Title
Web archiving: Policy and practice

Journal
Journal of Digital Media Management, 8(3)

ISSN
2047-1300

Authors
Christensen, Marlayna K.
Maches, Tori

Publication Date
2020-03-01

Peer reviewed

Web archiving: Policy and practice

Received (in revised form): 24th October, 2019

Tori Maches
is the Digital Archivist at UC San Diego. Her work includes responsibility for UC San Diego's web archiving efforts, as well as capturing born-digital materials stored on legacy media, and other aspects of digital preservation for born-digital archival materials. Her previous roles include assistant archivist with Concordis LLC, and digital archives programme scholar at UCLA. She is also the Vice-Chair of the Society of American Archivists Web Archiving Section, Vice-Chair of the UC Libraries Born-Digital Common Knowledge Group, and a member of the UC Libraries Web Archiving Common Knowledge Group.

UC San Diego Library, University of California, San Diego, 9500 Gilman Drive #0175-S,
La Jolla, CA 92093-0175, USA
Tel: +1 858 822 0838;
E-mail: vmaches@ucsd.edu

Marlayna Christensen
is the University Archivist at UC San Diego. She is responsible for liaising with and collecting select records from campus and affiliated groups. Her previous roles include Reference Librarian for Interlibrary Loan at New York University; Head of Interlibrary Loan and Assistant Head of Circulation Services at Yale University; and Director of Access Services at University of California, Santa Barbara. Over her 30-year career, she has served on a number of professional committees with the Society of American Archivists, the Society of California Archivists, the American Library Association, and within the University of California.

UC San Diego Library, University of California, San Diego, 9500 Gilman Drive #0175-S,
La Jolla, CA 92093-0175, USA
Tel: +1 858 534 8605;
E-mail: mkchristensen@ucsd.edu

Abstract The UC San Diego Library has been collecting and providing access to archived web content since 2007. Initial collections were created on an *ad hoc* basis, with no high-level plan to identify websites and content of interest, and there was little documentation of how early collection decisions were made. As time passed, the library's web archiving efforts increased in scale, and outgrew this informal approach. Efforts were made to standardise web archiving processes and policies via collection request forms and standardised metadata, eventually culminating in the creation of a web archive collection development policy, and collection and quality control workflows and tracking. This article outlines the process of creating these tools, including establishing institutional needs and concerns, evaluating the wider landscape of web archiving policies and norms, and considering sustainable use of available resources. The article also discusses future areas of work to ensure that web content of research and historical interest is captured in full, preserved responsibly, and made accessible even when the original websites have changed or disappeared.

KEYWORDS: web archives, collection development, policy, process, workflows

INTRODUCTION

Creating an institutional web archiving programme can be an intimidating prospect. Capturing and preserving websites and other content in web archive collections allows libraries and archives to provide these resources long after the original website has changed or disappeared; facilitate stable citation links and effective fact-checking; and guard against the loss of content with historic

value. However, the degree of institutional commitment required, combined with the lack of documentation on best practices, can make it difficult to know where to begin. Likewise, changing the scope, scale and mandate of a web archiving initiative from a pilot programme or *ad hoc* collecting effort to a fully realised part of institutional collecting can be similarly daunting. Web collections created on an *ad hoc* basis may not be structured in an intuitive or efficient way, making management, discovery and future collecting more challenging.

This article explores these and other considerations related to web archive collection-building, management and access. The writers will discuss challenges and strategies related to web archive collection development, description and management, using case studies from their own institution as a lens through which to explore more generally applicable solutions.

WEB ARCHIVING AT UC SAN DIEGO

The UC San Diego Library began archiving web content in 2007 as a collecting effort through the California Digital Library's Web Archiving Service (WAS), and captured content sporadically until 2011, when web archive collecting largely came to a halt due to staffing changes. Initial collections were created on an *ad hoc* basis (Figure 1); there was no high-level plan to identify websites and content of interest, and web archiving collection decisions were not documented for future reference. Web archiving resumed in 2015, after UC San Diego migrated to the Internet Archive's Archive-It subscription service with the rest of the University of California (UC) system after WAS shut down in 2015.[1] Archive-It includes curation, collection and access services.

Web archiving also became more systematic with the recruitment of a digital archivist. She implemented a standardised form to collect essential metadata for new and existing collections, and this metadata was added to the collection information on Archive-It's public portal, and used to create a catalog record with a link to the archived pages.

Due to this initial work to standardise web archiving metadata and formalise collecting, the number of web archive collections in institutional holdings began to expand, and this continued after the first digital archivist's successor was recruited in 2018. By this time, the programme had increased in scale and outgrown this approach to collecting, and creating a web archive collection development policy became a priority. In 2019, a project management process and workflow were developed to track the status of and next steps involved in creating and monitoring web archive collections (Figure 2).

INSTITUTIONAL WEB ARCHIVE POLICY NEEDS

As of autumn 2018, regularly scheduled captures included the library's website, campus news and local government websites. Archived web content also included collections created in response to issues and events of local impact, such as natural disasters, on-campus incidents and regional political activity. According to internal account statistics generated via Archive-It, as of summer 2019, the library has captured 5.6 TB of data since its migration from WAS to Archive-It, with approximately 3 TB of said data captured since 2017.

Given this sharp increase in web archiving scope and scale, the digital archivist was tasked with writing a collection development policy for web archives that would codify collecting practices developed organically as web archiving activity grew, and guide the library's web archiving work through future periods of growth.

Anecdotally, this is a common way for institutions to begin. The external colleagues contacted while the digital archivist began work on the policy were

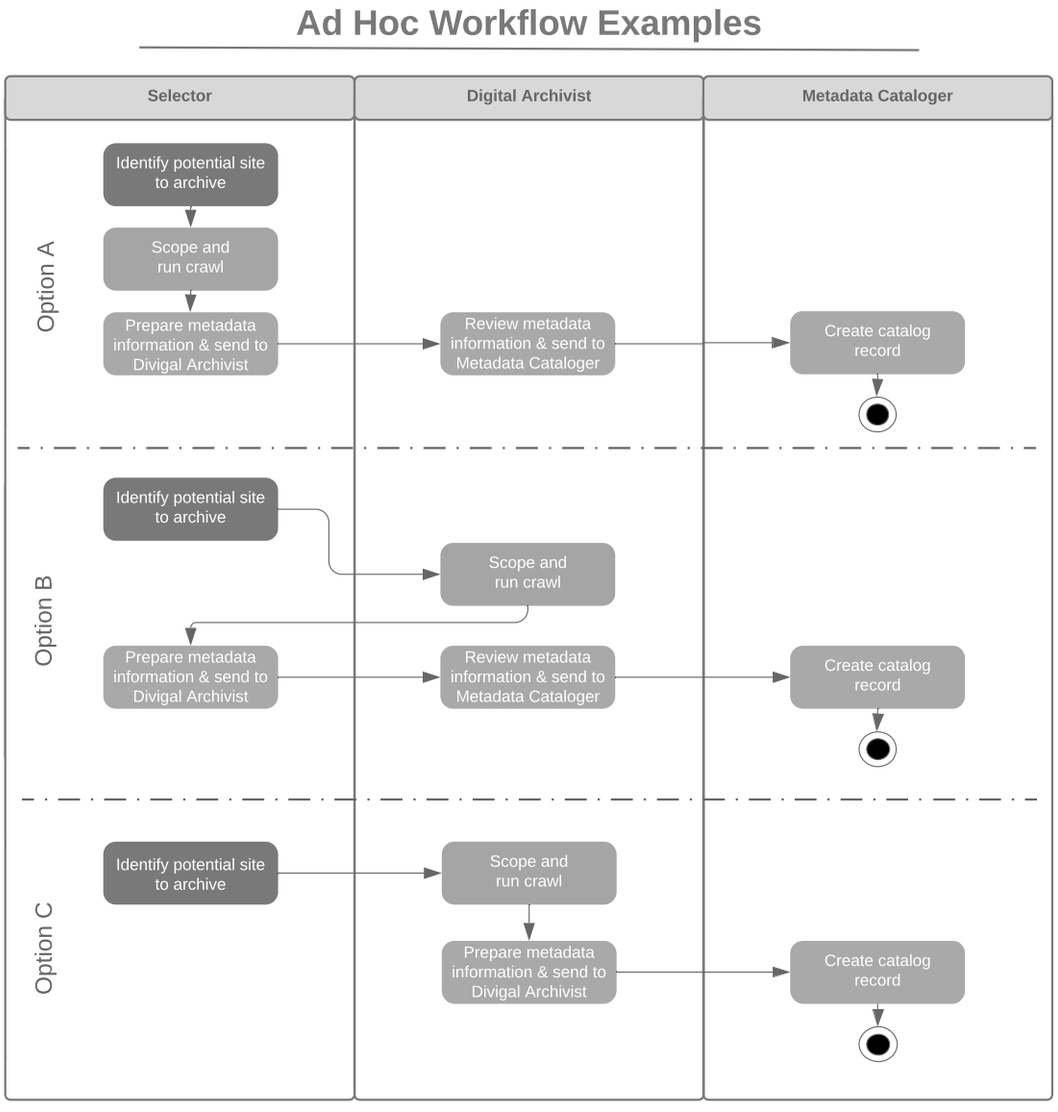

Figure 1: Prior web archiving workflow

largely in similar circumstances; their institutions began collecting web content without an official collection development policy, and either codified policies later, or relied on draft policies to guide web archive collecting. Research supports this assertion; the results of the National Digital Stewardship Alliance (NDSA) 2011 Web Archiving Survey indicated that 32 per cent of the institutions actively engaged in web archiving lacked a collecting policy for this work.[2] Additionally, 56 per cent of the institutions with web archiving programmes in the testing or planning stages lacked a collection policy that involved web archiving.[3] Reports from later iterations of the survey, which were conducted in 2013, 2016 and 2017, devote less focus to institutions with emerging web archiving programmes; respondents to these surveys had more mature programmes, so this question was no longer relevant.

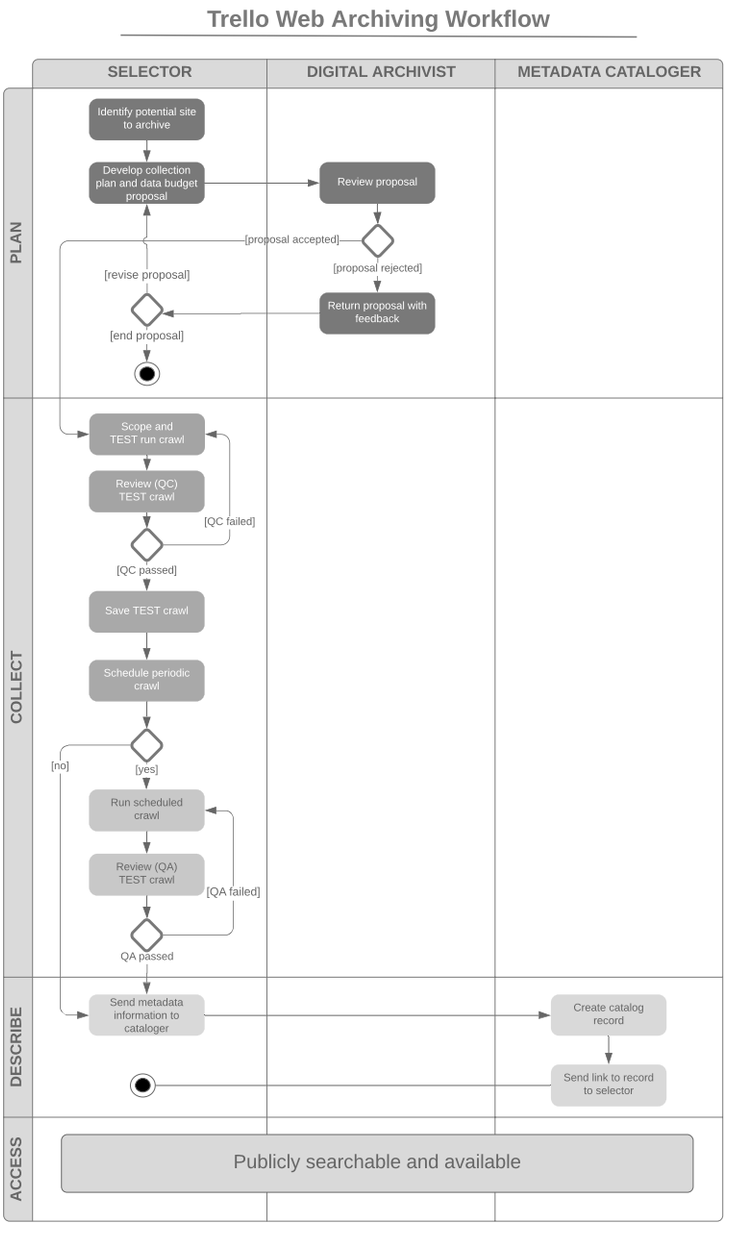

Figure 2: Trello web archiving workflow

Before beginning work on the policy, the digital archivist conducted research to determine institutional priorities and learn about common practices in web archiving. The first step in this research was to establish a list of internal stakeholders, consisting primarily of library administrators and subject selectors responsible for or interested in curating web archive collections. Meeting with these stakeholders would allow the digital archivist to understand the history of the institution's existing web archive collections, and determine priorities for the institution's future approach.

The digital archivist built awareness of the web archiving project and collection

development policy via one-on-one conversations, and by attending relevant committee meetings to describe the library's web archiving work and progress made on the policy. These meetings represented an opportunity for outreach about past and current web archiving efforts, to ensure that subject selectors knew that they could pursue web archiving for websites related to their subject areas, and to solicit feedback to guide future work on the policy. Once a draft was complete, attending committee meetings allowed the digital archivist to present the draft for comments, determine the extent of revisions that would need to be made, and, following these revisions, receive administrative approval to implement the policy.

After meeting with these stakeholders, a list of primary institutional concerns and priorities was developed. This list included:

- a need for more awareness among selectors that web archiving tools were available;
- a need to streamline the collection proposal process to facilitate the timely capture of at-risk or otherwise ephemeral content;
- a need for better documentation of the history of and decisions related to each web archive collection;
- a need to remain within a limited annual data budget for new web archive collecting;
- a need for more established and transparent processes regarding resource allocation, including establishing a list of individuals and entities responsible for resource decisions related to web archiving.

Generally, there was institutional interest in web archiving at all levels. Selectors wanted the processes of creating, augmenting, and maintaining web archive collections to be streamlined, transparent and widely acknowledged throughout the library. Meanwhile, administrators wanted assurance that enhanced collecting and increased awareness remained in balance with a limited annual budget.

ENVIRONMENTAL SCAN OF WEB ARCHIVING COLLECTING POLICIES

In addition to these internal conversations, the digital archivist conducted an environmental scan to understand the landscape of web archiving work, as well as common practices that could be used to shape the policy document. This environmental scan sought to discover how policy approaches change over time, how quickly existing policies became outdated, and the information typically included in collection development policies for web archives.

The environmental scan included publicly available policies from institutions of different sizes and types, policies created by other UC institutions, and a high-level view of trends and common practices in the field. 'Common practices' are distinct from 'best practices', as there are not yet codified best practices for web archive collection development policies, but there are widely adopted norms. However, as these practices are not standardised, the concept of web archiving 'best practices' can change dramatically in a relatively short time. This became evident after reviewing older policies, as well as observing how responses to NDSA web archiving surveys changed over the years.

Publicly available policies were obtained via a search engine, and nine representative policies were selected for inclusion in the environmental scan: three from public universities, four from private universities, one from a small private college and one from a government library. Unpublished draft documents from UC San Francisco and UC Davis were also included in the scan.

After gathering the policies, the digital archivist noted each policy's approach to different aspects of collecting (eg selection, acquisition, access, notification

of or obtaining permission to collect web content, takedown policies, deselection or deaccessioning, institutional priorities), to identify similarities. In the absence of accepted web archiving best practices, comparing policies to identify trends in structure or content allowed for a greater sense of web archiving norms.[4] Out of the many different topics addressed in these policies, only selection criteria appeared in all of them. Issues of access, permissions and institutional priorities appeared in many.

Many of the web archiving policies dovetailed with other collection development policies at their respective institutions. Compared with these more traditional collecting policies, the web archiving policies devoted more attention to subject-agnostic criteria, such as uniqueness; whether the web content is at risk of being lost; whether it complements existing collections; or if the content represents voices and communities that are not well represented in the institution's existing holdings. They also addressed issues specific to web archiving, such as whether capture is feasible with current technology and approaches to clearing legal and technical permissions. Some policies had not been updated in several years, and thus reflected outdated attitudes about web archiving approaches, or relied on defunct tools or methods. These older policies not only showed how the web archiving landscape changed over the last five or ten years, but emphasised the importance of revisiting policies regularly to ensure that they remain useful and relevant. This observation informed the decision to embed a revision schedule into the policy, to prevent the policy from becoming inaccurate or outdated as tools and practices change.

DRAFTING THE POLICY

After the environmental scan, a policy was drafted and shared with internal and external colleagues for comment. The policy was also presented at the Society of California Archivists' Annual General Meeting in April 2019, to inform the broader archival community of the work accomplished and to solicit feedback. The policy was edited to reflect comments and input, and the subsequent draft was approved for use in July 2019.

The policy itself is divided into five sections, each reflecting a set of considerations related to web archive collecting: Collecting, Robots.txt, Access, Preservation and Takedown.

The Collecting section describes how to create a collection, or group of one or more related URLs that address a topic of interest, and explains the library's selection criteria for web content (said criteria are drawn from existing internal policy for selecting materials for digitisation). Both policies address materials that encompass a variety of subjects and formats, and for which the university does not necessarily own the copyright. Selection criteria include: degree of likely research interest; whether the website is UC-owned or affiliated; whether the website or content is at risk; and whether the material addresses an existing collection gap or missing perspective. These criteria currently serve as guidelines to direct collecting, with collections that meet more criteria receiving higher priority, but the library has not yet determined a specific process for final collection decisions. Reusing the library's existing selection criteria for digitisation provided a model intended to be as transparent and easily understood as possible. The criteria were simple, subject-agnostic, already in use, and similar to the web archiving selection criteria employed by peer institutions.

This section also outlines responsibilities and considerations of the library, digital archivist, content selector and, where applicable, content creator. It also outlines the process of requesting that an Archive-It collection be created, notifying content owners of the library's interest in capturing

their content and requesting that they deny permission within a specified period of time if they do not wish for their content to be captured, and the technical capabilities and limitations of Archive-It's web-crawling technology. For example, Archive-It is best-equipped to capture static content, and content that requires little or no user input, so the policy informs selectors that dynamic or heavily interactive content would not be a good candidate for capture.[5] Given these capabilities and restrictions, the Collecting section also states that although the library will make a good faith effort to capture requested websites, it cannot guarantee that this will be possible for every website. This section of the policy serves to manage expectations; by outlining these responsibilities and capabilities at the start, those involved in capturing a website will understand both expectations and commitments.

The Robots.txt section details the library's policy for capturing websites that employ robots.txt exclusions. Robots.txt exclusions prohibit web crawlers (programs whose functions may include search engine indexing, web archiving, or malicious capture of personal information) from functioning for a particular website.[6] These exclusions may be included to block malicious programmes, but, as Archive-It uses similar technology, it is blocked as well.[7]

Archive-It can be directed to ignore robots.txt exclusions during capture of a specific website, so it is important to consider how to address these exclusions in order to capture desired content.[8] The library chose to ask permission before capturing sites with a robots.txt exclusion unless the exclusion is part of the platform hosting the content, or otherwise not under the content owner or creator's control. This policy is slightly more conservative than common practice — according to the 2017 NDSA web archiving survey, 70 per cent of institutions surveyed did not seek permission or notify content owners before capturing content, while 91 per cent of institutions surveyed had never been asked to take down or cease crawling content they had captured without permission.[9] However, the policy does not involve asking for permission to capture all content with a robots.txt exclusion; content such as YouTube videos, for which Archive-It documentation recommends ignoring robots.txt exclusions due to the platform's automatic restrictions on crawling videos and page styling, would not require permission to capture.[10] This way, websites employing robots.txt exclusions can be captured, but the wishes of content owners who do not wish for their content to be preserved are still an institutional consideration.

The Access section describes access decisions that Archive-It requires at the time of capture, in order to ensure that collections and individual websites have the correct metadata. This section of the policy explains the circumstances under which a collection might be listed as 'public' in Archive-It, that is, available to users via Archive-It's public access portal, instead of available only to staff with an Archive-It login.[11] Collections are marked 'private' if they are incomplete, undergoing quality control, under embargo, or otherwise not ready for public access. All other collections or websites are marked 'public'.

The Preservation section acknowledges that Archive-It does not provide local backup or preservation options, and commits to exploring solutions for preserving web archive files.[12]

The Takedown section reuses the existing institutional takedown policy for publicly available digital content. That is, if a content owner believes that their content is available in violation of fair use, or does not want their content archived, they can contact the institution to request that it be removed from public access. The decision to rely on and reuse these policies for web archiving, where applicable, not only underscores the connection between web archiving and existing institutional collecting work, but also

provides a familiar template for parts of the policy that were not web archives-specific. This is intended to forestall potential qualms with archiving copyrighted material, especially as the language in this portion of the policy has already been approved for use within the library.

FUTURE POLICY WORK

As of now, the policy has been approved for use as an approved draft. The library will continue to respond and adapt to its experiences as the policy is implemented, and continue to refine the document during this trial period. Issues to be addressed during this period include determining responsibility for collection decisions, ensuring that the collection proposal and creation process functions as planned, and evaluating the selection criteria to determine both whether they accurately reflect institutional priorities for new collecting, and how each criterion should be weighted in relation to the others.

Overall, the library's experiences highlight general guidance for future work on and revisions to the policy, and to other institutions interested in creating web archive policies of their own. For example, a crucial lesson was the importance of considering a policy's longevity; web archiving policies should be created with consideration of how professional approaches and institutional resource commitments might change over time. Web archiving philosophies change quickly and frequently. Technological affordances, prevalent tools and resource availability are all subject to change, both at an institutional and profession-wide level, and websites themselves can be ephemeral. This means that it is especially important to be able to review and revise a policy regularly, to address changing resources and keep the policy relevant (eg due to reliance on defunct tools, or resource and other changes that affect projected timelines for work completion).

MOVING FORWARD

The new collection development policy provides structure for decisions throughout the current web archiving process, and allows the library to identify and prioritise content to capture and make continuously accessible to users, even after the original website has changed or disappeared. After websites of interest have been identified and approved for collecting, the formal web archiving workflow begins. This process is divided into four steps: plan, collect, describe and access, reflecting the major types of work involved in creating web archive collections and beginning to capture content. Tracking each collection as it moves through this process allows the library to manage and maintain a group of collections that will continue to expand over time. It also ensures that captured web content is complete by facilitating communication between collection stakeholders, and quality control review of collections with different needs and technical problems.

PLAN

The first stage of UC San Diego's workflow involves creating a collection plan, which outlines the steps and states the objectives for capturing the targeted content. Questions asked and answered during the planning phase generally include:

- Does the collection meet the selection criteria detailed in the collection development policy?
- How much data must be captured?
- What type of crawl would work best (eg single page, single page plus external links, entire domain, entire domain plus external links)?
- How often should the crawl occur?
- How many times will a given URL be crawled? If crawls are ongoing, when will the decision to crawl the URL be re-evaluated?
- Who are the stakeholders in this collection? Site owners, subject selectors?

- What rights does the institution hold?
- Will the captured collection be immediately accessible to the public? If not, how long will it be private, or when will it be reviewed for release?
- What resources are available for capturing and maintaining the collection (eg staffing, time, money)?
- Are there limitations which should be placed on the collection in order to use resources effectively (eg impose a data cap, or exclude specific types of documents or links)?
- How will users find and access the collection?
- Will the captured content resemble the original website when users access it?

Trello was implemented for staff engaged in web archiving work to track and manage workflows, communicate with each other, and record decisions made by the selector, the digital archivist and metadata librarian. Trello (https://trello.com) is an online project management tool that is free to use with limited functions, or fully available with a subscription. Trello was selected because it provided a visually intuitive, flexible, shared environment to monitor the progress of active collections and facilitate communication and collaboration between subject selectors, metadata librarians, the digital archivist, and other staff assisting in the process. Trello's interface is based on project boards that are used to track stages of a project, and are subdivided into sections, or lists, representing each step and collecting associated tasks, or cards (Figure 3). The institutional Web Archiving Project board is shared with staff engaged in web archiving work.

The Web Archiving Project board is divided into lists corresponding to each phase of creating a new collection. Each list is divided into cards detailing outstanding tasks, questions, issues and decisions. Once a new collection is proposed, a corresponding card is added to the Potential Collections list on the board. On the card is recorded the basic collection information needed to create the collection within Archive-It, set parameters for the crawl, and create collection-level metadata. The standard descriptive information collected at UC San Diego for a new collection includes collection title, a brief description of the collection, creator name(s), date range of

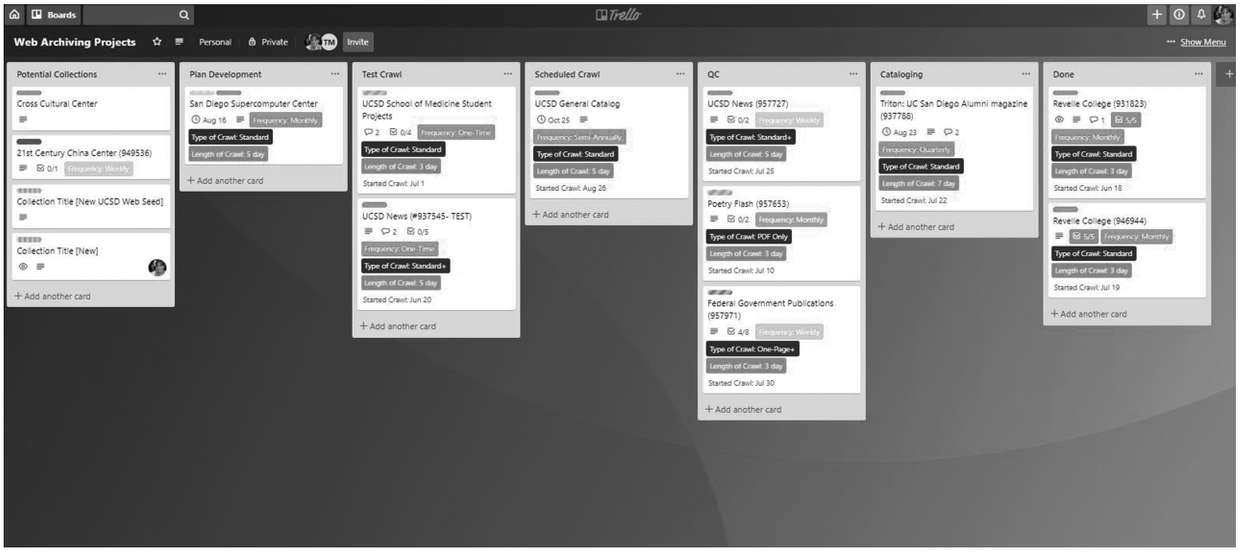

Figure 3: Trello project board

content, starting date of capture, language(s) of the website, publisher, rights and access information (will the collection be publicly accessible, or only available to staff with an Archive-It login), geographic and topical subjects, name of the librarian or department that requested the collection, and the seed (ie starting point) URLs to be crawled (Figure 4). Additional crawl information recorded on the card includes crawl frequency, crawl extent, and either an end date for crawling, or a date to re-evaluate an ongoing crawl.

As a new collection moves through the process, staff move the corresponding card from one list to the next to monitor the

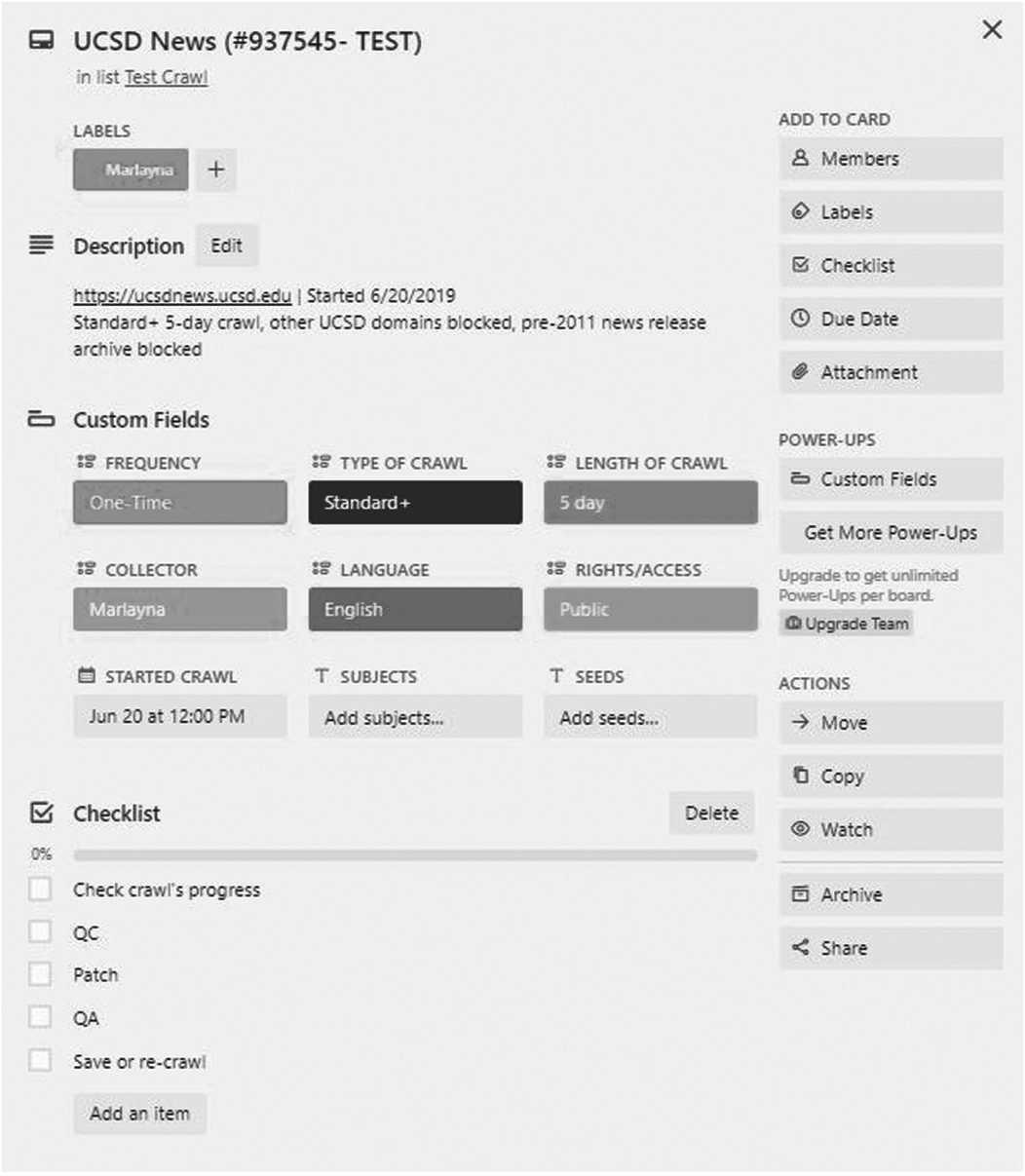

Figure 4: Trello card

collection's status and make any changes clear. When more granular tracking is needed, staff add checklists to cards to show work completed.

COLLECT

The Collection phase includes deciding the scope of the content to be captured, testing to ensure that this content is captured as expected, and quality control or assurance review.

Scoping

Defining which content is included in or excluded from capture, or scoping the capture, is an essential part of shaping the outcome. As the library primarily uses Archive-It, which employs a web crawler, options for scoping a capture can include how far the crawler will go from the original URL (eg capturing a single page or an entire domain), how frequently captures occur, and limits on data captured and time spent crawling. Limiting data captured, and carefully considering time spent crawling, are particularly important when scoping captures for social media sites and pages with embedded media. A lack of careful scoping can result in capturing large amounts of unwanted data.

Most of the crawls at UC San Diego are scoped to capture only the immediate domain URL. When the site included related content beyond the organisational domain, an expanded level crawl was used. In the instance of capturing the campus news site, which linked to a number of local news sources, the institution expanded the crawl to include pages external to the campus.

The frequency at which a URL is captured is also a factor to consider during scoping. Questions used to determine frequency include how frequently content of interest changes or disappears from the page, and whether it is important to capture every change to the available content. For example, in the case of federal government websites, daily or weekly changes may be significant. On the other hand, the general catalogue of campus courses does not change significantly during the school year, and can be captured less frequently.

Testing

In addition to typical crawls, Archive-It provides a test crawl option, which allows staff to capture new URLs on a trial basis without impacting the data budget. After a test crawl is complete, staff can review the crawl and decide whether to delete or save it, depending on whether the crawl captured the desired content. Once saved, the crawl is counted against the annual data budget. If left unsaved for 60 days, the crawl will expire, and be deleted automatically. If a test crawl captures too much or too little information, the scoping rules can be adjusted before beginning another test crawl.

When starting a new collection or adding new seeds to an existing collection, the library employs test crawls to determine whether the planned scoping rules and crawl parameters are adequate. Once a test crawl is complete, staff perform quality control to determine whether the test crawl produced the anticipated outcome. If so, the test crawl is saved, and, for ongoing crawls, an automated crawl is scheduled to run at the desired frequency.

Quality control/quality assurance

When a crawl is complete, staff review the resulting crawl report and captured content to determine whether the crawl was successful. Due to the shifting landscape of web content and diversity of intended outcomes, there is currently no standardised or recommended web archive quality control checklist available based on best practices. UC San Diego developed a basic checklist of elements to review and measure the effectiveness of captures, which is added to

each collection's Trello card and used to track quality control progress. Occasionally, items are added to the standard quality control checklist if a collection has content that is difficult to capture. After a test crawl of a new URL or collection, staff review the capture to determine if there was adequate time allowed for the crawl, if the content is complete, and if links function as expected. UC San Diego's checklist includes:

- Did the crawl stop too soon due to an insufficient time limit? If so, what types of content were not captured?
- Does the capture look like the original pages?
- Is any expected content missing?
- Do links connect as expected?
- How much data was collected? Does this match the apparent size and complexity of the page or website?
- Which host websites or file types represent the largest proportion of captured data? Is this appropriate or expected?
- What types of documents were collected? Appropriate or expected?

When a crawl is not complete or content is not captured as expected, staff review the report and determine which scoping parameters to adjust, remove or add before beginning another test crawl. Changes are noted on the corresponding Trello card.

The crawl report details elements that may need attention. For example, one crawl regularly captured 17–20 GB of data weekly, yet never finished the crawl in the allotted time. On first review of the output, the captured pages displayed correctly, but many documents from the domain URL were still queued for capture. Initially, it appeared that the website simply included many large media files. Upon closer review, staff discovered that much of the captured data represented unrelated social media and external source content (*New York Times*, *Wall Street Journal*, etc). Additionally, Facebook and Twitter content linked from the original page had been captured in several languages, which was well beyond the intended scope of the crawl. In response, the scope was adjusted to limit social media capture for this collection. With these changes in place, the adjusted crawl began to capture between 500–800 MB weekly. However, while parameters for capturing social media links were changed, parameters for external content were not. As some external publisher content was expected in the crawl, refining the parameters to allow some, but not all of this content would be challenging. In this instance, as the amount of data captured was significantly reduced, the decision was made to allow the extra external content rather than spending time to identify and exclude unwanted content.

Quality control on completed crawls is necessary to ensure proper collection. Following a quality control checklist and reviewing crawl reports to confirm that all expected content was collected can be time- and labour-intensive, but ensures that, if any content is missing, there is still an opportunity to recapture the missing information before the original website changes. The content captured early in the library's web archiving project did not go through a rigorous quality control review. Later, it was discovered that many of the captures were only partially complete. As the captures were several years old, the websites in question had changed or disappeared, and the missing content could not be recovered.

Fortunately, for recurring captures, quality control becomes less burdensome over time. While initial crawls consist mostly or completely of new data, and therefore require careful and complete review, the amount of new data per crawl decreases as the static portions of a page are captured. Once a recurring crawl is more established, and very little new content is captured in each crawl, quality control or quality assurance can become less frequent, and consist of spot-checking the captured website instead of reviewing every page. However, some degree of quality assurance

is still necessary to ensure that sudden or drastic changes to a website (eg URL changes, website redesigns) do not affect crawl success. For this reason, the library found that spot-checking established crawls remained necessary to confirm that Archive-It was still collecting the appropriate content.

DESCRIBE

Once a web archive collection has been created and made publicly accessible, descriptive metadata help users to discover it. In the 2018 OCLC Research Perspectives report, 'Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group', the authors state 'the context within which a website or collection was archived is of vital importance to end users for understanding potential uses of the content'.[13] The report identifies essential descriptive elements, defines them, and provides a crosswalk with other descriptive standards commonly used for archival collections. The library adds seed- and collection-level metadata to content on Archive-It to improve discoverability through the public user interface, which allows users to search on descriptive metadata. These metadata are also used in the local online catalogue.

ACCESS

The library currently provides access to web archive collections via Archive-It's public user interface, as well as by including links to archived web content in its existing local resources. The library creates local catalogue records for each collection with subject headings and other pertinent tracings (see, for example: http://roger.ucsd.edu/record=b9626849~S9). Library cataloguers also add links to web archive collections in related finding aids. That is, if permission to archive a donor's website is included in the donation, purchase or transfer of analogue materials to Special Collections and Archives, the archived site is included in the finding aid. Other web archive collections are included in associated finding aids with the EAD tag. Linking web archived collections in the field (see, for example: https://library.ucsd.edu/speccoll/findingaids/sac0001.html) helped establish a high-profile location for users to look for archived websites. Another potential access point for web archive collections could be links included in relevant subject and course-specific resources developed by campus librarians. The library will continue to evaluate access to and use of archived web content, and look for additional ways to bring such content to researchers' attention.

CONCLUSION

Policy and practice offer a framework for structuring web archiving work, but should remain flexible to allow institutions to adapt to changes in content, technology and user needs. This perspective will continue to inform the library's web archiving efforts and next steps. Progress has been made toward addressing all of the institutional concerns identified during work on the web archive collection development policy, but none of these considerations have been resolved fully. Instead, the library is using this initial period after the policy's approval as an opportunity to evaluate how the current version of the policy works in practice, and determine changes that may be required over time. For example, the library's existing collecting activity is within the 2019–2020 data budget, with room for new collecting. As outstripping the current budget is not an immediate problem, the library has time to observe the policy's effects on new collecting, and evaluate whether the approach to data budget allocation outlined in the policy is reasonable and sustainable in practice. This and other 'wait and see' elements

of the policy represent opportunities to take incremental steps toward resolving institutional concerns, and ensure that the library is not without a policy while these issues are considered in more depth.

In addition to ongoing work to address institutional concerns and priorities, and respond to changes in how the internet is used and understood, the library will explore ways to preserve captured content locally to avoid reliance on backups outside institutional control. Currently, UC San Diego is wholly reliant on Archive-It to host and maintain captured content, but the library is considering other options, which could include depositing backups in Chronopolis, the institutional digital preservation repository. As the library moves toward maintaining local copies of its web archives, additional types of data will be captured and added to collections, including technical and administrative metadata, some of which may be captured automatically by the harvesting tool. Local preservation would not replace functions or services provided by Archive-It, but would provide a backup in case of emergency.

Another area for future work is ensuring that the library can commit sufficient time and resources to support both current and future web archiving activity. As the library's web archiving activity continues to grow, there will be a need for increased staffing support, diverse harvesting tools to capture a variety of technologically challenging content, and innovative approaches to anticipate and respond to change. These represent substantial institutional commitments, and will require flexibility and creativity to ensure that these commitments do not outstrip institutional capacity.

References

1. Lack, R. (2015) 'Announcing a new partnership: California Digital Library, UC Libraries, and Internet Archive's Archive-It Service', available at: https://www.cdlib.org/cdlinfo/2015/01/14/announcing-a-new-partnership-california-digital-library-uc-libraries-and-internet-archives-archive-it-service/ (accessed 28th June, 2019).
2. NDSA Content Working Group (2012) 'National Digital Stewardship Alliance Web Archiving Survey Report', available at: http://www.digitalpreservation.gov/documents/ndsa_web_archiving_survey_report_2012.pdf (accessed 10th December, 2018).
3. *Ibid*.
4. Maches, T. and Stine, K. (2019) 'Web Archiving CKG "Birds of a Feather"', available at: https://wiki.library.ucsf.edu/display/UCLCKG/2019-05-21+UC+DLFx+Birds+of+a+Feather (accessed 28th June, 2019).
5. Lohndorf, J. (2019) 'Known web archiving challenges', available at: https://support.archive-it.org/hc/en-us/articles/209637043-Known-Web-Archiving-Challenges (accessed 5th August, 2019).
6. Praetzellis, M. (2018) 'Avoid robots.txt exclusions', available at: https://support.archive-it.org/hc/en-us/articles/208001096-How-to-avoid-robots-exclusions (accessed 5th August, 2019).
7. *Ibid*
8. *Ibid*.
9. Farrell, M., McCain, E., Praetzellis, M., Thomas, G. and Walker, P. (2018) 'Web Archiving in the United States: A 2017 Survey', National Digital Stewardship Alliance, available at: https://doi.org/10.17605/OSF.IO/3QH6N (accessed 10th December, 2018).
10. Praetzellis, M. (2019) 'Archiving YouTube videos', available at: https://support.archive-it.org/hc/en-us/articles/208333753-Archiving-YouTube-videos (accessed 5th August, 2019).
11. Blumenthal, K.-R. (2019) 'Controlling access to your web archives', available at: https://support.archive-it.org/hc/en-us/articles/208334003-Controlling-access-to-your-web-archives- (accessed 20th July, 2019).
12. Blumenthal, K.-R. (2019) 'Partner guide to downloading Archive-It Data', available at: https://support.archive-it.org/hc/en-us/articles/209643793-Partner-Guide-to-Downloading-Archive-It-Data (accessed 20th July, 2019).
13. Dooley, J. and Kate, B. (2018) 'Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group', available at: https://doi.org/10.25333/C3005C (accessed 9th August, 2019).