Archive-It and Online Auction Catalogs

A report on the functionality of Archive-It
as a tool for harvesting data from auction house websites

Sean Leahy
February 3, 2011

Introduction

The Archive-It pilot project for the NYARC libraries has aimed to test the functionality of the tool developed by the Internet Archive for capturing online auction catalogs. This report will attempt to summarize the process and the findings from this project, as well as point out the challenges posed by the Archive-It tool. The project began at the Frick Art Reference Library in October 2010 in collaboration with the Thomas J. Watson Library at the Metropolitan Museum of Art to help begin a collection of online auction catalogs.

Auction house websites have been offering complete and downloadable auction catalogs for more than ten years; however, libraries that have historically had strong collections in auction catalogs have yet to seriously begin harvesting this rich (and often free) digital content. In some cases, all of these online catalogs are archived by the auction house and accessible online; however, there is no guarantee that this will always be the case. This project opens the door for the NYARC libraries to be pioneers in the world of auction catalog collections by providing the first publicly and perpetually accessible repository for online auction catalogs.

This report will assess the results of the initial crawl that yielded 500,000 files from eleven auction house website. The following advantages and disadvantages to using Archive-It have been determined.

Advantages to Archive-It:

- Fast and automated downloading of content
- Automatic storage and access to archived content
- Auction catalog captured and presented in a variety of ways
- Simple to use
- Great potential for collaboration

Disadvantages to Archive-It:

- Overwhelming amount of web data makes refined scoping very difficult
- Access to certain content can be restricted, but now content can be deleted
- Errors in downloaded content
- Incomplete capture of website content
- Upkeep and description of archived content requires a lot of time
- All content is hosted by the Internet Archive (though institutions can pay to have content sent to them)
- Inefficiency of searching the archived data

What is Archive-It

Archive-It is a subscription service available from the Internet Archive "that allows institutions to build and preserve collections of born digital content." Begun in 2005 as a means for Internet Archive users to create more dynamic and subject-specific collections Archive-It now has more than 170 partner institutions and hosts more than 1300 collections. The majority of partner institutions are the libraries and archives of universities, but range from federal institutions to independent researchers. While the Internet Archive conducts "complete captures" of the web every two months, Archive-It allows for captures to occur on a continual basis under the watchful eye of a particular institution. To date, well over two billion URLs have been captured; some collections provide a snapshot of the online response to a current event while others attempt to maintain a complete archive of their institution's online presence.

Archive-It offers its users an open-source software tool named Heritrix to "crawl" selected websites (or "seeds") in order to capture and preserve content. There is great versatility for users when choosing what to crawl and how often to crawl it. In exploring the collections of other institutions, frequency ranged from "one-time" crawls (observed in numerous collections), to monthly or quarterly crawls, as observed in the Montana State Archives collection and Columbia University's Human Rights collection. In a small number of cases, such as Archive-It's WikiLeaks 2010 Document Release Collection, crawls were conducted daily, a frequency suited for a collection focused on a current event.

Crawl Versatility

In the course of this project, I employed multiple crawl techniques to help determine the best means for capturing auction catalogs. The parameters for a crawl are flexible: Archive-It allows the user to choose how many seeds to crawl, the length of each crawl, the maximum amount of data crawled, and the frequency of repeated crawls. This versatility is a chief asset of Archive-It because the unpredictability and variability of online content necessitates a specifically scheduled and "scoped" crawl for each seed website. Archive-It allows the user to "scope" content in a variety of ways, permitting the user to limit what is crawled in order to create a more focused collection. This would include length and data limits, although for FARL's purposes, these are not useful scoping methods. However, instructing the tool to scope one part of a website allows, in this instance, an opportunity to avoid extraneous pages (such as contact information, bidding forms, or "About" pages). As well as this, "host constraints" allow a user to block any pages that contain a specific phrase in the URL.

Before looking closely at the results of this project, it is important to keep in mind that the goal of this project differs from the majority of collections hosted by Archive-It. While other institutions have been motivated to make a "complete capture" of a particular set of websites, gathering as much material as possible related to a particular topic, the goal for FARL is to create a collection of online auction

catalogs. As catalogs can be accessed only through auction house websites, the challenge will be to trim away all the extraneous URLs that the crawler will want to capture. Relevant content includes price results, auction calendars, lot listings, lot details, press releases, and catalogs released in more than one language; these pages are mixed in with -- and link to and from -- myriad irrelevant content: contact information, FAQ, bidding forms, site maps, and services pages, among many more. Once this extraneous content is archived, it cannot be deleted, which presents one of the major drawbacks of Archive-It and one of the major obstacles in creating a curated collection. The end result, ideally, will be a collection of URLs devoted only to particular auctions or items up for auction that could be continually added to will little maintenance. Creating a collection of seeds, each tailored to focus on auction catalog content only while avoiding irrelevant pages, may not be a reachable goal; however, as Archive-It improves its tool, more refined harvesting will be possible.

Seeds

For the pilot project, eleven auction house websites were selected for crawls. These "seeds" varied greatly in content, quality, and depth; and their variety provided a good test of the capabilities of the Archive-It service. The seeds selected were:

- Dreweatts (www.dnfa.com) -- An auction house based in Great Britain. Catalogs are available online up to one month before the auction takes place. The website has an extensive archive of previous sales and results, reaching back to January 2000.
- Bonhams (www.bonhams.com) -- This auction house is also based in Great Britain with salerooms worldwide. The website has an archive of auction catalog similarly as large as Dreweatts.
- International Auctioneers (www.internationalauctioneers.com) -- This website differs from the other seeds in that it compiles the auction catalogs of eight auction houses based on both sides of the Atlantic.
- Tajan (www.tajan.com) -- Tajan offers PDF downloads of all of its online auction catalogs, dating back to 2001.
- Nagel Auktionen (www.auction.de) -- Nagel, one of Germany's leading art auctioneers, has a small archive of online catalogs (with none available from 2009 or earlier) and a short listing of auction results.
- Hosane (www.hosane.com) -- Hosane, an auction house based in Shanghai, has about fifty catalogs available online, dating back to 2006.

- R.W. Oliver's (www.rwolivers.com) -- The smallest auctioneer from this list, R.W. Oliver's is also the only example from the eleven seeds that puts its auction catalogs and information exclusively online

- Heritage Auction Galleries (www.ha.com) -- This website posed different challenges than the rest because it acts more as an aggregator rather than treating each auction individually. The website itself is difficult to navigate, and the results from our crawl mirrored many of those difficulties.

- Auction.fr (www.auction.fr) -- This dynamic websites culls together auction information from numerous sources in France and contains a searchable archive of more than two million objects.

- Pandolfini (www.pandolfini.it) -- Based in Florence, this Italian auction house has been a fixture on the antiques market for several decades. The website offers PDFs of all of its auction catalogs and auction results.

- Gunther Kunstauktionhaus (www.dresden-kunstaucktion.de) -- This small website offers up to date information and catalog listing for this German auction house. However, information is not archived for very long on this site.

Because of the vast amounts of data that was retrieved, it was difficult to look comprehensively at the results of each seed. During the early investigations of the results, it was clear that certain seeds produced more interesting and telling results than the others. For example, Dreweatts' extensive archive provided a huge amount of well-captured data to look more closely at. Others, such as Bonhams' website and the International Auctioneers' website, also produced a great deal of data and became the focus of much of my investigations into the crawl results.

Initial Crawl

Following a tutorial from representatives at Archive-It, the initial crawl for our collection was conducted on October 7th, 2010. The document limit was set at 500,000 pages, a number that was reached after about forty hours of crawling. Reaching this limit left another 450,000 pages "queued," which reflects the number of pages that were in scope but were not able to be crawled. The 500,000 pages amounted to 9.9 gigabytes of data. This was by far the largest crawl that was conducted during the project.

The seeds that produced the most results (in terms of URLs captured) were International Auctioneers, Dreweatts, and Bonhams. The seed that returned the most data was Tajan. This was due to the fact that much of what was captured came in PDF format, greatly increasing the size of each file that was archived. Bonhams accounted for nearly as much data as Tajan because of its numerous high-quality

image and Flash features. The Gunther Kunstauktionhaus turned up the fewest results, with only seventy-seven URLs captured.

HTML files accounted for more than half of the data archived, with 366,000 files in total. Image files (jpeg) and PDFs both accounted for about 2.5 gigabytes of data. More than 90,000 images were captured, in comparison to the barely more than 1,000 PDFs that were captured.

As the numbers show, Archive-It performed a very thorough crawl of these eleven seeds. It is difficult to tell exactly how much of the captured data is directly related to the catalogs, but searching the results suggests that a majority of it is. While ten gigabytes is a sizable amount of data, better scoping would help reduce this number on subsequent crawls. Along with better scoping, "data de-duplication," which I will detail later, also ensures that the amount of data archived in subsequent crawls would be reduced further. It is very difficult to predict how much new content Archive-It will discover among the seeds it has previously crawled; however, if these eleven seeds were crawled once a month, it reasonable to expect between four and six gigabytes of new data would be archived.

Looking Closely at the Results

In the following sections, I will discuss some of the functionality challenges that I encountered following the initial crawl. These challenges range from downloading errors to extraneous content; and while few if any are major obstacles in the harvesting of online auction catalogs, they do represent the shortcoming of Archive-It as it relates to creating a curated collection.

Missing Catalogs

A majority of these auction houses—such as Tajan, Gunther, and R.W. Oliver's—had only a small amount of auction catalog content available online. For these seeds, seemingly all of this content was captured. However, for Bonhams and Dreweatts, whose archives reach back eight and eleven years respectively, it appears that Archive-It had difficulty discovering catalogs that were more than two or three years old. According to the Crawl reports, these pages were not blocked by robots.txt or out of scope. This was also borne out by using the "crawl one page only" function, which allowed me to capture these older pages one by one. It is possible that either the time limit or the document limit is responsible for the absence of these auction catalogs, suggesting that a content-rich site such as Dreweatts would be a challenge to capture completely. This instance of uncaptured content also shows that even when archived and accessible on the web some data is not guaranteed to fall within the purview of the Archive-It tool.

Extraneous Content

While the results of this crawl were very edifying, any future crawls would not be conducted in such a way. There were no host constraints and the only method of scoping was the document limit (which, at the time, we did not think would be reached); therefore, while the crawl captured most of the content relevant to auction catalogs, it gathered a massive amount of extraneous or useless material, as well. Generally, the biggest culprit on these websites was pages related to news updates or press releases. While contact information and historical information is unlikely to change often, news about upcoming sales and sales results appears can appear on a daily basis on some sites. (On others, such as Gunther or R.W. Oliver's, news updates may appear only a handful of time per years.) These press releases generally appear as a single HTML files, but can also be presented as PDFs, as was the case with Bonhams' online magazine.

Another important aspect of this project is that it is looking at auction catalogs for every kind of sale. FARL's collection of auction catalogs is extensive but it is not pervasive. The catalogs reflect Henry Clay Frick's collecting tastes and generally work within that scope. Therefore, auction catalogs for the sale of wine, coins, automobiles, sports memorabilia, jewelry, movie posters, clocks, Asian art, rare books, scientific instruments, maps, and dolls do not form part of the Library's collection. Oftentimes, sales for items such as these are more prevalent than an sale of Old Master paintings. Even if the Watson Library at the Metropolitan Museum of Art, with its far more inclusive interests, were incorporated into this collection, there would still be hundreds of catalogs that would be of little use to researchers and librarians at either institutions. While scoping would help avoid non-auction catalog material, the tool does not appear to developed enough to perform the task of capturing specific auction catalogs on a website.

Duplicated Content

As subsequent crawls are completed in order to capture new auction information, there will inevitably be some information that is repeated or duplicated. Fortunately, Archive-It uses a tool that performs "data de-duplication," which helps prevent identical content from being stored more than once. Archive-It records when a URL has been crawled and compiles a list of dates (that a user may view) that reflects the number of crawls. If content on a page has changed, an asterisk will denote that the updated page has been stored

While it is reassuring to know no identical content will be captured, there still abounds thousands upon thousands of pages that simply *appear* to be identical. For example, a user may choose which portal to use while on the Bonhams' website based on location: Australia, United Kingdom, United States, or Hong Kong. Yet, regardless of location, the auction catalog content is identical; the only change is found in the URL, which includes the portal (/aus/, /usa/, /uk/, /asia/) the user has chosen. In cases such as this,

Archive-It captures the page from each portal, thus created four identical pages. On top of this, almost indistinguishable changes (such as a change is size or shape of a button) will appear even at the same URL. This redundant content is certainly a cause for concern and appears to be an inevitable drawback when dealing with online content

Another source for duplicate pages is a crawler "trap." A trap is, according to Archive-It, "a set of webpages that creates an infinite number of URLs for the crawler to find." The most common trap found on websites is a calendar page, and the initial encountered such a trap on the International Auctioneers page. These calendar pages automatically generate links the "next month" or "previous month," creating endless reiterations of essentially the same content. Following the first crawl, the report showed that IA had the most captured URLs, which was not a complete surprise given that the website is comprised of auction information from eight auction houses. However, the number of pages still in the queue (366,000) dwarfed any of the other seeds. Upon closer look, I learned that the crawl had generated an enormous amount of duplicated content: Searching for a print entitled "Davos-Platz" by Rudolf Dickenmann produced 36 results; some inspection reveals that all thirty-six results are more or less identical. The only discrepancy between results was in the URL, most likely the result of the calendar that is found on all lot listings for this website.

The crawler traps pose a legitimate problem for a collection's data budget, while duplication -- like the kind seen in the results from Bonhams -- is less of an issue. Each html file amounts to around 100 kilobytes, a negligible amount of data; however, any content that is captured cannot be deleted, and after months or years those small files would add up. How much of a concern this is would depend on the estimated size of the collection. However, it is important to note that the size of the data budget is flexible and can exceed well beyond the 256 GB that were allotted for this pilot project.

Duplicated pages also make the process of searching more difficult and frustrating. If the first ten or twenty results appear to be identical, a user might find it futile to dig deeper into the remaining search results. Careful analysis of Crawl reports and search results will help narrow the scope for subsequent crawls and hopefully diminish the prevalence of duplicate pages. However, this is a time consuming process that yields varying level of success.

Capture Issues

Issues with page formats and dead-end links go hand in hand in this project. Ultimately, both of these issues relate to a researcher's encounter with the collection. Because we are archiving web pages, a researcher may expect that the results look like live pages. Archive-It often has difficulty capture Java script and some XML content. When this occurs, the website's format (as well as some of its buttons) will look noticeably off. There was also an issue in capturing content that ran using Adobe Flash. Though this

problem was only encountered once, it does pose a legitimate problem to the capture of online auction catalogue content. Fewer and fewer auction houses are utilizing the PDF format for offering access to catalogues online, instead favoring more dynamic applications such as Flash that will allow users to "flip" through pages of lot listings.

As well as this, robots.txt can also wreak havoc on downloaded content. A robot allows a webmaster to block crawlers from specific parts of a website. Permission must be obtained from the webmaster to allow Archive-It to crawl the blocked files. In addition to this, Archive-It has a feature that allows partners to bypass the robot.txt files. This feature is used by about 20% of Archive-It partners. During the pilot project, the robots.txt were not bypassed, therefore there were some occasions when relevant content was not captured. The best example of this came from Dreweatts' website, where many of the captured pages are missing banner images and a majority of the fonts do not correspond with what is found on the live site. This may reflect a concern over copyright infringement. In both cases, running a Quality Assurance check on a crawl's results will help determine the cause of certain capturing issues.

Capture issues related to JavaScript and XML or two robots.txt create numerous dead-end links. Though these did not commonly appear when exploring the results (the best example came from the Bonhams' pages), they could be jarring or frustrating for a user. Familiarity with the results and some creative searching technique helped me overcome the obstacle posed by dead-end links. However, can it be expected that a researcher will be as patient? Will the researcher treat the archived pages as though it were a live website, and will dead-end links frustrate and ultimately turn researchers away from the collection?

Passwords and Permissions

The final issue concerns password protected sites. Tajan, Heritage, and Pandolfini each require users to register in order to look at auction catalogs or auction results. While none of these websites ask for a fee in order to register, obtaining access to these websites could be difficult. Similar to the robots exclusion, bypassing the password control would require permission from the site's webmaster, this time allowing the crawler's IP address access to content that is protected.

The problem of running into password barriers, along with the blocked images (in the case of Dreweatts' website), speaks more broadly to the issue of permissions. Many institutions are currently experimenting with capturing and archiving web content. Some of these efforts are on a small enough scale that gaining permission may not be an issue; however, there are also programs, such as the UK Web Archive hosted by the British Library, that are collecting thousands of websites. Obtaining expressed permission from every site in the Web Archive would be impossible. The British Library's policy is to ask permission from every website, and if there is no response, the site will be archived. If the

administrator of the website does not want the website available on the Archive once it is there, they have a "Notice and Takedown Policy" that respects intellectual property rights and data privacy.

Columbia University, for its Human Rights collection on Archive-It, has even gone a step further by enlisting the help and support of Dr. Kenneth Crews and the Copyright Advisory Office at the University. This not only will ensure that all privacy and rights are respected, but may help produce a permissions model for the archiving of digital content. And while FARL will not need to go to those lengths, it will be important to keep track of and report on all issues or successes associated with permissions, and to share that knowledge with other art reference libraries.

Subsequent Crawls

After spending some weeks exploring the results of the first crawl, I began to perform subsequent on a few selected seeds, including Bonhams, Dreweatts, International Auctioneers, and Tajan. The goal of the subsequent crawls was to assess the amount of data that was being duplicated or captured twice. I began by crawling the Dreweatts website everyday for one week, and the results were telling. Each day, approximately 650 megabytes of data was crawled. On day one, more than 550 megabytes of the crawled data was determined to be new by Heritrix (meaning the content of the pages had changed in some way) and was archived. By week's end, the amount of new data archived had shrunk to less than 250 megabytes, showing that duplicate data was not being archived. However, it is important to note that the amount of new data shrank quickly in the first few days of crawling, but seemed to level towards the end of the week, suggesting that the amount of new data archived was not likely to shrink to zero. Other subsequent crawls showed similar results, even when the number of seeds was increased.

Test crawls

I later began relying more and more to rely on "Test crawls," which perform essentially the same task as a regular crawl but do not store any data, and useful information can be gleaned from the Crawl report. The report for a regular crawl specifies the amount of new data that is archived (from which we can extrapolate whether or not there is a large amount of duplicated data); and while the Test crawl report will not show this, it will list the "new URLs" that are captured. Ideally, multiple Test crawls would be performed per new seed to help determine the scoping for the website that is to be archived. This way, unlike with our initial crawl, data storage space would not need to be used up in order to find out what is captured. This process would be tedious and would require a great deal of time for each seed: The list of new URLs from each Test crawl more often than number in the tens of thousands. It is impossible to investigate these lists one URL at a time, but the URLs are grouped together in such a way that helps clarify what content or portion of the website is being gathered. Putting in time to investigate these lists

closely would result in a well-scoped seed and saved storage space, though it cannot be expected to do away with irrelevant files altogether. At the outset of the pilot project, Test crawls were not relied on because we were not aware of how beneficial they would be for our needs; however, it should be stressed that this aspect of the tool is essential for creating a collection and for efficient use of the data budget. Test crawls are useful at any stage in the duration of a collection, and in order to set up any new seeds, numerous Test crawls would need to performed for each seed.

Setting Up New Seeds

The time required to refine each seed before conducting a crawl in earnest could vary greatly. The first consideration is for the size of an individual website, which can includes only a small number of pages or can be extensive. The second consideration is for the amount of extraneous content that is desired. Since extraneous content is impossible to avoid, it would be important to determine what level of scoping will be needed. Reviewing the Test crawls and testing host constraints is a time consuming process, and it does not always result in well-scoped seeds. If the goal is to capture content quickly and across a wide variety of auction houses, then less time should be put into scoping; if the focus is on only a handful of seeds, then more time could be devoted to conducting Test crawls and defining the host constraints.

Access and Discovery

The last issue to address for this project is the means of access that Archive-It gives its users, for which there are multiple options. All archived content is hosted by Archive-It, and if an institution chooses to make it collection publicly available, the easiest portal of access is through the Archive-It website (www.archive-it.org). On the homepage, users can select which institution's collection they would like to view. Once a collection has been selected, users then have may either browse by seed or search all the results for the collection. Both basic and advanced search options are available, with the advanced search allowing users to search within one seed only.

Archive-It also gives instructions for adding a search box to its partners' websites, which appears to be the most popular way for an institution to link its users to the archived collection. The search box is for basic searches only. Once a search term is entered, the user is shown results on Archive-It's website. The content is entirely hosted by Archive-It, so all results will be viewed using Archive-It's interface. This, unfortunately, is unavoidable, but the search box allows partners at least one access point to a collection. Other partners have opted to input metadata into their own records (rather than using Archive-It's DC fields), and have chosen to link to the archived content from their own OPAC. This may belie a

lack of faith in Archive-It's search mechanism and its metadata capacities; however, this may also prove the most useful way to connect FARL patrons with the auction catalog collection.

Improvements and Updates to Archive-It

In December 2010, Archive-It released an update that included these URL lists. Prior to this update, it was not possible to view the URLs; only the captured pictures, accessed using the Wayback Machine, could be viewed. This was a considerable improvement and made accessing each crawl and making decisions about scoping much easier. As well as this, the metadata (using Dublin Core standards) for each page is now searchable. Many partners with Archive-It (the Frick included) expressed a desire to be able to export metadata from their collection to other cataloging systems, and in the latest release, Archive-It states it is committed to improving this aspect of their tool. For the time being, they have made available documentation (courtesy of the Montana State Library) that describes how to export the Dublin Core metadata and transform it into a MARC record using the MARCEdit tool.

These changes would prove to be very beneficial for the sort of curated collection that FARL is interested in, and points to continuing refinement of the tool on Archive-It's end. In fact, reading over the grant proposal Columbia University put forth to begin their Archive-It collection in 2009, it is clear that, even in the last two years, Archive-It has made many essential improvements. With an increasing number of institutions becoming involved, each with their own set of expectations, questions, and suggestions, it is likely that the Archive-It will continue to enhance their tool in the coming months and years.

The goal of this pilot project was to test Archive-It as a viable tool for archiving online auction catalogue content. However, how this experiment would play out was not initially clear. Prior to the pilot project, general goals were discussed by both FARL and the Met, after which the project was set up with Archive-It. Based on what we have learned, it is clear that this pilot project would have benefitted if more specific goals had been articulated at the outset. Because this did not happen, at times, the Archive-It team seemed to have a limited understanding of what FARL was looking to accomplish with this tool, though they were always quick to respond to questions and suggestions. In-depth communication prior to initiating a collection, as well as throughout the management of it, is a necessity when developing a collection.

Conclusion

This report has sought to outline the basic operations of Archive-It, to assess the service's functionality regarding FARL's goals, and to describe what challenges FARL will face in its attempts to capture online auction catalogs using this tool. We are in the nascent stages of digital archiving tools such as Archive-It, and there are no doubt many hurdles to clear before the technology is robust enough to

meet our every demand. It is important to keep in mind that the collection that FARL is interested in creating differs from others that are currently employing Archive-It. Beginning with this collection will hopefully allow the Frick to pioneer how cultural institutions capture and archive online digital content, beyond just auction catalogs.

Finally, there are a few vital points that will require continued consideration as the Frick moves ahead with its strategy for capturing online auction catalog content:

1) **Crawl Efficacy** -- Ensuring the efficacy of each crawl is the biggest challenge posed by this project. At this stage, the Archive-It tool cannot reliably perform the narrow crawls that would be ideal for the Frick's online auction catalog collection. The most time-consuming aspect of this collection will be related to creating efficient crawls: Setting up Test crawls, scoping each crawl, and thoroughly investigating the results will require a good deal of time and effort.

2) **Crawl Frequency** -- Determining the frequency of subsequent crawls will also be a challenge, as each auction house works on its own varied schedule. This can be accomplished by studying the calendar or schedule for each auction house and using that to gauge the frequency of crawl for each particular seed. Although, in most cases, crawling once a month or bi-monthly appears to be sufficient.

3) **New Formats for Catalogs** -- Some of the seeds used for this project allow users to search on an item or object level, aggregating results from all auction available on the website. These aggregators present a model for an online auction catalog that differs greatly from the traditional printed catalog. More and more, these websites are doing away with PDF surrogates of their printed catalogs and are instead introducing a more dynamic means of interacting with the auctions items. This presents unforeseen challenges for archiving online auction catalogs because the form of auction content online in the future is impossible to predict.

4) **Access and Discovery** -- Access to stored content on Archive-It can happen in a number of ways and can be updated to better reflect users' wishes or suggestions. Linking from FRESCO to directly to Archive-It may help in avoiding the sometimes messy search for content. Because of the obstacles noted above, a search can return a dizzying number of seemingly similar results, and an online tutorial would be very helpful.

5) **Data Budget Management** -- As this collection grows, storage space management will be an important consideration. However, running numerous Test crawls for all potential seeds will help create carefully calibrated crawl scopes, which will in turn lead to efficient use of the storage space allotted by Archive-It.

6) **Metadata** -- Metadata appears to be more and more of a priority for Archive-It, as reflected in recent updates to the tool and responses to partners' suggestions. For the time being, Archive-It only

provides instructions for transforming the DC metadata into a MARC record. However, it is likely that institutions will continue to demand improved integration of the role of metadata. And, as more institutions provide ideas and input regarding this aspect of the service, the more useful it will become.