THEME 7

The Web Archiving Life Cycle Model

by **Kristine Hanna**
Director of Archiving Services at the Internet Archive
https://archive.org/about/bios.php

INTRODUCTION

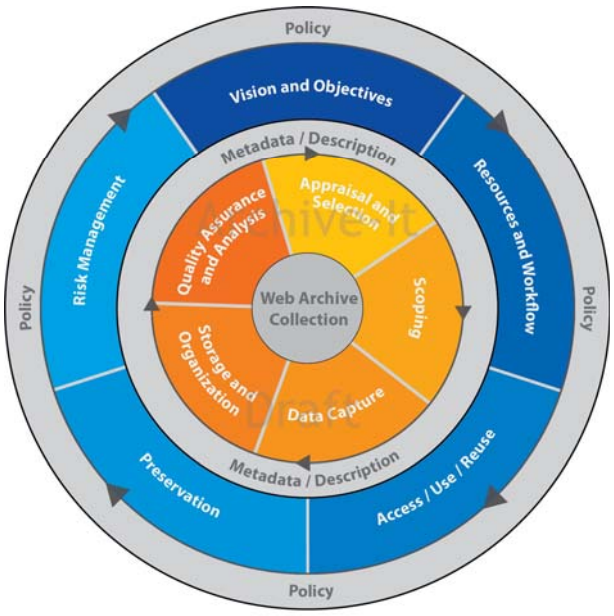
The technological tools for archiving the web have been evolving steadily for more than a decade. However, best practices and a common model of web archiving have yet to emerge. The Web Archiving Life Cycle Model is an attempt to incorporate the technological and programmatic arms of web archiving into a framework that will be relevant to any organization seeking to archive the web. Archive-It, the leading web archiving service in the community, developed the model based on its work with memory institutions around the world.

The Internet Archive has been archiving the web since 1996. In 2002, the Internet Archive released Heritirx, the open source web crawler. In 2009, the Heritrix crawler's file output, the WARC file, was adopted as an ISO standard for web archiving, demonstrating both the prevalence of active web archiving programs and the importance of the web crawler itself. In early 2006, the Internet Archive launched the Archive-It web archiving service (www.archive-it.org) with 13 pilot partner institutions. Archive-It is a subscription web archiving service that helps partner organizations harvest, build, and manage born digital collections. The partner base has steadily expanded since its launch, with 237 partners in 46 U.S. States and 15 countries, as of January 2013.

Despite growth in the number of web archiving programs, many institutions still struggle with developing best practices and methodologies to accomplish their goals. This difficulty partially stems from constantly evolving web technology, which can make it difficult to archive certain types of content effectively. Conflicting and evolving policy decisions from various stakeholders as well as shifting organizational structures and job responsibilities pose further obstacles to establishing best practices. Additionally, some organization stakeholders have not fully adopted the belief that web archiving is crucial to their digital preservation activities; and as a result, funding remains limited or non-existent.

Archive-It is a Web archive service that follows the Web Archiving Life Cycle Model

In order to address the lack of standard best practices and to increase awareness of the importance of web archiving as fundamental to digital preservation, the Archive-It team developed the Web Archiving Life Cycle Model (WALCM). This model is based on the team's experiences as well as lessons learned by countless partner institutions, including in-depth case studies from six of those institutions. The WALCM is an attempt to represent common workflows and create a measurable model for organizations to reference in order to create or improve their web archiving programs.



WEB ARCHIVING LIFE CYCLE MODEL

DEVELOPING THE WEB ARCHIVING LIFE CYCLE MODEL

The Archive-It team developed the model organically, using feedback and lessons learned from organizations archiving the web. The majority of these organizations use the Archive-It service to archive web content for their organization. These partner institutions provide feedback based on their use of the service, communicated to the Internet Archive through email, phone calls, and in-person conversations at conferences and partner meetings.

Additionally, more formal feedback comes through partner presentations at conferences, surveys designed by Archive-It staff, as well as partner presentations at conferences and documentation relating how they and their colleagues meet the challenges of web archiving.

The Archive-It team drafted the first iteration of the Web Archiving Life Cycle Model. This preliminary design was circulated to a subset of Archive-It partners who provided feedback on missing or superfluous elements and on the model's visual presentation. Next, the Archive-It team incorporated this input into a more graphically pleasing model that was sent to all Archive- It partners for general feedback. This feedback shaped a further re-design and the resulting version of the model discussed in this paper. The information in this paper is also based on in- depth email exchanges and phone interviews with six Archive-It partners between April and July 2012. These institutions are: Columbia University, University of Alberta, Montana State Library, State Library of North Carolina, State Archives of North Carolina and Creighton University. Information in this paper also comes from a survey of Archive-It partners conducted in August 2012.

THE MODEL EXPLAINED

The model is an attempt to distill the different steps and phases an institution experiences as they develop and manage their web archiving program. Although the model is broken down into individual steps, each action is not discrete. Archive-It considers the steps and phases to be related, with a significant amount of overlap between them.

The shape ofthe rnodel is circular to suggest the repetitive nature ofthe stepe inthe life cycle. As users move through each step, theyeventuallyfmd themselves back at the beginning, or repeating certain stepe, depending on their tasks. For example, the process can restar!when an institution adds new websites toan existing collection or ereatesan entirelynew collection. The rnodel includes circles within circles lo suggest these repetitive cycles within the bigger process.

The outermost level of the life cycle is the policy band. Almost every aspect of web archiving involves some sort of policy decision. These policy decisions may involve developing a new policy specific to web archiving or the adaptation of an existing policy to new situations. By encompassing the life cycle steps with a policy band, the model visually represents the ever- present nature of policy making. In a second band, the model similarly represents metadata and

description. Archive-It chose to incorporate metadata as a band rather than as a segment of the wheel to emphasize that creating, importing, and exporting metadata can be done as part of a number of other activities in the lifecycle.

The blue circle just inside the policy band represents the high-level decisions an institution faces as it sets up and manages its web archiving program. The individual steps are briefly defined as follows and will be discussed in more depth later in this paper.

- Vision and Objectives: here institutions clarify the goals of their web archiving program.
- Resources and Workflow: institutions review their available resources including financial, expertise, staffing, potential collaborators and others in order to determine how to proceed with developing or changing their web archiving program.
- Access / Use / Reuse: institutions make decisions about whether and how to provide access to their collections and monitor how the content is used by their patrons.
- Preservation: institutions make decisions about how they want to preserve the data they collect in their web archiving activities. This includes WARC files, metadata, and X.
- Risk Management: When institutions consider their approach to risk in creating a web archiving program, they look at copyright and permissions as well as access.

The inner orange circle describes the day-to-day tasks involved in the business of archiving the web. These tasks include the following.

- Appraisal and Selection: institutions decide specifically which websites they want to collect.
- Scoping: institutions may opt to archive portions of a website, whole sites, or even entire web domains.
- Data Capture: Here, institutions fine-tune how they want to capture their data through decisions about crawl frequency and types of files to archive or not archive. The scoping and data capture phases of the lifecycle often overlap as they involve similar activities and decisions.
- Storage and Organization: This step includes a temporary or long-term storage plan for the archived data. For some institutions, the storage and organization phase of the lifecycle might also constitute their preservation activities.
- Quality Assurance and Analysis: Here, institutions review what they have archived and the level to which the resulting collection satisfies the goals they set out at the beginning of the life cycle.

At the center of the lifecycle is the collection itself, the archived web content. This data is the end result of all preceding steps, and it is what will be preserved. Capturing and preserving collections of data is at the heart of all web archiving activities and is therefore the center of the model.

WEB ARCHIVING LIFE CYCLE MODEL: THE OUTER CIRCLE

1. The Outer Circle

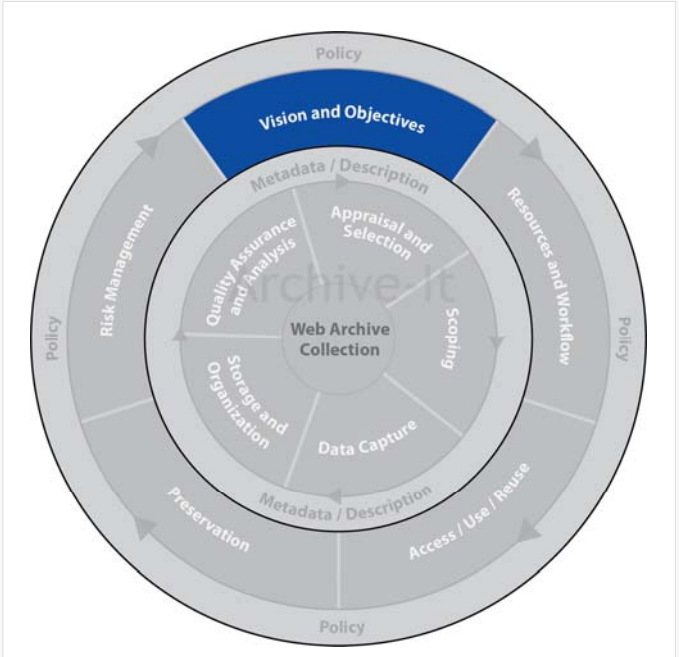
1a. Vision and Objectives

To determine a vision and objective for web archiving, an institution must ask itself why it is choosing to archive the web, what it wants to accomplish in

doing so, and how these steps relate to the institution's broader mission. This step in the cycle primarily occurs as institutions initially plan their program; however, institutions do tend to re-visit and re-define their web archiving objectives throughout the life of the program. These periods of re-examination may result from a specific stimulus, such as a change of resources, or may be an ongoing question considered along with and in relation to their other collection policies.

Memory institutions choose to archive the web for many different reasons, depending on their own institutional mandates as well as the objectives of their stakeholders. Some institutions choose to archive the web because they believe that specific web content is at risk of disappearing and therefore needs to be captured and kept accessible; particularly in the case of rapidly changing spontaneous events, like natural or manmade disasters, political uprisings, and memorials for public figures. Other institutions have mandates to archive specific publications that are only available in digital formats, such as university course catalogs or state agency reports and publications. Additionally, some institutions have legal mandates to archive all official records produced by the institution within their domain, constructing an historical record of their institution's web presence over time. Still other institutions view web archiving as an extension of their overarching collection development policy or their digital preservation programs, and they may archive web content that enhances or supplements the topics alfeady being emphasized in their traditional collecting activities. Researchers and academics realize the important of creating a thematic/topical web archive on a specific subject or topic that includes different perspectives and social commentary from the increasing influence of social media sites, including tweets, blogs, posts and comments. Some institutions have a set of different goals and as a result set up multiple collections to achieve each objective. Regardless of the specific vision for each web archiving program, this vision shapes many of the policies and decisions made in later steps of the web archiving lifecycle.

As one example, Columbia University Library has been working with Archive-It since 2008. The library collects web content in several areas. First, the library captures the Columbia University web domain in coordination with University Archives. Second, the library has several other collections built around specific themes and topics. These topics include global human rights, historic preservation and city planning, and New York City religious institutions. These born-digital collections complement and supplement the library's existing physical collecting activities. Columbia describes its overarching goal in web archiving as "believ[ing] that freely available web content [is] an increasingly important source of content necessary for current and future research that [is] not yet integrated into academic library collection development models." (personal correspondence with Alex Thurman and Tessa Fallon, May, 2012).



THE OUTER CIRCLE - VISION AND OBJECTIVES

Similar to Columbia University, University of Alberta also realized that the university was not capturing born-digital material and that it needed to include web archiving in its digital preservation strategy. However, the university did not start out with such a clear vision. Originally, the University of Alberta inherited over 80 websites from a non-profit organization that lost its funding. Realizing that hosting these websites would be resource intensive, the university took an "archiving" approach, which they felt would be a more sustainable way to take custody of the content. University of Alberta thus began using the Archive-It application to complete this project. Their first year with Archive-It (2009) was largely based on the websites inherited from the dissolved non-profit organization (personal correspondence and conversation with Geoff Harder, June 2012).

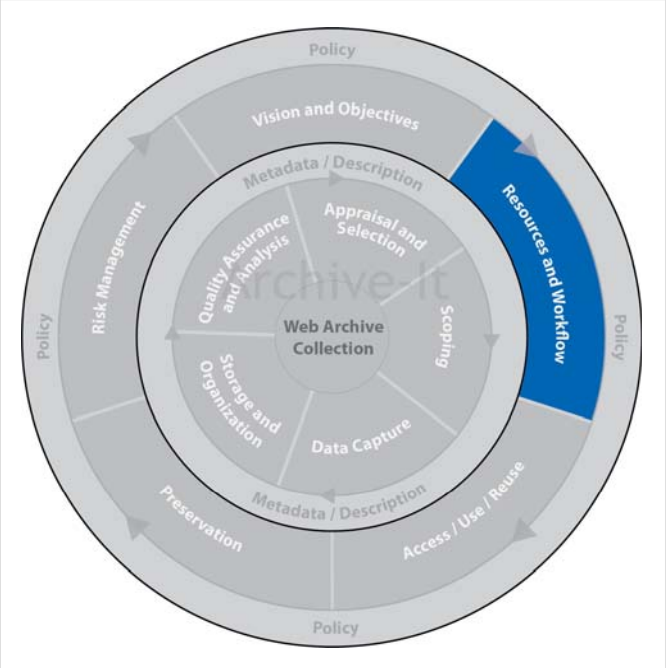
Starting in 2010, the University of Alberta began using Archive-It as a broader collection development tool. The development of national web archiving programs is not as strong in Canada as it is in some other countries. To help fill this gap, the university library has begun collecting in earnest in several areas, including but not limited to: Canadian prairie politics and economics, government documents, grey literature for business and health sciences, circumpolar studies, and provincial education curriculum materials. In this way, the vision of their Archive-It program matches their collection development policy for their non-digital collections. Two of their big issues moving forward relate to refining their discovery strategy and improving the visibility of their collections. They are particularly interested in out how to most effectively provide access to their web archives alongside other digital collections Because the university is concerned with digital scholarship, they want to make sure researchers are able to use their web archive collections just as they now use other resources (personal correspondence and conversation with Geoff Harder, June 2012).

Institutions archive Web content for many different reasons and aims according to their digital conservation strategy

Montana's state library offers an example of a different institutional vision. The Montana State Library (MSL) web archive seeks to archive state documents, which are now often only available online. Their objective is to "meet the information needs of state agency employees, provide permanent public access to state publications, support Montana libraries in delivering quality library content and services, work to strengthen Montana public libraries, and provide visually or physically handicapped Montanans access to library resources" (personal correspondence with Beth Downs, James Kammerer and Chris Stockwell, May, 2012). A Montana State Library staff summarizes the library's reasons for archiving the web: "With the precipitous decline in the submission rate for print publications and an inverse, exponential rise in the rate of web based publishing, Archive-It has completely supplanted the historic state depository library tradition of acquiring and distributing print state publications one at a time" (personal correspondence with Beth Downs, James Kammerer and Chris Stockwell, May, 2012). At the beginning of their subscription in 2007, Montana set up one policy to govern most aspects of their web archiving program, including selection criteria for what to archive, crawl frequency, and outreach. Interactions between Archive-It and MSL since 2007 indicate that this approach has been successful and is meeting the objectives of the state library.

1b. Resources and Workflow

The resources and workflow phase of the lifecycle can be interpreted in several ways. In the context of the model's outer circle, institutions examine the resources and workflows that can be leveraged to create or maintain an entire institution's web archiving program. In this way, resources and workflow can be considered similarly to "policy", as they can be applied in multiple areas of the web archiving life cycle model. Resources and workflow should also be considered as general program management terms that can be applied to each of the elements in the model's inner ring. In this context resources and workflow become part of the day-to- day activities



THE OUTER CIRCLE - RESOURCES AND WORKFLOW

of web archiving. For example, how much time can an institution spend reviewing their crawls or how many people should add websites to the Archive-It application? Subsequent sections of this paper will discuss specific management workflows in depth.

One of the key resources organizations have at their disposal is their staff. In-depth discussions with several Archive-It partners in the spring and summer of 2012, as well as a survey conducted by Marquette University reveal some comprehensive data regarding the staffing models in place at a wide range of Archive-It partner institutions (Sweetster, 2011).

Of the 37 institutions that responded to the Marquette survey, one-third have two or more individuals involved with Archive-It, and over 25% have four or more individuals involved. The survey also found that half of the responding institutions spend less than 1 hour per week working with their Archive-It accounts, and 44% spend 1-5 hours per week working with the application. The Marquette survey also asked respondents to describe the types of individuals working within Archive-It. Table 1 displays these findings; please note that respondents could select more than one staff grouping, so results do not sum to 100%.

Discussions with the six Archive-It partners highlighted in this paper revealed similar results to the Marquette survey. The partners provided details about their Archive-It staffing, including the number of staff and nature of their work. The results are summarized in Table 2. These results share another similarity with the Marquette survey results. Most of the staff tend to come from the library or archives (this author is inferring that subject specialists and metadata curators are part of a library staff), with additional involvement from information technology staff and students.

In addition to staffing, the resources and workflow in this model also encompass how institutions manage other resources. For example, Columbia uses an internal database to track any information that cannot be included in the Archive-It application, such as administrative information and permissions data from sites they have contacted. Another example is the decision to collaborate and divide management of the web archiving program between the State Library of North Carolina and the State Archives of North Carolina. The two institutions manage a single collection of state government agency websites. In dividing up the day-to-day work, the two agencies have several well-established workflows, which they have developed since they first began using Archive-It in 2005. The state library and archives alternate responsibility for conducting the crawls, and both institutions perform quality control of the data harvested. The individual staff members have turned over throughout the years; however, despite this

Archives Staff	64%
Library Staff	42%
Digital Projects Staff	30%
Information Technology Staff	8%
Other (such as students or "web team")	8%

TABLE 1:
Type of staff at an institution working with Archive-It

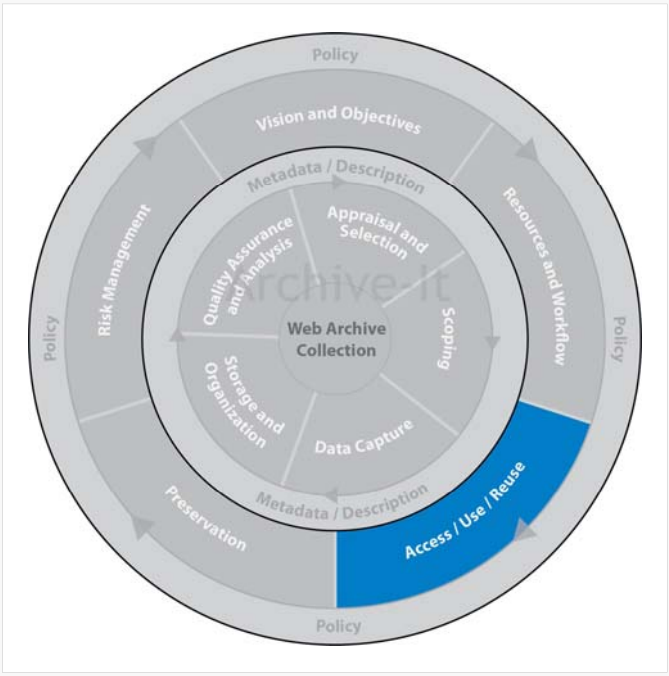
INSTITUTION	NUMBER OF STAFF INVOLVED	STAFFING DETAILS
Columbia University	1 + some involvement from other staff	Currently (2012) one web curator runs crawls, scopes seeds and manages the Archive-It account, although they have had two web curators in the past. Students, the metadata curators and web programmer also use different parts of the application on a more limited basis.
Creighton University	1	Creighton has one full-time Archivist, and one of his responsibilities is to administer Archive- It; he also gets a small amount of help from others at the Library
University of Alberta	1 lead technical person, with up to 40 people actively logging into the application	Alberta has a very large network of individuals actively using Archive-It, many of whom are subject specialists.
Montana State Library	3	The most active users are the state publications librarian (who oversees the program), the metadata cataloger, and the library systems programmer/analyst, who handles technical issues.
North Carolina State Library and Archives	4	Management of Archive-It is evenly split with two representatives from the state library and the state archives.

TABLE 2:
Number and type of staff working with Archive-It

turnover, the institutions have found that their partnership has been an "easy collaboration to maintain" (personal correspondence with Kelly Eubank, Lisa Gregory, Kathleen Kenney, and Rachel Trent, June 2012).

Of the six Archive-It institutions highlighted in this paper, the University of Alberta has the largest web archiving program in terms of staffing. Alberta began using Archive-It with a small team of several individuals in 2009, and the team has since grown to over 22 people actively contributing to the program. They have also incorporated a number of subject specialists into their work. Additionally, the team has a government documents librarian and a metadata librarian involved in the application. A representative from information technology supports these individuals and filters their questions to Archive-It staff at Internet Archive. At a higher level, the library has a "born digital working group" composed of staff from around the library. This group, composed mostly of individuals from collection development, helps shape web archiving policy in general and use of Archive-It in particular. Additionally, an Archive-It users group, which has a broad membership base, builds and shares knowledge about Archive-It.

Unlike the University of Alberta, Creighton University only has one archivist who manages the university's Archive-It subscription and also initially championed it as a necessary resource. David Crawford learned about Archive-It at the 2008 Society of American Archivists conference and worked to build support for setting up an Archive-It subscription at Creighton. Eventually, he received a donation from a board member to initiate their web archiving program by funding a subscription to Archive-It. Using a tool like Archive-It allows Crawford to accomplish his goal of archiving the university's web presence, which he would not have been able to do on his own due to a lack of in-house expertise (conversation with David Crawford, July, 2012). Crawford's experience of having to build support for web archiving on his own seems consistent with interactions Internet Archive has had with other small institutions like Creighton. Smaller institutions often take longer to get their program up and running due to fewer staffing and fiscal resources. Some smaller colleges and universities have formed consortiums to support their web archiving programs in order to expand their pool of resources for web archiving (see for example the Tri-College Consortium of Bryn Mawr, Swarthmore and Haverford http://www.archive-it.org/organizations/74) one of the original Archive-It pilot partners.



THE OUTER CIRCLE - ACCESS/USE/REUSE

1c. Access/Use/Reuse

Establishing access, use, and reuse policies is vital to a successful web archiving program. Institutions consider whether and how they want to provide open access to their web archives, if and how to promote the collections, as well as how to govern public use of the material. Managing these processes is the primary goal of the access / use / reuse phase of the web archiving lifecycle.

Part of the creation of an access policy will include choosing the specific technology or tool to provide access to the archived web pages. However, for the purposes of this model, we instead consider the higher level policy decisions around access. This is in part due to the fact that all of the individuals interviewed for this project access web archives using Wayback software, the open-source viewing tool that allows the public to browse archived web pages just as they would experience a live web page.

The majority of Archive-It partners have their archived content publicly available; although an increasing number are requiring some content to be kept restricted for a period of time – either a specific URL, an individual collection or their entire account with multiple collections. And the Archive-It team is starting to see more requests for content to be restricted by IP address to enable reading rooms in university libraries to have more flexibility around access. (Note: the service expects to have this capability in April2013).

Archive-It partners can refer their patrons to the Archive-It website for collection access http://www.archive-it.org or they can link to their collections from their own site through a search box or links to the Wayback software. Both approaches work for partners depending on

their access needs. For example, the State Library of North Carolina and the State Archives of North Carolina provide access to their Archive-It collections from their own website. They have created a robust portal, which provides information about web archives for the public and information professionals, as well as instructions for using the web archives (http://webarchives.ncdcr.gov/). Creighton University, on the other hand, has taken a different approach. They refer their patrons to the Archive-It website for access to the collections and do not provide access from their institutional website. In David Crawford's words, they prefer their patrons to be "self directed" (conversation with David Crawford, July, 2012).

Institutions need to analyse how to take advantage of their resources and funds to create or maintain their Web archiving programmes

Like the State Library of North Carolina and the State Archives of North Carolina, Montana's State Library also created a portal on their own website that provides access to their Archive-It collections (http://msl.mt.gov/For_State_Employees/connect/default.asp).In addition to providing access to data collected using the Archive-It service, Montana State Library extracted older web pages dating back to 1996 from the Internet Archive's general web archive. These web pages are accessible from the portal along with the Archive-It data, which dates back to 2006. The

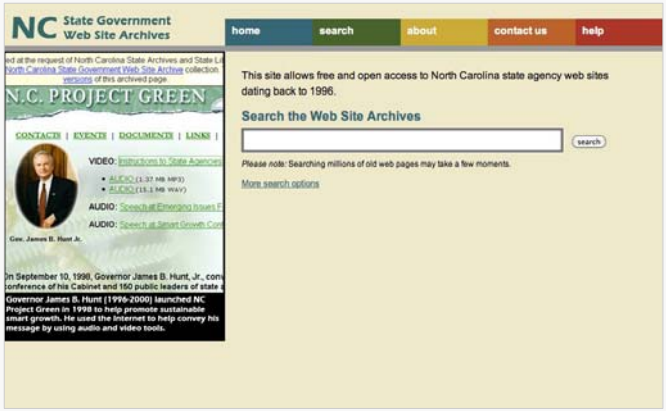


IMAGE 1:
Home Page of the NC State Government Web Site Archives, http://webarchives.ncdcr.gov/

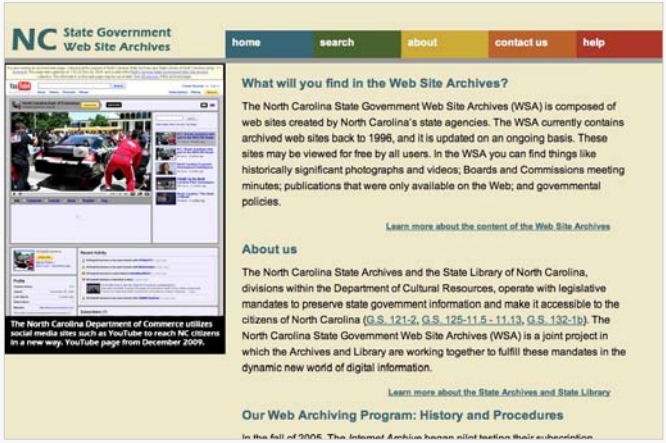


IMAGE 2:
"About" the NC state Government Web Site Archives, http://webarchives.ncdcr.gov/about.html

1d. Preservation

Data gathered in preparation for this paper suggests that preservation is an evolving issue for institutions that archive the web, which goes hand in hand with the evolving nature of digital preservation and the development of digital repositories. The Archive-It team found that their partners tend to employ several different preservation strategies. Many institutions that work with the Archive-It service rely on the Internet Archive for storage and preservation of their WARC files and associated metadata. There are several partners that also transfer their data to a local hard drive or download their WARC files directly from Internet Archive servers. A few partner institutions are working to incorporate WARC files into their local digital repository, although these projects are still in their infancy.

library's goal for providing access through their own website is to "create a single identifiable brand that will be associated with state government information" (personal correspondence with Beth Downs, James Kammerer and Chris Stockwell, May, 2012). Montana has also found other innovative ways to draw attention to their web archives. All Montana State Library webpages contain a "page history" link in the footer. These links direct visitors to archived versions of the web page so they can see how it has changed over time. For example the "page history" on the state library's home web page http://msl.mt.gov/[1] directs the visitor to a list of easy to browse capture dates for that web page: http://wayback.archive-it.org/499/query?type=urlquery&url=http://msl.mt.gov/&dates=

IMAGE 3:
Montana State Library borne page, http://rnsl.rnt.gov/

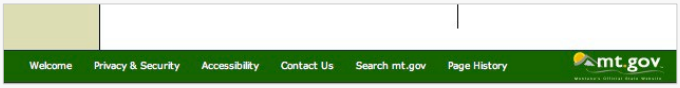
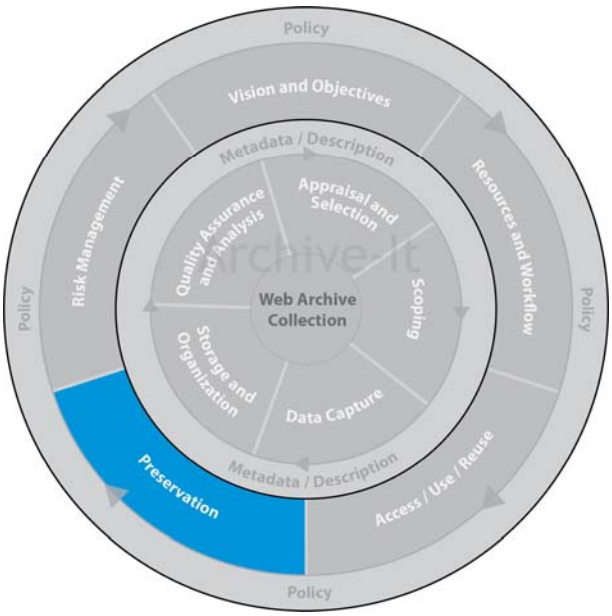


IMAGE 4:
Detail of Montana State Library borne page footer



THE OUTER CIRCLE - PRESERVATION

Based on a recent survey completed by Archive-It partners, partners do want to preserve their data and have multiple copies of their data in multiple locations. However, they are grappling with how to get there. In the survey, 56% of respondents answered that they would like to archive their data in their own local repository (regardless of the platform they use).

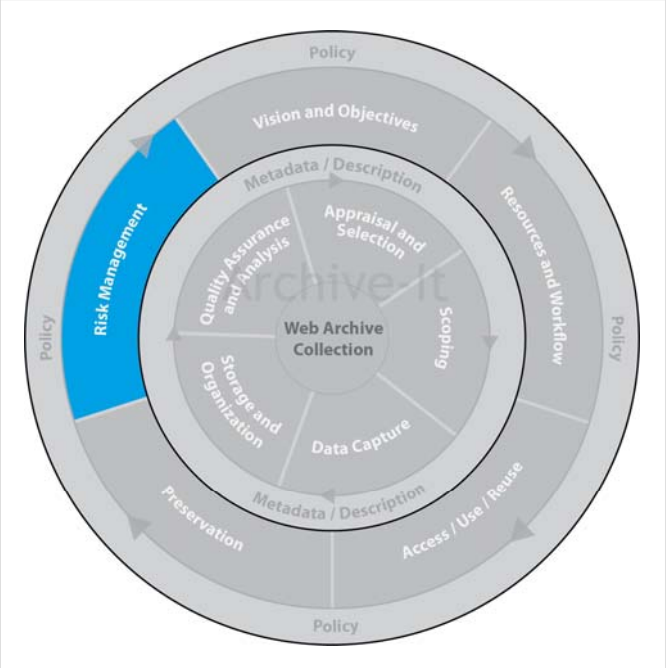
However, 31% of partners reported that they prefer to store their data at the Internet Archive, either because they are satisfied with that strategy or do not have the means to preserve the data elsewhere. Approximately 60% of respondents do not yet have a local digital repository. The two highest cited reasons for not having a repository are "unsure of our needs" and "weighing which system to choose". These results along with anecdotal information gathered over the years from Archive-It partners strongly suggest that partners are grappling with issues of how to preserve the data they collect from web archiving, and one can expect substantial developments in this area of the model in the coming years.

1e. Risk Management

In developing a web archiving program, many institutions consider the level of risk related to copyright they are willing to accept and how they will manage this risk. Whether and how institutions decide to seek permission from site owners before archiving is one of the clearest examples of risk management policy making in action. The Archive-It service has long used robots.txt (an easy-to-use technological solution) as a permissions management tool, which provides an automatic way for site owners to exclude their sites from the archiving process. In addition to the robots.txt protocol, Archive-It partners sometimes seek out website owners to get written permission before beginning to harvest.

For example Columbia University contacts site owners directly, and formally asks permission to archive websites before they begin their harvests. This is a multi-week process in which the site owner is contacted twice. If there is no response to the first contact after three weeks, the Columbia team sends a follow up message. If they still do not hear anything after an additional three weeks, they proceed with the harvest. Overall, Columbia's response rate is 52%: of 783 sites contacted, 400 responded and granted permission, 378 did not respond, and only 5 site owners have responded negatively asking that their sites not be archived (personal correspondence with Alex Thurman, February, 2013). Similarly, the University of Alberta selectively asks permission for sites they archive. This decision was based on discussions with their legal department who gave them a "risk threshold" to follow, and they ask permission when they feel the need to stay within this threshold (personal correspondence and conversation with Geoff Harder, June 27, 2012).

Risk management decisions can also be seen in the choices institutions make when deciding which sites to archive. Originally, the State Library of North Carolina and the State Archive of North Carolina collected only state agency websites. However, in 2009, they started collecting the feeds of state agencies on social networking sites like Facebook, Twitter and Flickr. Despite the fact that the content



THE OUTER CIRCLE - RISK MANAGEMENT

was on a third-party website and not controlled by a North Carolina state agency, the archivists and librarians made the decision to move forward with the archiving after weighing the potential risks and outcomes (personal correspondence with Kelly Eubank, Lisa Gregory, Kathleen Kenney, and Rachel Trent, June 2012).

Not all organizations ask for permission before capturing content; and many organizations are clear that as an archive and/or a library, their organizations has the right and the mandate to capture publicly available content on the live web. Fair use and fair game are two phrases the Archive-It team hears from partners when deciding to capture publicly available web content. In many cases an organization's mandate extends to include ignoring robots.txt on CSS and style sheets so the archived web page renders completely. And in other cases this policy includes researchers and historians capturing documents and/or websites to be able to present an accurate and comprehensive portrayal of a subject matter, with is increasingly including publicly available content on social media sites.

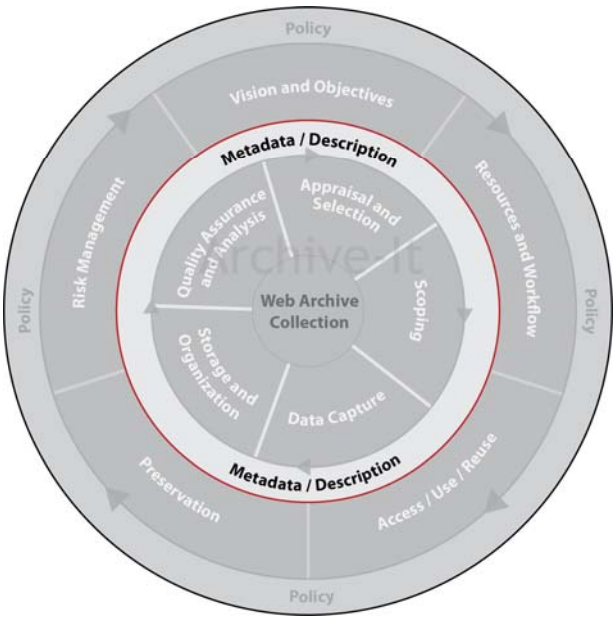
Risk can be managed and mitigated preemptively; and sometimes institutions may also need to address potential issues that come up after archiving of content has taken place. At Creighton University, a photographer became upset that his website had been archived, despite the fact that the site was part of the larger university web space and was therefore crawled per University records management policy. Creighton decided to remove the website from the archive and worked with the Archive-It team to handle the issue, and the content was removed within hours. Since then, Creighton has decided that if there is a risk of embarrassment or litigation, they will remove content from the web archive (conversation with David Crawford, July, 2012).

Note: The Archive-It service does not take a stand on copyright; and follows the Oakland Archive Policy, established in 2002, striving to work collaboratively with content providers. The service will honor requests to remove content from public access.

2. Grey Band

2a. Metadata and Description

Based on information from partners, the Archive-It team concluded that the metadata and description part of the web archiving cycle, like policy, overlaps significantly with other steps of the cycle. Therefore, the decision was made to present metadata and description as an encompassing band of the model rather than its own discrete part of the process. As with most aspects of web archiving, best practices are evolving regarding the use and creation of metadata and descriptive trends for web archives. However, the Archive-It team can make some conclusions based on how institutions use the metadata and description functionality in Archive- It. Data gathered internally by the Archive-It team in 2011 shows that over 70% of Archive-It partners generate collection level metadata, over 60% generate seed metadata, and 10% generate document level metadata. Seeds are the starting point URLs for web crawls and documents are the individually archived web pages. Additionally, this same data showed



GREY BAND - METADATA AND DESCRIPTION

that 60% of partners create both collection and seed metadata. Some partners, such as Columbia University, generate a significant amount of metadata for their Archive-It collections and work with Archive-It to change and expand the application's metadata functionality. While past statistics on metadata generation are not available, the Archive-It team believes based on anecdotal evidence that the rates at which partners are creating metadata have grown. The Marquette survey corroborates these findings. The survey asked how Archive-It partners use the descriptive features of the application. Key findings from the survey include:

- 35% of respondents prepare metadata at the collection level beyond the required description field; 35% do not.
- 81% of respondents do not prepare metadata for individual documents captured by Archive-It crawls.
- 75% of those who do prepare metadata for individual documents generate it manually as opposed to scraping it from the site.
- A majority of survey respondents do not catalog Archive-It content at any level within a catalog record (collection, seed, document) (Sweetster, 2011).

Overall, the Marquette survey authors believe that one of the major outcomes of their work is the suggestion that Archive-It partners are not generating metadata for their collections in the Archive-It application itself. Sweetster offers three possible reasons for this: "organizations just haven't yet gotten around to preparing metadata in Archive-It and are still in their infancy in terms of their web archiving efforts. Organizations do not believe that metadata is warranted or useful to be created [and] organizations are focusing their metadata creation practices in areas outside the Archive-It platform" (Sweetster, 2011).

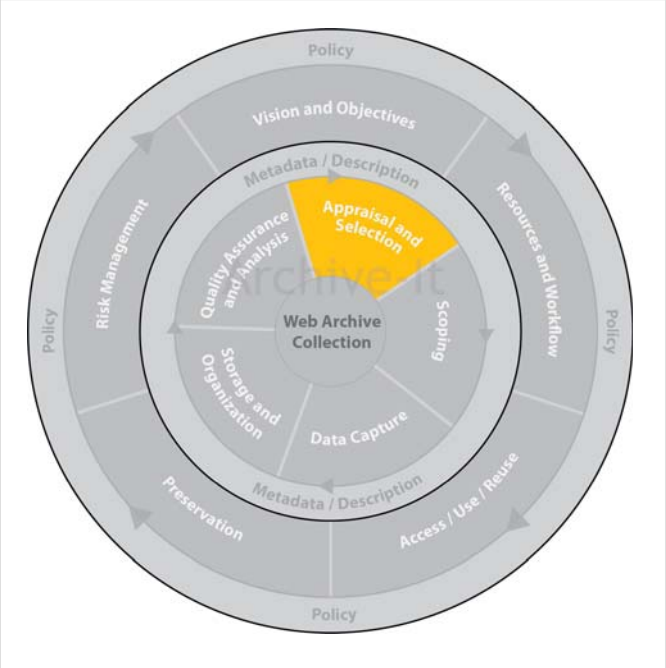
3. The Inner Circle

The preceding life cycle phases have been part of the outer circle of the model, which relate to the broader questions around creating and defining an institutional web archiving program. The remaining phases of the model, or those in the inner circle, describe the day-to- day activities of managing a web archiving program.

3a. Appraisal and Selection

The appraisal and selection phase of web archiving involves choosing specific websites for capture. This step involves more granular, specific decision points than the broader "vision and objectives" policy phase of the lifecycle. In creating policy, institutions envision overarching plans for the entire program, such as what subjects will be included in the collecting activities. In the appraisal and selection phase, however, institutions choose the specific URLs they will archive. (10) As the forthcoming examples indicate, these choices can be made in a variety of ways, with different types of individuals contributing.

State archives and libraries, for example, typically focus their web archiving efforts exclusively on state agency websites and collect those URLs. This is true of Montana State Library, the State Library of North Carolina and the State Archives of North Carolina. However, in the case of North Carolina, they also



THE INNER CIRCLE - APPRAISAL AND SELECTION

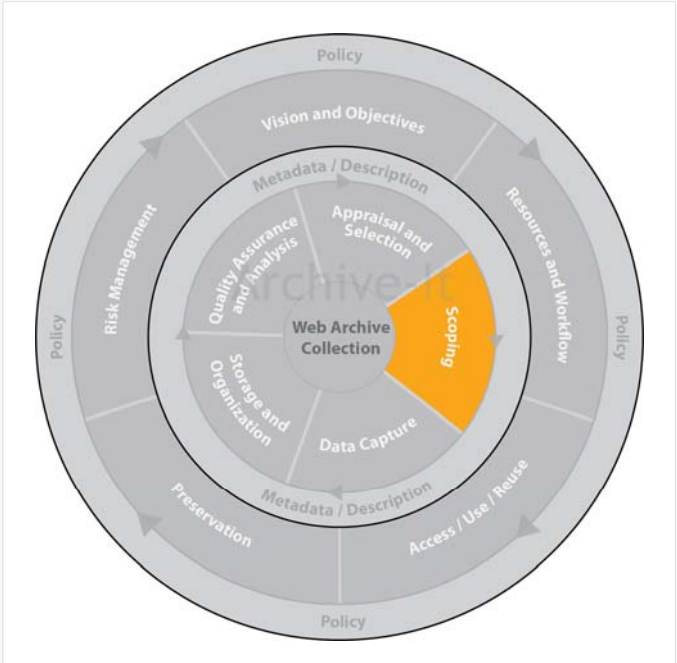
archive social media feeds generated by state agencies on Facebook, Twitter and Flickr, because they see these feeds as extensions of the official web based records. This policy decision is further described in the risk management section of this paper.

Universities that archive the web sometimes take a different approach to site appraisal. They tend to archive the university web presence or create collections based on specific themes. For example, the major topic areas of Columbia University and the University of Alberta web archive collections include human rights issues and Canadian industry and culture, respectively. Translating the institution's major objectives into a list of sites to crawl is the goal of the appraisal and selection process. To do so, the University of Alberta for instance works with subject liaisons to choose URLs. Appraisal and Selection is an evolving area and one we hope to learn more about from our partners.

3b. Scoping

After choosing what sites to archive, institutions must decide if they want to archive entire websites or portions thereof. This can be done before the first page is captured or after content is harvested as part of the overall collection quality review. This part of the lifecycle can be quite technical depending on the tools an institution uses.

The Archive-It service gives institutions several ways to adjust the scoping of their crawls. First, partners can limit what they crawl by listing only part of a website as the starting point for the crawl instead of the entire website. For example, an institution could choose to archive http://www.ncgov.com/government/index.aspx instead of http://www.ncgov.com and would only capture pages nested under that URL. Archive-It also includes other tools that can limit how much of a site is crawled. Recent survey results show that 73% of respondents report that they use a host-constraining tool at least sometimes, This tool allows partner to block specific hosts, or sub-sections of a site, from being archived. For example, an institution may not want to collect third party images that may be hosted on a target website. Limiting the duration of a crawl through



THE INNER CIRCLE - SCOPING

time limits is the second most used tool, as reported by 64% of respondents.

Currently 27% of Archive-It partners run some crawls that capture only PDFs, and we expect to see this percentage increase as PDF's become more prevalent on the web and increasingly the only record available. The Archive-It service is researching adding this capability for other types of file formats. As social media sites become an increasingly vital component to collecting activities, the service is exploring singular ways to provide capture and access solution to social media. Primarily Facebook, Twitter, Facebook and You Tube, as of December 2012.

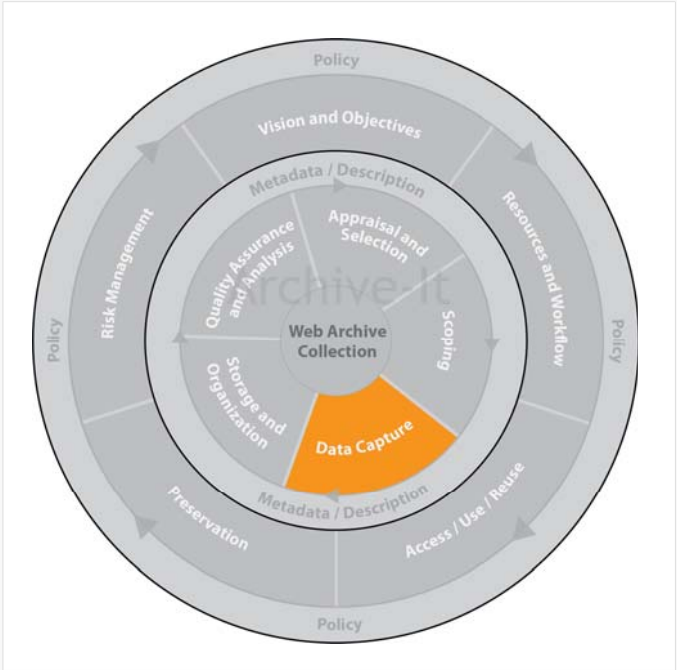
As mentioned above, the scoping process can be quite technical. The complexities involved in effective crawl scoping were a surprise to the team at the University of Alberta. They have found that they need to re-adjust their policies as they crawl sometimes adapting to the kind of data they actually can collect (personal correspondence and conversation with Geoff Harder, 2012). Similarly, Creighton has also found that scoping a crawl involves some extra work; David Crawford finds that he often needs to educate people on campus about the web space, and he tries to work with web programmers to request that they consider crawling needs when making changes to sites in the future (conversation with David Crawford, July 2012).

3c. Data Capture

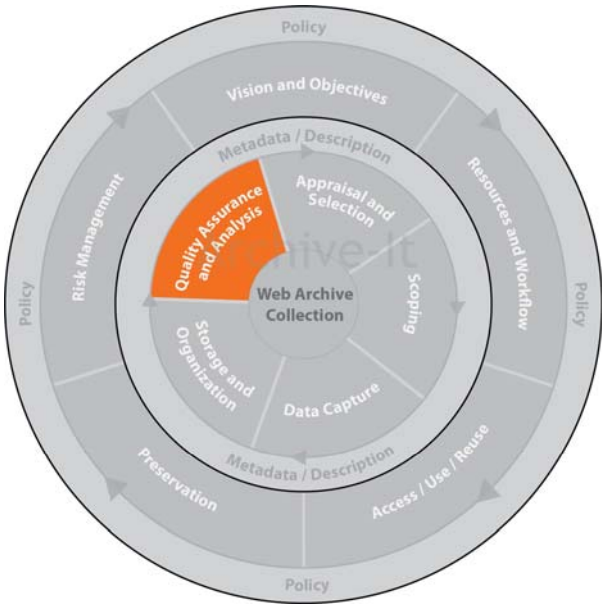
Once institutions have chosen what websites to capture and how to do so, they put their plans into action in the data capture phase of the process. Here, they will deal with the nuts and bolts of the crawling software. They will determine the frequency and timing of their crawls and when to cut-off long crawls, and then they will set their crawls to begin. The Archive-It application includes features that allow partners to make adjustments to the frequency and duration settings in the open source web crawler (Heritrix).

Scheduling crawls for ongoing and reiterative data capture is an area where institutions using Archive-It exercise a lot of control over their crawls. Data gathered in 2011 showed that 78% of all Archive-It partners use more than one crawl frequency. In other words, they do not crawl all of their sites at one interval, they use different schedules for different collections and websites. At the time the data was collected, the most popular crawl frequencies were one time, monthly and quarterly.

Given how diverse websites are in terms of their structure and construction, the data capture step of web archiving can produce a number of surprises. For example, a site can be much bigger than anticipated and therefore exhaust storage resources. Similarly there are ways for web masters to keep their sites from being archived, which can require technological intervention or negotiation between the parties involved. For example, David Crawford from Creighton University experienced issues archiving



THE INNER CIRCLE - DATA CAPTURE



THE INNER CIRCLE - QUALITY ASSURANCE AND ANALYSIS

websites that were preventable by the webmasters, and he was surprised in talking to the webmasters how little they new about the inner workings of their websites (personal correspondence and conversation with David Crawford, July 2012). To try and prevent data capture surprises, Archive-It allows partners to use a test crawl feature that produces reports on data captured without actually capturing any data. This option allows institutions to see what they would have archived without using their resources unnecessarily. The recent Archive-It partner survey shows that 69% of respondents always or often run test crawls when adding new seeds or starting a new collection.

3d. Quality Assurance and Analysis

After institutions capture data from their desired sites, they should review what they archived and assess its quality and completeness. This can be done through reports generated by crawlers or by clicking through the archives themselves by way of an access tool like the Wayback software. The process of web archiving can include trial and error. Like most aspects of web archiving, no single best practice for quality assurance has emerged among institutions that archive the web. However, there are some common trends among Archive-It partners in terms of the types of crawl information they review.

Archive-It survey data shows that a majority of partners often or always review their post-crawl reports generated as part of the service. This is due to the fact that institutions tend to be interested in how much material and exactly what kind of material they are collecting when they start a web archiving program. Findings from the 2012 summer survey of Archive- It partners show that 68% of responding institutions review their host reports on a regular basis. Only 11% rarely or never do so. Reviewing reports can take time, and reviewers need to know what anomalies to look for. Three survey respondents said that the lack of staff/resources make it difficult to analyze reports after every crawl. In 2011 the service implemented a QA tool and the ability to run a patch crawl on top level Url's that had not captured completely the first time around. The response has been positive and the service has been working on extending the QA tool capabilities. At the time of this writing there is little anecdotal knowledge about exactly how Archive-It partners perform quality assurance on their crawls; and it is one of our objectives to learn more about this area as partner's needs become more tangible.

CONCLUSIONS AND NEXT STEPS

The web archiving life cycle model is one step on the road to creating a set of best practices for creating and maintaining a web archiving program. After more than seven years of running the service and working with forward thinking partners, it is clear to the Archive-It team that the web does remain "a mess" and that it is in all of our best interests to continue to work together to find solutions to capturing and displaying web content. As technology continues to develop and as information is increasingly published exclusively online, more institutions of all sizes will need to be archiving web content. Many of the Archive-It partners have been pioneers in web archi-

ving, and enjoy sharing what they have learned. And even as we share our knowledge in this paper, we know that the web and best practices for web archiving will continue to evolve. The Archive-It team anticipates that this model and the institutions that work with it are flexible enough to grow and evolve side by side with the web they are trying to archive.

NOTES

1 Due to upcoming platform migrations, Montana State Library's URLs may change in the near future

CONVERSATIONS/EMAIL

University of Alberta: Conversation with Geoff Harder, June 27, 2012

Montana State Library: Correspondence with Beth Downs, James Kammerer and Chris Stockwell, May 29 2012

NC State Lib/Archives: Correspondence with Kelly Eubank, Lisa Gregory, Kathleen Kenney, and Rachel Trent, June 8, 2012

Creighton: Conversation with David Crawford, July 17, 2012

Columbia: Correspondence with Alex Thurman and Tessa Fallon, May 17, 2012 and February 21, 2013

AIT stats from 2011: correspondence with Kristine Hanna, September 2012

WORKS CITED

Sweetser, Michelle (2011). *Archive-It Metadata Usage Survey Results* [Powerpoint slides]. Retrieved from: https://webarchive.jira.com/wiki/display/ARIH/Archive-It+Meeting+Presentations+2011