WEB RESOURCES COLLECTION PROGRAM DEVELOPMENT

The Columbia University Libraries/Information Services

SUMMARY

The Columbia University Libraries seeks $715,573 for a three-year project to develop and implement a program for incorporating web content into the Libraries' collections. Building on the results of a planning grant completed in 2008, the Libraries will put into production procedures for selecting, acquiring, describing, preserving, and providing access to freely available web content, specifically in the subject area of human rights. Over the course of the three-year project, 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. Throughout the project, Columbia will promote its discoveries by reporting on activities through blogs, listservs, and presentations at professional meetings. During the final months of the project, Columbia will host an invitational conference of major research libraries to promote discussion of this model and identify ways to promulgate its use. Columbia will also create and share a best practices document outlining recommended procedures, to ensure that results are available for wide distribution.

During the first year, the project will focus 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 will continue this process and will focus on use of the collected content in scholarly research, teaching, and learning. During this phase, the methods developed will be integrated into Columbia's routine processes of collection development, description, and access. At the end of the project, it is expected that the cost to continue and expand this program will 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. (A full account of the planning grant activities and findings is given in the appendices.) 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. This proposed implementation project will encourage development of such consensus, but will also be responsive to shifts in community norms.

Over the next three years, tools for web archiving and for the presentation and re-use of archived resources will continue to develop, and that development will in turn shape scholars'

understanding of how these resources can best be made available. This project will employ two complementary approaches—using the Internet Archive's harvesting software and storage model as well as a locally managed harvesting tool such as Web Curator Tool—that will explore different avenues to the problem of archiving and providing access to archived content.

We embrace the importance of harvesting and archiving full web sites by preserving as much context and original presentation as possible. At present this approach is more scalable than selective approaches. But we also believe it is important to be able to present, index, and access some types of document-type material from different web sites singly and in combination with other material in a way that is integrated with other related resources—electronic, print, and paper archives.

Drawing on the expertise of its Center for Human Rights Documentation and Research, the Columbia Center for New Media Teaching and Learning, and the Center for Digital Research and Scholarship, Columbia will test models for describing and organizing web resources in relation to related print and archival collections and for making this content available for use, implementing those found most fruitful for discovery, research, and teaching.

Further, Columbia University Libraries/Information Services has a unique resource in the Copyright Advisory Office (http://www.copyright.columbia.edu/). Dr. Kenneth Crews will actively contribute to the project to help navigate the complex copyright and intellectual property issues involved in web content capture and preservation.

To be successful and extensible, a program for web content collecting must have a global outlook, with benefits accessible to libraries, archives, scholars, and practitioners throughout the world. This proposed project will provide opportunities for all of these constituencies to contribute their expertise and to demonstrate a framework that can be adopted (and adapted) by other institutions to build collections of web content for additional subject areas.

The resulting program will reveal the transitions in organization, skills, and collaborative action that libraries need to undertake as non-commercial content moves from print to digital distribution, to support full life-cycle management of web resources and implement mainstream web site and document collection development into the work of the library. By demonstrating practical, effective means of integrating digital, archival, and print collections, the project offers the potential for transformative impact on the ways libraries perform these core functions

The proposed project dates are July 1, 2009 to June 30, 2012, with a requested grant end date of December 31, 2012 to ensure administrative tasks are completed within the grant time frame.

WEB RESOURCES COLLECTION PROGRAM DEVELOPMENT
The Columbia University Libraries/Information Services

A. BACKGROUND

Columbia University is an independent, privately supported, nonsectarian institution of higher education. Founded in 1754 as King's College by royal charter of King George II of England, it is the oldest institution of higher learning in the state of New York and the fifth oldest in the United States. From the beginning, the institution's goal was defined as "the Instruction and Education of Youth in the Learned Languages and Liberal Arts and Sciences." This mandate has not essentially changed, even with the transformation of King's College into Columbia, one of the world's foremost research universities.

The University is committed to preserving the quest for knowledge as more than simply a practical pursuit, through its broad range of innovative multidisciplinary programs and through the earnest exploration of difficult questions. It seeks to make significant original contributions to the development of knowledge, to preserve and interpret humanity's intellectual and moral heritage, and to transmit that heritage to future generations of students.

Columbia University Libraries/Information Services is one of the top five academic research library systems in North America. The collections include over 10 million volumes, over 100,000 journals and serials, as well as extensive electronic resources, manuscripts, rare books, microforms, maps, graphic, and audio-visual materials. The services and collections are organized into 25 libraries and various academic technology centers. The Libraries employs more than 550 professional and support staff.

The services of the Libraries extend well beyond the university. Access to digital resources is provided through the Libraries' web site (http://www.columbia.edu/cu/lweb). Onsite access to the physical collections is available to anyone affiliated with members of the SHARES program under the auspices of OCLC and of the New York Metropolitan Reference and Research Agency. The Libraries also fills thousands of interlibrary loans through cooperative arrangements with OCLC, RAPID, the Regional Medical Library Center of New York, and others.

The Libraries actively seeks support from external sources and has successfully secured funding for a wide range of projects from organizations including the Andrew W. Mellon Foundation, the Carnegie Corporation, the Getty Foundation, the Henry Luce Foundation, the National Endowment for the Humanities, the National Historical Publications and Records Commission, and the Starr Foundation.

B. RATIONALE

Building Research Collections

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.

Transition to Digital Formats

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. With support from the Andrew W. Mellon Foundation, Columbia will develop and demonstrate the value of this work for the wider scholarly community—and the need and potential for broad, collaborative action.

C. PROJECT DESCRIPTION

The objective of this project is to create a continuing program of web content collecting as an integral component of the Libraries' collection building practices. Upon its completion, the Columbia University Libraries will have established the technical, organizational, financial, and human resources needed to continue this collecting and to extend into new subject areas.

While the primary objective is to establish a model for collecting web content produced outside libraries, a truly holistic model needs to place these resources within the context of print, archival, and digital collections—and to recognize that those distinctions are increasingly problematic as print and archival materials become available in digital form. During the second and third years of the project, we will focus on the technical and metadata development needed to bridge the differences in practice for describing and presenting print publications, paper archives, and digital representations of documents and collections. The final result will be a model for ongoing mainstream web site and document collection development as part of the work of the library.

Project Oversight and Personnel

The project will be directed by Robert Wolven, Associate University Librarian for Bibliographic Services and Collection Development for Columbia University Libraries.

The primary work of acquiring, organizing, and describing collections will be done by two full-time Web Collection Curators, to be hired. The two Web Collection Curators will: design and implement a web-based tool to gather input from librarians, archivists, and scholars on the relevance and importance of specific web content from candidate organizations and web sites; solicit nominations for additional sites; secure permission for archiving from selected organizations and content owners; establish, test, and monitor parameters for web harvest of the selected sites; enhance extracted metadata to create finding aids for archived collections and catalog records for selected documents; organize and conduct usability studies; and work with digital technology and preservation staff to incorporate web content into Columbia's evolving digital infrastructure.

While the two Web Collection Curators will work closely with each other, they will report through separate administrative lines. One Curator will be based in the Rare Book and Manuscript Library and will focus initially on those organizations for which Columbia already holds paper archives. The second Curator will be based in the Global and Area Studies Division (which also includes the Center for Human Rights Documentation and Research) and will focus on organizations based outside the United States and Western Europe. This arrangement will

allow the program to draw on a broad range of language, subject, archival, and metadata expertise, as we refine our understanding of the skills needed, and work to extend these activities into other units.

During the second and third years, we will hire a Digital Library Analyst/Developer to develop and refine a strategy for integrating archived web resources into our campus search and discovery environment. He or she will work closely with Columbia technical, archival, and public services staff. While current planning is based on continued use of Internet Archive for archival storage, we will also explore deeper integration of resources stored in that manner with Columbia's Fedora-based repository services, using OAI-ORE along with other resource-linking and integration tools.

During the second and third academic years covered by the project, three graduate student interns will also be hired to assist with organization and description of the selected content, based on the model established by the Mellon-funded project *Graduate Internships in Primary Source Collections.*

Throughout its duration, the project will draw on the resources of Columbia's Libraries/ Information Services. Area Studies librarians will select content for inclusion, foster contacts with human rights organizations, and provide liaison with faculty. Archival and metadata curators will assist with resource description and organization. The Assessment Librarian will develop and conduct usability studies, and staff from the Center for New Media Teaching and Learning and the Center for Digital Research and Scholarship will evaluate the use of selected web content in Columbia's teaching and research programs.

Further, the Libraries/Information Services has a unique resource in the Copyright Advisory Office (http://www.copyright.columbia.edu/). Dr. Kenneth Crews will help navigate the complex copyright and intellectual property issues involved in web content capture and preservation. Dr. Crews will also assist with agreements regarding rights and permissions and act as liaison with the University's General Counsel.

Project oversight will be provided by a steering committee consisting of the Project Director, the Directors of the Rare Book and Manuscript Library and the Global and Area Studies Division, and the Director of the Libraries Digital Program Division. A faculty committee will be established to advise on content development and presentation and to assess potential impacts on teaching and research. Global and Area Studies librarians will provide coordination with related programs, such as those at the Center for Research Libraries and the University of Texas at Austin.

Subject Focus

The proposed project will focus 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 will be made to ensure that all relevant academic programs are engaged in the selection and evaluation of content. As the project progresses, subject librarians and scholars will be encouraged to recommend content extending beyond human rights to related political, social, and environmental interests.

Project Plan

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 will work in iterative fashion, refining procedures as the web content collection expands and the available tool set evolves.

1. Selection

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 will work 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.

2. Permissions

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 will work 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.

3. Harvesting and Archiving

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 open-source 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 website 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 open-source Web Curator Tool, and shared their written evaluation of the program with us (see Appendix 2 of the planning grant Final Report), 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 e-mail correspondence. While Web Curator Tool offer certain advantages in tracking permissions and selective harvesting of individual documents, its use for these purposes is labor-intensive and less suitable for full website harvesting.

Having now crawled over 80 seed sites using Archive-It, we are familiar with its advantages and shortcomings and remain confident that Archive-It is the best available option and is 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 in the next 4-6 weeks) that will allow re-crawls of a given seed to harvest only its new and/or changed content, saving storage space; and, later in 2009, 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 potentially other tools in the context of our 2nd and 3rd year work towards integrating selected Web content into our local environment (see section **6 Making Materials Available for Use**).

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.

4. Description and Organization

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 document 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, brief MARC records for all selected sites will be generated 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, website-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 website 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, in order to allow effective integration with existing library collections, and to facilitate reference linking. (See also under section 5, **Disclosure**, below.) 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.

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 approaches 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. Even at this stage little of the work will be duplicative, however; rather, sets of metadata will be repurposed and recombined for different presentation. 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.

5. Disclosure

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, 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.

6. Making Materials Available for Use

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 on continued use of Internet Archive for web site archiving, we will also explore the selective 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.

- **Targeted Harvesting and Local Repurposing.** 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.

- **Site Maps and Resource Maps.** 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 XML-based 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.

- **Merged Resource Maps.** 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.

7. User Input and Assessment

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 del.icio.us.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 website 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 cause and to refine the frequency with which content should be harvested.
- The time required to set up and monitor each website 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.

The Libraries' Assessment Working Group will provide advice and assistance in designing and interpreting these and other assessment measures.

Summary of Project Timeline

The proposed project dates are July 1, 2009 to June 30, 2012, with a requested grant end date of December 31, 2012 to ensure administrative tasks are completed within the grant time frame.

Year 1
Two Web Collection Curators will be hired to review work completed during the planning grant; develop and refine procedures and criteria for selecting, harvesting, and describing web resources; and begin acquiring, organizing, and describing web content.

Using the list of sites identified during the planning grant, they will: generate and review MARC records from delicious.com metadata; work with selectors to establish criteria for describing and/or archiving sites and their content; review results of the web crawls and develop a quality-assurance checklist; work with selected human rights organizations to secure permission to archive content and develop model agreements for further use; create finding aids for representative sites and MODS records for selected documents; and propose appropriate coding to OCLC for entry of MARC records into the Registry of Digital Masters.

During the latter part of the year, the Curators will develop and implement a web-based tool allowing librarians and scholars to evaluate candidate sites and nominate additional sites for collection. The Curators will also seek archiving permission from additional organizations and begin to harvest and describe additional high-priority web sites.

Year 2
The Curators will continue to seek archiving agreements, harvest new sites, and create finding aids and MODS records for selected content, refining procedures developed during Year 1. During the academic year, the Curators will hire and train student interns (with skills to complement the Curators' own language expertise) to perform repeat harvests of archived sites, update finding aids and create additional metadata, and incorporate newly selected content.

A Digital Library Analyst/Developer will be hired and will begin to test Web Curator Tool (or other similar tool) to harvest selected document-type content from three to four major human rights organizations and deposit the content in Columbia's Fedora-based asset repository, working with the Web Collection Curators and other Columbia Libraries/Information Services staff to generate and/or create appropriate metadata. The Analyst/Developer will test the feasibility of creating RDF linkages to versions of the documents in context as stored in the project's Internet Archive collection.

Working with the project Curators and with advice from Area Studies librarians (and through them, end users), the Digital Library Analyst/Developer will develop a prototype "Human Rights Electronic Reference Collection," re-exposing the locally stored content in a coherent, searchable

presentation. Project staff will work with the Center for Human Rights Documentation and Research to begin to assess this presentation and identify desired enhancements.

Year 3

The Web Collection Curators and student interns will continue to add new sites and content and create and update resource descriptions. As Archive-It adds new capabilities, the Curators will test the possibility of improving earlier harvests by re-crawling sites with new parameters to obtain additional content.

The Digital Library Analyst/Developer will test the creation of XML site maps and resource maps for harvested sites and continue work to improve local presentation of the "Human Rights Electronic Reference Collection." As these methods are developed and tested, project staff will work together and with the Libraries' Metadata Librarian to refine the models used to describe harvested content and, if appropriate and feasible, apply new models to metadata created in early phases.

The Curators will document procedures for identifying and selecting new content, criteria for determining treatment, and methods for harvesting and describing web sites, and they will train other Libraries' staff in their use. The exact staff to be trained will depend on future developments in the Libraries' organization and on lessons learned during the project, but will most likely include selectors, electronic resource librarians, archivists, and catalogers. The Project Director will ensure that these workflows are tested as applied to additional subject areas, and by non-project staff to additional human rights content, to ensure a smooth continuation of the program after the project's end.

During the final months of the project, Columbia will host an invitational conference of major research libraries to promote discussion of this model and identify ways to promulgate its use. Columbia will also create a best practices document outlining recommended procedures, to ensure that results are available for wide distribution.

D. EXPECTED OUTCOMES AND BENEFITS

This three-year project aims to achieve three objectives:

First, the accumulation of a substantial collection of web resources in the field of human rights, together with the infrastructure, tools, and procedures needed to sustain and grow this collection over time.

Second, development of a framework for resource discovery, access, and presentation that allows the collected web resources to effectively meet the resource needs of scholars and human rights workers at Columbia and throughout the world.

Third, integration of procedures for collecting web resources with the routine work of selectors, e-resources staff, archival curators, and catalogers, such that the program will continue and grow without additional external support.

Perhaps more important, the program embodied in these three outcomes 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. Throughout the project, Columbia will promote its discoveries by reporting on activities through blogs, listservs, and presentations at professional meetings. During the final months of the project, Columbia will host an invitational conference of major research libraries to promote discussion of this model and identify ways to promulgate its use. Columbia will also create and share a best practices document outlining recommended procedures, to ensure that results are available for wide distribution.

E. INTELLECTUAL PROPERTY ISSUES

The project will make use of existing software to harvest and archive content with Internet Archive. Columbia does not propose to create content through the project and will not acquire any intellectual property rights for the content collected from the web. Any local ingest of content will use existing open-source software.

Software tools developed during the course of the project will use open standards to the extent possible and will be made available under a Creative Commons license.

Please see page 8 of this proposal for information on the role of our Copyright Advisory Office and the Section 108 Study Group.

F. LONG-TERM SUSTAINABILITY

Columbia is committed to implementing and sustaining a program of collecting freely available web content, as part of its ongoing mission to support research and teaching and to maintain collections for future users. With a three-year investment, Columbia will be able to rapidly build, test, and refine the tools and procedures needed to support such a program, while collecting and describing a body of content in the subject area of human rights.

The program in human rights will then be easily sustainable because growth in content will be incremental. The scale of web collecting at Columbia beyond the project period and beyond the field of human rights will depend to some degree on developments at other research institutions. At the minimum, Columbia will target those areas in which its collections have national or international prominence, continuing to build centers of excellence that other libraries can depend on. Ideally, Columbia will also make use of, and contribute to, web collecting programs elsewhere by recommending resources and integrating descriptive metadata to enrich its own collections. Where gaps continue to exist, Columbia will collect at a more modest level, as with print resources.

The Libraries will not seek new investment from external sources to sustain this program, but to a combination of transformative action and strategic redirection of resources. The Libraries' strategic plan calls for a shift of resources from the acquisition, description, and maintenance of published print resources towards a greater investment in electronic resources, special collections and archives, and digital materials. Several measures have already been implemented that have allowed a 200% increase in staff devoted to licensed electronic resources and a doubling of staff devoted to metadata for special collections. Further changes are planned so that this reallocation can continue even as we face growing fiscal constraints. These changes will be assisted by replacing some of the labor-intensive work of collecting individual print documents with more efficient procedures for collecting this content from the web.

A key component of the final year of the project is redistributing tasks from project staff to other units within the Libraries, to allow a smooth transition from concentrated activity and rapid program growth to a model of distributed action and gradual expansion. This transition will be facilitated by re-use in other contexts of techniques developed for the web collecting program. These include tools and policies to support automated notification and selection of new content; recommender tools to broaden the community engaged in collection development; metadata extraction tools to assist in cataloging online resources; and metadata practices that replace some item-level description with higher-level organization.

The metadata and technical models developed for this project will be broadly applicable to other forms of digital (and digitized) content. As special and archival collections are digitized, and as new archival collections encompass individual and organizational source materials in electronic form, the techniques applied to web content may be extended to these resources. Thus the outcomes of this project play an important part in the Libraries' future strategic directions. In the staff of its Library Digital Program and Center for Digital Research and Scholarship, Columbia has and will apply the resources needed to continue this work.

G. REPORTING

The proposed project dates are July 1, 2009 to June 30, 2012, with a requested grant end date of December 31, 2012 to ensure administrative tasks are completed within the grant time frame. We propose the following reporting schedule:

Reports	**Due date**
Interim narrative and financial reports covering July 1, 2009–June 30, 2010	September 30, 2010
Interim narrative and financial reports covering July 1, 2010–June 30, 2011	September 30, 2011
Final narrative and financial reports focusing on July 1, 2011–Dec 31, 2012 (but presenting a comprehensive report of the project)	March 31, 2013

Robert Wolven, Project Director, and Karen Kapp, Grants Officer, will submit the narrative reports to the Foundation. The financial reports are prepared and issued by Columbia's Sponsored Projects Finance, part of the Office of the Controller.

The structure of our reports will follow the Andrew W. Mellon Foundation's Reporting Instructions document, when it becomes available.

H. BUDGET NARRATIVE

Columbia will contribute the time of several employees to the project. Robert Wolven, Associate University Librarian for Bibliographic Services and Collection Development, will direct the project. Michael Ryan, Director of the Rare Book and Manuscript Library, and Pamela Graham, Acting Director of the Global and Area Studies Division, will provide day-to-day supervision of the two Web Collection Curators. Stephen Davis, Director of the Libraries Digital Program Division, will direct the work of the Digital Library Analyst/Developer. Advisory group members and other staff, including Kenneth Crews of the Copyright Advisory Office, will also contribute their time in-kind.

Columbia and other conference attendees will incur the costs of an invitational conference of major research libraries, to be held at the end of the project.

The project proposes to hire two Web Collection Curators. These will be three-year positions. (Columbia proposes to hire the Project Manager of the 2008 planning grant to fill one of these positions, allowing a carry-over of expertise and a rapid start to the project.)

The project proposes to fund a Digital Library Analyst/Developer for two years, commencing in the second year of the project.

At the end of the grant period, these three positions will be discontinued, with the continuing work having been transferred to other positions in the Libraries as a permanent part of their responsibilities.

The budget also includes funding for three student interns during each of the second and third years of the project. These academic-year positions will spend a total of approximately 1,500 hours assisting with the acquisition, organization, and description of selected web content.

The budget includes funding to extend and expand the Libraries' current contract with Internet Archive for use of Archive-It software and storage of digital materials harvested during the course of the project. This software is necessary for the proposed project and will be used primarily for its execution.