Lorena Ramirez-Lopez
Assignment 1
Digital Preservation Fall 2014

The International Internet Preservation Consortium (IIPC) is an organization made up of over 50 different international archives, libraries, and institutions that collaborate in order to improve the tools, standards and best practices of web archiving for current and future access and use. Amongst their missions and goals, IIPC "is to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations (netpreserve.org)". In order to achieve its mission, IIPC requires its members "to enable Internet content from around the world to be preserved in a way that can be archived, secured, and accessed over time; to foster the development and use of common tools, techniques and standards for the creation of international archives; and to encourage national libraries everywhere to address Internet archiving and preservation and develop web archives (netpreserve.org)". These web archives are saved for not only current research, but also for future research and cultural heritage. René Voorburg from the National Library of the Netherlands (and current member of IIPC) sums up the dire need for preserving and actively doing web archiving: "[if] a researcher of fifty years time wanting to study our time to see how we experience, how we live, without having internet…it is completely impossible…You can't study our life and our time without internet so we must preserve it."

IIPC is an organization based on the membership of national, university and regional libraries and archives. Originally, only the National Library of France and 12 other participating charter institutions were allowed to fund and participate in projects. However, since its inception in July 2003, libraries, archives, museums and cultural heritage institutions around the world can inquiry about membership. There are roughly 50 members from over 25 different countries that are currently part of IIPC. These members are expected to work collaboratively within its country's own legislative framework to identify and develop the selection, preservation, and access of Internet content. Despite the fact that IIPC is a global organization, there is a severe lack of representation from the southern hemisphere of the world, specifically from South America and Africa. However, this discrimination may be due to participation fees and unclear frameworks rather than a lack of collections.

The scope of the materials for these collections all focus on the born-digital content created and collected from the Internet. There are materials that have been scanned and digitized from scholarly publications, works of art, documents, news, and other source materials, but the access and display of these objects are all done via web pages. "Web archiving is the process of collection portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use (netpreserve.org)". Today, the selection process revolves on: selection, harvest, preservation, and access. Web archives are selected to compliment existing collections of institutions while serving different goals. This content is harvested via broad and large crawls from software that downloads code, images, documents, and other necessary files in order to reproduce the web site *at the time of capture.* These web crawlers will also collect metadata about the conditions of the process *during the time of capture.* Now already the issue is that there needs to be various "snapshots" of the web site in order to preserve and recreate our use and experience. However, these crawlers can only work while the web site is still accessible and running. And then there is the issue of *what*

should be collected and preserved from this vast array of content. These web pages are constantly changing and in order for future generations to understand the Internet within our present context and have access to the same materials, the web must be captured in real-time. As a consequence of the Internet's rapid change, IIPC's focus shifts throughout the years. Between its initiation and 2006, the development of access tools, content management, and researchers requirements had to be established. Between 2007 and 2010, access, harvesting, preservation, and standards were IIPC's newest goals. The question "'what should we preserve?' is directly addressed by the Researchers Requirements Working Group from IIPC. It recognizes that, because of the huge volume of material on the Internet, it is inevitable that not everything can be collected. This means that the decisions that we make about what to collect now will have an enduring impact on what is available to researchers of the future. This working group, which consists not only of members but also of invited researchers in the area of Internet studies, is aiming to define a common vision of what needs to be collected." This is why standards and techniques must be developed amongst collaborating members. These collaborations have led to development of important standards, tools, and techniques for countries to adapt when preserving the web. Within terms of acquisition, there have been enhancements to the Heritrix crawler by Internet Archive, a member of IIPC. A Heritrix is an open-source extensible, web-scale, archival-quality webcrawler project where its interchangeable components can be plugged to a specific website. The setup involves the configuration of specific components: scope, frontier, and processer chains that define a crawl. The Scope determines what URIs (websites) are included in this crawl. The Frontier tracks which URIs are to be scheduled for download and which URIs have been already collected. And the Processor chains perform specific actions established by the programmer. There has also been the development of ArchiveFacebook, which is a Mozilla Firefox add-on for people who want to archive their Facebook. Another important accomplishment was developed for collection storage and maintenance: the WARC standard. WARC- Web ARChive film- actually specifies a method for combining multiple digital resources into a collective archival file with other related information. And tools like the Open source wayback for access and display. The Wayback Machine is a replay for web archives stored in ARC or WARC file formats in order to allow temporal navigation of archived web resources. Like the Heritrix crawler, Internet Archive developed the Wayback Machine. It is one of the key softwares developed thus far to serve the changing needs of the archiving community and its users.

Nevertheless, no one member in IIPC is faced with the same challenges and legal situations. Each institution has a specific collection development policy based on their countries legislative frameworks. In France, web archiving is a legal obligation required from its institutions for cultural heritage purposes. Within the Library of Congress, web archiving is a "*collection-based activity…* part of a *named subject, event,* or *themed-based collection* rather than an individual one-by-one archiving practice. Some websites also use a robots.txt file to provide instructions to crawlers, which can interfere with the archival of the actual site. Some sections of the sites or even entire sites can be blocked using robots.txt. Some IIPC members obey robots.txt except when it comes to inline images and stylesheets. Others are seeking permission in order to bypass the robots.txt so that the sites archived are as complete as possible. Site owners wishing to be archived

should inspect their robots.txt files to ensure that they are preservation-friendly and do not restrict archival crawlers from visiting. For IIPC members and also other organizations to understand how to make their sites more "preservation-friendly", resources have been gathered and made accessible to the public. In depth documentation is reserved for only members, but there are actually three accessible and useful blog posts that focus on making a web site that is more easily archived. "Designing preservable websites, redux" by Nicholas Taylor of the Library of Congress, "How to make websites more archivable" by Helen Hockx-Yu of the British Library, and "Five Tips for Desigining preservable websites" by Robin Davis of the Smithsonian Institution Archives. All post can be found via the IIPC's FAQs page.

International Internet Preservation Consortium. http://netpreserve.org

Bibliotheque nationale de France. “Digital legal deposit: four questions about web archiving at the BnF.” http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html

Library of Congress. “ Library of Congress collections policy statements supplementary guidelines.” http://www.loc.gov/acq/devpol/webarchive.pdf

"Sustainability of Digital Formats Planning for Library of Congress Collections." *WARC, Web ARChive File Format*. Library of Congress, n.d. Web. http://digitalpreservation.gov/formats/fdd/fdd000236.shtml

Miranda, João. “Web Harvesting and Archiving.” *Instituto Superior Técnico do Lisboa* http://web.ist.utl.pt/joaocarvalhomiranda/docs/other/web_harvesting_and_archiving.pdf

Phillips, Margaret E. “What should we preserve? The question for heritage libraries in a digital world.” *Library Trends* 54.1 (2005) 57-71. http://ezproxy.library.nyu.edu:2311/journals/library_trends/v054/54.1phillips.html

Hallgrímsson. “Internatinoal Internet Preservation Consortium: A short introduction.” National and University Library of Iceland. 2008. http://www.bl.uk/ipres2008/presentations_day2/39a_Panel_Discussion.pdf