Web Archiving Use Cases

Emily Reynolds
Library of Congress, UMSI, ASB13

March 7, 2013

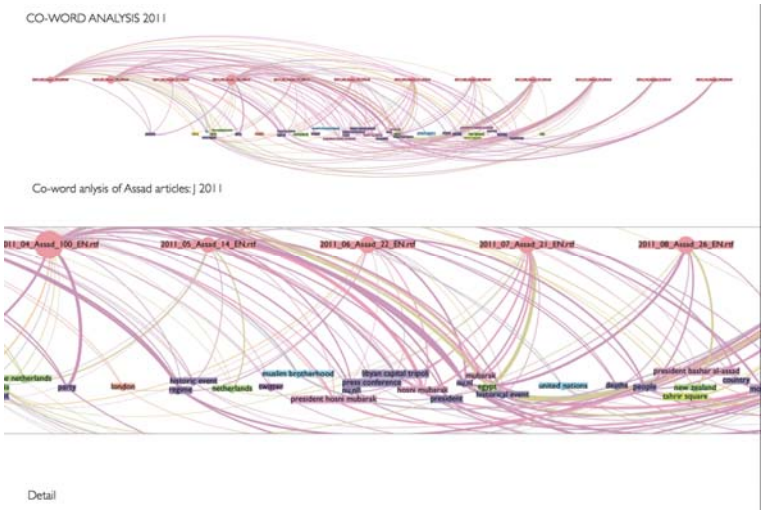
Table of Contents

1. Data mining and analysis

In their 2011 report Web Archives: The Future(s), Eric T. Meyer, Arthur Thomas, and Ralph Schroeder (of the Oxford Internet Institute) highlighted the need for web archives to be interoperable with research methods already being used by researchers studying the live web. These methods include visualization, access through powerful search tools, analysis of social networks and geographic data, and other cutting-edge approaches to studying Internet content.

As the technology supporting web archive collections becomes more developed and researcher data analysis methods more sophisticated, the possibilities for large-scale machine processing of web archive collections becomes more feasible. Tools like the visualization resources offered by the UK Web Archive (http://www.webarchive.org.uk/ukwa/visualisation) and the web datasets made available by the Common Crawl (http://commoncrawl.org/) and the Stanford WebBase Project (http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/) make web archive data more accessible to researchers, highlighting the value of archived websites for scientific analysis.

Text mining



Findings of temporal co-word analysis for query "Assad,"
https://wiki.digitalmethods.net/Dmi/Winter13SearchingTheArchive

Large-scale corpuses of captured websites offer the possibility for analysis of textual patterns and trends. Research projects studying the frequency of term usage or sentiment analysis have used web archive collections to extract, visualize, and analyze the language used in crawled websites.

This type of analysis can uncover relationships such as co-occurrence frequency between terms. Sentiment analysis can also be performed on large bodies of text to determine the emotions used when discussing specific topics. Much like digitized books can be mined for language usage patterns, websites show modern patterns of language.

Examples

N-gram Search, UK Web Archive (British Library)

http://www.webarchive.org.uk/ukwa/ngram/

> *"The N-gram search is a phrase-usage visualization tool which charts the monthly occurrence of user-defined search terms or phrases over time, as found in the UK Web Archive."*

Searching the (News) Archives, Web Archive Retrieval Tools (University of Amsterdam)

https://wiki.digitalmethods.net/Dmi/Winter13SearchingTheArchive **(findings 1 and 2)**

> *Demonstration of the possibilities of research with web archive search tools, focusing on a collection of a Dutch news aggregation website. Shows word frequency visualizations and analysis of term co-occurrence over time in relation to major news events..*

Sentiment Analysis and the Reception of the Liverpool Poets, Helen Taylor (Royal Holloway)

http://domaindarkarchive.blogspot.com/2012/11/sentiment-analysis-and-reception-of.html

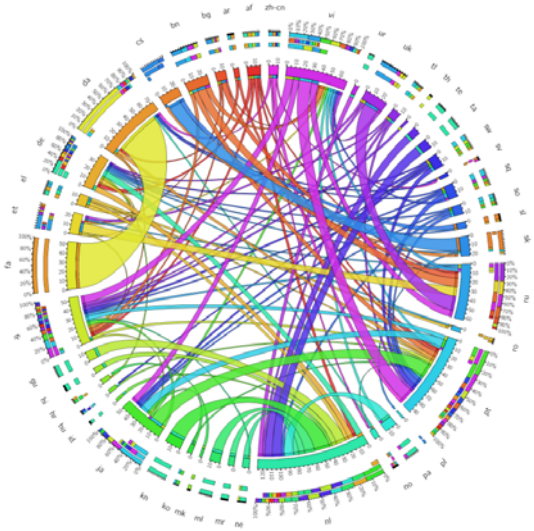
> *Proposal for using web archive collections to analyze online discussion of 1960s Liverpool poets, in comparison to data from newspapers and other formally published reviews.*

Footprint of the world top companies, Alexandre Marah and Ferenc Szabó (University of Tewnte)

https://github.com/norvigaward/naward05/wiki/Norwig-Award-:-Footprint-of-the-world-top-companies

> *Project using Common Crawl-captured websites to count mentions of the top 1000 companies on Forbes' list of The World's Biggest Public Companies*

Link analysis



Visualization of linking between websites of different languages, Babel 2012 Web Language Connections, https://github.com/norvigaward/naward25/wiki/Babel-2012---Web-Language-Connections

As large bodies of websites are captured, so are the links and connections between them. These networks of linked sites and data can be mined to observe the relationships between individuals, organizations, and ideas over time. Just as this kind of analysis is done on websites and social networks on the live Web, it can be used with web archive datasets to view changes over time or at points in the past.

Examples

Link Analysis, UK Web Archive (British Library)

http://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/linkage

> *"This visualisation shows an overview of how a subset of the sites in the JISC UK Web Domain Dataset (1996-2010) are interlinked. For each year, the corresponding chord diagram shows the percentage of links between the different second-level or top-level domains, such as the percentage of links found in *.ac.uk pages that link to *.co.uk pages."*

Babel 2012 Web Language Connections, Hannes Mühleisen (CWI)

https://github.com/norvigaward/naward25/wiki/Babel-2012---Web-Language-Connections

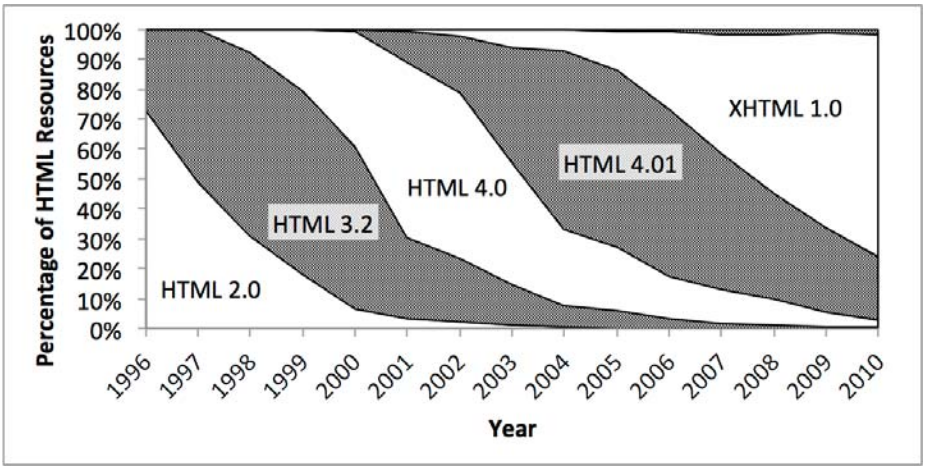
> *Project using Common Crawl-captured websites to discover and visualize links between websites in different languages*

Study of ~1.3 Billion URLs: ~22% of Web Pages Reference Facebook, Matthew Berk (Zyxt Labs, Inc.)

http://zyxt.com/post/26851542949/study-of-1-3-billion-urls-22-of-web-pages-reference

> *Analysis of Common Crawl-captured website linkages to Facebook.*

Analysis of technology trends



HTML Version Usage Over Time, UK Web Archive,
http://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/fmt

The formats captured in web archive collections can serve as a timeline for the development of web technologies. Analysis on captured websites can show changes in usage of file formats, programming languages, markup, and other attributes over time. These datasets show the rise and fall of various web formats, perhaps highlighting formats that need preservation attention or showing trends in markup and formatting.

Examples

Format Profile, UK Web Archive (British Library)

http://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/fmt

> *"The dataset is a format profile, summarising the data formats (MIME types) contained within all of the HTTP 200 OK responses in the JISC UK Web Domain Dataset (1996-2010)."*

Web Data Commons, Christian Bizer et al. (University of Mannheim & Karlsruhe Institute of Technology)

http://webdatacommons.org/

> *"More and more websites have started to embed structured data describing products, people, organizations, places, events into their HTML pages using markup standards such as RDFa, Microdata and Microformats.*
> *The Web Data Commons project extracts this data from several billion web pages. The project provides the extracted data for download and publishes statistics about the deployment of the different formats."*

An analysis of the use of JavaScript libraries on the web, Dennis Pallett et al. (University of Twente)

https://github.com/norvigaward/naward18/wiki/Report

> *Project using Common Crawl-captured websites to discern and analyze the most common JavaScript files and libraries in use on the web.*

Geographic analysis & mapping

Geographic terms used in crawled websites can be used to plot websites or referenced places on maps. This kind of analysis can show what areas are most-discussed, or where hubs of Internet publishing exist. Geographic data from web archives can also be combined with other methods of geographic analysis.

Examples

GeoIndex, UK Web Archive (British Library)

http://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/geo

> *Dataset available for download with data about postcodes mentioned in archived websites, forming "an historical GeoIndex of the UK web."*

London French Geo-Indexing and Image Tagging, Saskia Huc-Hepher (University of Westminster)

http://domaindarkarchive.blogspot.com/2012/11/london-french-geo-indexing-and-image.html

> *Proposal to use web archives to "map out the areas of London with the greatest concentrations of French inhabitants on the basis of the post-codes associated with 'French' web sites/spaces," to support other research relating to the French population in London.*

---

2. Preservation and stability

The ability to preserve and provide long-term access to web content has long been a goal of web archiving initiatives. In their 2011 report *Web Archives: The Future(s)*, Eric T. Meyer, Arthur Thomas, and Ralph Schroeder (of the Oxford Internet Institute) discuss the importance of archived websites as an antidote to linkrot and "changing information that overwrites older versions without any way to see or revert to previous pages." Having the level of stability and accountability ensured by web archives counteracts these challenges, which can be challenges for researchers, scholars, and everyday users.

Obtaining access to prior versions of web resources is also the use of web archives that is likely most familiar to casual users. Sites that were thought lost, or older versions of still-existent sites, can be accessed through tools like the Internet Archive's Wayback Machine or web archive collections at other collecting institutions. These archived sites are also an important component of the historical record. As communication and information-sharing increasingly occurs online, capturing websites including social media channels means ensuring the preservation of the historic events (and day-to-day life) that they depict.

Persistent linking

Because websites and other digital content can disappear or change without warning, web archives offer the opportunity for users to provide links to specific, stable versions of the content in question. This may be achieved either with formal persistent identifiers assigned to each resource, or by means of a consistent and stable URL structure for accessing resources. Having access to a known version of a website at a specific location allows for users to reference this content and access it knowing precisely which version of the website is being used.

Examples

Collection Overview, Library of Congress Web Archive

http://lcweb2.loc.gov/diglib/lcwa/html/ss/ss-overview.html

> *Resources in the Library of Congress Web Archive are assigned a unique Citation ID, which redirects to the location of the archived website. These IDs ensure that even if the archive's standard URL structure should change, cited websites will still be able to be located.*

How do I cite Wayback Machine URLs in MLA format? (Internet Archive)

http://archive.org/about/faqs.php#265

The Internet Archive gives specific information about how best to cite archived websites in publication, based on the MLA citation standard.

Access to deleted or modified content

Geocities.com as captured on April 13, 2001,
http://wayback.archive.org/web/20010413160638/http://geocities.yahoo.com/home/

Web archives can provide access to sites that have since been deleted or changed, so that users can specifically access material that they are no longer able to access on the live web. This is perhaps the use of web archives most familiar to casual users, as the Internet Archive's Wayback Machine and similar services make the contents of web archives easily accessible based on URL and date.

Examples

Memento project

http://www.mementoweb.org/guide/quick-intro/

Tool allowing users to capture and access previous versions of websites.

GeoCities Special Collection 2009

http://archive.org/web/geocities.php

Internet Archive collection of Geocities sites captured prior to its 2009 shutdown.

Exploring the lost web, UK Web Archive (British Library)

http://britishlibrary.typepad.co.uk/webarchive/2012/10/exploring-the-lost-web.html

Blog post about preserving at-risk web content to prevent it from disappearing.

Accountability

Crawling websites over time allows for modifications to content to be observed and analyzed. This type of access can be useful in ensuring accountability and visibility for web content that no longer exists. On one hand, companies may archive their web content as part of records management practices or as a defense against legal action; on the other, public web archives can show changes in governmental, organizational, or individual policy or practices.

Examples

Public Health in Local Government, 2001-2012: web representations and practices, Martin Gorsky (London School of Hygiene and Tropical Medicine)

http://domaindarkarchive.blogspot.co.uk/2013/02/public-health-in-local-government-2001.html

Proposal to use web archives to analyze the web presence of public health organizations to observe changes in policies and practices.

Why Social Media Archiving Reduces Regulatory Risk, Mark Middleton (Hanzo Archives)

http://web.hanzoarchives.com/bid/90974/Why-Social-Media-Archiving-Reduces-Regulatory-Risk

Blog post describing the benefits of web archives in terms of accountability and authenticity, by providing "proof of authenticity" and "clear audit trails" for web content.

Policy for Responding to Legal Requests (Internet Archive)

http://archive.org/legal/

Legal FAQ (Internet Archive)

http://archive.org/legal/faq.php

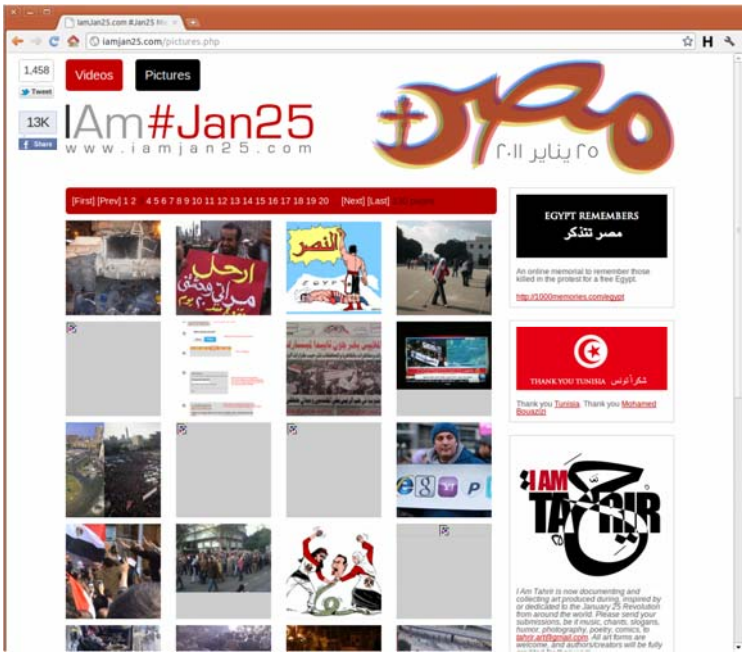
Internet Archive policy on authenticating web content found in the archive.

Blog posts from Nextpoint, Inc.

"The Four Corners of Social Media and e-Discovery," http://www.nextpoint.com/ediscovery/2714/the-four-corners-of-social-media-and-ediscovery/
"How Hard is Authenticating Social Media?," http://www.cloudpreservation.nextpoint.com/how-hard-is-authenticating-social-media/
"Can Courts be Friends with Facebook?," http://www.cloudpreservation.nextpoint.com/courts-and-facebook-cant-be-friends/
"How to Get Social Media Evidence Into Court," http://www.cloudpreservation.nextpoint.com/how-to-get-social-media-evidence-in-litigation-every-time/

Blog posts describing how web archiving can aid organizations in terms of accountability and recordkeeping, as well as issues with using web archive content as evidence in a legal context.

Historic preservation



Social media from 2011 Egyptian revolution, showing broken image links,
http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html

As more news reporting shifts onto rapidly-updating, informal social media feeds, a great deal of data relating to current events stands to be lost. Web archives provide access to this short-lived content, ensuring that these parts of the historical record will not be lost.

Examples

Japan Disaster Archives: Collaboration for successful web archiving, Lori Donovan (Archive-It)

http://blog.archive-it.org/2013/02/28/japan-disaster-archives-collaboration-for-successful-web-archiving/

Blog post about the Archive-It collection relating to the 2011 Japan earthquake and the range of websites captured as part of the effort to preserve records of that event.

End of Term Archive
http://eotarchive.cdlib.org/index.html

Collaborative web archive project collecting sites that document "the United States Government's World Wide Web presence during the transition between the administrations of President George W. Bush and President Barack Obama."

Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?, Hany M. SalahEldeen and Michael L. Nelson (Old Dominion University)

http://arxiv.org/abs/1209.3026

2012-02-11: Losing My Revolution, Web Science and Digital Libraries Research Group (Old Dominion University)

http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html

> *Blog post and journal article about work examining the quantity of lost web content relating to the 2011 Egyptian revolution. Overall, it was found that more of 10% of content was lost after a year, despite efforts to preserve it.*

Web Archiving and Special Collections: The Case of the Latin American Government Documents Archive, Trevor Owens (Library of Congress)

http://blogs.loc.gov/digitalpreservation/2012/06/web-archiving-and-mainstreaming-special-collections-the-case-of-the-latin-american-government-documents-archive/

> *Blog post describing changes in government websites during and after 2006 coup in Honduras.*

---

3. Outreach and education

Because the web is such a dynamic and visible part of human culture in the 21st century, it has naturally become an integral part of the services offered by educational and cultural heritage institutions. In a few instances, web archives have been used in physical museum exhibits and as components of online exhibits; there are also efforts to involve schoolchildren in the creation of web archive collections in order to engage them with history and highlight the importance of capturing this valuable resource.

Taking the web archive off the virtual shelf – presenting archived websites in a library exhibition, Maxine Fisher (State Library of Queensland)

http://blogs.nla.gov.au/australias-web-archives/2012/10/11/taking-the-web-archive-off-the-virtual-shelf-presenting-archived-websites-in-a-library-exhibition/

> *Blog post about the integration of archived websites in museum exhibits.*

K-12 Web Archiving (Archive-It & Library of Congress)

http://www.archive-it.org/k12/

> *Initiative in which 10 K-12 schools use Archive-It to select and capture web content for archiving.*