Center for Research Libraries
Middle Eastern Political Parties Web Harvesting
and other efforts

A paper presented to the "International Collections Development Workshop"
February 27, 2006
James Simon, Director of International Resources

The Center for Research Libraries recently participated in a pilot assessment of the Internet Archives' on-demand Web archiving service called "Archive-It" (http://www.archive-it.org). The pilot crawl occurred September–November 2005. The Center cooperated with a number of partner institutions to select and evaluate crawls of sites performed by the Internet Archives using their open source crawler, Heritrix. The following is our summary of the activities and assessment of the current capabilities of "Archive-It" as a tool for capture and presentation of Web-based research resources.

I. History & Background

Political Communications Web Archiving (PCWA) Investigation

In November 2002, the Center for Research Libraries received funding from the Andrew W. Mellon foundation for an investigation and planning effort on preserving politically-oriented Web sites from regions of the world that have relatively little infrastructure to preserve their own materials. Political communications–that is, Web sites of political and other non-governmental organizations–are vital primary sources for history, political science, and area studies. These materials tend to be produced erratically or even spontaneously, and disappear quickly. The research value of these materials depends on the preservation of the content, as well as certain evidentiary traits including the look and feel of the original material, metadata on the date and time of initial presentation, authorship and source of transmission, and other elements.

For many years, CRL has served as a framework for building shared university and research library collections in traditional formats. In its strategic plan for 2002-2006, CRL expressed its objective to promote and support action on the North American and global levels for cooperative preservation of print and digital scholarly materials. The Mellon-funded project was designed not to create an archive of materials *per se*, but rather to produce the general specifications for ongoing cooperative archiving of political Web materials.

The investigation, occurring over a period of 18 months, focused on three key aspects of Web archiving:

- Long-term resource management: To determine the organizational and economic framework necessary to support the archiving of Web-based political materials on an ongoing basis, and the persistent availability of those resources for long-term research use.
- Curatorship: To identify the optimal curatorial regimes, practices, and tools for ongoing identification, targeting, and capture of the various types of Web-based political communications; also to reconcile Web archiving and curatorial methodologies with traditional collection development activities.
- Technology: To identify and specify the most appropriate technology architectures, tools, and techniques for gathering and preserving Web-based political communications; to assess the associated costs, benefits, characteristics, and risk factors involved.

The investigation culminated in a final report, which may be found via CRL's Web page at: http://www.crl.edu/content/PolitWeb.htm. Some of the key findings of the investigation focused on characteristics and behaviors of producers and users of Web-based political communication, recommended curatorial regimes, technical considerations, and long-term sustainability issues. The report lays out a proposed service model for a distributed, sustainable, community-driven Web archiving program. The proposed model describes functional requirements, participants, cost factors, and accountability measures.

Archive-It Pilot Program
In March 2005, the Internet Archive approached CRL and several other organizations with plans to launch an on-demand Web archiving service. The Internet Archive is a not-for-profit corporation that has maintained a broad-based capture of a broad (if not deep) "snapshot" of the World Wide Web since 1996. The Internet Archive utilizes cached data provided by its commercial search engine partner, Alexa, and organizes and displays the data through its Web portal, the "Wayback Machine" (http://www.archive.org/).

The Internet Archive worked with the Library of Congress and other national libraries through the International Internet Preservation Consortium to develop an open source crawler capable of performing focused crawls of data for institutions seeking to capture selected sites. The open source crawler, Heritrix, is used by a number of national libraries in their domain-based crawls.

With "Archive-It," the Internet Archive intends to provide a service for subscribers with limited technological capability to capture and search "collections" of distinct Web archives. A prototype was demonstrated in May 2005, and a memorandum of understanding between pilot participants and the Internet Archive was signed in July of that year.

Pilot partners were encouraged to select up to three collections of 100 sites each, the collection parameters set by the tool (since the pilot, Archive-It has modified its service to allow for a single collection of up to 300 sites). CRL engaged a number of its constituents to identify up to 100 sites for archiving purposes, and to indicate the desired frequency for harvesting each during a two or three month pilot period. Participants were asked to assist in the analysis of results and advise CRL on the usefulness of the archiving for library purposes.

CRL took three separate approaches to the IA pilot[1]:

- Subject-based crawl – Web sites of Middle Eastern political parties
- Event-based crawl – Sites relating to Liberia's presidential election 2005
- Format-based crawl - Selected on-line newspapers from Africa, Latin America, Southeast Asia, and the Middle East

These three crawls reflect the types of resources the PCWA investigation had identified as "fugitive" materials–resources susceptible to frequent content update, change, or disappearance. The resources also corresponded to the range of materials identified as particularly important by users of Web-based materials in their research. Specifically, the User survey conducted on behalf of the PCWA identified the following resources as important resources to archive:

1. Government sites
2. Political party sites
3. Online newspapers and news sources
4. Sites related to political and non-governmental organizations (incl. IGOs, international human rights organizations).

II. "Archive-It" services

Archive-It was first introduced in 2005 by the Internet Archive, and has been under continual development since the initial approach to institutions. According to the Internet Archive, Archive-It is a "Web based service that allows partners to create, manage and search their web archives through an easy to use web interface." It was designed as a low-barrier entry to Web harvesting, particularly for institutions with limited technical infrastructure to construct their own harvesting program. Functions within Archive-It include Web site crawling, organization and data management, technical reports for crawl monitoring, an interface to input site metadata, and full-text searching.

[1] A fourth distinct crawl was undertaken based on recommendations made in the PCWA report–materials based on a pre-selected list of sites created through a curator-driven Web portal. In this case, we selected Southeast Asian human rights documentation listed in the University of Wisconsin's "Portals to Asian Internet Resources" (PAIR). A detailed study of these materials has not yet been undertaken.

Archive-It operates fully on open source tools: in addition to Heritrix, the search function runs on open source software called "Nutch," which builds on Lucene Java, adding Web-specifics, such as a crawler, a link-graph database, parsers for HTML and other document formats. Nutch has been modified to search Web archive extensions, such as the .arc files used by the Internet Archive. Crawled sites are displayed using the open source Wayback Machine platform, although "collections" crawled under the Archive-It service do not appear in the regular Wayback service until after a period of time.

Archive-It was designed for maximum configurability and flexibility. The crawler was designed to capture any material that can be "downloaded from the public Web without direct user intervention." It captures multiple file formats, text in any language or machine-readable script, and content in a variety of design styles and formats (e.g., static HTML, .php and .asp files).

The Heritrix crawler has a number of admitted limitations, as do most available crawlers. Archive-It does not crawl the "deep Web" or any materials in databases or pages that require users to enter form data. It also will not crawl password-protected sites, and it does obey "robot.txt exclusions." That is, the crawler will not capture sites that a site owner has requested not be crawled. A site owner objecting to material crawled can put a robots.txt exclusion in place, which will stop the site from being harvested and will eliminate access to any versions that may have been previously harvested using Archive-It.

Other limitations of crawling include difficulty with JavaScript elements, streaming media, server side image maps (like other functionality on the Web, if it needs to contact the originating server in order to work, it will fail when archived).

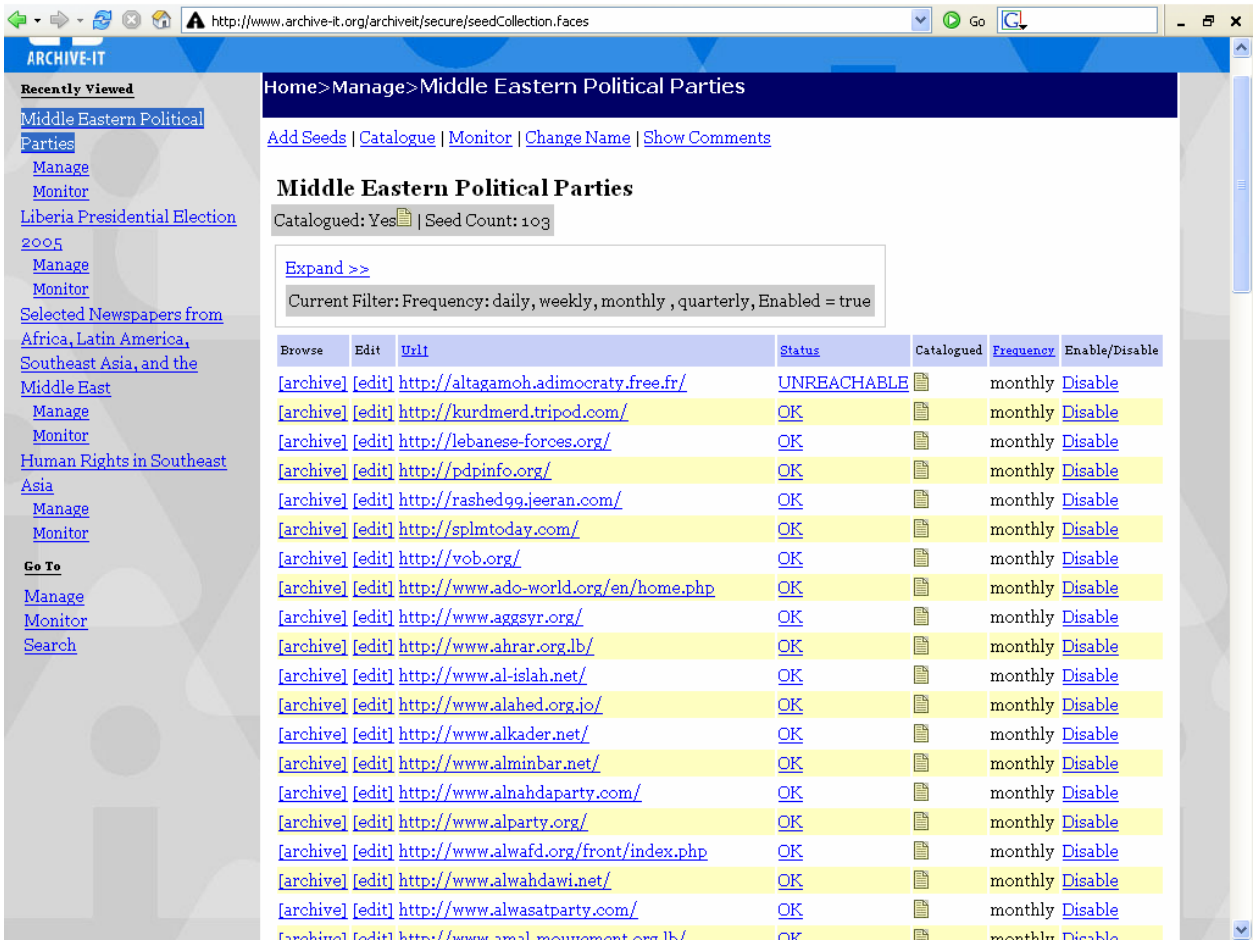

Figure 1 - Archive-It administrative interface showing the "Middle Eastern Political Parties" collection.

The administrative interface of Archive-It demonstrates the relatively straightforward, technologically simple nature of the tool. In order to set up a Web "collection," a subscriber must simply submit a list of "seed URLs" from which the crawler will begin. Once the collection is populated, a subscriber can add cataloging metadata (Dublin Core) to sites, specify frequency of crawls (from daily to quarterly), and edit collection-level information.

At the time of the pilot crawl, Archive-It specified that all crawls would get all documents associated with the root URL, as well as any site "one hop out of scope," or sites that are linked directly from the seed URL. This function has been discontinued for participants as of this writing, as crawling “one hop off” posed a number of complications for users.

The pilot beta application began crawling in September and ended November 18, 2005. Subscription costs were waived for pilot participants. Throughout the pilot program, participants were encouraged to give feedback to Internet Archives, and a number of improvements and upgrades were made in mid-program.

III. Pilot Assessment

Subsequent to the crawl period, CRL tasked bibliographers and specialists involved in the initial selection to perform an assessment on the crawl results to determine the efficacy of the crawl and the utility of the service. CRL asked its partners to assess the tool based on the following guidelines:

Quality of capture
- -Availability of all text, images, file types on a site
- -Satisfactory display of scripts
- -Errors in the capture (such as the current date appearing in frame)
- -Errors in navigation (such as linking externally to the live site instead of within the archived versions)

Depth of capture
- -All pages within a "seed URL" captured? If not, where is the error?
- -Pages outside of the "seed URL" captured? If so, how far down were they captured?

Frequency of capture
- -How frequently did the site change (i.e. how often do new versions appear in the archive)? Were the changes major or minor?
- -Did Archive-It accurately note the changing content of the site? That is, did the " * " accurately show when changes in content happened, and if not, how?
- -How frequently would you recommend this type of site or collection of sites be crawled?

Searchability
- -Does the search capture the term(s) you are using to seek?
- -Does the search support searching in non-roman?
- -Is the search interface confusing?

Overall evaluation
- -Based on your assessment, was this collection worth crawling?
- -Does this crawler meet your needs as a curator of archived sites? Would it meet the needs of your users?
- -What would the ideal display of this material be like? Are there particular pages or documents we would want to link to? Or would we leave the discovery of resources up to the researcher?

Assessments based on these criteria were submitted by the specialists and fed into CRL's further investigation of the benefits and drawbacks of Archive-It crawling and searching. Based on the submitted assessments, we concluded that in its current incarnation Archive-It is a useful tool in many cases, but also poses a number of limitations that need to be addressed before broad-scale implementation.

Some of CRL’s findings, including limitations of the service are as follows:

External links

Although the pilot parameters specified that the crawl would capture “one hop off” the seed URL, we found there was great inconsistency in this implementation. In some cases, crawls did capture links off of the seed URL, but in a majority of cases it did not capture these materials.

Crawled sites contain a variety of pages and content hosted within the seed URL, although in many cases they link to external resources, such as interviews and reports hosted by affiliated organizations such as *al-Jazeera*. Also typical, many sites host content on multiple network locations even though it is intended to be experienced as a single “place” by the user (also referred to as a “supersite”). Though technically the seed URL is different, the content is related and hosted by the same organization. "One hop off" should theoretically capture this material, though the crawler failed in a majority of cases.

Varying URLs on site

lebanese-forces.org → (captured)

www.lebanese-forces.net → (not captured)

bachirgemayel.lebanese-forces.org → (not captured)

www.lebanese-forces.org → (not captured)

Figure 2 - http://lebanese-forces.org/ Archived site (11/11/2005)

As mentioned above, Archive-It has currently suspended the “one hop off” feature. Participants need to be aware of the necessity of including additional seed URLs if the selector desires the content therein to be captured.

Internal links

In some cases, pages within sites were inconsistently captured. There did not seem to be a consistent identifiable reason behind this, save perhaps for the relative complexity of sites in which pages were missed. Sites using more complex generation of pages (.asp or .php linking to individual articles) demonstrated serious navigability problems in Archive-it. For example, *al-Quds al-Arabi* (http://www.alquds.co.uk/), an Arabic language newspaper published in the United Kingdom, utilizes relatively sophisticated .asp code to pull and deliver issue sections and individual articles. The capture crawled the index page and the second level for each news "section" (economy, politics, etc.). However, the crawl failed to capture any further links to specific articles in every edition. The crawl also did not capture any of the PDF versions of the title on the

site. Upon investigation, the Archive-It support team indicated that the site should have been crawled—data collection statistics showing the number of URLs crawled would seem to support this. However, either due to a flaw in the capture or in the Wayback display, much of the site's content is inaccessible to users. This is a serious deficiency in capture for this site.

Non-Roman script:
The most disappointing limitation of the crawl tool is its current inability to search archived Web materials utilizing non-Roman script. While the capture of Web sites, by and large, correctly capture and render Arabic, Tamil, Burmese, and other non-Roman scripts, searching utilizing the Archive-It tools is limited to Roman script only. This is a serious inadequacy for the applicability of the current service to international collections. At this writing, Archive-It reports it is investigating the possibility of modifying the search functionality to search non-Roman script.

IV. Administrative Tools Assessment

In addition to the curatorial evaluation of the sites crawled, CRL examined available technical and administrative tools as well as some of the data reports presented therein to assess whether administrators of the service are provided with sufficient data and capacity to manage crawling activities.

The administrative tools provided by Archive-It present a general overview of each crawl, indicating the start and end time of each crawl, amount of data collected (total MB), number of URLs crawled, and general rate of data transfer. Detailed reports provide:

a) A detailed breakdown of data collected per crawl, with URLs and bytes captured per seed URL;
b) MIME types per crawl, with a breakdown by type of file and size; and
c) A "seed status report" for each crawl by URL, documenting the success (or failure) of the crawl.

Figure 3 - Monitoring subsection of Archive-It.

The amount of data provided allows for a reasonable non-technical assessment of each crawl's performance. The data provides, for example, concrete information on the size and number of URL's captured per each seed. A specific example of one crawl report follows:

Middle Eastern Political Parties 103 Seed URLs

Summary Report

Finished (TL):	11/14/05 01:49 PM
Start Date	11/11/05 01:35 PM
Data Collected	32318.56 MB
Documents Crawled	701,697
Average Doc Rate	2.61/s
Average KB Rate	34.00 K/s

The number of "documents crawled" as reported is higher than the sum of total documents crawled from the selected seed URLs likely due to two factors: a) some seed URLs in the crawl were included in error (at the time of the pilot, selectors could not remove seeds from a collection once selected), and b) some, but not all sites "one hop out of scope" were captured in the course of the crawl (often, advertisements, site counters, or actual content links). Based on an examination of the individual seed URL counts, it is estimated that each crawl captured approximately 475,000 relevant unique URLs per crawl.

Top 5 Hosts/size

HOST	URL_COUNT	BYTES
www.marzeporgohar.org	130076	1953252352
www.ssnp.com	56542	1202561619
www.qudsway.com	44612	475391634
lebanese-forces.org	33764	1376461315
www.usfp.ma	25792	7325032982

By size, the 475,000 URLs captured took up approximately 25,056,447,160 bytes (23,896 MB). The average size of sites crawled was 5,000 individual URLs or documents, with a file size of about 250MB.

There are some limitations as to what one can do with the available data. For instance, the method of viewing reports does not currently allow administrators to easily combine reports or view data over a period of time, which might be a useful assessment of a seed URL's growth or change over time.

Additionally, the data provided does not tend to expose problems other than significant errors with the crawl (redirects, HTTP status errors such as code 404 [Not Found] or 500 [Internal Server Error]). Given the inconsistencies noted above, a more robust crawl report would be useful for proper feedback for and response by the participant.

At this time, technical or administrative metadata cannot be automatically derived from captured Web sites. That is, there is no mechanism to extract header information and metadata embedded in sites and pages. Thus, cataloging or description of the sites being crawled must be performed manually, a time-consuming, and hence costly, process.

V. Cost Assessment

The current pricing structure for the Archive-It service is set at $10,000 per year minimum, which allows the subscribing institution to capture up to 10 million unique URLs during a year. Archive-It characterizes the size of the subscription as allowing for roughly 50-100 sites, crawled once per month. Archive-It closely monitors the pages crawled and reports on the subscriber's "Annual Document Budget" to ensure they do not unintentionally crawl over the allotted amount (a subsequent 10 million site "budget" can be purchased for approximately $7,000).

Based on the sample data shown in Section IV, it is possible to extrapolate a base cost for crawling individual sites. At $10,000 for 10 million URLs, each URL costs $0.001 to crawl. Thus, an average site of 5000 URLs would cost approximately $5.00 per crawl. The largest site crawled in the pilot, www.marzeporgohar.org with 130076 URLs, would cost $130 each time crawled.

This amount is very economical on the one hand, considering the sheer amount of content provided at such a minimal cost. However, costs are dependent on the number of times a site is crawled and the amount captured each crawl. Crawls are repeated sequential full snapshots, not incremental, self de-duping harvests. That is to say, a crawl of an individual site will capture the entire site every time, even if many of the site's pages had not changed. For instance, the *Marze por Gohar Party* (an Iranian opposition party in exile) site contains information on the party, news and articles, political statements and manifestoes, photos, videos, audio clips and more. Certain portions of site changed daily during the pilot crawl, but the majority of pages on the site are static pages from past events. To capture this site daily would take in all the necessary changes, but would also grab the static archived pages. The annual cost of this, consequently, would balloon to nearly $35,000 annually for this one site alone.

One must also factor in the costs of selection, cataloging, review, quality control, and maintenance of the site in the overall cost of the project. An assessment of these costs was not undertaken as part of this analysis, but general cost conclusions of such activities were made in the PCWA report mentioned above.

VI. Archive-It as model for Web harvesting: an OAIS assessment.

The Political Communications Web Archive project specified an ideal distributed Web archiving model. The description included four elements for each activity or "layer" of the model:

1. Functional requirements: the activities, processes, and outputs of the activity or "layer."
2. Participants: the general characteristics, skills, and capabilities of the individuals or organizations undertaking the activity.
3. Cost factors and sensitivities: the general types of costs and the factors that influence the cost level for the activity and incentives for investment by participating organizations and entities.
4. Accountability and control: the organizations or constituencies to which the entity performing the activity is accountable, and which exercises control of the inputs and outputs of the activity.

Based on the PCWA specifications for the ideal Web-archiving regime, it can be said that Archive-It does provide certain core services for a trustworthy preservation and access model, although it requires that certain key activities and thus costs must be borne by the stakeholder community. Those activities are as follows:

a. Selection:

Functional requirements: Selection/Curation activities involve identifying authoritative and appropriate content, and determining the technical and curatorial standards for capturing and preserving that content.

Consistent with the PCWA model, Archive-It facilitates distributed selection: archives can be developed and maintained by local or specialized communities, or even by individual researchers, who possess specialized expertise on the area of archiving focus. Under this model selectors are responsible for searching Web content, specifying the content to be captured and preserved, determining the moment, frequency, depth, and scope of capture, annotating content and providing descriptive metadata.

Currently, the Archive-It service provides minimal automated technical information about sites. Technical reports include details on numbers of URLs and Bytes captured per site per crawl, the MIME-type files captured, and a basic crawl report and any problems encountered. It would seem it is capable of searching and capturing a certain amount of technical metadata, but there is no current implementation of automatic metadata generation.

b. Stewardship:

Functional requirements: Stewardship is a critical activity, upon which responsibility for continuous management of the archive's content and assets ultimately rests. Stewardship supports and monitors services and functions for the overall operation of the archiving effort; makes decisions and executes transactions pertaining to scope of the archives, participants, accessibility, and disposition of archives content and related assets; provides the nexus which gathers and pools the expertise and resources of diverse institutional and individual participants; establishes, formalizes, and monitors fulfillment of the terms for archiving activities, and ensures that standards and specifications for selection and presentation of content accord with User needs, as expressed in the collection development policies and governance.

The stewardship role is assumed by the participating institutions in the service. Each participant is individually responsible for maintaining their own monitoring activities of the Web archive. The institution is responsible for notifying hosts/producers on the means and terms of archiving, creating and administering policies regarding selection and access; authorizing or "accrediting" selectors, access providers, administration, and other participants per standards established by users or their proxies; monitoring and controlling ingest of new content to the archives and selector activity; and ensuring "authenticity" (chain of custody) of the archive's content.

This, of course, also applies to financial asset management as well, since the scale of archiving activities will be reliant in part on the flow of funds and other resources. In this role Stewardship ensures that the scale of archiving activities and archives content are in line with the supporting resources. The addition of content to the archives, level of functionality in the presentation of content to users, and other cost-generating activities will have to correspond to levels of income or investment provided by the user communities directly or through their libraries and organizations.

c. Ingest / Harvesting:

Functional requirements: Content is captured for archiving directly from the Web. Ingest activities include "pointing" or programming the Web crawler/harvester; undertaking the site crawls; receiving file data and capturing or generating the corresponding technical metadata, and notification of the Producer/Host about the archiving activities.

Archive-It performs most of these activities: crawling the sites and documents; storing and replicating the composite files according to Selection standards and specifications; documenting, and annotating content with, circumstances of capture (such as date, method) providing baseline documentation for "chain of custody" of archived content; and ensuring integrity of content and important metadata (the AIP descriptive information). Archive-It assumes the high-level programming expertise and functionality needed for centralized harvesting and ingest. Archive-It, however, does not notify producers of the Web sites harvested, requiring them to "opt out" of the archive by actively notifying Internet Archive or blocking access through "robots.txt."

d. Administration and Data Management:

Functional requirements: Administration activities involve monitoring and controlling data flows and auditing and certification of data and processes. Administration also monitors additions to the Repository, and provides feedback to Selection based on crawl results, changes in targeted materials, and changing factors in the crawl environment, to inform subsequent selecting activities and criteria. Administration preserves the functionality of archives content and migrates content to new platforms and formats as needed. It provides quality assurance of data by ensuring appropriate configuration and functionality of Ingest, Repository, and Access systems (hardware and software). It develops, procures and maintains tools and technologies for selection, annotation, and presentation of archives content, according to requirements established by Stewardship.

Archive-It and the Internet Archives fulfill this role. The Internet Archive sets policies and standards, including terms of use; Archive-It provides appropriate access, tracks changes in the targeted materials,

and maintains the tools for ongoing activity. Notification and feedback for subsequent selecting activities is not as robust as the PCWA investigation recommended, although this may improve over time.

As a service provider, Archive-It offers an alternative to development and hosting of such a service by libraries and other communities themselves. On the other hand it remains to be seen how responsive IA will continue to be to the needs of its community of users, i.e., the research libraries employing the service. Entrusting to a third party maintenance of the tools required to archive, manage and present the materials selected poses a significant risk to libraries and their constituents in the absence of specific mechanisms guaranteeing accountability of that party. (The Internet Archives, the organization offering the Archive-It service, is a privately controlled not-for profit organization.) Thus far, however, Archive-It has demonstrated great flexibility and willingness to tailor the tool to meet the needs of participating libraries.

VII. Conclusions

The efforts of the Internet Archive to offer a broader range of archiving services, especially for smaller institutions with less robust technical capability, are advancing the important issues of the technical and curatorial requirements for this type of activity. For many organizations, Archive-It will provide a useful tool for archiving both their own and external Web content. As an increasing amount of scholarly resources is being made available exclusively on-line, this tool will provide the opportunity to capture, and independently maintain and display that content even after it has disappeared from the Web. Several of the pilot projects have moved on to become full subscribers to the Archive-It service. The archives provide access to content from the Web not previously offered through many other "over-the-counter" programs. The range of participant programs will continue to demonstrate the effectiveness and long-range impact of Web harvesting.

The costs of Web archiving through Archive-It appear economical at first glance. The costs of capture and storage of the material are minimal in relation to the amount of data crawled. However, one must factor in additional costs in the process to see the full picture. The Internet Archives assumes only some of the roles required in the optimal model for a trusted electronic repository. It is up to the subscriber to maintain other functions, most of which are quite costly activities to undertake. Selection, management of the sites being crawled, technical assessment to ensure sites are captured correctly, content assessment and discoverability of resources, and long-term administration all involve costs that must be weighed in the process.

One additional potential cost to note is one that might have to be borne by libraries in the event of a discontinuation or failure of the Archive-It program. This is the cost of extracting harvested content and metadata from Archive-It and re-hosting it independently as either a primary or backup archive. It is unclear what costs this process would involve. But they are costs that participating libraries should factor into their evaluation of the Archive-It service.

Some remaining issues bear further attention:

- *Copyright*: The Internet Archive effectively side-steps the issue of copyright infringement by stating that it does not "own" the archive content, but rather stores the material on behalf of its clients. Potential participants in the service should be clear about the implications of this approach with respect to their legal liability for archiving copyrighted materials.
- *Discoverability*: Archive-It's search functionality is somewhat limited. As mentioned, it cannot currently search and discover resources in non-Roman scripts. The search features do not allow for advanced search capabilities (such as searching metadata) and cannot currently search across discrete collections. To enhance discoverability, participants have created their own interfaces to their collections, creating manual links to specific sites or pages within sites. This adds cost to the management of the resource.
- *Persistence*: Internet Archives stores the data captured by Archive-It on two separate servers and provides links with persistent URLs. However, it is not clear upon what terms the Internet Archives will guarantee the persistence of these resources in perpetuity. The migration of Web content that requires client-side plug-ins (applications such as Flash multimedia or QuickTime) or

proprietary fonts may be quite costly in the future, even if they are stored correctly by the Internet Archive. The rationale for distribution of the costs of such migration, and the basis for decisions on what functionality will be maintained for Archive-It captured materials over the long term need to be articulated.

- *Transparency*: There is not a significant amount of transparency in the governance, technical infrastructure, and data security of the Internet Archive. A more detailed assessment of the Internet Archive as a trusted digital repository is recommended.

It has been demonstrated that the Archive-It service, as currently offered, has certain limitations that need to be addressed before optimal capture is achieved. However, it is likely that these issues can be more usefully addressed through ongoing engagement with the providers of the service rather than lack of participation. The Center for Research Libraries will continue to assess the methodologies and technologies of Web archiving for its constituent international resources collecting and preservation programs.

Revised 8/31/06