Social Media Archiving at the Smithsonian

MARAC 2013
April 26, 2013
Lynda Schmitz Fuhrig
Smithsonian Institution Archives
Smithsonian’s First Home Page, 1995
SI–Q: What has given us water from Mars and daggers from India?
Facebook, Twitter, YouTube … Oh My

Growth of Social Media Accounts at the Smithsonian Institution

- Facebook
- Twitter
- YouTube

Oct. 2010
Sept. 2012
Apr. 2013
Smithsonian Performance Dashboard

Social Media

- 1,167,672 Facebook User Likes
- 1,634,734 Twitter Followers
- 25,018,387 YouTube Views

The compiled total taken from the two most prominent social media sites, Facebook and Twitter, of:
- The number of Facebook users who “liked” the Smithsonian units
- The total number of Twitter “followers” who are following the Smithsonian units via their (Smithsonian’s) tweets

Facebook follower count is likely to be lower than the Twitter follower count because Facebook requires account profiles to represent actual people and each person can only have one account. Twitter has no such restriction.
OWNEYtheDOG
@OWNEYtheDOG
I rode Railway Mail trains in the 1890s and prevented train wrecks. Now I'm at @PostalMuseum. I love #NationalTrainDay.
Washington, DC  http://www.npm.si.edu/owney

Follow OWNEYtheDOG
Full name
Email
Password
Have an account? Sign in.
Sign up

Tweets
OWNEYtheDOG @OWNEYtheDOG
Ahh!
Collapse ✅ Reply ✕ Retweet ✳ Favorite
8:50 PM - 24 Jun 12 via Echofon · Details

Libby Norcross @libbydoodle
@OWNEYtheDOG Woof!
Expand

John Ratnaswamy @JohnRSports
@libbydoodle @OWNEYtheDOG Bow wow!
Expand

Libby Norcross @libbydoodle
@JohnRSports @OWNEYtheDOG ruff!
Expand

John Ratnaswamy @JohnRSports
@libbydoodle @OWNEYtheDOG Yip, yip!
Expand

OWNEYtheDOG @OWNEYtheDOG
Can dogs tell time? Possibly not. In any case, Pontiac, IL gave me a watch fob on today's date, 1895. ow.ly/b4ppm
Expand

OWNEYtheDOG @OWNEYtheDOG
OW @MalindaAnn Hill Aw! Tell her she's the best.
View conversation
Can’t Save Everything

- More than 8 billion web pages
- Social media use jumped 356% in U.S. since 2006
- 100,000 tweets sent every minute
- 48 hours of video uploaded to YouTube every minute
- 152 million blogs on the web

<table>
<thead>
<tr>
<th>Service</th>
<th>Homepage</th>
<th>Person</th>
<th>Date</th>
<th>Platform</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smithsonian Institution Traveling Exhibition Service</td>
<td><a href="http://sites.si.edu/">http://sites.si.edu/</a></td>
<td>Heather Foster</td>
<td>Yes</td>
<td>6/20/2011</td>
<td>Heritrix</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cheryl</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.shoesigo.si.edu/">http://www.shoesigo.si.edu/</a></td>
<td></td>
<td>Yes</td>
<td>7/6/2011</td>
<td>Heritrix</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.blog.museumonmainstreet.org/">http://www.blog.museumonmainstreet.org/</a></td>
<td>Museum on Main Street - Road Reports Blog</td>
<td>Yes</td>
<td>7/6/2011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Twitter</td>
<td></td>
<td>9/30/2010</td>
<td>Searchastic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YouTube</td>
<td>Yes</td>
<td>09/30/2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>03/08/2012, 03/02/2011</td>
<td>PCF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>06/14/2012</td>
<td>Heritrix</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>03/12/2012, 03/05/2012</td>
<td>PCF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>09/30/2010</td>
<td></td>
</tr>
</tbody>
</table>
Past Web Archiving Procedures

- Files transferred from central IT
- Files transferred by webmaster
- HTTrack web crawler
- Scripts used to create XHTML preservation files but very manual and time-consuming
Crawling in Heritrix

Status as of Jul. 27, 2011 15:54:18 GMT  Alerts: no alerts

Admin Console  0 jobs pending, 106 completed  1406 URLs in 51m28s (0.45/sec)

Crawler Status: CRAWLING JOBS Hold

Jobs
- Running: marine-science-pearce_rcvr_site-no-ex
- 0 pending, 106 completed
- Alerts: 0 (0 new)

Rate:
- 0.45 URLs/sec (0.46 avg)
- 7 KB/sec (13 avg)

Time
- 51m28s elapsed
- 10m41s remaining (estimated)

Totals
- downloaded 1406
- 82% complete
- 292 queued
- 1698 total downloaded and queued
- 41 MB crawled (41 MB novel)

Memory
- 37553 KB used
- 48320 KB current heap
- 260160 KB max heap

Load
- 0 active of 5 threads
- 1 congestion ratio
- 292 deepest queue
- 292 average depth

Refresh

Identifier: org.archive.crawler.jmxport=8849.name=Heritrix.type=CrawlService_gui.port=8080.host=sia-websites
Social Media

- Appraise each social media account separately
  - Accounts are used differently, even in the same application
- Appraise often
  - Accounts may be used differently over time
- Some accounts include significant original content, others are primarily a vehicle for sharing links
Terms of Service

- Read carefully
- Change frequently
- Some forbid crawling of sites
Interactive Features and User-Generated Content

The Smithsonian provides certain interactive features, such as blogs, for the exchange of ideas and information by visitors in furtherance of its mission. Accordingly, the Smithsonian encourages you to interact with its sites subject to the following conditions:

- You are solely responsible for the content that you submit.
- By posting content, you are giving the Smithsonian and those authorized by the Smithsonian permission to use it for any educational, promotional or other standard museum purpose.
- You will not post any content that is off-topic, partisan-political, contains personal attacks or expletives, or is otherwise abusive, threatening, unlawful, harassing, discriminatory, libelous, obscene, indecent, pornographic or that infringes on the rights of any third party.
- You will only post content that is yours or for which you have received permission from the person or entity that owns the content and the related intellectual property rights.
- You will not insert your own or a third party’s advertising or promotional content into your posting.
- You understand that the Smithsonian may monitor or review discussions, chats, postings, transmissions, bulletin boards or any other postings from time to time.
- You understand that the Smithsonian assumes no responsibility or liability arising from the posting of your content or for any error, defamation, libel, omission, obscenity, danger or inaccuracy contained in any posted information or from any failure or delay in removing such content.
- You will not introduce any software viruses, worms or other programs designed to damage software, hardware or telecommunications equipment.
- You agree that you will not collect information about the users of this website or use such information for any purpose.
- You are not under the age of 13.

The Smithsonian reserves the right to monitor content or information provided by website visitors in its sole discretion, to edit or remove any content that, in the Smithsonian’s judgment, does not comply with the foregoing rules. The Smithsonian is not responsible for any user-generated content that is posted on the SI Websites or for any failure or delay in removing such content.

museums, research centers, and the national zoo. http://www.si.edu/

General Information

Welcome to our Facebook page! Please feel free to share thoughts about our posts, ask us questions, or tell us about your visit.

We hope you’ll contribute to this interactive forum and to our ongoing conversation about the work we do to further the Smithsonian’s mission to increase and diffuse knowledge.

While on-topic discussion is encouraged, we ask that you express yourself in a civil manner and treat other users with respect. The Smithsonian also monitors may remove posts consistent with its terms of use, as described at http://si.edu/TermsOfUse#user-gen.
How to Save

- Crawling is not always effective
  - Had worked in-house with some Twitter accounts in November 2011 and then Twitter made changes

- Some providers have their own export tool
  - Usually need account password
  - May be proprietary
  - May not export everything

- Third-party export tools exist for some account types
  - Often export as spreadsheet or XML
  - Service could easily disappear

- Screenshots are laborious but may serve as a last resort
  - Can document “look and feel”
Archiving tools

- Facebook do it yourself within personal account
- ArchiveFacebook (Firefox plugin)
- HootSuite/TwapperKeeper
- The Archivist, SnapBird, ThinkUp, SocialBro
- Recollect.com and Tweetbook (PDF e-book)
OWNEytheDOG
Jan 13, 4:39pm via HochSuite
Employees in Nashville invented the “tagger-stacker machine” to stack up mail before it went into the next
#mailmachine.

OWNEytheDOG
Jan 13, 4:42pm via HochSuite
#artemiscolland WA!

OWNEytheDOG
Jan 13, 4:46pm via HochSuite
Art RT @leelandemond: #FF for museum mascots:
@SabraSaberboth @Lupus_The_Hippo @SplotchusNH
@OWNEytheDOG @DiosTheDeer @Giga_a_dino

OWNEytheDOG
Jan 13, 4:47pm via HochSuite
@Iwa You guys wear tags just like me!

OWNEytheDOG
Jan 13, 4:47pm via HochSuite
@marleytiner Happy Friday to you, too!

OWNEytheDOG
Jan 13, 4:47pm via HochSuite
Nice! RT @mazzie: Hanging with two of my favorite dogs:
@the_big_dog and @OWNEytheDOG http://at.com/7MND478
Developed by Graz University of Technology in Austria as research tool
Grabeeter
Grab and Search your Tweets

Search Tweets from Smithsonian

Enter your search query (web, elearning, ...):

3396 available Tweets | Page 1 of 136

1 2 3 4 5

smithsonian @shannonnnmartin You might also love their twitter!
@amhistorymuseum

Wed, 18 Apr 2012 18:54:29

smithsonian @rmiofmaryland You can follow them @3D_Digi_SI!

Wed, 18 Apr 2012 17:02:22

smithsonian For dentist, the narwhal’s smile is a mystery of evolution

Smithsonian
We’re not a museum. We’re 19 of them! Plus 9 research centers & the National Zoo. Follow your favorites

- Twitter User: smithsonian
- Tweets on Twitter: 5843
- Indexed Tweets: 3396 (58 %)
- Links: 2439 (72 %)
- Friends: 127
- Followers: 587439
- Url: http://si.edu/Connect/twitter

Statistics
Twitter Archive
Archive-It Crawl of Twitter

You are viewing an archived web page, collected at the request of Smithsonian Institution using Archive-It. This page was captured on 20:48 14 Jan 14, 2013, and is part of the Smithsonian Institution Social Media collection. The information on this web page may be out of date. See All versions of this archived page.
WARC – Web ARChive file format

- International standard – ISO 28500:2009
- Extension of the ARC format in use since 1996
- Container format
Inside a WARC file

WARC/0.18
WARC-Type: response
WARC-Target-URI: http://www.smithsoniansociety.org/
WARC-Date: 2012-05-08T21:04:11Z
WARC-Payload-Digest: sha1:1f0NBVLJVK5WDZACI0ZCT6TRQGDGDA
WARC-IP-Address: 160.111.252.228
WARC-Record-ID: <urn:uuid:29f9ab6a-ae33-4763-af9a-7d058243e3bc>
Content-Type: application/http; msgtype=response
Content-Length: 12758

HTTP/1.1 200 OK
Content-Length: 12456
Content-Type: text/html
Content-Language: http://www.smithsoniansociety.org/index.html
Last-Modified: Tue, 20 Mar 2012 17:22:48 GMT
Accept-Ranges: bytes
ETag: "0e4c12be6cd1:4877"
X-Powered-By: ASP.NET
Date: Tue, 08 May 2012 19:49:41 GMT
Connection: close

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">!-- InstanceBegin template="/Templates/jss.dwt"
codeOutsideHTMLIsLocked="false" -->

<head>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<!--[if IE 6]><link rel="stylesheet" href="_css/screen.ie6.css" type="text/css" media="screen, projection" />
<link rel="stylesheet" href="_css/print.ie6.css" type="text/css" media="print" />
<![endif]-->
<!--[if IE 7]><link rel="stylesheet" href="_css/screen.ie7.css" type="text/css" media="screen, projection" />
<link rel="stylesheet" href="_css/print.ie7.css" type="text/css" media="print" />
<![endif]-->
<script type="text/javascript" src="js/swindow.js"></script>
<script type="text/javascript" src="js/jquery.js"></script>
<script type="text/javascript" src="js/jquery.dimensions.min.js"></script>
<!--[if IE 8]><script type="text/javascript" src="js/jquery草案.js"></script>-->
<script type="text/javascript">

</script>
Facebook Blast from the Past

Smithsonian in your Neighborhood | Facebook

Sign Up Facebook helps you connect and share with the people in your life.

Smithsonian in your Neighborhood

Museum/Art Gallery · Washington, DC

Wall

Everyone · Smithsonian in your Neigh...

Smithsonian in your Neighborhood

The Smithsonian Institution, MIT, and 17 Smithsonian Affiliates— including Aerospace Museum of California, Museum of the Rockies, and Kansas Cosmosphere and Space Center—collaborate to help kids decode clues in the interactive game VANISHED.

Smithsonian and MIT Partner to Turn Kids into Scientific Investigators
www.artdaily.org

B WASHINGTON, DC/8 The Smithsonian Institution and MIT announced the April 4 launch of VANISHED, an 8-week online/offline environmental disaster
The Reddit AMA link is live! Marc Fries answers your questions about flaming rocks from space beginning at 1:30.

IANA Smithsonian meteorite scientist. Ask me anything! : IANA

Hi there. I'm an expert in meteors and meteorites — technically, a research associate in mineralogy at the Smithsonian's National Museum of Natural History.

Today’s dramatic meteorite strike in Russia raises a lot of questions. We have the answers! Join us for our first live Reddit AMA (link to come.)

Keep watching the skies! (But don’t forget to duck)

www.e-torch.org

Feb 15: Keep watching the skies! (But don’t forget to duck) In a screen grab from video taken by a dashboard camera, a streak could be seen in the sky on a highway from Kostinbay, Kazakhstan to the Chelebinsk region in Russia. (www.ngt.kr, via Associated Press) A meteorite plummeted to Earth in western S...
Keep an Eye Out

WAIL
Web Archiving Integration Layer
One-Click User Instigated Preservation

What Is It?
Web Archiving Integration Layer (WAIL) is a graphical user interface (GUI) atop multiple web archiving tools intended to be used as an easy way for anyone to preserve and replay web pages. Tools included and accessible through the GUI are Heritrix 5.1.0, Wayback 1.6, and warc-proxy. Support packages include Apache Tomcat, phantoms, and pyinstaller.

WAIL is written mostly in Python and a small amount of JavaScript.

Screenshots
WARCreate

Google Chrome Extension

“Create WARC files from any webpage”

What is it?
WARCreate is a Google Chrome extension that allows a user to create a Web-ARCHive (WARC) file from any browsable webpage. The resulting files can then be used with other tools like the Internet Archive’s open source Wayback Machine. The tool is an evolving product with the end result pushing toward being a personal web archiving solution for those that wish to securely archive their metadata in a standardized way.

How Do I Install It?
Simply download the Google Chrome extension (.CRX) file and open the extension. Chrome should load it after confirmation but if not, just drag the file to a Chrome window and it should be handler properly.

What Are Its Features? What Features Will It Have?
The current goal of the tool is to allow a user to create WARC file from any page in a web browser. Future features include:

* Encoding/Encryption Support for archive content privacy
* WARC merging - allow multiple archiving sessions to exist in a single WARC file automatically
* Memento integration

Where Can I Download It?
WARCreate is currently in a very heavy development cycle from all of the suggestions given at JCOL. Our 1.0 publicly available release will be available soon. Please keep an eye on this page, the Web Science and Digital Libraries Research Group’s blog and/or e-mail Matt Kelly to be notified when the first feature complete version has been released.

I Found A Bug! Where Do I Report It?
The "Web Archiving Bucket" is an initiative launched by Aleph Archives, to preserve data and provide libraries and organizations with free-to-use web archiving tools and components.

This initiative was launched after studying the needs in data preservation as well as the challenges encountered during not only the harvesting process but also the quality of the archives and its access time.

The Web Archiving Bucket provides a set of tools to help archivists and professionals in their daily work.

**WSDK: Warc Software Development Kit**

WARC Software Development Kit (WSDK) represents a set of simple, compact, and highly optimized Erlang modules to manipulate the WARC ISO 28500:2006 file format.

WSDK allows you to quickly build robust Web Archiving softwares in no time.
Other Things to Consider

- Mobile websites - Some you can crawl
- Apps – TourML/screenshots
TourML was created to serve as a simple and portable representation for tour content. Having a common language for this content will enable content management systems, kiosks, and mobile applications to speak a common language. Additionally, users of TourML will be better positioned to share and sustain their content no matter how the technology that we use to deliver these tours changes.

LEARN MORE

TAP is a collection of free and open-source tools which support the creation and delivery of mobile tours. The tools also serve as examples of producing and consuming tour content using the TourML specification. Currently TAP consists of authoring tools built on top of the content management system Drupal, a native iOS mobile application, and a web-based mobile application built upon the jQuery Mobile library.

GET STARTED

What's on TAP?

TourML v1.0.0
Submitted by admin on Thu, 09/06/2012 - 11:56
The TourML specification has withstood a great number of tour applications, ranging from the quick and dirty to deep, complex narratives. Given the stability and maturity it has shown we have now released version 1.0.0.

Read more

TAP CMS beta3 Released
Submitted by admin on Fri, 05/18/2012 - 15:07
We are pleased to announce the third beta release of the TAP Content Management System. This release fixes a bug with the strongarm module.

Read more

TAP CMS beta2 released
Submitted by admin on Tue, 05/08/2012 - 12:16
We are pleased to announce the second beta release of the TAP Content Management System.

Read more
C. Maxx Stevens (Seminole/Muscogee) is a visual storyteller whose deeply personal, eclectic constructions tell stories about places and people from her past. This audio guide presents the artist's perspective on selected sculpture, installation and prints in this solo exhibition, House of Memory.

Stop Group: Wichita

Transcript
C. Maxx Stevens: House of Memory

Stop Group: Wichita

Transcript
C. Maxx Stevens: House of Memory

Stop Group: Wichita

Transcript
C. Maxx Stevens: House of Memory

Stop Group: Wichita

Transcript
C. Maxx Stevens: House of Memory
Do Something

SERIOUSLY AMAZING

Explore | Visit | Get Involved | Connect | Educators | Kids | Researchers | Volunteers | Members | Fellows Interns

Home | Museums and Zoo | Research Centers | Cultural Centers | Exhibitions | Events | Collections | Newsroom

Connect with the Smithsonian
The Smithsonian goes social! Participate with your Smithsonian through social networking, gaming, mobile, and more.

Smithsonian
Ask Smithsonian
Want answers? Just Ask

SMITHSONIAN E-NEWS
Latest updates on exhibitions, research, and events.
SIGN UP TODAY

Apps for Aged Smithsonian Orangutans using iPads for Enrichment

WAYS TO CONNECT

Blogs
Facebook
Flickr
Mobile
Pinterest
Podcasts
Twitter
YouTube
Virtual World

Smithsonian Seriously Amazing

Questions come alive at the Smithsonian.

The Smithsonian asks and answers questions about science, art, history and culture every day.
Connect with the curious at seriouslyamazing.si.edu »