Submitted on: 16 June 2015

World Sustainable Development Web Archive: Preserving and disseminating knowledge for sustainable growth

Steven W. Witt
International and Area Studies Library, University of Illinois at Urbana-Champaign, Urbana, Illinois, U.S.A.
E-mail address: swwitt@illinois.edu

Lynne M. Rudasill
International and Area Studies Library, University of Illinois at Urbana-Champaign, Urbana, Illinois, U.S.A.
E-mail address: rudasill@illinois.edu

Abstract:

The grey literature produced by Non-Governmental Organizations (NGOs) is considered to be some of the most ephemeral material available on the Internet. NGOs often operate on tight budgets and in opposition to governmental initiatives, a combination that makes their web presence particularly transient.[1] *In response to the ephemeral nature of NGO materials on the internet, the International and Area Studies Library of the University of Illinois initiated the World Sustainable Development Web Archive project to preserve web content published by NGOs that focus on environmental and economic sustainability. To ensure broad coverage of these movements, the archive actively collects materials in multiple languages and cultural groups. This paper describes the use of shared web archiving platforms to initiate and sustain web archiving activities that support scholars and enable small organizations to archive and preserve their historical content and discusses the challenges that have been encountered both with the technology and within a changing political landscape.*

Keywords: Web Archives; Non-governmental Organizations; Preservation; Access; Sustainable Development

[1] See Witt, S. W. & Rudasill, L. (2009). Non-governmental organizations and information. In Bates, M. & Maack, M. (Eds.) Encyclopedia of library and information sciences. (3rd Ed). (8 pages) London: Taylor and Francis.

Scholarly communication in all of its formats implies a certain degree of reliability. The reliability of the author's integrity related to his or her work, the reliability of the data that is used in the research that is being done and the reliability of the citations referred to within the work all play an important part in the progress of any field of study and are the basis upon which disciplines stand. Changes in scholarly communication from print to electronic format have challenged the user and the producer of the information in a multitude of ways that go beyond cost and the big deal into the growing area of grey literature – that which does not go through the normal peer review and publishing process. What is the impact of grey literature that is produced by non-academic experts who work in non-governmental organizations (NGOs), those who research the issues and problems that we face today in a global existence from the grassroots level and upward?

The number of NGOs listed in the Yearbook of International Organizations is approximately 67,000 this year. In their finest forms, these organizations provide us with information about problems that challenge us. NGOs are fairly ubiquitous as are the issues with which they deal. We know some of these civil society organizations quite well – OxFam, Amnesty International and Human Rights Watch come to mind immediately. But there are thousands of lesser known organizations. NGOs focus on issues that begin at the grassroots level and bring the challenges forward to the public usually with the hope that the problems they identify can be solved through public opinion and political pressure – local, national and international. With few exceptions, NGOs quickly discovered that promoting their causes via web access was a convenient and relatively efficient way of informing the general public and initiating action. In the process of developing Internet access, the groups have published press releases, videos, e-brochures, and myriad other forms of information sharing on their webpages. But unlike most publishers and libraries, there has often been a rather unorganized approach to the information the NGO provides.

Link Rot and Content Drift

The smallest groups often use volunteer labour to populate their pages which often results in lack of accessibility, link rot and what is referred to as content drift. These last two concepts are really quite prevalent in many websites. Link rot refers to the fact that URLs that once existed disappear entirely. Surely we have all encountered the dreaded "404 Server (or Page) not Found" message. This can be caused for a variety of reasons. Sometimes the server indeed disappears along with the organization that supported it. Sometimes the disappearance of the organization is related to lack of financial support, but in certain areas, the disappearance is the result of a government crackdown on criticism it might be receiving from the site. The idea of content drift is perhaps just as common. The item still exists somewhere on the site, but the original URL has changed. This can be due to reorganization of the website itself or the removal of the information for a variety of reasons. In addition, content drift relates to slight changes to the original document. The easiest representation of this would be the various editions of the Intergovernmental Panel on Climate Change. The report of this panel includes the preliminary report, the synthesis report, a summary for policymakers as well as the full report which did not appear until after the previous three.

Over the last fifteen years the phenomenon known as "link rot" has been increasingly studied, and become increasingly common. (Notess 2014) This disappearance of links to websites found in scholarly communication has created particular challenges to the librarian

from the problem of disappearing links in library guides (Tyler, D.C. &McNeil, B. 2003) to conference proceedings (Hughes, B. 2006) to legal research (Jackson 2013). We commonly tell our students to look at the references in the academic materials they are reading to find other, authoritative resources for their study. However, the change from print to electronic formats in publication, as well as the expansion from journal articles to websites to blogs to social media, has made it more difficult to say with assurance that what one reads today will still be there tomorrow. Note that this is just in the area of scholarly communication. There is a large body of grey literature that exists on the web that is occasionally used by scholars, but frequently used by the general public, policy makers and others relating to decisions that must be made in personal and public venues. A great deal of this information is used on a daily basis.

In a recent article in PLOS, joint researchers from University of Edinburgh and the Los Alamos Digital Library Research and Prototyping Team explored link rot and content drift in scholarly communication in the area of science publications in ArXiv, Elsevier and PLOS. (M. Klein, H Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou & R. Tobin, 2014) Their findings were stunning. They found that approximately one in five links cited in the corpus of scholarly literature they reviewed suffered from link rot. If the disappearance of web citations in the scholarly literature is so prevalent, what can be said about the disappearance of websites that belong to NGOs especially when they are under pressure from the authorities? This happens in many countries regardless of their level of development. For an excellent, if disturbing, list of sites that have come under pressure to be taken down from both government and private agencies go to the Electronic Freedom Foundation and visit the Takedown Wall of Shame. (https://www.eff.org/takedowns)

The Internet Archive through its Wayback Machine is the seminal resource for the preservation of websites throughout the world. Begun in 1996, this resource has been capturing web pages on a regular basis and now allows individuals to submit URLs of interest. It provides many challenges to effective use since one must know the URL of the site one is looking for, rather than providing an intuitive search. For example, a search for IFLA or the International Federation of Library Institutions and Associations produces nothing, but a search for https://www.ifla.org provides the first snapshot of the organization's web site on December 6, 1998. A note on the page indicates that an earlier site existed under http://www.nlc-bnc.ca/ifla for which we find a snapshot on July 5, 1997.

Scope and Purpose of the World Sustainable Development Web Archive

In 2013 members of the International and Area Studies Library at Illinois were approached by the University Archives to participate in a web archive services pilot study. The University Library began the project in collaboration with the California Digital Library Web Archive Service. First, a specific subject area was identified that all members of the unit would be interested in working on – Sustainable Development. The Library aimed to discover and archive websites produced by NGOs relating to sustainable development in any language and from outside of the United States.

The University of Illinois' World Sustainable Development Web Archive aims to preserve web content published by NGOs that focus on environmental and economic sustainability. By collecting and making these materials available to scholars, students, and the public, the International and Area Studies Library hopes to support interdisciplinary research and inquiry into both particular and global trends in worldwide advocacy for

sustainable development. The sites collected have a rich array of documentation, data, images, and media that preserve the diverse perspectives, activities, and practices of sustainability NGOs around the world. The sustainability archive is similar to other academic library projects to preserve NGO literature such as Columbia University's Web Archive, which "is a searchable collection of archived copies of human rights websites created by non-governmental organizations, national human rights institutions, tribunals and individuals" that began in 2008 (Columbia University, 2015).

To ensure broad coverage of sustainability movements, Illinois' archive actively collects materials from NGO's that represent multiple linguistic and cultural groups. Each site is selected by a subject specialist from the International and Area Studies Library. The relevance of site content, perceived stability, and organizational structure are considered when NGO sites are selected for the archive. Further, the archive prioritizes sites that appear to be in peril because of known political conflict or evidence of website neglect. Although some web archives seek permission from organizations to "crawl" and capture their sites, we made a conscious decision to ask for "digital" permissions by honouring each server's Robots.txt file which grants or denies webcrawlers the permission to archive a site. This practice is analogous with collection policies for print grey literature and ephemera gathered from NGO's by research libraries. Each site archived is made available 180 days after its initial capture to ensure that the archived copy is not confused for a mirror of the original content.

Initial Process to Seed Archive and Organize the Web Archive

Two steps were necessary to move the project forward. The library initially focused on accessibility issues to ensure appropriate meta-data existed for each organization and site. First, the subject specialist identified a small selection of NGO sites to capture. The initial listing of sites consisted of a total of 38 organizations representing East Asia, Eastern Europe, South Asia, Latin America, the Middle East, and Africa. Organizations within this group ranged from the highly localized to regional and international advocacy groups. In addition, the scope of the organizations varied, ranging in advocacy efforts from biodiversity to environmental education to water. To help organize these resources, librarians contributed both Library of Congress Subject Headings and open tags following Dublin Core standards to enrich access to the archived sites.

These subject headings and tags were used to create a taxonomy to enable future subject searching across the archive and regardless of each page's and organization's working language. Table 1 displays fifteen tags translated into Hindi, Arabic, and Chinese that were selected within this process for translation based upon their frequency and cross-relevance among sites and between subject specialists. All terms were translated into English, Spanish, Hindi, Arabic, Russian, Chinese, and Japanese to reflect the languages emphasized in the International and Area Studies library print collections. These translated terms will provide a means to search the archive within these topics across languages and regions.

Table 1: Translated Sustainability Terms for Web Archive			
English	**Hindi**	**Arabic**	**Chinese**
Environmental education	वातावरण शिक्षा	تعليم/تثقيف بيئي	环境教育
Energy	ऊर्जा	الطاقه	能源
Education	शिक्षा	تربية	教育
Indigenous peoples	आदिवासी	سكان الاصليين	原住民
Youth organizations	युवा संस्था	منظمات الشباب	青年组织
Pollution	प्रदूषण	تلوث	污染
Biodiversity	जैव विविधता	تنوع حيوي	生物多样性
Conservation	संरक्षण (भण्डारण)	محميات	保存
Environmental protection	वातावरणीय सुरक्षा	حماية البيئة	环境保护
Sustainable development	टिकाऊ विकास	التنميه المستمرة	可持续发展
Water	पानी, जल	ماء	水
Climate change	जलवायु परिवर्तन	تغيير المناخ	气候变化
Nuclear energy	परमाणु उर्जा	طاقة نووية	核能
Development	विकास	تنمية	发展
Agriculture	कृषि	زراعة	农业

Subject specialists also provided descriptions, geographical, and organizational information for each organization included in the archive. Table 2 displays a full site profile with the translated tags integrated into the record. As the project progressed, this meta-data was never fully utilized in the CDL Web Archiving Platform as the platform does not have robust system for making metadata available to end users or re-organizing and filtering of search results through an open API. These limitations, which are discussed later, limited the end-user experience and site accessibility.

Table 2: Example Meta-Data for Archived NGO Website	
URL	http://www.peace-forum.com/gensuikin/
Vernacular Title	原水爆禁止日本国民会議 (原水禁)
English Title	Gensuikin (Japan Citizens' Assembly against Atomic and Hydrogen Bombs)
Description	Gensuikin is an anti-nuclear advocacy group that has its organizational roots in the anti-nuclear armament movement in the 1950s. The group's focus has since expanded to include protest against the use of nuclear energy and promotion of clean energy.
LCSH	Antinuclear movement; Nuclear disarmament; Renewable energy sources; Energy policy

Languages	Japanese; English
Country	Japan
Tags	antinuclear movement; nuclear development; nuclear energy; clean energy
Combined Archive Wide Tags	nuclear energy; energy
Translations	反核運動; 核開発; 原子力発電; クリーンエネルギー

Web archiving commenced in April of 2013 with the initial sites selected by Librarians. These sites were each set-up to be archived once every six months. In addition, each site is embargoed for a period of 180 days prior to making the archived site available for public searching. Subject specialists continue to add sites to the archive, which currently contains 218 captured sites.

World Sustainable Development Web Archive Metrics:

- Size: 210 GB
- Captures: 441
- Files captured: 2,065,379
- Average capture duration: 10h 56m 39s
- Average files captured per capture: 4,683
- Average size per capture: 477 MB

Ongoing Archive Maintenance

Technical Infrastrucure vs. Access and Usability

Like many digitization projects, there is often a disconnect between the technical architecture and modes of access. The California Digital Archive WAS, which will be retired in the summer of 2015, presented similar challenges. Much of the meta-data available on the administration side of the archive was not available to end-users. In addition, the search interface didn't allow for the filtering of sites based upon the meta-data available on the back end.

The reporting system for the WAS is an example of an technical architecture focused nearly exclusively on the significant challenges of capturing, preserving, and rendering archived web sites.

Illinois is currently in the process of transfering the archive to Internet Archive's Archive It platform, which as a much more extensive search capabilities and access features. As Figure 1 displays, the reporting feature of the platform exclusively provides data related to the capture and archive process, yet doesn't provide usage statistics or data focused on access to the sites archives. When creating an archive that is to be the basis of research services, developing a complimentary access infrastructure is essential to project success.

Figure 1

Challenges to Maintaining NGO focused Web Archives

In addition to platform challenges, archiving NGO website presents ongoing maintaince issues. As sites are archived on a regular basis (in this case twice per year), it is essential to monitor the success of each capture and update, remove, or edit sites as organizations change. The case of the Crimea Republican Association provides an excellent example of the challenges of archiving NGO's. As Figure 2 displays, the site was successfully captured in September of 2014 with over 4,000 files in 187MB of data.

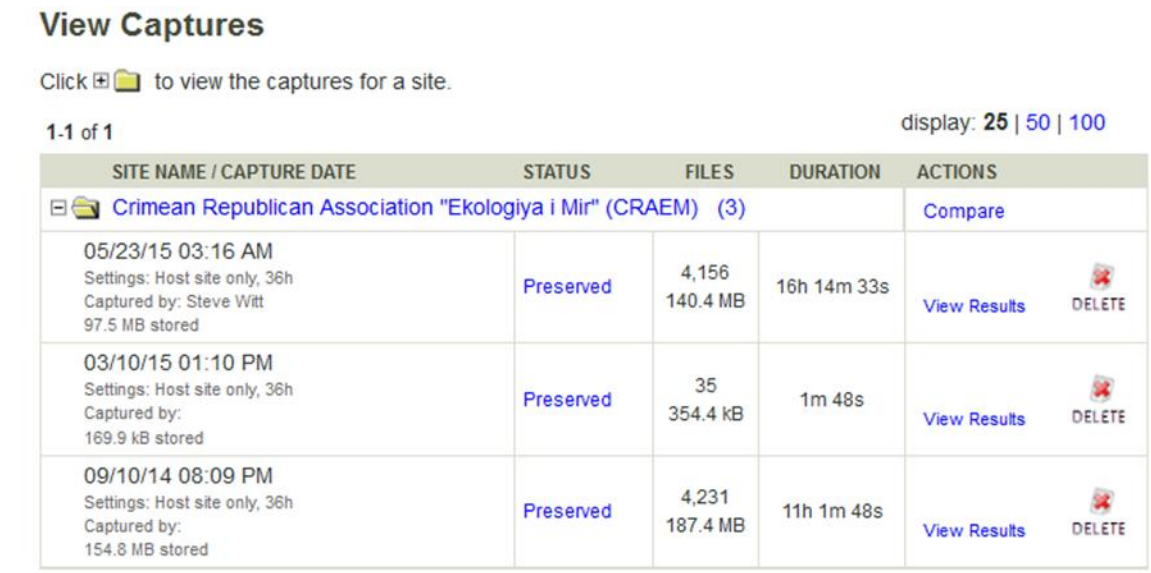

Figure 2

By March of 2015, however, the site was no longer available. In March of 2014, this region moved from the Ukraine to Russia. As seen in Figures 3 and 4, the September 2014 version of the site displays a Ukraine address for the organization dispite the annexation in March.

Figure 2

Figure 3

An internet search reveals that the organization's webiste moved from the http://www.ekomir.crimea.ua/ to a new URL http://ekomir.org/. The new URL, however, was registered to and listed in the Ukraine (80.245.112.25). Updating the capture setting for the new URL yielded the Crimea Republican Association's site and showed that the organization's address had been updated to Russia on the Russian version of the site while retaining the Ukraine address on the English version and a Ukraine IP address. Whether this is a political statement or negligence in updating the full site is unknown. It does, however, provide a glimpse into the challenges of maintaining ongoing web archiving amidst factors that include political change, organizational resources, and technical factors governing internet accessibility that are outside of the Library's control. Although these are problems similar to collecting print ephemera and grey literature, the ongoing development of web archives that attempt to collect and document ephemera are challenged to overcome the technical changes within web sites that include the adoption of new hosts and domains, changes in site technology, or moves to social media platforms such as Facebook through which content achieves an ambigious corporate ownership. This requires a time-commitment and knowledge of both web-based technologies and intellectual property regimes that goes beyond physcial collecting in many regards.

An example of technical challenges that includes the need to archive social media and sites simultaneously is seen in the Uganda Water and Sanitation NGO Network site. Figures 5 and 6 show two versions of the archived site. If a web archive needed to capture the site and content as a point in historic time of the organization or topic, web capture tools which are designed to grab computer code that is later rendered as a website create problematic

scenarios. Reviewing captures of the site in June 2014 and June 2015, one can see differences in the site structure, new reports, and listings of key events from the organization that would be of use to historians, funding agencies, other NGO's, and the organizations itself. If you look, however, at the Twitter feeds that are imbedded into the site, you'll notice that this is not archived content at all. The only portion of the feed that is archived is the code. As we can see in Figures 5 and 6, the Twitter feed itself displays the most recent post, regardless of when the page was archived. As organizations move increasingly to imbedding code and creating dynamic content that is derived from multiple dissemination platforms, this problem will increase if the capture technology does not keep pace with website technologies
.

Figure 4

Figure 5

Conclusion

The information explosion that we have been experiencing for the past two decades has often caused us to ponder whether all information is worth preserving, and surely there is some chaff that can fall to the wayside. However, the challenges of preserving that which is and might be useful to us are made very apparent in our exploration of web archiving and grey literature. When this type of information existed in our vertical files, we had the choice of reviewing and weeding items. With the advent of born digital grey literature, these options have disappeared to a large degree. The technology, however, has advanced to enable this type of capture in part because of the interest of librarians and in large part due to the interest of programmers in the challenges that the digital world presents. The archiving of various Twitter feeds is some indication of the possibilitites that exist. But as noted previously, the concurrent archiving of websites and social media needs further exploration.

What is needed now is even greater collaboration and cooperation between technologists and librarians, in fact between programmers and area and subject specialists, not only to identify and articulate the challenges of web archiving, but to provide responses to these challenges. The growing body of research into link rot and content drift should be coupled with experimentation into the responses that can be coordinated to the problems they present. The impact of these responses will go well beyond simply retaining information that is held on random NGO websites to enhancing the reliability of a great deal of scholarly literature as well.

As many organizations move to adopt social media platforms as enhancements and replacements to their websites, the challenges of capturing the information produced by NGOs will increase. The already extant pressures these organizations experience from their own constituents and governments alike will grown in the future. We need to combine the expertise that is available at a technological and social level to anticipate and react to the changes that these pressures will evoke to develop a robust and nimble approach to web archiving for the future.

References

Hughes, B. (2006). Link? rot. URI citation durability in 10 years of AusWeb proceedings. *AusWeb 2006: 12th Australasian World Wide Web Conference*. Retrieved from http://ausweb.scu.edu.au/aw06/papers/refereed/hughes__linkrot_/paper.html

Jackson, L. J. (2013). 'Link rot' is degrading legal research and case cites. *ABA Journal*, 98(December). Retrieved from: http://www.abajournal.com

Klein, M., Van De Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., et al. (2014). Scholarly context not found: One in five articles suffers from reference rot. *PLOS ONE*, 9(12). doi: 10.1371/0115253

Notess, G. R. (2014). Surviving rot and finding the online past. *Online*, 38(2), 65-67.

Tyler, D. C., & McNeil, B. (2003). Librarians and link rot: A comparative analysis with some methodological considerations. *portal: Libraries and the Academy*, 3(2), 615-632. Retrieved from https://muse-jhu.edu.

Union of International Associations. (2014). *Annuaire des organisations internationales = yearbook of international organizations* (51st ed.). Geneva, Switzerland: Société de l'Annuaire des Organisations internationales.