Submitted on: 03/07/2014

The integrity of research is at risk: Capturing and preserving web sites and web documents and the implications for resource sharing

James G. Neal
University Libraries/Information Services, Columbia University, New York, USA
jneal@columbia.edu

Abstract:

Born digital materials, for example websites and web documents, present particular challenges to academic and community libraries and their collection development, discovery and access, preservation, and resource sharing programs. This paper will explore the policy, workflow, legal, governance, financial, and service framework for capturing and preserving web content in the context of expanding collaborative collection development agreements among libraries. The paper will describe the program at the Columbia University Libraries in the area of Web Resources Archiving Collaboration.

Columbia has sought to situate its program within the broader mission of its Libraries, collecting content deemed important for current and future research, preserving the content for future scholars, but also providing access in novel ways to foster current use in teaching and research. The Libraries' program focuses on four main types of web resource collections:

- *Thematic collections of content relating to the University's academic programs;*
- *Web content from organizations and individuals whose archives are held by the Libraries;*
- *Significant content from the University's own website;*
- *Sites identified by researchers and library subject specialists as at risk.*

The Columbia Libraries are developing and testing models for collaborative action in four areas:

- *Developing the policy, procedural, legal, governance and financial framework to enable Columbia to provide web archiving services for a group of libraries in support of existing collaborative collection development agreements;*
- *Conducting a rigorous investigation of the use of web resources in 21st century scholarship in the area of human rights, and assessing what further development is needed in current web archiving efforts to support future needs in this area;*

- *Working with web content-producing organizations to provide tools and techniques that optimize websites for harvesting and preservation and to develop complementary methods and procedures for direct transmission of content that cannot be readily harvested from the web;*
- *Administering, through a multi-institutional advisory board, a series of competitive awards for innovative uses of web archives in research, and a second series of awards to support development of tools that facilitate the process of web archiving and/or the use of web archives.*

Keywords: born-digital content, web resources, scholarly integrity, resource sharing.

The integrity of research is at risk: Capturing and preserving web sites and web documents and the implications for resource sharing

Let's consider the following scenario. A scholar prepares a major research paper, and it is published in a prestigious journal. The paper includes in its references several web sites and web documents which are crucial to the arguments of the author. Another researcher reads the paper in the journal and disagrees with the conclusions and wants to consult the evidence. However, the web sites and documents are no longer retrievable and cannot be reviewed. The links are broken or the sites have been taken down or significantly changed. Is this an issue of scholarly integrity?

Or this second scenario. A researcher on the debate over climate change wants to locate and review the web sites and reports of organizations, government agencies and individuals that have presented data, arguments and recommendations over the past decade. The researcher finds that many of the web sites and documents no longer are available and thus cannot be consulted thus making it difficult if not impossible for the research project to proceed. Is this an issue of scholarly integrity?

Or this third scenario. A researcher completes an important research study and it is accepted by a leading journal in the field. The paper and the research data supporting the project will be available through the publisher repository. Because of the agreement that the author negotiated with the publisher, the researcher is allowed to make the work available in various open access sites. So the work also appears in a disciplinary repository, in an institutional repository at the researcher's institution, in a special academic department repository, in a research data repository, in a government repository because of the grant that funded the project, on the personal web site of the researcher, on the teaching/course web site of the researcher, in a national repository, and so on. Which version of this work will be discovered and used, cited, preserved without changes? Is this condition of repository chaos an issue of scholarly integrity?

The capturing, curating and archiving of web sites and web documents cuts across the full range of the core responsibilities of libraries. But our procedures, workflows, policies, and technologies are insufficient to the tasks of information selection, acquisition, synthesis, navigation, dissemination, interpretation, understanding, use, application and preservation. How do we confront the conditions of constant mutation, of waves of new software applications, of unfriendly copyright laws, of pressures to create the collective collection, of unrealistic user and researcher expectations? What are the implications for our library

resource sharing programs when we advance beyond the traditional delivery of analog publications and their digital surrogates?

Academic research libraries have long seen it to be part of their mission to build coherent collections of scholarly and research resources to support the needs of their institutions. To achieve and maintain this coherence, they select, acquire, describe, organize, manage, and preserve relevant resources—and, if only by default, they exercise lesser degrees of curation for resources deemed out of scope or of short-term interest.

For print (analog) resources, libraries have stable and generally well-supported models for building and maintaining collections. The roles and responsibilities of selectors, acquisition departments, catalogers, and preservation units are well understood and to a considerable degree interchangeable from one library to another. Specific procedures vary among libraries and change over time, but the basic model has remained the same.

For commercially published digital resources, models are emerging that diverge from past practice: resources are often purchased in large packages, rather than as individual titles; access is governed by license terms, rather than through physical receipt and processing; catalog records are increasingly supplied by intermediaries en masse, rather than created by the library. Still, the fact that business transactions are needed simply to provide access to basic resources ensures that these actions will be taken and that purchased resources will be managed as part of the library's collections.

As more and more non-commercial materials are available in digital form, this established concept of collection building is called into question. The role of any individual library in shaping collections is less clear when some digital materials are accessible regardless of the user's physical location or affiliations. "Acquisition" is not always necessary to provide access and may be insufficient to enable preservation. As retrospective collections are digitized from many libraries, locally developed print collections will lose coherence if they become disconnected from these emerging digital counterparts.

For non-commercial web resources, there is as yet no common understanding of what ought to be done to identify relevant resources, make them available, integrate access with other collections, and ensure that they will continue to be available for future users. Investigations during the 2008 Mellon-funded planning grant confirmed that individual aspects are being addressed in fragmentary fashion, with some attention given to selection, bibliographic description, and the technical and rights issues, but that such activity is largely confined within separate communities of selectors, catalogers, and digital library technologists. Few libraries have articulated a coherent end-to-end set of policies and procedures for "collecting" such content.

There are a growing number of international initiatives created "to foster the development and use of common tools, techniques and standards that enable the creation of international [Internet] archives," to quote from the mission statement of the most prominent such group, the IIPC (International Internet Preservation Consortium). IIPC members include over 30 international libraries and the Internet Archive, and it has working groups devoted to Standards, Harvesting, Access, and Preservation. A newer consortium, LiWA (Living Web Archives), comprising eight European organizations, is explicitly focused on technical advancements, promising to "extend the current state of the art and develop the next

generation of web content capture, preservation, analysis, and enrichment services to improve fidelity, coherence, and interpretability of web archives."

The work of these and other similar groups to develop and improve web archiving standards and tools eases the technical development burdens facing individual projects. Web archiving projects are increasingly numerous—major national library efforts include PANDORA in Australia, Minerva at the Library of Congress, and the UK Web Archiving Consortium—yet even these are in part still considered experimental, designed to gain experience with the processes and technology of web archiving, and often devoted to collecting a set of resources relating to a specific event of limited duration.

A much smaller number of programs have taken on a mission to collect and preserve web resources on a continuing basis. Typically, such programs focus on an organizational mandate such as collecting documents produced by a state's government, or web sites emanating from within the country served by a national library. The North Carolina State Government Web Site Archives is a particularly robust program of this sort. Generally, preservation of web content is the raison d'etre of these programs; few, if any, have made web content an integral component of the library's collecting activities.

These existing web archiving programs are in many respects complementary to the proposed program at Columbia. Even the programs which fall closest to the scope of Columbia's human rights web collection effort, such as the University of Texas's exemplary LAGDA (Latin American Government Document Archive), do not share our focus on at-risk NGO-produced content. A handful of other Internet Archive partners have assembled test collections of single crawls of selected regional NGOs or environmental NGOs. In some cases they are collecting and preserving human rights content that falls within the subject scope of Columbia's interests. In the future, other collections may match our scope more closely and lessen the need for Columbia to collect the same materials. The LOCKSS model (http://www.lockss.org/lockss/Home) is instructive, however, in suggesting the value of distributed work and the risk inherent in relying on single digital copies; if more than one library were to acquire the same important web content, the overall goal of preservation would only be enhanced.

During the course of its 2008 planning grant project, Columbia explored several possible models for each component of a web collecting program and identified methods suitable for a sustainable, scalable, continuing program. It now remains to test these methods in a production environment, apply them to the large body of relevant content identified during the planning grant, and embed these procedures in appropriate parts of the Libraries.

The Columbia Libraries views web content collecting as central to the mission of any research library, and it intends to make it an integral part of its collection building practices. The Libraries are putting into production procedures for selecting, acquiring, describing, preserving, and providing access to freely available web content. The Libraries will test and refine procedures and the tools used to implement them, and adjust the model to take advantage of technology improvements and changes in community understanding of best practices for web archiving.

Columbia's work will serve as a model for other libraries to use, adapt, and improve in their own web collecting activities. Our goal is to model the life cycle process of web content as part of a research library's collection development best practices that can be shared and

discussed with the wider communities of research libraries and scholars. During the final months of the project, Columbia hosted an invitational conference of major research libraries to promote discussion of this model and identify ways to promulgate its use. Columbia also created and shared a best practices document outlining recommended procedures, to ensure that results are available for wide distribution.

During the first year, the project focused on the retrospective collection of human rights content that has appeared on the web over the last decade, while developing tools to support an ongoing program. The second and third years continued this process and focused on use of the collected content in scholarly research, teaching, and learning. During this phase, the methods developed were integrated into Columbia's routine processes of collection development, description, and access. At the end of the project, it was expected that the cost to continue and expand this program would be incremental and sustainable.

A planning grant funded by the Andrew W. Mellon Foundation and conducted jointly by the Columbia University Libraries and the University of Maryland Libraries in 2008 demonstrated that it is feasible to implement a holistic model for incorporating web content in research library collections, but also showed that the field of web archiving is immature. Tools and procedures exist to support each component of a collecting program, but there is no commonly accepted body of best practices or agreement on objectives and desired outcomes.

The next phase of the project focused on the multi-disciplinary subject of human rights. As expressed in the Universal Declaration of Human Rights, this concept includes such commonly recognized areas as freedom from torture, slavery, and arbitrary arrest, but also embraces social, cultural, and economic rights, freedom of movement and assembly, the right to work, and more. During the 2008 planning grant, seventy distinct thematic areas were identified within the selected web sites. The content originates from non-governmental organizations, international bodies, government agencies, grass-roots advocacy groups, and personal blogs, and includes news bulletins, reports, case studies, audio, video, images, and maps.

Analysis conducted during the 2008 planning grant shows that the field of human rights provides a fertile starting point for web content collecting. A survey of holdings in WorldCat and in Columbia's print collections demonstrates that publications from human rights organizations have an important place in library collections. Of the 538 organizations surveyed, some 41% are included in Columbia's print collections, with holdings ranging from a handful of titles to well over one hundred. Nearly 70% of these organizations have had authority records created by libraries.

Despite this importance, print collecting from many of these organizations has been only marginally successful. In numerous cases, fewer than half of an organization's print publications have been collected by any library, the titles that have been acquired are not widely available, and holdings of serial titles are often incomplete. Interviews with library selectors make it clear that this spotty record is not a reflection of the importance of the materials, but of poor distribution and unavailability of these publications through standard acquisition channels.

Further analysis shows that the web content produced by these organizations is even less likely to receive notice in library collections. Over 20% of the surveyed organizations have no records in WorldCat, despite having significant publications available online. For the

remaining organizations, online content is typically represented, if at all, only by a single journal or the organization's annual report.

At the same time, this content is receiving increased scholarly notice. Recent articles in scholarly journals devoted to human rights frequently cite online reports, news stories, case studies, and documents. While the output of major organizations such as Amnesty International and Human Rights Watch is most frequently cited, sources also include many smaller organizations from Africa, Asia, Latin America, and the Middle East.

At Columbia, this content is important not only to the 57 departments, centers, and institutes studying human rights, but to the programs of Columbia's Earth Institute in such areas as poverty, global health, environmental hazards, and sustainable development. During the course of this project, every attempt was made to ensure that all relevant academic programs are engaged in the selection and evaluation of content. As the project progressed, subject librarians and scholars were encouraged to recommend content extending beyond human rights to related political, social, and environmental interests.

The components of this proposed program derive from work completed in the planning grant. They include:

1. Selecting appropriate content and describing and organizing the selected resources
2. Seeking permission to archive
3. Harvesting and archiving content
4. Describing and organizing content
5. Disclosing actions and intent
6. Making material available for use
7. Assessing results

While these processes are largely sequential, project staff worked in an iterative fashion, refining procedures as the web content collection expands and the available tool set evolves.

We have identified and characterized 600 human rights web sites through tags on delicious.com at http://www.delicious.com/hrwebproject. A sub-group from these sites has been evaluated by Area Studies librarians and used to gain experience with Archive-It software. These sites were selected based on such factors as the importance and nature of the content; country of origin; type of organization; overlap with print collections; and perceived risk that the content may disappear or be removed from the web. This sub-group will form an initial set of content for further development—refinement of harvesting scope; seeking permission to archive; and description, organization, and disclosure.

While this work proceeds, methods for further selection will be put in place, initially using the remaining sites from the tag group. A web-based form will be used to solicit input on these sites. The form will be circulated through listservs of librarians and scholars interested in human rights (such as H-Human Rights, the listserv developed by the Human Rights Section of the International Studies Association, http://www.h-net.org/~hrights/). These groups will also be encouraged to nominate additional sites for consideration.

Beyond this direct selection, several methods will be tested to identify new sites of interest. RSS feeds from delicious.com will identify sites newly tagged with appropriate terms. As harvesting progresses, new links appearing in harvested content will be examined for possible

selection. Project staff will work to establish connections with other institutions maintaining data on Non-Governmental Organizations, such as Duke University's NGO Research Guide, the Minnesota Human Rights Library, and the database underlying the IGO/NGO Search provided through the GODORT Section of the American Library Association (http://wikis.ala.org/godort/index.php/IGO_search).

As new sites and content are identified, standard criteria will be used to determine appropriate treatment. In general, web sites based in countries with strong national archiving programs and those emanating from government agencies and research universities will not be given priority, in order to focus on content more likely to be "at risk." For these sites, Columbia will focus on creating or enhancing metadata to ensure appropriate access.

Input from selectors will be used to identify important characteristics of each site, and those characteristics will guide decisions about harvesting, such as the importance of linked sites, frequency of capture, and the depth of content analysis required. As we gain experience, general policies will be developed to minimize the need for explicit analysis.

During the second and third years of the project, the Web Collection Curators worked with Columbia's Collection Development Office to document these policies and procedures and promote their use in additional subject areas, to make web content an integral part of collecting responsibilities.

Explorations with several human rights organizations during the 2008 planning grant suggest that many are willing to grant permission to archive their web content, so long as the process does not place burdens on the organization's work and the archived content is not restricted.

Accordingly, Columbia will attempt to develop formal agreements for archiving whenever feasible. The Web Collection Curator based in the Rare Book and Manuscript Library will work to develop explicit agreements with organizations for which Columbia holds paper archives, such as Human Rights Watch and Amnesty International. The Curator based in Global and Area Studies will develop a generic Memorandum of Understanding for web harvesting and will work through the Area Studies Librarians to secure agreements with selected organizations in other world regions. Initially, these agreements will be modeled on those developed and tested by other web archiving programs, such as the PANDORA permission letter templates (http://pandora.nla.gov.au/manual/general_procedures.html#formlet).

When it is not feasible to establish contact with a web site owner and the content is considered "at risk" of disappearing, the Curators will document attempts made to secure permission. In such cases, web sites will be harvested by non-intrusive means following the principles recommended by the Section 108 Study Group in its discussion of Preservation of Publicly Available Online Content (Section 108 Study Group Report—An Independent Report sponsored by the United States Copyright Office and the National Digital Information Infrastructure and Preservation Program of the Library of Congress (www.section108.gov/docs/Sec108StudyGroupReport.pdf) and practices developed by other web archiving programs, including: respecting robots.txt files; framing harvested content to clearly indicate its nature; linking to the original site; and removing harvested content upon request by the owner. The Curators will make these policies publicly available and will continue to monitor both legal requirements and best practice in this area, consulting with Columbia's Copyright Advisory Office.

During the second and third years, the Curators worked with Columbia's Collection Development Office and with the Continuing and Electronic Resources Management Division to document procedures for seeking and recording permission to archive, and to develop a routine workflow parallel to that for license review of electronic resources.

Existing web harvesting tools are primarily of two kinds: commercial-hosted services that combine crawling and archiving, and commercial or open-source locally run tools that allow more flexible crawling but require more local technical support and do not address archiving.

Chief among the commercial hosted services are: the Archive-It web application offered by the Internet Archive, currently used by several dozen academic and government partners; the Hanzo Archives, focused on records management and corporate clients; and the OCLC Web Harvester, which attempts to be a hybrid service in that it requires bundling with OCLC's locally run CONTENTdm digital management software. The most evolved and widely adopted of the opensource locally run tools are the IIPC-developed Web Curator Tool and the Danish NetarchiveSuite.

During the 2008 planning grant period, Columbia initiated a 30-day free trial of Archive-It, ran several crawls, and then entered into a one-year contract. OCLC provided a demonstration of their Web Harvester, but the restrictions limiting use to CONTENTdm made this an unsuitable choice. With respect to locally run tools, we were initially intrigued by the Web Archiving Workbench, an OCLC project that we were disappointed to learn had been discontinued and subsumed into the Web Harvester.

We also experimented with the PC-based WebCopier Pro, a commercial software product, in order to evaluate the functionality it provides for local harvesting and processing of web site content. While WebCopier Pro has many good features, we had questions about its robustness for large-scale harvesting efforts and concern about basing our ongoing strategy on a commercial product from a small company (Maximumsoft http://www.maximumsoft.com/) with a single product line.

Meanwhile our grant partner, the University of Maryland, downloaded and tested the opensource Web Curator Tool and shared their written evaluation with us, and project managers from Columbia and Maryland discussed their respective experiences with Archive-It and Web Curator Tool on the phone and in follow-up email correspondence. While Web Curator Tool offers certain advantages in tracking permissions and selective harvesting of individual documents, its use for these purposes is labor-intensive and less suitable for full web site harvesting.

After crawling over 80 seed sites using Archive-It, we were familiar with its advantages and shortcomings and remained confident that Archive-It was the best available option and was actively improving, through development of new features driven in part by partner feedback. Recent new features include the display of seed-level metadata on partner pages and the inclusion on the same pages of a link to an automatically extracted video collection derived from a partner's archived content. (See Columbia's partner page at: http://www.archive-it.org/collections/1068.) Most promising among forthcoming features are: a de-duplication component (expected to be released soon) that will allow re-crawls of a given seed to harvest only its new and/or changed content, saving storage space; and the possibility of adding document-level metadata.

While our large-scale harvesting will be handled through Archive-It, we will more fully test WebCopier Pro, Web Curator Tool, and possibly other tools in the context of our second and third year work towards integrating selected web content into our local environment.

If Archive-It diversifies from the best-available service for whole-site archiving to also enable more flexible document-level organization and access, then our current plan to migrate our archived content in 2–3 years into a locally hosted environment to maximize its discovery and use could become less pressing. In the meantime, Archive-It will be used to acquire all content deemed of potential interest, subject to the technical limitations of web crawlers, and respecting all robots.txt restrictions.

Based on how frequently the roughly 80 sites that we have crawled update their content, many sites could be crawled as little as once or twice a year. Fewer sites (including the large NGOs whose physical archives are housed at Columbia) are updated often enough to justify quarterly or even monthly crawls. We may also harvest sites thought to be at greater risk of loss more frequently. The budget requested to support use of Archive-It allows storage of up to 15 million documents and 1.5 terabytes.

The Curators will also be responsible for regular quality assessment of crawls. During the first months of the project, the Curators will develop a standardized checklist comparable to that used by the North Carolina State Library and Archives (http://webteam.archive.org/confluence/download/attachments/3979/Crawl_Verification_Steps_2007_03_30.pdf) to ensure adequate and consistent quality control.

Some of the content available from human rights web sites corresponds to publications also (or formerly) issued in print, while the site as a whole often resembles an archival collection, with a great deal of ephemeral content and minor documents grouped into related series—news reports, press releases, images, case studies, etc. A multi-faceted approach to providing access is necessary at present to take advantage of the different venues used by researchers for discovery. As we gain experience with the techniques described below and are able to assess their effectiveness, and as techniques for integrating access across different types of records continue to improve, we will simplify description to those methods found to be most cost-effective and sustainable.

Initially, building on analysis completed during the planning grant, we will generate brief MARC records for all selected sites from delicious.com metadata via an automatic process with limited manual review. The resulting records will follow a model for access-level records established by the Library of Congress and since applied effectively by Columbia for Internet resources. Through this technique, web site-level access to a large number of organizations can be made available immediately in Columbia's online catalog and through OCLC's WorldCat.

For more complex web sites and for groupings of content, the Curators will create finding aids, as described in "An Arizona Model for Preservation and Access of Web Documents" (R. Pierce-Moses and J. Kaczmarek: An Arizona Model for Preservation and Access of Web Documents. Dttp: Documents to the People. 33:1. P. 17-24, 2005). With web-based resources, a finding aid can provide multiple ways of organizing the same material virtually—by format, topic, etc.—within a single web site's content, across multiple sites, and in relation to Columbia's print, archival, and other electronic human rights collections.

This approach has been successfully applied at the National Archives, London, and in the Matthew Shepard Web Archive at the University of Wyoming.

During the first year of the project, finding aids will be created for individual web sites. These finding aids will be made available through Columbia's web site in a presentation modeled on our Archival Collections Portal, through OCLC's ArchivesGrid, and via the web site of the Center for Human Rights Documentation and Research. As the collection grows, the Curators will test models for cross-organizational finding aids, with series highlighting specific topics, regions, or genres.

For selected serials and documents, the Curators will create MARC catalog records according to prevailing library standards for formulating names and identifying titles, in order to allow effective integration with existing library collections, and to facilitate reference linking. For many sites this will not be necessary, as the individual documents are less important than groups or themes, best described by other means. For others, interviews held with selectors during the planning grant suggest the types of documents for which separate catalog records are deemed important: serials and reports analogous to those that have been collected in print. Many of these resources have standard identifiers (ISSN or ISBN), and catalog records will facilitate their discovery through Open URL links. These records will follow the same Library of Congress access-level model discussed above, but applied to individual documents rather than entire organizational web sites.

Initially, costs to create these individual catalog records will be low, as existing records for print counterparts can often be repurposed with little modification. As the program develops, these types of documents will no longer be collected in print, and the work of cataloging the web versions will be largely substitutional, and thus sustainable.

During the second and third years, three student interns will be hired to assist with the creation of metadata for newly added sites and to update the initial set of catalog records and finding aids as web sites are re-crawled and new content added to the archive. During this same period, the Curators will work with the Digital Library Analyst/Developer to develop and test models for presentation of the finding aids, cross-collection searching, and linking to content at various levels.

These different types of finding aids and records are all necessary at present because approaches to describing web resources must still be considered experimental. During the early stages of the program, portions of the same content may be described and exposed in different ways, in order to ascertain which methods are most effective for access. Assessment of costs and usage will help determine which access paths and descriptive methods are most effective in guiding users to these materials. These multiple approaches can be undertaken efficiently because even at this experimental stage little of the work will be duplicative; rather, sets of metadata will be repurposed and recombined for different presentation. MARC catalog records will in many cases be derived or adapted from existing records describing printed serials and documents. Authority work for names will be done once (if needed at all) and re-used in multiple contexts. Sections of finding aids for an organization describing groups of related documents will be extracted, used to generate MODS records (see below) or recombined to create new finding aids organized around a topic or region. Once these techniques have been applied to a broad body of content, the results will be evaluated for effectiveness, and only those approaches found to be most useful will be continued.

For any web collecting program to be effective, its results must be transparent to the wide library community. A significant problem with current activities is the difficulty of determining whether a particular web site or resource is being captured, and if it is, with what degree of continuing commitment. In the absence of any commonly accepted standards for describing web resources, finding resource descriptions through the open web is largely a matter of guesswork.

For these reasons, the project will use several approaches in disclosing its work beyond Columbia's local discovery systems. In order to relate the collected resources to corresponding and analogous print collections, the Curators will create standard library catalog records for selected serials and documents—that is, descriptions that follow prevailing cataloging practice for identifying titles and forms of names—exposing those records in WorldCat and registering those that have been archived in OCLC's Registry of Digital Masters. Such disclosure will allow other libraries to harvest these records and to both substitute for and supplement their collecting of print materials from human rights organizations.

The Curators will also generate collection-level and series-level records from finding aids using the Metadata Object Description Schema (MODS) maintained by the Library of Congress. While methods and details differ, this level of cataloging is finding increasing favor in web archiving programs such as those at the Library of Congress and the National Archives in the United Kingdom. The resulting records are not yet available in one place, but this common approach offers the potential for record sharing through WorldCat, the European Web Archive, and similar aggregations.

During the second and third years, we will develop and refine a strategy for integrating archived web resources into our campus search and discovery environment. We will explore such approaches as: using archival finding aids to describe both archived and live web sites (as described above) including those harvested iteratively over time; integrating individual components of web sites (such as working papers and other documents analogous to print publications) into the Libraries' catalog and OCLC; linking archived web sites and site components to related resources at Columbia and elsewhere; presenting archived web site components within the overall context of Columbia's electronic resources, specifically in conjunction with the resources of Columbia's Center for Human Rights Documentation and Research.

This work will be supported by a Digital Library Analyst/Developer who will work closely with existing Columbia technical, archival, and public services staff. While current planning is based local archiving of the document-like content within Columbia's Fedora-based repository environment to provide the basis for better discovery, access, and management of these key resources.

We will also build on Columbia's relationships with several human rights organizations that have deposited their print archival collections at Columbia; namely, Human Rights Watch, Amnesty International USA, and the Committee of Concerned Scientists.

- In addition to using Archive-It for overall web site harvesting, we will use alternative harvesting technology, such as the Web Curator Tool, (http://webcurator.sourceforge.net/), to archive significant document-oriented content from the three targeted human rights organizations' web sites. (This strategy is based

in part on discussions with faculty and researchers who have emphasized the importance of archiving document-type material—e.g., area reports, thematic reports, annual reports—for their work.)

Document-type content from the three target organizations' web sites will then be deposited in our Fedora-based asset repository with metadata created through a combination of human and machine-assisted techniques. We will also explore the feasibility of creating RDF (Resource Description Framework data model) linkages to the original versions of documents in context as stored in the Internet Archive. These locally stored document copies will then be re-exposed as part of an evolving "Human Rights Electronic Reference Collection" hosted on Columbia's Center for Human Rights Documentation and Research web site. Having local copies of the documents will improve searching and indexing and allow their content to be accessed in conjunction with other related documents.

Finally, these harvested electronic document series will be selectively cataloged in CLIO and pushed out to OCLC with links to the locally stored copies. This will be especially useful in cases where the Libraries has already been collecting the same titles in printed form (e.g., the "Human Rights Watch world report"), and the electronic versions can then be presented bibliographically as a continuation of the print publications.

- In order to more fully expose the content of harvested sites, we will also explore tools and techniques for generating basic XML resource maps of fully harvested sites that can serve as the basis for creating both human and machine-actionable representations of the sites' content. The creation of simple geographic and thematic taxonomies will be accomplished by using a combination of metadata and content embedded in web pages, directory paths, and some human intervention.

  Once a site has been crawled and analyzed in this way, it should be possible to generate an XML resource map of a harvested instance of a site stored in the Internet Archive. This information would be exposed more broadly using, e.g., the OAI-ORE resource map protocol along with other newer techniques for describing aggregations in machine-processible ways. Further, it should be possible to formulate the XMLbased resource map information in such a way that it can be searched and displayed in conjunction with the standard EAD finding aid for the site, acting as the functional equivalent of an archival "container list," but with direct links to the corresponding archived content.

- To the extent we are successful in creating resource maps that reflect geographical and thematic content of the three target organizations' web sites, we should then also be able to create a composite resource map for the three sites with geographical and thematic content correlated and linked using appropriate RDF syntax.

  This merged resource map could be externalized as a searchable and actionable entity, while at the same time allowing us to create a more effectively integrated presentation of human rights documentation within the context of the evolving "Human Rights Electronic Reference Collection."

Once we have harvested, mapped, and indexed a targeted corpus of human rights content, we will be positioned to explore the use of additional technologies such as automated metadata extraction, contributive/collaborative taxonomy building, and semantic web approaches. The staff of Columbia's Center for Digital Research and Scholarship will provide guidance in this area. During the course of the project, we will work to implement approaches that can continue to operate beyond the end of the grant, to sustain and grow the value of the virtual human rights library we have created. Library selectors and curators will continue to be able to recommend the harvesting of web sites and, where appropriate, deeper harvesting and integration into our locally maintained collections. The Web Curator Tool (or whichever tool we settle upon during the project) will be added to the suite of supported tools we use to grow our digital collections. These new tools and strategies will allow us to continue working with human rights archivists and professionals affiliated with Columbia's Center for Human Rights Documentation and Research and elsewhere to address their collecting, research, and teaching needs more effectively and creatively.

We will also be able to respond more effectively in other domains as the needs arises for better access to current and historical web-based content.

Input from scholars, librarians, archivists, practitioners, and representatives of human rights organizations is essential to our model. Project staff will work with faculty and students at Columbia associated with the Center for the Study of Human Rights (CSHR) and the Law School's Human Rights Institute (HRI), along with those affiliated with the broad array of human rights-related programs, courses, and regionally oriented institutes on campus. Local input will ensure that our project aligns with the needs of those most actively using our collections.

Key stakeholders beyond Columbia are another source of input; this group includes scholars, librarians, archivists, advocates, and practitioners, especially those based in human rights NGOs.

Specifically, our proposed project requires user input in two key areas: selection of content for archiving, and usability of content presentation. First, we will structure opportunities for suggesting sites for capture, building on the extensive list of web sites Columbia librarians have tagged on delicious.com. During the initial stages of the project, a nomination form similar to the University of North Texas' tool [http://digital2.library.unt.edu/nomination/] will be distributed through listservs dedicated to human rights and selected listservs focused on area and regional studies. (For example, the H-Net's H-Human Rights Discussion Network, managed by the International Studies Association's Human Rights Section, the HR_Archives-L list for archivists and librarians, and the International Council on Archives Human Rights Working Group.) We will explore creating ongoing methods of soliciting nominations as the project advances.

Gathering input in the second key area, usability of content presentation, will be an important task during the second and third years of the project. In conjunction with library subject specialists, project staff will work with faculty, students, scholars, and NGO representatives to assess the types of access and presentation needed to optimize the use of archived web content in research and teaching. Methods for gathering this input can include a combination of group discussions, individual in-depth interviews, and targeted surveying. The collection of harvested sites on Internet Archive, the documents collected via the Web Curator Tool and made available locally, and collections of related material within other web archiving projects

will allow users to compare and contrast different presentations and identify desirable features.

Further, we will form a content and use advisory group and recruit Columbia faculty, one or two scholars from other institutions, an archivist or librarian specializing in human rights from outside of Columbia, and representatives from U.S. and internationally based NGOs. This group will provide guidance on broad questions of the project's development and execution, complementing the specific feedback solicited though the efforts described above.

In addition to gathering input from users, several other metrics will be applied to assess the effectiveness of the program.

- The costs to create metadata using each of the proposed approaches (MARC records, MODS records, finding aids) and methods will be carefully measured and compared with alternative measures. (This type of assessment was applied to Columbia's model for semi-automated cataloging of internet resources, as discussed in "Kate Harcourt, Melanie Wacker, Iris Wolley. (2007). Automated Access Level Cataloging for Internet Resources at Columbia University Libraries. Library Resources & Technical Services, 51(3), 212-225.")

- Usage statistics will be compiled from Columbia's web site and catalog, and used both to evaluate the effectiveness of the metadata and to assess the relative importance of different types of organizations and content.

- A link checker will be used to identify the frequency of, and reasons for, broken links in the metadata records created. (This technique has again been applied to the internet cataloging program cited above, with results yet to be published.) The results will help to define ongoing maintenance needs and to refine the frequency with which content should be harvested.

- The time required to set up and monitor each web site crawl will be tracked for varying levels of depth and quality assurance, to find the most cost-effective means of harvesting.

- Selectors will be surveyed to gauge impacts on related activities, such as identification and selection of print resources, compilation of subject guides, and reference assistance.

In September 2012, The Andrew W. Mellon Foundation awarded a new grant to Columbia University to develop and test models of collaboration with other research libraries, with scholars, with web content producers, and with other web archiving programs. The goal of the project is to extend the effectiveness of Columbia's web resource collecting program and of the collective web archiving work within the US. The project builds on the web collecting program established over the past four years, allowing Columbia to exercise a leadership role situating its web collecting program in the broader national and international web archiving framework. Four work areas will be targeted by the project.

This three-year program aims to achieve four objectives.

First, developing and testing a framework that will enable Columbia to provide web archiving services for a group of libraries in support of existing collaborative collection development agreements. The final outcome will include a model agreement specifying the roles and obligations of Columbia as a service provider and of participating libraries, the legal, intellectual property, and governance terms of the agreement, and the basis on which costs will be assessed and shared. This framework will then allow Columbia and its collecting partners to incorporate web content into their collections through a shared infrastructure and common services, rather than building redundant capacities and expertise that may be underutilized. The resulting program will provide a model that might be adopted by other library consortia.

Second, investigating and adopting practices in collecting web content that better serve the needs of scholars, based on a detailed investigation of the use of such content in the field of human rights. As an outcome of this investigation, Columbia will adjust the scope of its web collecting program to ensure that the types of content most useful to scholars are fully represented, and will adjust its metadata practices and interface design to improve discovery and usability of the collected content. While the investigation will focus on human rights scholarship, the methods and findings should be adaptable to other subject areas.

Third, developing techniques to improve the collection of web content by working directly with content producers to promote the adoption of low-cost features that optimize websites for harvesting, identifying and testing methods for direct transmission of content types that can not be acquired by standard harvesting methods. As an outcome, Columbia will adopt these techniques to improve the efficiency and effectiveness of Columbia's own program. In addition, we will develop a set of "best practice" guidelines for content providers and web archive administrators that will allow other web collecting agencies to achieve similar benefits.

Fourth, improving both the process of web archiving and the usability of web archives by extending the array of tools available for these purposes through a series of monetary awards to developers and scholars. As an outcome, applications developed through this project will be made broadly available through opensource software licenses, with code deposited in recognized open-source code repositories. Summaries and links to applications will be posted on Columbia's web collecting program website. By making the resulting tools broadly available, the project will benefit not only Columbia's program but other web collecting initiatives, and will help to stimulate further research and development.

While many librarians are concerned about the preservation and future accessibility of web content, few libraries have developed the capacity and expertise to support a robust web archiving program. Moreover, widespread replication of these capacities would appear to be redundant and inefficient. Through its Human Rights Web Archive, (http://library.columbia.edu/indiv/humanrights/hrwa.html) Columbia has developed tools to support community engagement in identifying important web resources – e.g., a webbased form used by librarians, scholars and others to nominate websites for collection. The Human Rights Web Portal, currently in development with a public launch planned for October 2012, will provide a public interface allowing the collected content to serve users worldwide. Columbia is well placed to provide a web archiving infrastructure that can serve the interests of many research libraries. Still lacking, however, is an organizational framework to support

such a role, and a detailed assessment of the governance, financial, policy, and operational implications of such collective action.

Columbia participates in a network of formal and informal collaborative collection development agreements, through 2CUL (a broad collaboration with Cornell University Libraries), the Manhattan Research Library Initiative, MaRLI (including New York University and New York Public Library) and the BorrowDirect consortium of research libraries in the northeast U.S. While most of these arrangements have been informal until recently, several within 2CUL and BorrowDirect have now been documented through detailed memoranda of understanding (MOUs). These agreements and relationships provide a starting point from which to build a framework for collaborative web collecting.

We will begin by convening a meeting of senior staff from our collection development partners to assess their level of interest in the collection of web content, to discuss potential models for collaborative action (including options for centralization or distribution of functions) and to identify issues (including intellectual property issues, such as policies for seeking permission to archive) and policy questions that should be addressed in the course of the project. Project staff will conduct a detailed assessment of web content related to current areas of collaborative collecting, including contemporary composers, U.S. local history, and Slavic, South Asian, and Southeast Asian studies. This assessment will include identification of websites with relevant content, test crawls and quality review of those sites to identify any technical issues and to estimate the scope and scale of a continuing program. Assessment of candidate sites will follow a procedure and template successfully applied to a range of human rights websites in the course of the Libraries' current web collecting grant. Throughout this process, the Web Archiving Librarian dedicated to the project will engage collection development staff from partner institutions via conference calls, site visits, and exchange of data and documentation in consideration of operational and policy questions.

Using the results of the content assessment and interim discussions, we will propose one to three models for formal, continuing collaboration, covering governance and policy formulation, costs and financial models, service level agreements, and methods of work. We will then convene a second meeting of senior staff from our partners to evaluate the proposed framework and assess the feasibility of a formal agreement.

Ideally, decisions about the types of web content to collect, the ways in which content is made available, and the functionality that must be preserved, should all be based on the uses made of web archives by scholars and teachers. Thus far, little is known about such uses, in part because web archiving efforts have focused more on technical aspects and on building repositories that are large enough to offer promising avenues of research. The collections built by Columbia over the past three years now offer such a body of material, and connections with scholars at Columbia and elsewhere may be used to explore current and potential use cases in depth.

We will conduct a rigorous investigation of the use of web resources in 21st-century scholarship in the areas of human rights and historic preservation, and assess what further development is needed in current web archiving efforts to support future needs in this area. The Web Archiving Librarian, assisted by the Libraries' Data Analyst, will compile a list of citations to web resources in scholarly publications over the past 5-10 years. These citations will be analyzed to identify the types and characteristics of sources cited. Those sources will

be compared to the websites currently archived by Columbia, and where appropriate added to Columbia's collection.

Project staff will then search specific citations in the Human Rights Web Portal and in Columbia's Archive-It collections to assess the ability of the web archives to function as sources of information with respect to both content and functionality, and to track citations should the cited content disappear from the live web. In particular, interoperability with tools developed through the Memento project (www.mementoweb.org) and compliance with its protocols will be tested and verified.

Based on the information compiled through the literature review described above, the Web Archiving Librarian, in coordination with the Libraries' subject specialists, will interview selected authors on their use of web information sources and on their current and potential use of web archives. Based on these findings, our collecting and access practices will be modified to better serve users' actual needs.

In collaboration with Columbia's Department of Computer Science we will work to engage faculty and graduate students in conducting research involving data mining or other techniques that begin to exploit Columbia's large and growing test bed of harvested and fully-indexed websites. In addition, we will encourage the development of tools and techniques that could be used to enhance access to content available in our new Human Rights Web Portal. Faculty in our Department of Computer Science have expressed preliminary interest in developing graduate student projects with us and have also indicated that other funding opportunities may be available to support faculty and graduate student research in this area.

During the first months of the project, the Libraries Digital Program Division (LDPD) Director and the project Analyst/Programmer will meet with Computer Science faculty to identify specific topics suitable for graduate student research projects. Preliminary discussions have focused on building enhanced search and analysis functionality for the Human Rights Web Portal; developing or applying visualization tools; experimenting with new approaches for extracting content and meaning from harvested Web pages; taxonomy development; and addressing some of the problems of HTML indexing (e.g., heterogeneity of content types, varieties of standard and non-standard conformance to HTML protocols). During the 2013-2014 academic year, the project Analyst/Programmer will assist selected graduate students in completing projects by providing extracts of data and content, advising on optimal techniques for working with the content, and assisting in answering questions and resolving problems.

We anticipate that this engagement with Columbia's Computer Science Department could lead to an ongoing set of research and development initiatives that not only benefit Columbia Libraries' web collections but also lead to advances generally in this domain.

Columbia's experience in building its web collections has shown that many site owners are willing and even eager to have their content preserved and made accessible. However, sites are rarely optimized for harvesting, making it difficult to reliably capture all desired content and functionality. While it is unlikely that site owners will spare much attention for website re-engineering, a few simple techniques available now can ease the process and improve results. Columbia has also begun collecting and archiving electronic records from

organizations whose Web sites are also being collected, providing the opportunity to work with candidate organizations to optimize archiving of both source files and websites.

Over the past three years, the Libraries' web archiving team has performed several test crawls of portions of the Columbia University website, and two production crawls. Working with University Archives and with the University-wide Columbia Web Steering Committee, we have identified priorities for the types of content to be collected. Many of the sites within Columbia's domain have been successfully harvested, but some issues and problems have also been identified. These issues are of two types: first, a site may include features that impede successful harvesting of some content, such as over-extensive use of a robots.txt file; second, some key content may exist only in a form not amenable to harvesting, such as a relational database.

Contacts that have been developed within the Libraries/Information Services structure can be used to explore and test ways to improve archiving of Columbia's web content. The Center for Digital Research and Scholarship (CDRS) maintains close contact with many academic departments, and has worked with faculty and department administrators to ingest content into our Fedora repository, Academic Commons. The Director of CDRS, also chairs the University's Columbia Web Steering Committee. Members of this Committee have expressed interest in promulgating and implementing guidelines on configuring websites for effective harvesting. In addition, a new position has been created within CUL/IS to provide oversight for data management.

Working within this framework, project staff will identify specific Columbia sites that have proved problematic for harvesting. They will then work with site owners to propose and test changes to improve archiving. This will be an incremental process; experience gained during the early stages of the project will help us understand what techniques, guidelines and documentation are most effective in working with sites of varying degrees of complexity and with site owners of varying levels of expertise. The end result will be a set of procedures that can continue to be employed as issues are encountered, and written guidelines that can be widely distributed.

In parallel with this work, project staff in consultation with University Archives will identify specific content of importance that cannot be acquired through harvesting. The staff will then work in collaboration with CDRS, with the Libraries Digital Assets Archivist, and with the relevant academic departments or administrative units to develop procedures for transferring files and associated metadata for ingest into the Fedora repository.

During the second year of the project, the techniques developed in connection with Columbia's own web content will be tested with external sites that are part of the Human Rights Web Archive. Three types of sites will be used in this test:

i. a site for which Columbia also holds the institution's non-digital (paper) archives (such as Physicians for Human Rights);

ii. a U.S.-based site with diverse content and complex structure (such as the Committee to Protect Journalists);

iii. a Latin-American site from an organization with which we have an established contact (such as Abuelas de Plaza de Mayo).

Using these three types of sites will allow us to assess the effectiveness of these techniques under different conditions: when a working relationship already exists with the site owner; when no such relationship exists, but communication can be readily established and a relatively high level of technical expertise can be expected; and, where communication may be relatively complicated and conducted through a third party.

Through this work, we will develop methods that can be broadly applied to improve the outcome of web collecting initiatives. Many of the institutions represented at the May web archiving summit noted the need for better tools to improve various aspects of the web archiving process or to enhance the usefulness of web archives. Such tools need not be complex, and development could often take advantage of existing programs and applications. Examples that have been suggested include a WordPress plug-in that would make a site available for automated detection and harvesting, and visualization techniques to allow easier understanding of changes within a website over time. In addition, those at the summit repeatedly noted the need to call attention to the potential of web archives for research, and to encourage scholars to consider innovative approaches making use of this content.

Although development of such tools and applications need not require expensive and long-term investment, it does require time and effort. The individuals and institutions most capable of undertaking such work have many other priorities. They may be engaged in related development without realizing the potential application to web archives. Those contemplating such work may be conceiving it within a specific local context, which could limit the broad applicability of any software developed. For all of these reasons, attendees at the May summit believed that a program of financial incentives could stimulate development of tools that would have a broad impact and lasting benefit. Through this project, Columbia will act as the coordinating agency for such a program.

As a first step, we will assemble an Oversight Panel comprised of individuals with expertise in technical aspects of web harvesting and archiving and use of archived content, from separate institutions. Potential members of the Oversight Panel might include lead Memento researchers, and administrators of web archiving programs. Informal nominations for the panel will be sought from attendees at the May web archiving summit, and we will consult directly with representatives of the California Digital Library's Web Archiving Service, the Internet Archive, and the Library of Congress. The Director of the Library Digital Program Division at Columbia will serve ex officio as the Panel's convener.

The International Internet Preservation Consortium (IIPC) is currently soliciting proposals for a series of projects to articulate needs for development in specific areas such as automated quality assurance and archiving of online games. While this IIPC program does not currently sponsor technical development, the program's outcomes will help to inform this component of the proposed project. Columbia has recently become a member of the IIPC and will maintain close contact with this program and with technical developments in the international web archiving community.

The Oversight Panel will hold an initial one-day meeting at Columbia to determine key elements of the incentive program, including: description of the types of tools and applications sought; eligibility criteria for awardees; submission requirements; and evaluation criteria. Prior to this meeting, we will survey participants in the May 2012 web archiving summit to identify specific examples of desirable software developments, to assist the Panel in shaping the request for proposals, and as illustrative examples for potential applicants.

Initially, proposals will be solicited for developments in two areas: the first, for tools and applications that assist the process of collecting web resources and rendering the archived content; the second, for tools and applications that facilitate and enhance the discovery and use of web archives.

Proposals will be accepted for incentive awards in amounts up to $25,000. The number of awards made will depend on the quality of the applications and the amounts requested, but will be no fewer than eight and no more than twenty. Proposals will be required to comply with certain minimal criteria: all work must be completed and submitted within the time frame of the project; and, all software developed must be fully documented and made available under one of the open-source software licenses approved by the Open Source Initiative (preferably a GNU General Public License) with code deposited in an open-source repository such as GitHub or Sourceforge.

Following its initial organizational meeting, the Oversight Panel will meet virtually (by videoconference, Skype, conference calls, and email) to review progress and receive and evaluate proposals. The submission and evaluation process for an initial round of proposals will be completed within the first six months of the project. Given the innovative nature of this program it is impossible to predict the number and cost of worthy proposals that will be received. In addition, considering the rapid evolution of web archiving technology, it may be desirable to reserve some funding to support a second round of awards at the beginning of the project's second year, allowing sufficient time for work to be completed by the end of the project. The project budget has been structured to support such a two-stage process. The plan for timing and distribution of awards is discussed in more detail in the Summary of Project Timeline below. Ultimately, the timing and distribution of specific awards will be determined by the Oversight Panel.

Although the bulk of the incentive program will take the form of competitive awards for software development, some existing open-source software has already become so widely used that a different approach may have greater benefit. As an example, both Columbia and the Internet Archive have begun to use SOLR software to index web content. Our experience has shown that while SOLR is widely used to index large text databases, it has not thus far been optimized to deal with the peculiarities of the WARC files on which web archives are based. Improvements in SOLR indexing as applied to web archives have great potential to reduce the collective system and staff resources needed to provide access to web content.

In order to ensure that such critical software components receive appropriate attention, $25,000 of the project budget will be reserved for commissioned software development and/or consultation. The Oversight Panel will determine whether and how to commission development of specific tools or consultation from known individuals or institutions. This decision will be made in the final months of the first year of the project. By preference, the Panel would specify the work to be completed and solicit competitive bids from known parties. However, there may be instances in which a specific institution or individual is known to be best positioned to complete the required work in the time available, and might be approached directly to submit a proposal.

At the completion of the project, a listing of awards with links to documentation of the resulting tools and software applications will be made available through the website for Columbia's web collecting program, and distributed to the International Internet Preservation Consortium and other interested agencies.

Over the past three years, Columbia has developed and implemented detailed policies and procedures for requesting permission from website owners to archive and make available the content for which they are responsible. These policies are in accordance with those described in our 2009 grant proposal, with procedures refined through experience. The basic policies described below are discussed in more detail in the document, "Columbia University Libraries Web Archiving Policies" available from the web archiving summit website (https://webarch.cul.columbia.edu/?page_id=13) and in the FAQ for site owners available on the website for Columbia's web resource collecting program (https://library.columbia.edu/content/librarywebsecure/bts/web_resources_collection/faq.html).

Our basic policy is to request written permission from the site owner before any content is harvested and archived. Requests are sent by email, using a standard format, translated into the principal language of the site when possible. If no response is received after two weeks, the request is repeated, with a notice that we will proceed with archiving unless the site owner has any objection. We also provide a notice and mechanisms for discontinuing archiving and removing content if an objection is registered later, by the site owner or the rights holder for specific content. All communications relating to permissions are retained in their original form and recorded in a database.

To date, these policies have been quite successful. Particularly in the area of human rights, we are often seeking permission from organizations that may no longer be active, or for content that may be at risk, during periods of unrest or upheaval. Nevertheless, we have received explicit permission from well over half of the organizations we have attempted to contact, and have been denied permission by only four organizations. Questions raised through this process have caused us to refine these policies, adding provisions for removal of specific archived content even after permission has been granted, if so requested by the site owner in response to intellectual property or legal concerns. On rare occasions we have rejected a site as a candidate for archiving because the site's own documentation states that permission to download content must be obtained from individual rights holders.
These policies are designed to ensure respect for intellectual property rights, and will continue to be applied to any content collected in the course of the proposed project. If, in the course of the project, any content is to be collected by a party other than Columbia University Libraries, that party will be required to agree in writing to follow and enforce the same policies.

As the project proceeds, it is possible that we will encounter new issues related to permissions. Some issues have been anticipated, but have not yet arisen; for example, we have not yet determined a course of action if an organization agrees to have content archived, but only with access severely restricted. New tools and refinements of existing software may present technical opportunities or challenges that are not covered by existing policies.

Issues relating to intellectual property and permission to archive will also be addressed in developing the model for collaborative web archiving. Policies in this area will be discussed in the meetings of collaborative collection development partners, and participants in a collaborative program will be required to formally agree to abide by approved policies.

Throughout our current project, Kenneth Crews, former Director of the Libraries' Copyright Advisory Office, has provided advice and guidance on issues related to permissions for

archiving and intellectual property rights. The Office continues to advise on new issues as they are encountered and will continue to provide advice and consultation throughout the course of the project.

In summary, the web archiving workflow at Columbia includes several key elements:

- selection of nomination of the site or document
- use of Archive-It for scoping, crawls, access and quality assurance
- cataloging workflows/output
- permissions policy and workflow
- project management tools for growing number of collaborations
- content discovery strategy
- content use/application strategy
- long-term preservation plan

Copyright and legal deposit laws, affecting the right to harvest web sites and provide access, vary specifically from country to country:

- No legal deposit requirement for websites or explicit library exception to allow harvesting websites (USA)
- Legal deposit of websites but no access (Norway)
- Legal deposit of websites but access is restricted to terminals at national libraries (Austria, France)
- No legal deposit requirement but national domain crawls conducted and full online access provided (Portugal)

The Internet Archive and the Wayback Machine are the most important sources for content from the public web back to 1996, with nearly 300 billion URLs, over 10 petabytes of data, and over 1,000 visits per second. The IA collection is vast and indispensable, but not comprehensive. The frequency of crawls, the depth of content capture and indexing, and the impact of blocked or restricted sites, all argue for complementary services through individual and groups of libraries.

Curated web archives present some important benefits: precise selection of desired web resources, fuller website capture, control of frequency of website capture, metadata enrichment, full-text searching, and data mining. In the area of description and access for archived websites, the Columbia Libraries use the following strategies:

- Archive-it.org site-level metadata (All thematic collections, DCMI, copied from MARC records if possible)
- Local catalog (CLIO) collection-level MARC records
- Local catalog (CLIO) site-level MARC records
- MARC records and holdings in WorldCAT
- Document-level MARC records
- Thematic archive portal on Libraries website

Archive-It, first deployed in 2006, is a subscription web archiving service from the Internet Archive that helps organizations to harvest, build and preserve digital content. Archive-It partners can focus on curation, cataloging and collaboration building. Archived content has

24/7 access and full text search capabilities for library users. Content is hosted and stored at the Internet Archives data centers.

In April, the Columbia University Libraries announced the winners of the Web Archiving Incentive Awards, funded as part of Columbia's 2013 Mellon Grant for Collaborations in Web Content Archiving.

The goal of the incentive awards program is to support the development of tools and techniques to improve web archiving and enhance the functionality of web archives. The five award-winning proposals were selected from a pool of 11 applications by a multi-institutional panel convened by the Libraries.

The award recipients and respective projects are:

Jimmy Lin, University of Maryland, "Warcbase: Building a Scalable Web Archiving Platform on HBase and Hadoop."

Zhiwu Xie and Edward A. Fox, Virginia Tech University, "Archiving Transactions Towards Uninterruptible Web Service"

Kim Dulin, The Harvard Library Innovation Lab, "Perma.cc: Mitigating the Pervasive Problem of Link Rot in Scholarly Works and Preserving Online Content"

Michele Weigle, Old Dominion University, "Visualizing Digital Collections of Web Archives"

Michael Nelson, Old Dominion University, "Tools for Managing Seed URLs"

The selected projects vary widely in scope and include work towards: applying modern, big data principles and technologies in order to scale up web archiving; effectively capturing local institutional web content while at the same time improving website availability; allowing authors and journals to directly initiate archiving of web pages when citing them in articles, thus preventing future link rot and ensuring the ongoing availability of cited content; developing automated ways of detecting changes in archived web pages over time and providing smart visual displays to help view the evolution of websites; developing a tool to help determine whether the topic of a web page being archived has gone significantly off-topic and thus needs to be reviewed to see if it has been converted to "web spam" or is otherwise no longer in scope for collecting.

Columbia is advancing a collaborative web content collection experiment with partner institutions in the Borrow Direct/Ivies Plus group. This includes work on a contemporary composers web archive, on an architectural preservation/urbanism/sustainability web archive, and on a climate change web archive. The Libraries program focuses on four main types of web resources collections:

- Thematic collections of content relating to the University's academic programs;
- Web content from organizations and individuals whose archives are held by the Libraries;
- Significant content from the University's own website;
- Sites identified by researchers and library subject specialists as at risk.

There are hundreds of libraries, archives, historical societies, museums and other organizations around the world advancing web content collecting and archiving programs. National libraries have provided essential leadership. The Web Archiving Service (WAS), a service of the University of California Curation Center, assists libraries in the web archiving process and has helped create a rapidly expanding set of public archives, unique collections of websites organized around a variety of diverse topics, and historical and current events, thus providing students and researchers with lasting access to ephemeral web content. The goal at Columbia and across this expanding community of interested and active organizations is to build an efficient, coherent, and scalable national and international framework for collecting, archiving and using web content.

As Columbia has proceeded with these projects, there have been several persistent questions:

- Do we truly understand what users expect to find in web archives? and how will the sites/documents be used?

- What web content has persistent value for education and scholarships? and how are collection development decisions to be made?

- What are the key barriers to the collection, use and archiving of web content?

- What are the preferred strategies for discovery and accessibility in a library?
- How is quality assurance to be maintained?

- How will the technologies and tools evolve? what investment is needed to advance open source options?

- How will success and value be measured and effective use assessed?

- How will collaborative efforts be structured and good governance assured?

So how does a paper on web content collecting and archiving fit into a discussion on the future of library document delivery and resource sharing services? Clearly, as licensed, digitized, born-digital and web content expands as part of library collections and user expectations and requirements, the nature and flow of resource sharing is redefined. What and how a library acquires or enables access to resources as part of a collective collection will be revolutionized. How a library describes a work and contributes a record to support discovery and access is increasingly diffused and chaotic. How a library takes responsibility for long-term preservation will be shaped by institutional, regional and national digital archiving strategies. The protocols for sharing of resources through interlibrary loan and document delivery will be replaced by new standards for shared, authenticated and open access. The role of staff and mediated services is being rethought in the context of user driven and managed information requirements. The physical delivery of objects disappears. And resource sharing becomes much more focused on staff expertise, software/applications, technology infrastructure, space, new services, and research and development, for example. The key question for us: Will resource sharing 2025 be defined by library leadership or library irrelevance?