation."[20]

In the last few years, fresh guidance specific to born-digital accessioning has emerged that encompasses this in-the-round perspective. These publications function more as open frameworks that allow informed decisions, flexing, and creativity than prescriptive digital "to-do" lists. A fear of losing records through inaction has spurred acceptance of diverse approaches. In 2018, Erin Faulder et al. wrote the *Digital Processing Framework* to encourage minimum standards and more consistent practice with born-digital records.[21] Undergirding *Framework* is acknowledgment that though there are many steps in stewarding digital content, the steps allow varied tiers of effort and may be completed nonlinearly. In fact, though "processing" is in the title, the authors specifically state that "'processing,' for the purpose of this framework, concerns activities that may overlap with other traditional archival functions including accessioning, preservation, and arrangement and description."[22] Accessioning-related tasks are included in many of the twenty-three steps, but "accessioning" itself is not a step. This adaptability and encouragement of a multifaceted approach is built into the framework's design. In 2020, Alexandra Chassanoff and Colin Post wrote *OSSArcFlow: Guide to Documenting Born-Digital Archival Workflows* to assist archivists with born-digital records documentation, proposing a model of thirteen steps with recommended actions.[23] Many of these

steps emphasize a holistic, goal-focused approach from the beginning, stating, "if the institution creates transfer, preservation, processing, or access plans for a collection, these plans are usually developed [prior to acquisition]."[24] They also examine the concept that conflicting gaps and overlaps naturally occur in archival workflows. Because archivists use a variety of tools to manage digital collections, workflows can get hung up on gaps between technologies. Conversely, using diverse tools and software can also create overlaps or redundancies in tasks, which can cause confusion about hand-offs and efficiency.[25] This issue is acknowledged and reflected in the *OSSArcFlow* recommended steps—there are gaps and overlaps between them, and this is a fact to navigate, not a flaw.

Similarly, in 2021, Monique Lassere and Jess M. Whyte expanded on Redwine et al.'s recommendations for accessioning born-digital materials to include a greater focus on the experiences and needs of donors and records subjects, as well as acknowledging the ways in which concepts identified as best practices (i.e., creating disk images of storage media) can hinder responsible and ethical stewardship.[26] They provide an overview of radical empathy in archives and use this to interrogate practices such as disk imaging at acquisition, which they acknowledge "engenders trust, authenticity, and integrity but also potentially discloses a multitude of hidden and sensitive information to the recordkeeper, who must decide on the proper course for review and redaction (or not)."[27] Furthermore, disk imaging "serves as a deferral method, a way to hold content in stasis until some future date when a more thorough review of its contents can be completed," which "may not come at all."[28] They present recommendations in six areas, while acknowledging that it may not be possible for a repository to responsibly steward all types of born-digital materials.[29] The focus is on principles and concepts for practitioners to consider in their own contexts and an ethics of care surrounding records creators and subjects, rather than on a checklist to follow.

Finally, another recent publication that employs a framework model is *Levels of Born-Digital Access*, a report by Shira Peltzman, Brian Dietz, and others from a Digital Library Federation working group.[30] It includes a useful table presenting five areas pertinent to enabling access to born-digital archival materials, as well as action advice across three levels of effort. We note that although the *Levels* report focuses on access, the interweaving of all archival functions, including acquisitions and accessions, is clearly called out. The report states, "one point that we heard repeatedly throughout the feedback process [for this report] is materials must be brought into a repository in a way that explicitly supports future use."[31] An eye toward access must be considered even when completing what *Levels* refers to as "prerequisite steps," or acquisition functions. This guidance builds on the trends supported in the other publications mentioned here, which emphasize a gradual shift away from rigid checklists and toward work that occurs in overlapping circles within a more holistic vision for iterative care of digital collections. This literature inspired our team to

be less afraid and more pragmatic and nimble in how we collect, accession, and process digital records. As we acquire and accession records, we are looking ahead to how researchers will discover and use them. We will present examples of our work within two common accessioning scenarios: being under time-sensitive pressures and acquiring web-native content.

Catch It While You Can: Proactive Approaches to Time-Sensitive Acquisitions

University Archives is often called upon to collect digital records in a moment of urgency, when a campus office is struggling to support or salvage content. This typically occurs when offices conduct server migrations, change commercial online publishing services (such as ISSUU or SoundCloud), or abandon a communication tool. Previously in situations like these, we would have done the simplest, most direct form of collecting: accepting a hard drive with the content copied to it, creating an accession record, and placing the media in a box "for later." However, saving media carriers long-term is no longer an acceptable practice in SC&A, where we recognize that the best strategy is file-based storage with active management.[32] Two useful examples of time-sensitive record acquisitions where accessioning was synched with processing are: campus notices and flyers, and a Health Sciences Tumblr account.

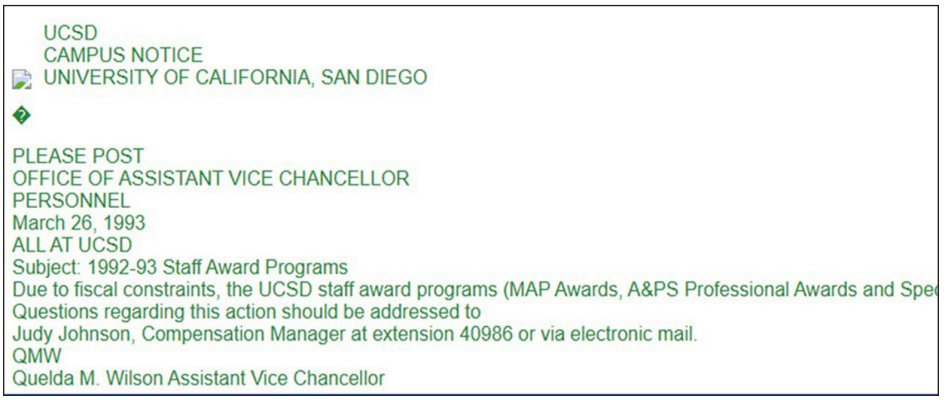
The campus Policy and Records Administration (PRA) office, the main unit responsible for collecting and managing the active records of the chancellor's office, and SC&A have worked closely over the decades to ensure that documentation of top campus administrative functions is preserved as mandated by the University of California. Until the early 2010s, PRA exclusively sent paper records to the archives. When the university archivist reestablished connection with this office, the two units decided that digital files would remain digital (not printed prior to transfer), and SC&A would develop methods to manage and preserve incoming digital files.

In 2023, PRA migrated to a new internal electronic records management system. In addition to moving records from its old repository to the new one, the office planned to remove records from its public website and store them exclusively in the new database, where they would no longer be publicly available. These records included campus notices and flyers dating from 1993 to the present that alerted the campus community to employment-related information, policies and regulations, events and activities, administrator appointments, safety and security matters, and general information. PRA contacted the university archivist asking if the notices would be of interest for the archives. They provided a rich source of documentation for thirty years of campus life, activities, and administrative goals, so preserving them before their removal from the public Internet was crucial. After reviewing the web pages, the university and digital archivists determined that crawling the notices would not be an appropriate capture method. Archive-It's full-text search can give

uneven results, and the original HTML pages were hard to read due to the existing formatting and text colors of green and blue.

Combined, those factors made a web crawl undesirable. We wanted the notices to be key-word searchable and meet accessibility standards, which would best be accomplished in the DAMS. With that in mind, we asked PRA to provide a spreadsheet of URLs for the objects and basic harvested descriptive metadata for each notice, so that we could capture the notices as files and ingest them into our DAMS. While we used software and command-line tools to accomplish this process, it did require both strategy and human labor to work through the technical challenges. In total, there were nearly 9,000 notices. UA created two accession records in ArchivesSpace for this material to reflect two content acquisition streams. Each batch of notices required a different workflow due to differences in page layout and source code. In the accession records, we documented the total gigabyte count of each accession and the estimated number of files. For the newest notices (2020–2023), we used a Python script and the print to PDF function in the command-line version of the Chrome browser to generate a PDF from each page URL. These notices were printer-friendly HTML, so no further intervention was required.

For the older notices (1993–2020), the process of bringing the files in was more complicated. These notices, as seen in Figure 1, were old HTML pages that were not legible when printed directly to PDF; there were no line breaks, so text was cut off at the page margins; most of the text was green or bright blue (a coloring code to distinguish different types of notices), which is not suitable for maximum readability; and special characters displayed incorrectly.

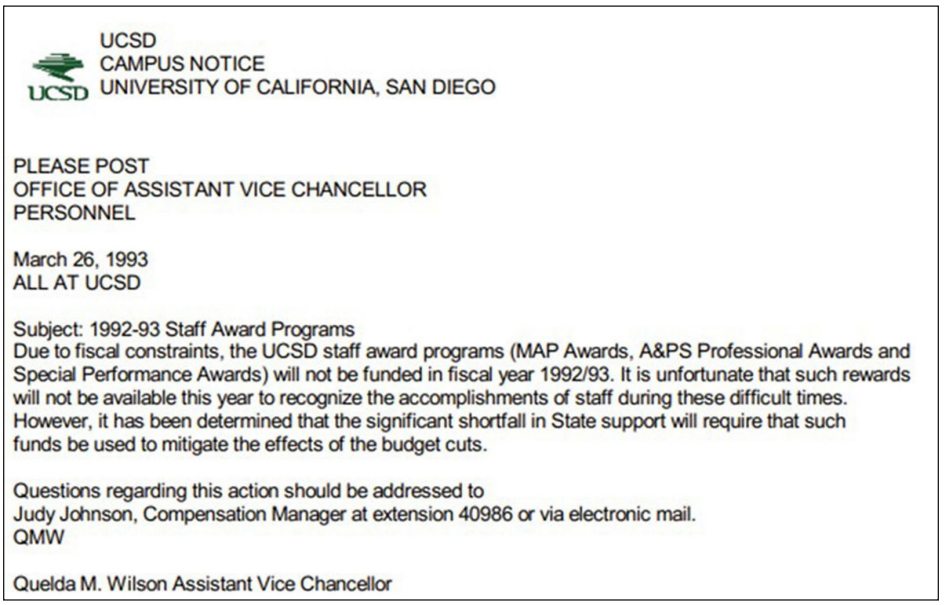

UCSD
CAMPUS NOTICE
UNIVERSITY OF CALIFORNIA, SAN DIEGO

PLEASE POST
OFFICE OF ASSISTANT VICE CHANCELLOR
PERSONNEL
March 26, 1993
ALL AT UCSD
Subject: 1992-93 Staff Award Programs
Due to fiscal constraints, the UCSD staff award programs (MAP Awards, A&PS Professional Awards and Spe
Questions regarding this action should be addressed to
Judy Johnson, Compensation Manager at extension 40986 or via electronic mail.
QMW
Quelda M. Wilson Assistant Vice Chancellor

FIGURE 1. "1992–93 Staff Award Programs." Unedited HTML version of notice (preservation copy), original state before interventions. *Courtesy of Special Collections and Archives, UC San Diego Campus Notices and Flyers collection.*

Additionally, the HTML formatting was inconsistent. It changed over the years, so there was a range of potential layouts, header images, and attendant problems depending on when it was created. In almost all cases, the formatting issues were consistent across a given date range (i.e., 1993–1998 would need one set

of interventions, 1999–2003 a slightly different set) and likely corresponded to the site's server framework used at the corresponding time. We created a Python script to harvest the HTML, address these issues, and generate PDFs from the resulting HTML files. While they required labor to create at the point of acquisition, the PDFs solved many of the problems inherent in the old website presentation, as seen in Figure 2.



UCSD
CAMPUS NOTICE
UNIVERSITY OF CALIFORNIA, SAN DIEGO

PLEASE POST
OFFICE OF ASSISTANT VICE CHANCELLOR
PERSONNEL

March 26, 1993
ALL AT UCSD

Subject: 1992-93 Staff Award Programs
Due to fiscal constraints, the UCSD staff award programs (MAP Awards, A&PS Professional Awards and Special Performance Awards) will not be funded in fiscal year 1992/93. It is unfortunate that such rewards will not be available this year to recognize the accomplishments of staff during these difficult times. However, it has been determined that the significant shortfall in State support will require that such funds be used to mitigate the effects of the budget cuts.

Questions regarding this action should be addressed to
Judy Johnson, Compensation Manager at extension 40986 or via electronic mail.
QMW

Quelda M. Wilson Assistant Vice Chancellor

FIGURE 2. "1992–93 Staff Award Programs," access copy. *Courtesy of Special Collections and Archives, UC San Diego Campus Notices and Flyers collection, [https://library.ucsd.edu/dc/object/bb9772123f].*

As the notices were previously publicly available, we felt it was a step backward if a researcher had to do a mediated search for this content on-site at the library or navigate their way to it via a baseline collection-level record. We wanted a public presentation of the content on the Digital Collections website. Following acquisition of the PDFs, and after making our accession records, the next key processing component was to prepare the files for DAMS ingest. A lot of heavy lifting had been done with the PDF conversion, leaving the primary tasks of adding our standard notes to PRA's descriptive metadata and removing duplicative or extraneous material, such as years of identical daily COVID-19 exposure notifications from the 2020–2023 notices. Once this was done, we ingested the searchable PDFs as access copies, along with the HTML originals, into the DAMS.

While this was a lot of work, it helped us iron out plans for the present and future of campus notice collecting: we now capture campus notices on an annual basis, preventing a new backlog and reducing the risk of technical debt from obsolete technologies piling up. The first accessions were challenging and took many months, as the Python scripts took time to develop and getting readable

PDFs took a lot of trial and error. However, with this learning curve under our belts and a growing library of scripts, ongoing accessions of new notices can now be completed within a day. The content is available on our Digital Collections site and includes campus notices up to July 2024.[33] As we initially responded to PRA and planned this acquisition, we envisioned where we wanted to go: while we could have set a web crawl using Archive-It, we knew that wasn't our big-picture goal for this material. We wanted to create readable and searchable copies for the public and felt it was best to adapt to that plan as part of accessioning, rather than set it aside as a "processing problem" for the future. Though this entailed a lot of early formatting and normalization work during accessioning, once that was completed, processing and presenting the content was straightforward.

In a different example, UA has a long history of collecting publicity materials and photographs from the Health Sciences Communications (HSC) office. Recently, there was a new opportunity: HSC had stopped using a Tumblr blog to share stories and images and planned to take the site down. Staff asked UA if the site's content, dating back to 2011, could be preserved. While UA does use Archive-It to capture campus websites and Tumblr content can be captured through web archiving tools, we deemed it inappropriate in this case as Tumblr requires a login to see more than a handful of posts. Since there were more than 1,100 posts, this approach was not feasible. In addition, Archive-It's text search can be challenging to use. A researcher would be unable to browse, find specific posts, or key-word search within the site.

Finally, web archiving is most effective in cases where contextual "look and feel" is a priority. HSC had primarily used its blog to post short articles about health-related topics and updates about Health Sciences' activities, rather than to interact with other blogs and users. This meant that we could focus on capturing post text and images instead of "likes," comments, and other details specific to social media. Because the content was more important than the blog's interface or structure, we determined the best acquisition method for preservation and access was to capture the blog as a collection of PDFs which could be ingested into our DAMS and made available through our Digital Collections site. We used a similar method to capture HSC's press releases, so when we explained the concept to HSC and presented our path forward, staff had a clear idea of what to expect and were receptive to our plan.

Acquiring this content and completing its accessioning while preparing it for online access and display was a multistep process. Tumblr includes a large "create an account to see more" banner on every page if a user is not logged in, and this obscures blog post content when we use the "print to PDF" process employed with campus notices. Knowing this, the first step was to determine whether we could remove the banner and similar artifacts before generating PDFs. As with nearly all "automated" processes, a significant investment of human time and expertise was needed up front. The digital archivist reviewed the source code for some sample blog posts to identify the HTML tags governing these artifacts, then downloaded

the posts as HTML files and manually edited them to remove this code. When we generated PDFs from these test files, they were fully readable, and the clutter was gone. When we shared the cleaned-up test PDFs with HSC staff as a proof of concept, they were happy with the results and began to compile a list of blog post URLs for us to capture. While they prepared the list, we began work on a Python script that would remove the blog's problem tags and automate the harvest for acquisition. This collaborative approach, with clear communication and shared problem-solving between UA and HSC, ensured our partners understood our approach to the records and had a visual of what the outcome of our work would be. They could see the human time investment and how library workers would ensure the successful preservation and future accessibility of their valuable digital content.

After running the script against these URLs, we had a collection of PDFs that included post content, blog formatting, and HSC's subject key words and tags, but no "log in to see more" banners or "follow this blog" buttons that would distract from or obscure the posts themselves. We were then able to extract descriptive metadata for each file based on publication dates and title headers and ingest the PDFs and accompanying metadata into our DAMS. Each reconstructed post included local notes stating that the PDF was a surrogate generated from the HSC Tumblr blog and created from a corresponding HTML page.

Acquisition, therefore, required not just capturing the files, but planning a migration and metadata harvest as a precustodial step. All posts are publicly available with full-text search on our Digital Collections site and are also linked within the finding aid for the hybrid HSC collection.[34]

The key to this project was that we didn't simply crawl the Tumblr blog and create an accession record pointing to an Archive-It link. That would not have been appropriate in this case due to Tumblr's access restrictions and the limitations of full-text search in Archive-It's platform. We communicated with HSC to understand what they had created, what they wanted to preserve, and how a user might try to access the content. Accordingly, we worked quickly to acquire and accession the blog as searchable PDFs based on a series of harvesting steps. While that front-loaded some work, it made it possible to seamlessly process and republish the collection before HSC pulled it offline.

Addressing Collecting Gaps and Facilitating Discovery of Web Native Materials

Some important university records are only available online. While we use Archive-It to capture university web pages and create catalog records to encourage discovery and connection to the content, the Archive-It interface is not intuitive. Similarly, born-digital records on websites or intranets, such as meeting minutes and agendas for important campus committees, do not necessarily offer easy

discoverability or keyword searching. They also lack a preservation layer. As described in the previous examples, campus offices also inevitably upgrade their servers or change technology tools, and a long run of content broken or presented differently over time can frustrate researchers. Two examples of very different web content the archives has accessioned and repackaged for researcher use are the records of the Academic Senate and student podcasts created by the university. The Academic Senate, which directs the educational function of the university and provides faculty advice to the University of California Board of Regents and the campus administration, maintains a website with information about the senate and its committees. As one of the three branches of shared governance at the university, it is critical for UA to document. The archives has a strong foundation of paper records from the senate's early years, but when documentation switched to electronic formats, the archives developed a major coverage gap. To fill the holes in the official record, the university archivist needed to collect all missing agendas and minutes for the representative meetings, many of which were online on the website. The goal was to acquire these materials, integrate them intellectually into the existing paper-based senate collection finding aid, and present the files online. Using the "Digital Collections" page, a user could keyword search across the content, which could not be done on the Academic Senate site. The files would also be ingested into the library's preservation system, which would provide long-term security and run checksum functions.

Initially, UA contacted the senate office to see if we could facilitate a direct transfer of digital files via file transfer protocol or a similar process. However, due to senate office staff changes, an office move, and the disruption of the pandemic, they could not readily identify the "final" digital versions of agendas and minutes on office servers. Office staff were confident that the best and final versions of documents were those posted to the website.

Therefore, our only real acquisition option was to take the files directly from the Web. In a task that blurred the line between precustodial work and acquisition, a library student employee was tasked with downloading several hundred public files, one for each posted meeting. As they did so, they prepared a spreadsheet, capturing filenames, meeting titles, dates, and other easily identifiable information to include as metadata for the eventual digital objects. The spreadsheet became the key documentation for the accession record, similar to a highly detailed box list or inventory. We also knew we could leverage it during the DAMS ingest to build the descriptive metadata for the digital objects.

Once acquisition was complete, the downloaded copies of the content were stored on library servers, and the accession record was created, the university archivist reviewed the materials and discovered that some of the downloaded files were several hundred pages long. These files were too big and clunky for an average person to download and easily browse. To make the documents more accessible to

users, during object-build (the processing step prior to DAMS ingest), large files were broken into subcomponent parts or "chapters" with a clear relational structure within the object. Now that the content is stored in our DAMS and presented online,[35] the entire corpus of documents is fully key-word searchable, and each item has a stable identifier for citation.

The meeting minutes from the Academic Senate are now available through two discovery points, which may serve different types of users. The Academic Senate office continues to post content online, where faculty expect to be able to sign in and find key documentation. The UA's job is not to replace that website, but rather to save the same content in a different bucket under archival stewardship for the public. Digital objects in UA's care benefit from the library's preservation layer and enhanced searching capabilities. Furthermore, if needed in the future for a researcher or the senate office, the library could facilitate a batch download of the PDFs instead of a one-by-one document download. If the senate moves to a new web server or must redesign the architecture of its site, staff can rest assured that the files are already backed up and available elsewhere. Another benefit of copies of the records being stewarded in UA is that we can present them in context with related materials. Researchers can discover the digital records through the traditional finding aid for the senate collection, where links to online content are embedded in the file list and also in the catalog record for the collection. When a user searches for a key word across DAMS content, terms may pop up in Academic Senate files as well as related university record collections, connecting thematic dots.

We believe it is good that researchers can use these different discovery paths seamlessly through web browser searches without getting bogged down in circles of links, mediation, and firewalls. Because of our work, Academic Senate officers know their records are safe, secure, and available through the library. In this way, we are fulfilling a service function with clear benefits to the originating office. In the past, contributors may have thought of the UA collections as hidden and opaque, locked in closed stacks and only available in person. The digital presentation of content, however, is immediately visible and shows its value to the creator.

While this enhanced visibility is a significant achievement, it also highlights the ongoing need to develop comprehensive strategies for the preservation and access of all forms of university digital content. Other types of university content native to the Web (newsletters, annual reports, unit histories, etc.) are important too. Sound recordings and slides from university podcasts make up an entirely different category. Podcasts were a way for campus offices and leaders to communicate with broad audiences starting in 2020 during the COVID-19 pandemic. At the start of COVID-19 shutdowns, Health Promotion Services (HPS) and Student Affairs (SA) released podcasts about health, wellness, and navigating the shift to remote learning: *Live Well, Be Well*[36] and *Triton Tools and Tidbits*.[37] These series ran during

the first year of the pandemic and represented an important component of campus messaging to students.

University Archives wanted to capture and preserve these podcasts, but acquiring them was more challenging than it first appeared. The podcasts were hosted on an external website. Embedded multimedia and dynamic content were challenging to capture, and even with high-touch crawl setups, episodes did not get consistent capture with reliable playback capabilities. In addition, limited text search abilities in web archives collections made it difficult for users to locate specific, individual episodes. This example illustrates the innate overlap between acquisitions, accessioning, and processing tasks that can appear with web-accessible and born-digital content.

At first glance, acquiring the podcasts could be as simple as a web crawl. But this approach would limit our ability to move forward with iterative processing and provide enhanced access. On the other hand, if we simply downloaded them, what would we do with all the podcast content? A finding aid list of podcasts could work, but didn't make much sense, either. We had to envision what mounting the podcasts on the Digital Collections page (after DAMS ingest) would look like, and how to get there, at a precustodial moment.

Since the content of the podcasts, rather than their contextual look and feel, was our primary focus, we decided to capture copies of the audio and visual files, create transcripts, assign subject terms, and rebuild each episode as a complex digital object in our DAMS. This would enhance searchable metadata far beyond what was available on the original sites. We also knew the sites were vulnerable. Indeed, one of the sites was taken down just a couple of years after we captured the podcasts. To acquire the episodes, the university archivist reached out to the creators and set up Google Drive folders shared exclusively between podcast producers and the archives. The podcasters deposited their audio files and any slides in the folder, where archives staff could access and upload them to Otter.ai for transcription. For metadata, we collected episode summary descriptions and participant names directly from the podcasters; if they couldn't give us content, we mined the transcriptions for this data.[38] The acquisition workflow facilitated new relationships with campus partners, created the opportunity to secure permissions, and made UA a key cocreator, adding value with the transcripts. The methodology was carefully documented in the accession record for the collection. This was not a speedy process, as it required building trust and collaborating with creators to share materials. Part of establishing this new relationship was to help the creators and participants (primarily students) understand what archives are and what archivists do to provide access and long-term preservation. As we met and discussed our processes and how we could preserve their work and highlighted existing digital collections, they became more excited about the prospect of being included in the archives.

When we planned for this acquisition, we visualized the type of user experience we hoped to provide: one-stop discovery via "collection landing pages" that can be found with any web browser and linked to the library catalog. With this goal in mind, the need for direct outreach to creators became clear. If we hadn't had time to build the objects and ingest them in the DAMS, a reasonable post-accessioning status could have consisted of the files sitting on the server and a collection-level record. In this situation, we could call up a podcast if we received a direct inquiry from a researcher. Instead, we set up a processing workflow that could be worked on consistently, as time allowed, over several months—generating transcripts, reviewing them for accuracy, and building the objects. This work could be done from home, which was good for remote workers during the pandemic. Again, UA positioned itself as a valuable campus collaborator that cares about campus messaging and communications.

Accessioning the COVID-19 podcasts through a file-sharing process with creators made us confident we could deliver value-added packages of audiovisual content long-term. This worked well and led to a regular workflow we currently apply to other podcast series. Additionally, materials processed in a timely manner as an outcome of smart, engaged accessioning leads to good relationships with campus partners who see the results online. Fostering a collaborative vision of building online collections, which offices may not even know is possible, grows confidence and trust that the creator's records will be treated with respect and carefully managed. Sharing records transparently on the Digital Collections website helps UA show its value and takes "archives" from an abstract concept to a visible product.

Conclusion

In a campus environment where the creation of paper records is close to extinction, the time to actively pivot toward new, nimble ways of acquiring and accessioning born-digital records is now. As our field evolves, an important component of working with born-digital content is clear articulation of the possibilities of what that work can look like and how we outline the myriad steps. Resources discussed in the literature review presenting guidance frameworks such as the *OSSArcFlow Guide,* the AIMS Born-Digital Collections white paper, *Levels of Born-Digital Access*, and Faulder et al.'s *Digital Processing Framework* can all assist practitioners to form decisions about levels of processing and intensity of effort. If an archivist is unsure how to describe or envision an end goal for processing a born-digital collection, the frameworks can point to different outcomes and accompanying levels of effort and complexity across functional areas. Just as archivists have grown more comfortable with the concept of iterative and flexible processing for traditional collections, the same principles apply to born-digital collections. We cannot be so focused on the technological challenges that we forget iterative

processing with varying approaches can also be suitable for digital collections, with the added bonus that digital text provides powerful new opportunities for search and discovery. Task checklists are important, but they ought not hold us up if every task cannot be completed equally well. We must be willing to work with donors in new ways, push appropriate content online instead of into boxes, and design workflows to wrangle the records outside of a rigid linear task progression.

Acquisition and accessioning details can vary from project to project. Over the last few years of work at UC San Diego, we have found the context of the records, and their importance, whether based in relationships, resource type, accessibility, or accountability, helps guide the decisions. Early migration and proactive object-modeling, or visualizing users engaging with the records online, is a good way to develop, test, and flex effective workflows. If archival workers remain unwilling or too intimidated to get hands-on with digital content even before acquisition, or to consider taking on tasks like migration, they lose the opportunity to generate some momentum while conversations with donors and contributors are fresh.

Every practitioner must understand their risk environment and fully comply with their repository's legal obligations responsibly. However, with license to do more than simply save files to media or a server and create accession records, tackling some work typically viewed as "processing" at the point of acquisition can be enormously helpful if the archivist documents their practices. If resources are available that allow creativity and flexibility with digital acquisitions—while following basic protocols and policies in place for authenticity and security, of course—each project will teach new concepts that can then be applied to future work. Time is the greatest resource, as digital projects with deliverable results that can be viewed by the public can take an enormous amount of human intellectual labor. However, investing the time to solve a puzzle will help make the archivist or their team more efficient when a similar problem appears in a new collection. Each digital processing project may be slightly different, but past successes and challenges inform the approach to the ones waiting around the corner.

As we illustrated in our examples, during preaccessioning surveys, it is important to think beyond the immediate accessioning steps to the archives' ultimate goal for the content (that is, beyond acquiring it). Can these records or publications be searched by the public and downloaded? Or are they restricted in some fashion and need to be limited to on-site serving or mediated requests? This question helps us decide on the tools we will use to acquire the records and how they will be saved. Are the file types such that format migration, to be documented in the accession records, is appropriate—or are assertive interventions truly damaging for preservation, security, and authenticity concerns? We also look at existing metadata, both descriptive and technical, that we can repurpose or leverage throughout every step of our stewardship, from creating accession records (and attaching automated reports using software such as TreeSize) to building objects for DAMS ingest. Even at the

precustodial stage, we consider how the digital records will ultimately be presented to researchers: what level of cataloging will suffice, or what type of finding aid or online collection page may be involved?

As a team, if we can't develop even a partial plan for the level of access we desire and outline a processing method to attain it, that is a red flag that perhaps we are not ready to acquire the collection. Once content is accessioned, UA is responsible for the care of the materials and the legal implications that go along with it.

The pivot point of our recent work evolution has been to acquire and accession content while articulating early on what discovery will look like and what elements of processing need to happen to get there. In the past, acquisition and creation of an accession record seemed like appropriately siloed, dead-end first steps. If we could accomplish those things, following our checklists, we felt we plateaued at a reasonable stopping point and were taking good care of incoming born-digital collections. Today, we feel that is a false premise. It is dangerous to acquire born-digital records simply to have them, without engaging in a holistic plan before, during, and after accessioning regarding intent and being transparent about the vision for their care.

Given our limited bandwidth at UC San Diego, we prioritize content that can be publicly available. Materials with restrictions or originating from offices that are not engaged in active conversations with UA about clear expectations or scheduling (such as planning around anniversaries or events) may be processed on a different timeline. Postprocessing, we contact creators to share links to their collections and related records. These conversations lead to strong relationships with key contacts in campus offices and a positive reputation for university archives. In all cases, we document the plans we discuss at acquisition in the accession record and troubleshoot technical issues assertively. We do this to avert being stymied by format challenges or pondering from scratch how to proceed if a collection hits a lull in active management for a few years. University archives is not just about boxes and media carriers anymore—we are advancing a robust online presence for university records that demonstrates our value to the campus community.

Notes

[1] "Policies for Administration of University of California Archives," University of California Libraries, captured at https://perma.cc/QNB6-MPCV.

[2] "UC San Diego Digital Asset Management System," UC San Diego Library, https://library.ucsd.edu/research-and-collections/digital-collections/dams.html, captured at https://perma.cc/KQW2-U59G.

[3] "UC San Diego Library Digital Collections," UC San Diego Library, https://library.ucsd.edu/dc.

[4] Dictionary of Archives Terminology, Society of American Archivists, s.v. "acquisition," "https://dictionary.archivists.org/entry/acquisition.html, captured at https://perma.cc/T8SY-D6XG, and s.v. "accession," https://dictionary.archivists.org/entry/accession.html, captured at https://perma.cc/D75M-JRY7.

[5] Heather Ryan and Walker Sampson, *The No-Nonsense Guide to Born Digital Content* (London: Facet, 2018), 53, https://doi.org/10.29085/9781783302567.

[6] Gabriela Redwine et al., *Born Digital: Guidance for Donors, Dealers, and Archival Repositories* (Washington, DC: CLIR, 2013), https://www.clir.org/wp-content/uploads/sites/13/pub159.pdf.

[7] Redwine, *Born Digital*, 4.

[8] AIMS Work Group, *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship* (2012), https://matienzo.org/storage/2012/2012-Jan-AIMS-WhitePaper.pdf.

[9] AIMS Work Group, *AIMS Born-Digital Collections*, vii.

[10] Two examples are *Personal Archives Accessible in Digital Media (PARADIGM) Project* (Universities of Oxford and Manchester, 2012), https://ora.ox.ac.uk/objects/uuid:116a4658-deff-4b06-81c5-c9c2071bc6d0, and Matthew G. Kirschenbaum, Richard Ovenden, and Gabriela Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Alexandria, VA: CLIR, 2010), https://www.clir.org/pubs/reports/pub149.

[11] AIMS Work Group, *AIMS Born-Digital Collections*, 2.

[12] AIMS Work Group, *AIMS Born-Digital Collections*, 46.

[13] Cyndi Shein, "From Accession to Access: A Born-Digital Materials Case Study," *Journal of Western Archives* 5, no. 1 (2014), https://doi.org/10.26077/b3e2-d205.

[14] Laura Uglean Jackson and Matthew McKinley, "It's How Many Terabytes?! A Case Study on Managing Large Born Digital Audio-Visual Acquisitions," *International Journal of Digital Curation* 11, no. 2 (2016): 64–75, https://ijdc.net/index.php/ijdc/article/view/11.2.64/452.

[15] Jackson and McKinley, "It's How Many Terabytes?!," 64.

[16] See Megan Barnard and Gabriela Redwine, Module 15, "Collecting Digital Manuscripts and Archives" and Erin Faulder, Module 16, "Accessioning Born-Digital Archives," in *Appraisal and Acquisition Strategies*, ed. Michael Shallcross and Christopher J. Prom, Trends in Archival Practice Series (Chicago: SAA, 2016), https://www2.archivists.org/publications/epubs/case-studies-trends-in-archives-practice.

[17] Josh Schneider and Daniel Hartwig, "Case Study 3: Stanford University Archives," in Module 15, "Collecting Digital Manuscripts and Archives," in *Appraisal and Acquisition Strategies*, ed. Michael Shallcross and Christopher J. Prom, Trends in Archival Practice Series (Chicago: SAA, 2016), 116, https://www2.archivists.org/sites/all/files/Module_15_CaseStudy3_Schneider-Hartwig.pdf.

[18] Itza A. Carbajal, "The Politics of Being an Archival Donor: Defining the Affective Relationship Between Archival Donors and Archivists," *Journal of Critical Library and Information Studies* 3, no. 2 (2021):17, https://doi.org/10.24242/jclis.v3i2.114.

[19] Carbajal, "The Politics of Being an Archival Donor," 18.

[20] Carbajal, "The Politics of Being an Archival Donor," 18.

[21] Erin Faulder et al., *Digital Processing Framework*, Cornell University Library eCommons, 2018, https://hdl.handle.net/1813/57659.

[22] Faulder, *Digital Processing Framework,* 1.

[23] Alexandra Chassanoff and Colin Post, *OSSArcFlow: Guide to Documenting Born-Digital Archival Workflows* (Educopia Institute, 2020), https://educopia.org/wp-content/uploads/2024/09/OSSArcFlow_Guide_FINAL-1.pdf.

[24] Chassanoff and Post, *OSSArcFlow*, 11.

[25] Chassanoff and Post, *OSSArcFlow*, 7.

[26] Monique Lassere and Jess M. Whyte, "Balancing Care and Authenticity in Digital Collections: A Radical Empathy Approach to Working with Disk Images," *Journal of Critical Library and Information Studies* 3, no. 2 (2021): https://doi.org/10.24242/jclis.v3i2.125.

[27] Lassere and Whyte, "Balancing Care and Authenticity in Digital Collections," 22.

[28] Lassere and Whyte, "Balancing Care and Authenticity in Digital Collections,"22.

[29] Lassere and Whyte, "Balancing Care and Authenticity in Digital Collections," 18–21.

[30] Shira Peltzman and Brian Dietz (project co-leads and editors) et al., *Levels of Born-Digital Access* (Alexandria, VA: Digital Library Federation, 2020), https://osf.io/hqmy4.

[31] Peltzman and Dietz, *Levels of Born-Digital Access*, 4.

[32] Per recommendations of the Digital Preservation Coalition: "Storage media can decay over time, leading to corrupted files. Storage media may become obsolete and unsupported by contemporary computers and the software that understands and provides access to them. The bits may be ignored,

abandoned, accidentally deleted or maliciously destroyed . . . " "Digital Preservation Handbook," 2nd ed., Digital Preservation Coalition, 2015, "Preservation Issues" and "Legacy Media" chapters, https://www.dpconline.org/handbook.

[33] "UC San Diego Campus Notices and Flyers," UC San Diego Library Digital Collections, https://library.ucsd.edu/dc/collection/bb4584364r.

[34] Health Sciences Communications Tumblr search, https://library.ucsd.edu/dc/search?utf8=%E2%9C%93&q=hstumblr. Tumblr posts and other digitized content are linked from the parent collection finding aid: UC San Diego Health Sciences Communications Public Relations Materials. RSS 6022. Special Collections & Archives, UC San Diego Library, https://library.ucsd.edu/speccoll/findingaids/rss6022.html.

[35] "UC Academic Senate. San Diego Division Records," RSS 901, UC San Diego Library Digital Collections, https://library.ucsd.edu/dc/collection/bb6428774p.

[36] "*Live Well, Be Well*," UC San Diego Library Digital Collections, https://library.ucsd.edu/dc/collection/bb2465491s.

[37] "*Triton Tools and Tidbits*," UC San Diego Library Digital Collections, https://library.ucsd.edu/dc/collection/bb1885278c.

[38] Elements of this creator-engaged workflow are outlined in Best Practice 3.3, "Solicit source-provided description of potential acquisitions, "in *Archival Accessioning Best Practices (ABP)* (National Best Practices for Archival Accessioning Working Group, 2025), https://accessioning.gitbook.io/archival-accessioning-best-practices/methods-and- practices/pre-custodial-considerations#bp3-3, captured at https://perma.cc/XY6P-YQT6.

ABOUT THE AUTHORS

Laurel McPhee is the supervisory archivist of Special Collections and Archives, UC San Diego Library, where she has worked since 2014. She leads and collaborates with a team of librarians, manuscript processors, and specialists to make collections of personal papers, organizational records, unique visual resources, and digital objects preserved and accessible. McPhee earned her BA from Harvard in Cambridge, Massachusetts, and her MLIS from the University of California, Los Angeles.

Tori Maches is the digital archivist of Special Collections and Archives, UC San Diego Library, where she has worked since 2018. Her work includes capturing born-digital materials, web archiving management, and other aspects of digital preservation for born-digital archival materials. Her previous roles include assistant archivist with Concordis LLC and digital archives program scholar at UCLA. Maches earned her BA and MLIS from the University of California, Los Angeles.

Marlayna Christensen is the university archivist in Special Collections and Archives, UC San Diego Library. In this role, she focuses on acquiring, preserving, and making university records publicly accessible, and assisting with campus history inquiries. She enjoys exploring novel approaches and technologies to improve archival workflows. Christensen has worked in the library since 2002, and prior to her 2016 appointment as university archivist, she led the Access Services and Reference units. She earned her BA and MLIS from Brigham Young University.