The Center for Research Libraries
GLOBAL RESOURCES NETWORK®

The Center for Research Libraries

16 April 2010

Human Rights Resources Profile

Web Ecology Project

By Sarah B. Van Deusen Phillips, Ph.D., Project Coordinator for Human Rights, Center for Research Libraries

Overview

The Web Ecology Project (WEP) is an independent and interdisciplinary research group located in Boston, Massachusetts, that explores the flow of culture and community on the Web. To accomplish this, WEP builds tools and utilizes Application Programming Interfaces (APIs) that permit large-scale data mining of social media. These tools have useful potential for human rights archivists seeking to capture and preserve new primary resources that appear on the Web via a variety of social media platforms (e.g., Facebook, Twitter, blogs, MySpace, etc.). The data collected is captured as text, which is then organized and analyzed in massive database files that can be used for both qualitative and quantitative analysis. Web Ecology's analytical tools allow researchers to capture and archive content from the Web and submit it to rigorous quantitative analysis to identify and characterize patterns of movement in content, culture, and online communities. More importantly for preservation, Web Ecology archives all of the data it collects and makes these data available to interested parties upon request. Web Ecology also hopes to develop a freely available "tool kit" of open source tools to allow researchers and organizations to harvest, analyze, and archive text-based Web content. These tools would allow archivists to capture and preserve important but fleeting digital documentation and evidence for human rights.

History & Mission

History
The Web Ecology Project was launched in Boston in spring 2009 by a loosely affiliated group of researchers interested in fast-paced Internet research that generates real-world outcomes. The investigators in the Web Ecology Project group have a common interest in cutting-edge methods to study the Internet and come from a variety of organizations and industries in the Cambridge area, including Harvard's Berkman Center, the Center for Future Civic Media, Convergence Culture Consortium, Harvard Graduate School of Education, as well as various marketing, digital media, and startup companies.

Web Ecology's work began with intensive research of Twitter and has branched out to other social network platforms. This work has been receiving positive feedback and input from the general Cambridge research community, including MIT and Harvard. Most recently, Web Ecology has been involved in adopting and adapting a number of existing methodologies—ranging from social network theory and economics to qualitative and ethnographic research methods—to delve deeply into online communities

and to better understand how online social networks form, as well as how content and culture shift and move dynamically across the Web.

Mission
Web Ecology aims to develop methods for analyzing and interpreting the structure of the Web to understand and contribute to the movement of information via social media on three levels: 1) the development of social media platforms, 2) how content travels through these platforms, and 3) how users produce and disseminate that content. Web Ecology attempts to "unify contemporary research and practice under a common focus, set of principles, and general approach" to gain insight into—and create means of exchange among—the various disciplines currently studying the Web from disparate and often unconnected points of entry.[1] The group hopes to "lay the groundwork for a more vibrant, more dynamic, and more useful field of research and community researchers."[2]

Chief Activities

WEP primarily engages in large-scale data mining to support analysis of online culture and the flow of information (or content) across social media. It aims to build tools that enable interested parties to engage in data-driven research of social media backed by network science. Related to this goal is the creation of an archive of data collected from the Web to support continuing analysis of Web content and social media use. As stated in the online article "Reimagining Internet Studies: A Web Ecology Perspective,": "The establishment of specific, dedicated archives benefits the maturation of related scholarship. [. . .] Accordingly, Web Ecology stresses efforts to curate data for potential analysis in . . . related areas of research."[3] To this end, Web Ecology engages in massive harvesting of text-based content from social media platforms and experiments with a variety of quantitative analytical models for discovering, tracking, and characterizing the movement of content and community on the Web.

Participants/Collaboration

At this point, the Web Ecology Project's independent group of investigators conducts research with an eye toward developing innovative methods and approaches for understanding the dynamics of Internet use. They are not currently collaborating with any other groups or individuals; however, they hope to build collaborations with interested parties in the future. The group views human rights as a potentially rich locus of collaboration, and ideas have been informally floated concerning the possible nature of these collaborations. For example, Web Ecology could do custom work for a human rights organization to develop a new tool for harvesting social media. The tool could be released to the organization for beta testing and Web Ecology could manage and support the program for an established number of years based on each group's needs. Though the Web Ecology team does not currently engage in this sort of collaboration, some of the investigators would like to see these types of partnerships in the future.[4]

Governance

The Web Ecology Project is an unaffiliated research group made up of independent researchers in the Boston area interested in the social processes of the Internet and social media. Members pool their expertise and resources to conduct relevant research into social media trends and to purchase infrastructure such as servers.

Services/Technology Resources

The Web Ecology Project offers or is developing the following services and technology resources:

---

[1] See: http://www.webecologyproject.org/2009/08/reimagining-internet-studies/#more-7
[2] *Ibid*
[3] See http://www.webecologyproject.org/2009/08/reimagining-internet-studies/
[4] Special thanks to David Fisher of the Web Ecology Project for describing the potential collaborations elaborated in this report.

- Archive of text-based content gathered from social media platforms and stored in searchable CVS files. Data for potential research or other partnerships are available upon request from Web Ecology and are delivered as database files that can be opened in standard database programs such as Microsoft Excel (see the database sample in Appendix A). Requests should be directed to contact@webecologyproject.org. Web Ecology hopes to add "value-added" annotation to address particular analytical goals or concerns.

- Development of open source technology for harvesting social media data
    - WEP created its first tool, the Google language python module, to identify the language that text content was created in and translate that content to English. The report "Code Release: Language Detection and Translation" contains a description of this tool and is available for download at www.webecologyproject.org.[5]

- A variety of social media reports demonstrate how the tools work and are published online at www.webecologyproject.org

Example: Twitter harvesting

To demonstrate how Web Ecology gains access to Web content, the following is a description of how the group harvests material from Twitter. (See Appendix A for a sample of harvested material).

Web Ecology has devised a workable solution for harvesting Twitter tweets, which it is extending to harvest data from other social media sources (e.g., Facebook, blogs, Flicker, YouTube, etc.).[6] By using readily available server technologies, working with Twitter's established data access interface, and drawing on the skills of trained programmers, the WEP research team collects, stores, and archives massive numbers of Twitter tweets.[7] The tweet-harvesting set-up is straightforward and can potentially be implemented by any organization wishing to gather similar materials from Twitter, as long as they have access to a programmer who can help manage the process.

To collect and archive tweets, WEP first gains access to Twitter's Application Programming Interface (API) by following a standard application process Twitter established for permitting access to data. An API serves as a common access point that allows various programs and platforms to "talk" to each other through shared variables, even if they do not share the same programming language. Basically, the API allows programmers to build applications that share information between platforms (for example, the ability to post Twitter tweets via Facebook or Facebook updates via Twitter).

With API access secured, data can be captured and downloaded from Twitter's database. WEP's programmers accomplish this by writing code that requests data from Twitter's servers via the API. The code instructs Twitter's server to harvest data that meet specific search criteria contained in the code request—typically key words or phrases that appear in tweets about the event or topic of interest. For example, if researchers wished to collect tweets related to the 2009 Iranian presidential election, they would submit search terms such as: #iranelection, Neda, Ahmadinejad , etc. When Twitter's data server receives the code command, it pulls all tweets containing any of the requested terms, bundles them as a data packet, and sends the packet back to WEP's server.

Once the data arrive in WEP's server, the tweets pour into a massive database program as individual text files accompanied by relevant metadata (time and date tweet was created, Twitter user name, and location, if available). The database is essentially a meta-form of an Excel spreadsheet organized in rows and columns-- the sort of chart that can be created when establishing a server's architecture. Once the

---

[5] http://www.webecologyproject.org/2009/09/code-release-google-language-tool/

[6] Special thanks goes to Dharmishta Rood of the Web Ecology Project for explaining the data harvesting and archiving process described in this report.

[7] Copyright on all tweets belongs to Twitter users; however, Twitter encourages users to contribute their tweets to the public domain (see http://twitter.com/tos for details on terms of service and copyright). Tweets submitted as such fall under fair use rules for copyright.

tweets are grouped and stored in this database, they are searchable and sortable, so both qualitative and quantitative analyses can be run. Most importantly, the information can be easily archived and shared because a database, a fundamental type of programming, does not change much over time, so the content will still be readable in the future.

Though the Web Ecology Project's request and delivery process is rapid and efficient, this process contains a few important limitations. First, once a code request is sent, harvest and delivery of data is automatic, however, the request process itself is *not*. The current iteration of the code must be handwritten and manually sent, which can complicate archiving tweets for the duration of an important event. Typically, Twitter users responding to events send out tweets for a few days, so that data need to download for the duration of the event to capture as much relevant material as possible. Since the WEP programmers have not yet written code to serve as a means of sending automated requests to Twitter, they have to manually resend requests for a particular set of terms at regular intervals over the course of several days as they follow a trending topic on Twitter. Second, although Twitter shares its data freely, stipulated time limitations on harvesting exist. At the time of this writing, only data up to five days old can be collected in response to a code request (although Twitter does maintain a database of all of the tweets ever posted since it came online in 2006). However, these limitations should not hinder harvesting if a researcher or archivist diligently begins requesting data shortly after an event begins to trend on Twitter and then regularly resends the request until the event dies down.

These exceptions aside, the process described above provides a model for establishing and maintaining archives of fleeting, first-person, digital documentation of key events produced through social media platforms. Though the Web Ecology Project team established this process for collecting and archiving Twitter data, other social media platforms--such as Facebook, MySpace, or LinkedIn--also use APIs to integrate their functions with other social networking platforms so that users can work seamlessly between their various social presences on the Web. Therefore, WEP's Twitter research process would also apply to collecting and archiving digital documentation from a variety of social media sources.

, WEP researchers make the data they collect and archive available to interested parties when and where appropriate, within the limitations of legal restrictions with Twitter and Twitter users. If you are interested in learning more about data availability, email WEP at contact@webecologyproject.org. Dataset availability is dependent upon WEP research; the group can only make data available that it originally collected for its own research interests. At the time of this report, WEP researchers state that they plan to store all of the databases and archives they create indefinitely as a resource to future investigators. For more information on the goals and objectives of the Web Ecology Project, see the mission statement at www.webecologyproject.org.

Challenges

Funding
As an organization, the Web Ecology Project is still evolving and its business plan is shifting.

Changing technology
Legal access requirements at various social media platforms are constantly changing, which impacts Web Ecology's ability to harvest Web data and can limit the ways that it can analyze or use those data once collected.

Structural challenges
The Web Ecology Group is currently revising its goals and research objectives as it responds to feedback and interest from the broader field of social media research.

Comparative Landscape

The emerging study of social media contains a variety of developing tools that potentially will allow archivists and activists to harvest and preserve relevant Twitter tweets, as well as other forms of social media such as blog posts, Facebook data, SMS data, and the like. These various efforts are able to

accomplish such harvesting through API access agreements with the target platforms. Though a wide variety of projects related to capturing information from social media for research purposes exist, listed below are a few representative efforts geared toward harvesting social media for analysis or archiving purposes that may be of specific interest to human rights organizations and archivists.

Automated Harvesting Tools and Resources:

JISC PoWR
http://jiscpowr.jiscinvolve.org/

- Offers a Twitter harvesting program, but not archived data. The program was created to harvest the tweets associated with a local conference and is available to download from the Web site.
- Offers links to online services that do some tweet harvesting (highly constrained)
- Information is available through a series of blog posts without a comprehensive report or easily discoverable program download.
- Offers information specific to archiving social media data for professional archivists.
- Offers the JISC-PoWR handbook "Preservation of Web Resources" (created in 2008 and not updated). Available at: http://www.scribd.com/doc/7760433/JISC-PoWR-The-Preservation-of-Web-Resources-Handbook

Tweetdoc
http://www.tweetdoc.org/

- Free online program that allows archiving of specific events followed in Twitter
- Uses hashtags or specific search terms to identify and collect all tweets related to the named topics
- Can save by date and time
- Produces a .pdf document of the collected tweets that can be downloaded and saved to a computer or server.

Manual Harvesting Projects:

ArchivePress
http://archivepress.ulcc.ac.uk/

- An initiative headed by JISC and the British Library to provide open-source software that will allow users to harvest and archive a variety of online social media
- Rather than using a Web-crawling approach, the platform experiments with RSS feeds and API access to blogs as a means of gathering blog content, including associated comments.
- Content stored using instances of Wordpress (a blogging platform available at http://wordpress.com)
- Hopes to achieve reliability and authentication of sources as well as citable content with persistent identifiers.

Columbia University Library Human Rights Web Archive
http://www.columbia.edu/cu/lweb/indiv/humanrights/hrwa.html

- Initiative funded by the Andrew W. Mellon Foundation to identify, preserve, and provide access to Web-based human rights materials
- Particular focus on Web sites at risk of disappearing in a short period of time to save these unique sources of information to the documentary record of human rights practice and research
- Subject specialists identify candidate Web sites and a Web archivist harvests the site's information using Delicious, a freely available Web-indexing program.
- Delicious entries are tagged by hand and submitted to ArchiveIt for preservation.
- Archived Web pages are catalogued in the Columbia University Library online catalog and WorldCat.

UK Web Archive
http://www.webarchive.org.uk/ukwa/

- Site that archives Web content from sites that publish research of interest to Archive participants
- Users can nominate sites to be harvested and archived
- Underpinning infrastructure provided by the British Library
- British government has passed legislation allowing six research libraries legal right to harvest and archive Web content (as long as they gain the Web site owner's permission)
- Freely accessible to the public as part of the national archiving initiative

See also for information supporting online archiving:

European Commission on Preservation and Access
Draft Report for UNESCO: "Preservation of Digital Heritage" (Yola de Lusenet
March 2002)
http://www.knaw.nl/ecpa/PUBL/unesco.html

- This report offers a discussion of the value of digital media as heritage and scholarly material.
- The "Approaches to digital preservation" section offers suggestions for harvesting and archiving a variety of online materials.
- Discusses legal issues related to collecting and preserving online and other forms of digitally produced documentation.

Appendix A:
Database Sample containing Twitter Tweets from the June 2009 Iranian presidential election protests*

id	created_at	text	source	from_user	from_user_id	in_reply_to_user_id	in_reply_to_status_id	iso_language_code
2########6	6/19/09 4:49	RT @ixxxxxx_student RT @ixxxxx RT @LxxxxNxxx# Please people keep going it on! spread the word! - http://bit.ly/F3Wrm #iranelection	web	ixxxx##	2######4			en
2########1	6/19/09 1:05	What is the link for the "1 click green" to support the efforts in IRAN?	TweetDeck	rxxxxxxx	2#####9			en
2########9	6/20/09 22:17	TxxxxCxxxxx Ten Iranian Videos http://bit.ly/2JVzr	web	AxxxCxxxx	1######0		1######8	es
2########7	6/14/09 19:54	Wow. Tanks on the streets of Tehran now	web	bxxxxxxxx	5###4			en
2########1	6/20/09 13:19	RT @xxxxxxxxxxxx: Lenin said: elections always won by party that counts the votes. Mao: Power comes from a gun. Iran combines the two!	TwitterFon	txxxxx	1#####9			en
2########6	6/19/09 14:06	Cxxxxxxx/Axxxxx: you are only strong with armory!You are far away from Allah but you know it because hell is your home! #iranelection #gr88	web	Shxxxxx##	2######5		2######3	en
2########8	6/19/09 12:55	[BBC FEED] Protest at Iran's 'evil UK' claim: The Foreign Office is to protest to the Iranian ambassad.. http://tinyurl.com/mbtkny	twitterfeed	gxxxxx	4#####2			en
2########9	6/21/09 22:23	But I can say this the Iranian gov't is very afraid of this power and exposure. When you are afraid you are actually losing power.	web	pxxxxxx	1#####1			en
2########7	6/22/09 12:50	RT Exxxxx Mxxxx newspaper (pro Karroubi) offices attacked by militia - #Iranelection confirmed	Twittelator	Gxxxxxxxxxxxx	3#####5			en
2########1	6/21/09 14:12	To All Iranian Protesters & regular Iranian citizens-I am an ordinary person who stands with you and prays often for your safety&well-being!	web	HxxAxxBxHxxxxx	1######7			en
2########5	6/23/09 19:18	Daschle's favorite line was "I'm concerned"-Obama on Iran "I'm concerned"-Reid is concerned Pelosi is concerned-about what too much freedom?	web	txxxxxxxxxxx	7###3			en
2########9	6/21/09 7:58	Canadian embassy refusing wounded tehran protesters? http://www.aworldwithoutspin.com/?p=576	web	ExXX_Bxxxxxxx	2######6			en
2########8	6/16/09 10:23	RSSupdate Twitter schort onderhoud op wegens Iran: WASHINGTON (ANP) - De berichtenwebsite Twitter heeft.. http://tr.im/oE4E	twitterfeed	gxxxxxxxx	2####0			nl
2########6	6/22/09 11:50	RT Iran clerics struggle bursts into the open http://bit.ly/2OZsOX #IranElection	web	dxxxxxxx#	1######7			en

Key: Id = identification number assigned by Twitter
Created at = date and time a tweet was posted to Twitter
Text = text content of the tweet
Source = posting platform (e.g. Twitter, Tweetdeck, Facebook, etc.) from which a tweet was posted
From user = the screen number of the user that posted the tweet
From user id = the unique user id number of the user that posted the tweet. This number is assigned by Twitter when an account is opened
In reply to user id = the unique user id number of the person to which a user's tweet is a reply (if this it is the case that the tweet is a reply)
[Need "In reply to status id"?]
ISO language code = international standard short code identifying the language in which the tweet was posted.

* Original data file supplied by Web Ecology Project. The data in this sample were masked to protect copyright licenses with Twitter and Web Ecology Project, and were taken from a larger file consisting of 50 randomly selected tweets from WEP's Iran election database. Such data are available from the Web Ecology Project upon request by emailing contact@webecologyproject.org.