End of Term Archive – 2008, 2012, 2016

Mark Phillips, Associate Dean for Digital Libraries, UNT Libraries
@vphill | mark.phillips@unt.edu

May19, 2017

it all began a long, long, time ago, in a far away

https://flic.kr/p/4N2jHU

https://flic.kr/p/4JNkLE

original end of term web archive partners

INTERNET ARCHIVE

LIBRARY OF CONGRESS

University of California
CDL
California Digital Library

NORTH TEXAS

GPO

for 2008/2012/2016 - all IIPC & NDIIPP/NDSA partners

extant gov web archiving efforts

Capture, Preservation, & Access

- LOC: .gov, election, other
- GPO: agency sites, often ephemeral
- NARA: congressional web harvest every 2 years
- IA: global & curated crawls
- Agency-level: NIH/NLM, DOE, DOL, HHS, CMS, others, using AIT or comm tools
- UNT & Others: Topical .gov collecting

Community Efforts

- Federal Web Archiving Group
    - most of those at left plus other feds
- Research Initiatives
    - academic
    - NGO or watchdog
- Citizen Driven
    - grassroots efforts
- End of Term
    - focused but large-scale multi-institutional project

goals of the end of term project

UNITED STATES CENTRAL COMMAND
Petraeus pays visit to troops in Anbar
United States Central Command – Sep. 16, 2008

United States Central Command
Sept 16, 2008

U.S. Department of State Official Blog
Feb 13, 2013

Healthcare.gov Twitter
Feb 15, 2013

□ work collaboratively to preserve public U.S. Government websites

□ document federal agencies' presence on the web during the end of Presidential terms

□ enhance the existing research collections of the partner institutions

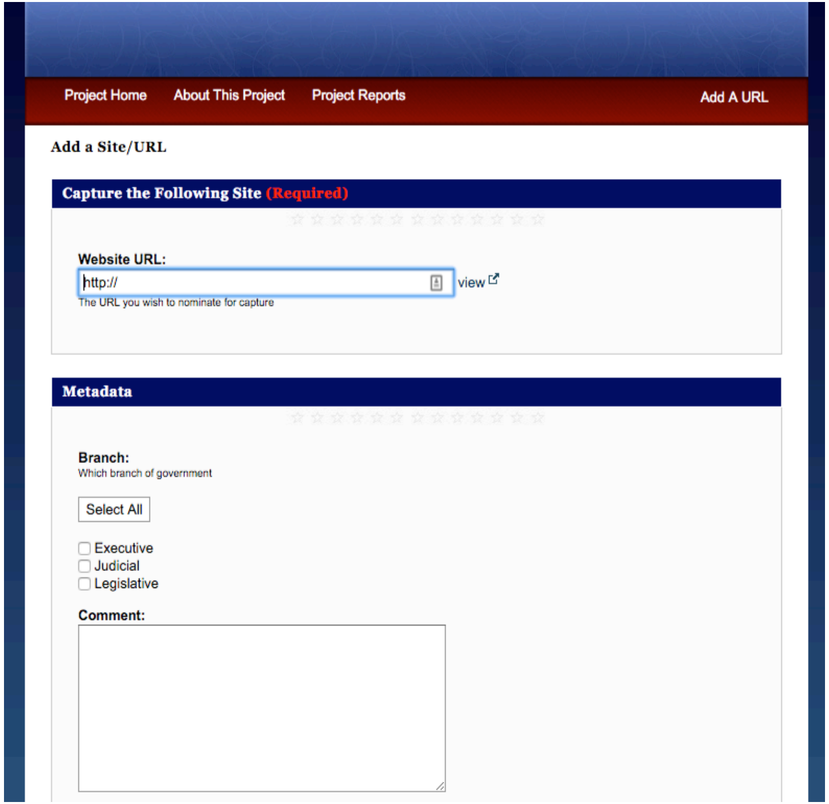
□ raise awareness about the need for preservation

□ engage with researchers and subject experts

eot collaborative distribution of work

- IA: crawling, preservation, access, full-text search
- LC: crawling, preservation, data transfers
- UNT: nomination tool development, crawling, nomination mgmt, preservation, access
- CDL: web portal, metadata
- GPO: URL nomination, outreach
- All: URL contributions, outreach, project management
- Others: URLs, education

Some variance of roles between 2008 & 2012 (and for 2016)



Add a Site/URL

Capture the Following Site (Required)

Website URL:
http://
The URL you wish to nominate for capture

Metadata

Branch:
Which branch of government

□ Executive
□ Judicial
□ Legislative

Comment:

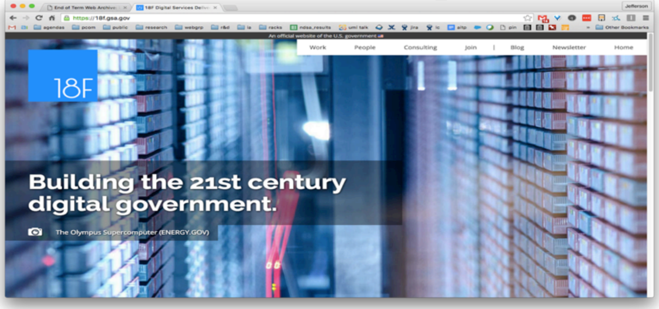
major funding brought to you by….

https://flic.kr/p/8uMXjb

no one

https://flic.kr/p/8uMXib

defining the “government web presence”



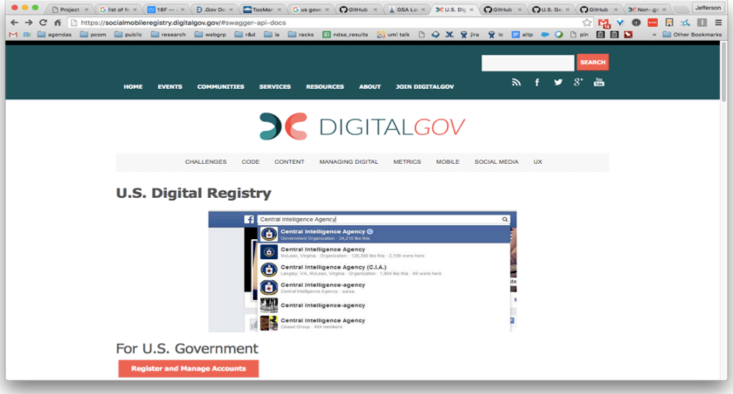
https://18f.gsa.gov

An official website of the U.S. government

Building the 21st century digital government.

The Olympus Supercomputer (ENERGY.GOV)

USA.gov



https://socialmobileregistry.digitalgov.gov/#swagger-api-docs

U.S. Digital Registry

Central Intelligence Agency

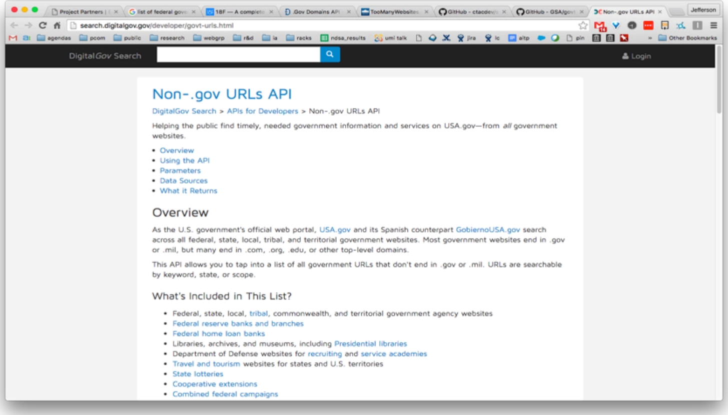
- Central Intelligence Agency
- Central Intelligence Agency
- Central Intelligence Agency (C.I.A.)
- Central Intelligence-agency
- Central Intelligence-agency
- Central Intelligence Agency

For U.S. Government

Register and Manage Accounts

Stanford WebBase Project

DATA.GOV



search.digitalgov.gov/developer/govt-urls.html

Non-.gov URLs API

DigitalGov Search > APIs for Developers > Non-.gov URLs API

Helping the public find timely, needed government information and services on USA.gov—from *all* government websites.

- Overview
- Using the API
- Parameters
- Data Sources
- What it Returns

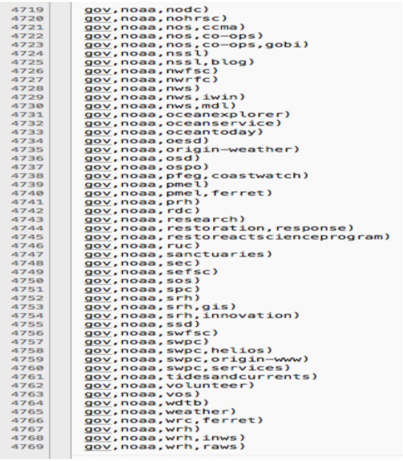
Overview

As the U.S. government's official web portal, USA.gov and its Spanish counterpart GobiernoUSA.gov search across all federal, state, local, tribal, and territorial government websites. Most government websites end in .gov or .mil, but many end in .com, .org, .edu, or other top-level domains.

This API allows you to tap into a list of all government URLs that don't end in .gov or .mil. URLs are searchable by keyword, state, or scope.

What's Included in This List?

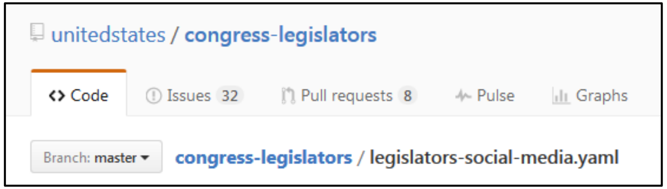
- Federal, state, local, tribal, commonwealth, and territorial government agency websites
- Federal reserve banks and branches
- Federal home loan banks
- Libraries, archives, and museums, including Presidential libraries
- Department of Defense websites for recruiting and service academies
- Travel and tourism websites for states and U.S. territories
- State lotteries
- Cooperative extensions
- Combined federal campaigns



```
gov,noaa,nodc)
gov,noaa,nohrsc)
gov,noaa,nos,ccma)
gov,noaa,nos,co-ops)
gov,noaa,nos,co-ops,gobi)
gov,noaa,nssl)
gov,noaa,nssl,blog)
gov,noaa,nwfsc)
gov,noaa,nwrfc)
gov,noaa,nws)
gov,noaa,nws,iwin)
gov,noaa,nws,mdl)
gov,noaa,oceanexplorer)
gov,noaa,oceanservice)
gov,noaa,oceantoday)
gov,noaa,oesd)
gov,noaa,origin-weather)
gov,noaa,osd)
gov,noaa,ospo)
gov,noaa,pfeg,coastwatch)
gov,noaa,pmel)
gov,noaa,pmel,ferret)
gov,noaa,prh)
gov,noaa,rdc)
gov,noaa,research)
gov,noaa,restoration,response)
gov,noaa,restoreactscienceprogram)
gov,noaa,ruc)
gov,noaa,sanctuaries)
gov,noaa,sec)
gov,noaa,sefsc)
gov,noaa,sos)
gov,noaa,spc)
gov,noaa,srh)
gov,noaa,srh,gis)
gov,noaa,srh,innovation)
gov,noaa,ssd)
gov,noaa,swfsc)
gov,noaa,swpc)
gov,noaa,swpc,helios)
gov,noaa,swpc,origin-www)
gov,noaa,swpc,services)
gov,noaa,tidesandcurrents)
gov,noaa,volunteer)
gov,noaa,vos)
gov,noaa,wdtb)
gov,noaa,weather)
gov,noaa,wrc,ferret)
gov,noaa,wrh)
gov,noaa,wrh,inws)
gov,noaa,wrh,raws)
```

NATIONAL ARCHIVES

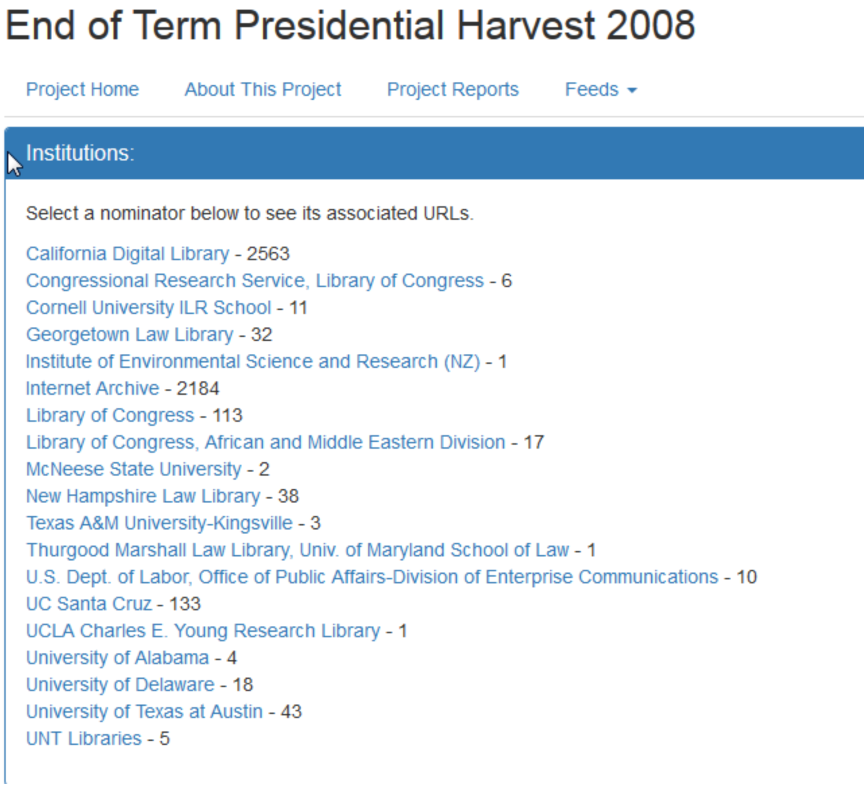
2004 crawl list of URLs



unitedstates / congress-legislators

congress-legislators / legislators-social-media.yaml

and people like you!

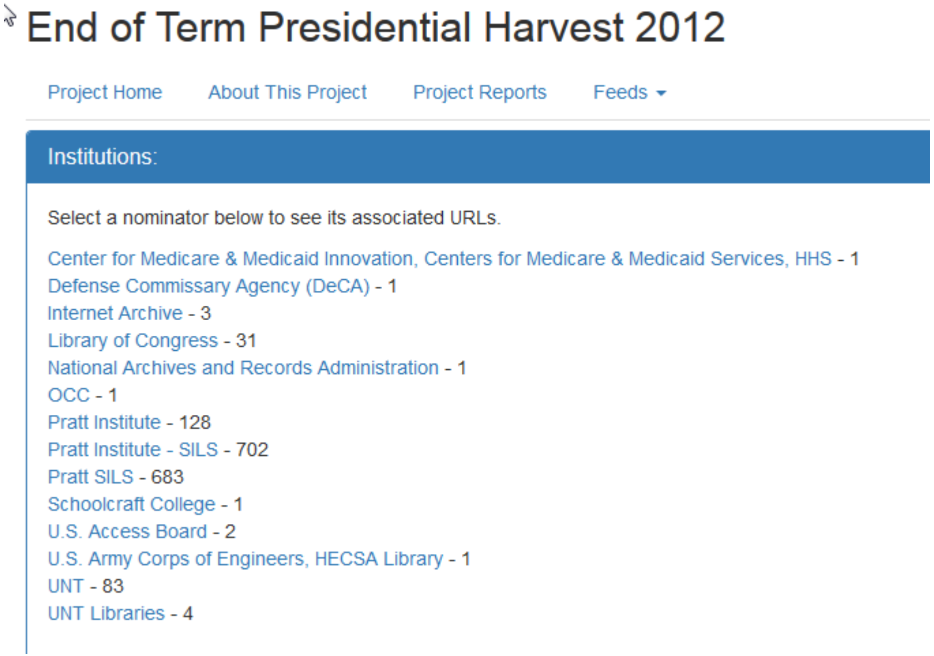


End of Term Presidential Harvest 2008

Institutions:

Select a nominator below to see its associated URLs.

California Digital Library - 2563
Congressional Research Service, Library of Congress - 6
Cornell University ILR School - 11
Georgetown Law Library - 32
Institute of Environmental Science and Research (NZ) - 1
Internet Archive - 2184
Library of Congress - 113
Library of Congress, African and Middle Eastern Division - 17
McNeese State University - 2
New Hampshire Law Library - 38
Texas A&M University-Kingsville - 3
Thurgood Marshall Law Library, Univ. of Maryland School of Law - 1
U.S. Dept. of Labor, Office of Public Affairs-Division of Enterprise Communications - 10
UC Santa Cruz - 133
UCLA Charles E. Young Research Library - 1
University of Alabama - 4
University of Delaware - 18
University of Texas at Austin - 43
UNT Libraries - 5



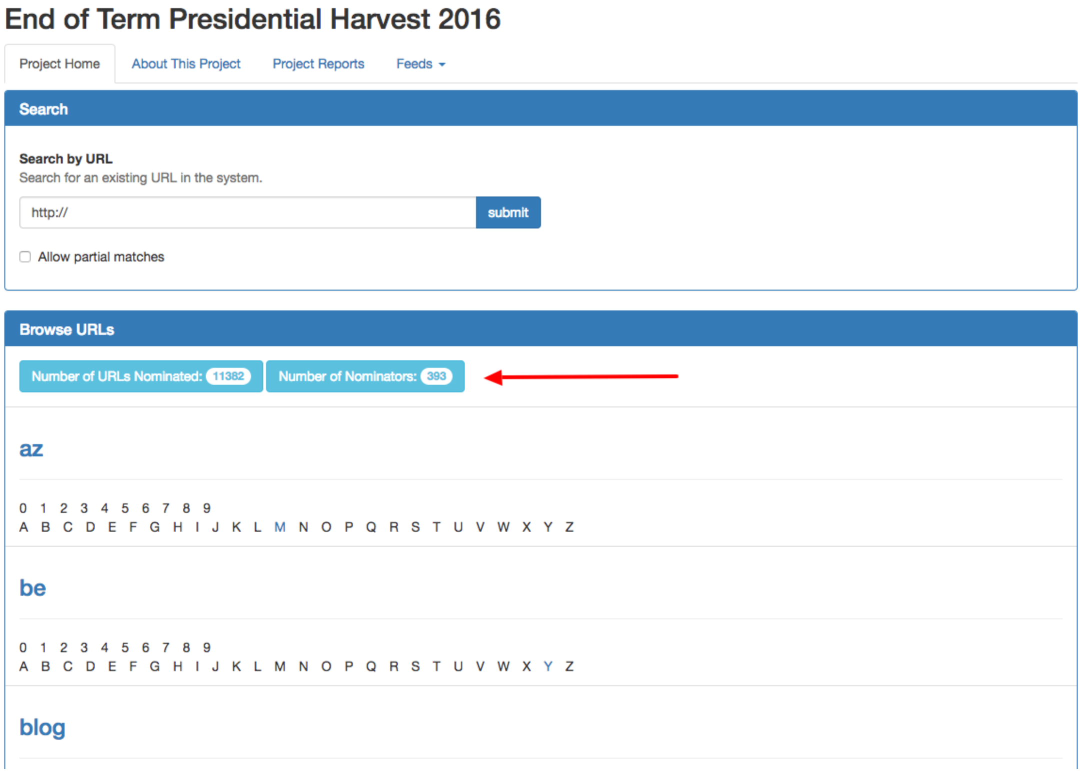
End of Term Presidential Harvest 2012

Institutions:

Select a nominator below to see its associated URLs.

Center for Medicare & Medicaid Innovation, Centers for Medicare & Medicaid Services, HHS - 1
Defense Commissary Agency (DeCA) - 1
Internet Archive - 3
Library of Congress - 31
National Archives and Records Administration - 1
OCC - 1
Pratt Institute - 128
Pratt Institute - SILS - 702
Pratt SILS - 683
Schoolcraft College - 1
U.S. Access Board - 2
U.S. Army Corps of Engineers, HECSA Library - 1
UNT - 83
UNT Libraries - 4

2016 Was Crazy



End of Term Presidential Harvest 2016

Project Home
About This Project
Project Reports
Feeds

Search

Search by URL

Search for an existing URL in the system.

http://

□ Allow partial matches

Browse URLs

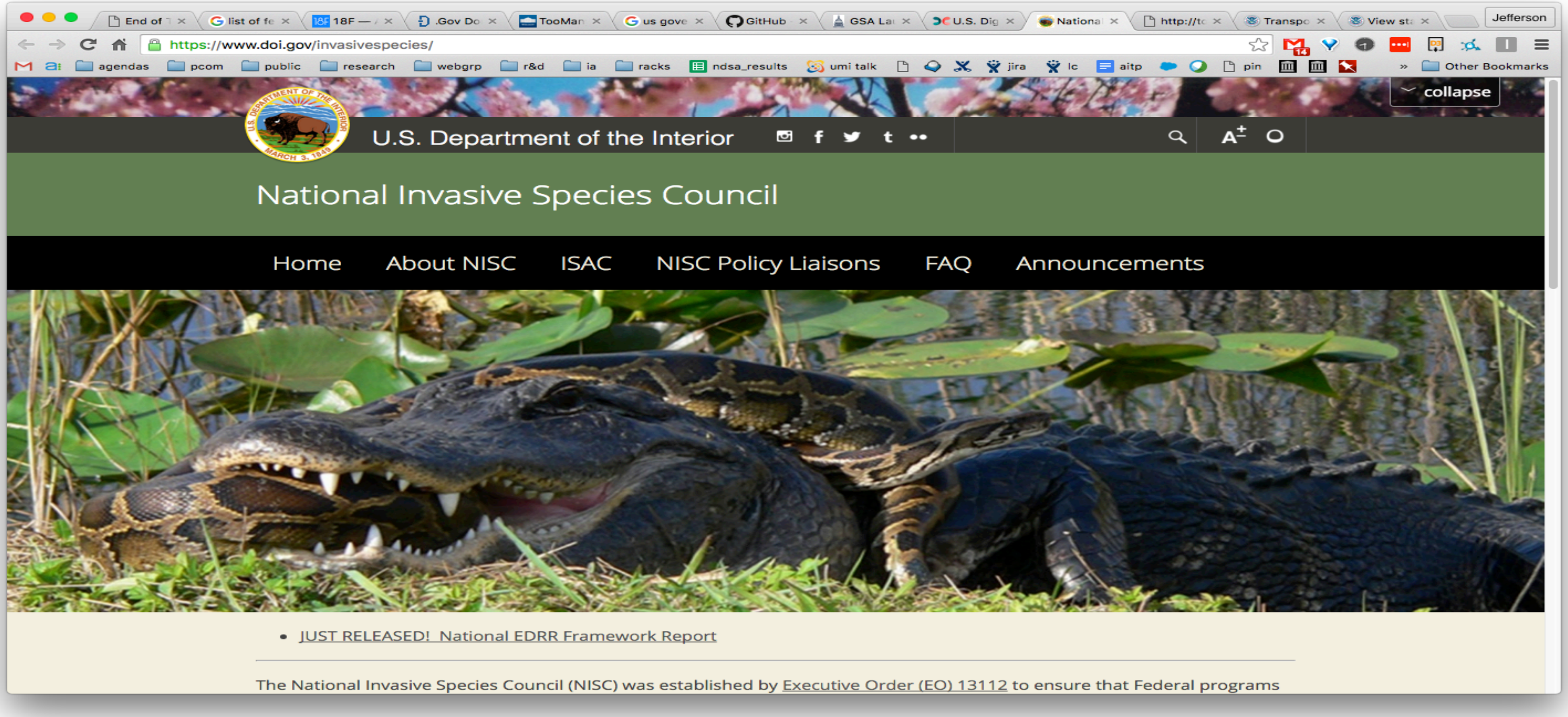
Number of URLs Nominated: 11382

Number of Nominators: 393

az
0 1 2 3 4 5 6 7 8 9
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
be
0 1 2 3 4 5 6 7 8 9
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
blog

Some challenges of the .gov Web

.gov websites proliferate like invasive species

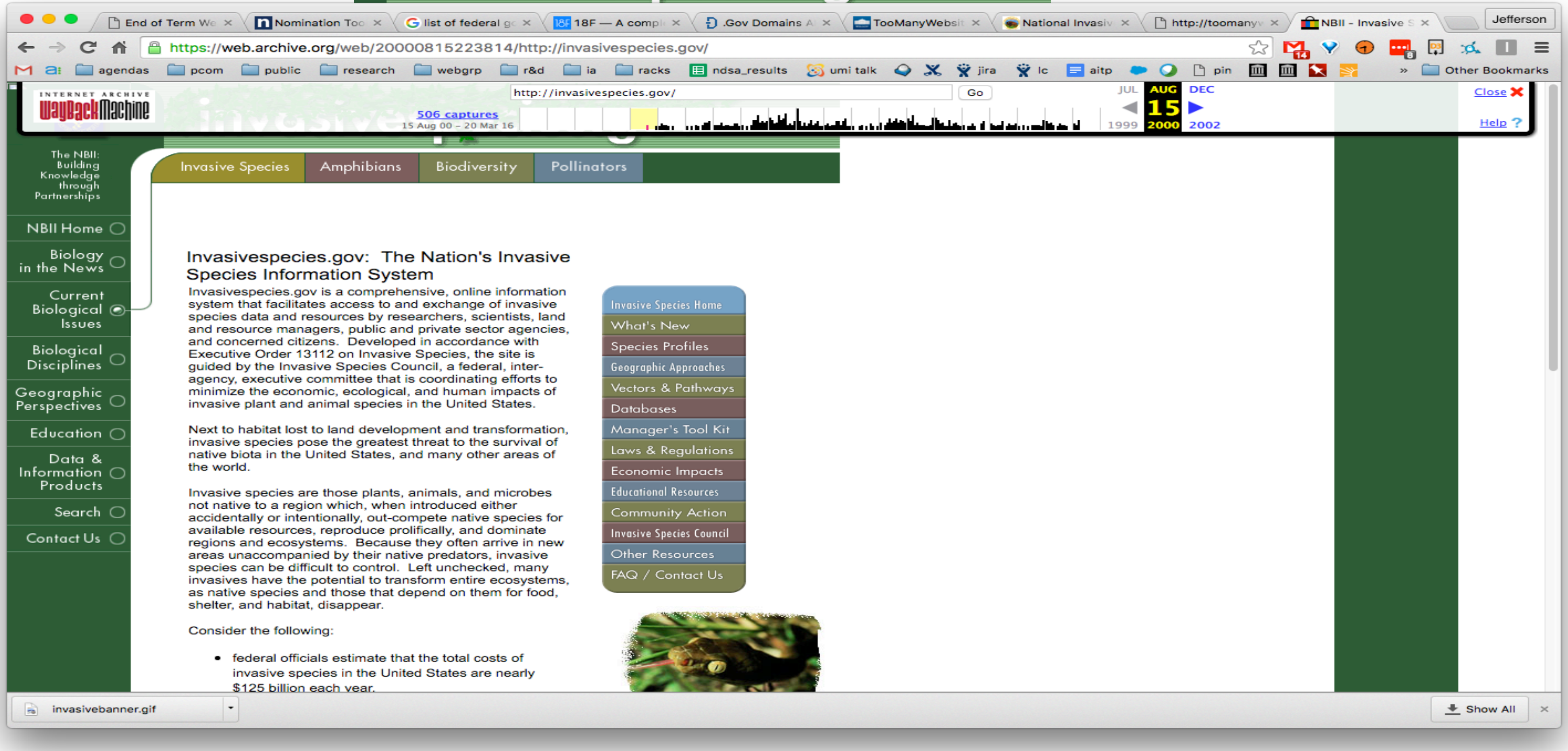


- JUST RELEASED! National EDRR Framework Report

The National Invasive Species Council (NISC) was established by Executive Order (EO) 13112 to ensure that Federal programs

and yes, invasivespecies.gov once existed

invasivespecies.gov



Invasivespecies.gov: The Nation's Invasive Species Information System

Invasivespecies.gov is a comprehensive, online information system that facilitates access to and exchange of invasive species data and resources by researchers, scientists, land and resource managers, public and private sector agencies, and concerned citizens. Developed in accordance with Executive Order 13112 on Invasive Species, the site is guided by the Invasive Species Council, a federal, inter-agency, executive committee that is coordinating efforts to minimize the economic, ecological, and human impacts of invasive plant and animal species in the United States.

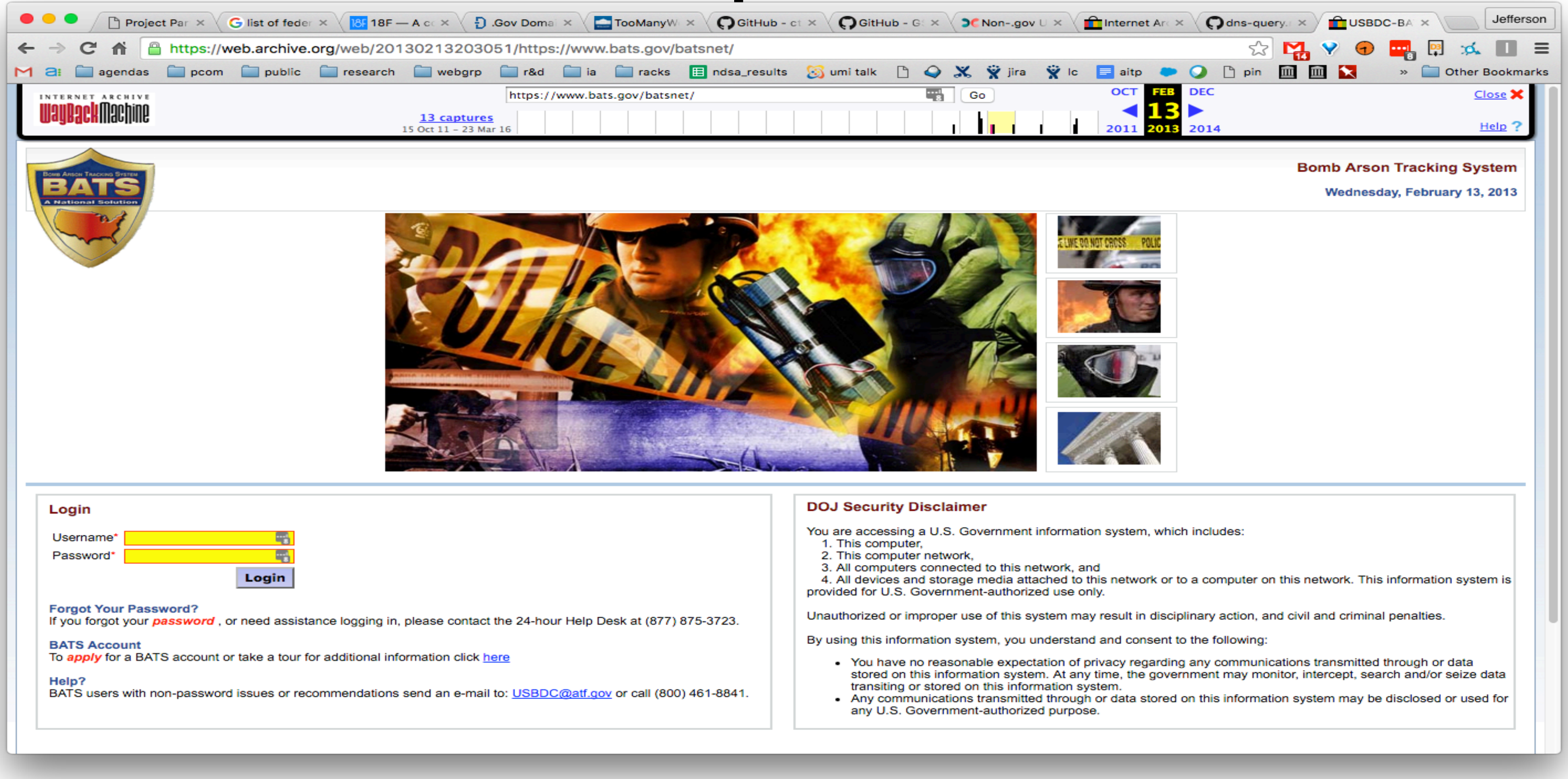
Next to habitat lost to land development and transformation, invasive species pose the greatest threat to the survival of native biota in the United States, and many other areas of the world.

Invasive species are those plants, animals, and microbes not native to a region which, when introduced either accidentally or intentionally, out-compete native species for available resources, reproduce prolifically, and dominate regions and ecosystems. Because they often arrive in new areas unaccompanied by their native predators, invasive species can be difficult to control. Left unchecked, many invasives have the potential to transform entire ecosystems, as native species and those that depend on them for food, shelter, and habitat, disappear.

Consider the following:

- federal officials estimate that the total costs of invasive species in the United States are nearly $125 billion each year.

some are non-public or unlisted



Bomb Arson Tracking System

Wednesday, February 13, 2013

Login

Username*

Password*

Login

Forgot Your Password?
If you forgot your *password* , or need assistance logging in, please contact the 24-hour Help Desk at (877) 875-3723.

BATS Account
To *apply* for a BATS account or take a tour for additional information click here

Help?
BATS users with non-password issues or recommendations send an e-mail to: USBDC@atf.gov or call (800) 461-8841.

DOJ Security Disclaimer

You are accessing a U.S. Government information system, which includes:

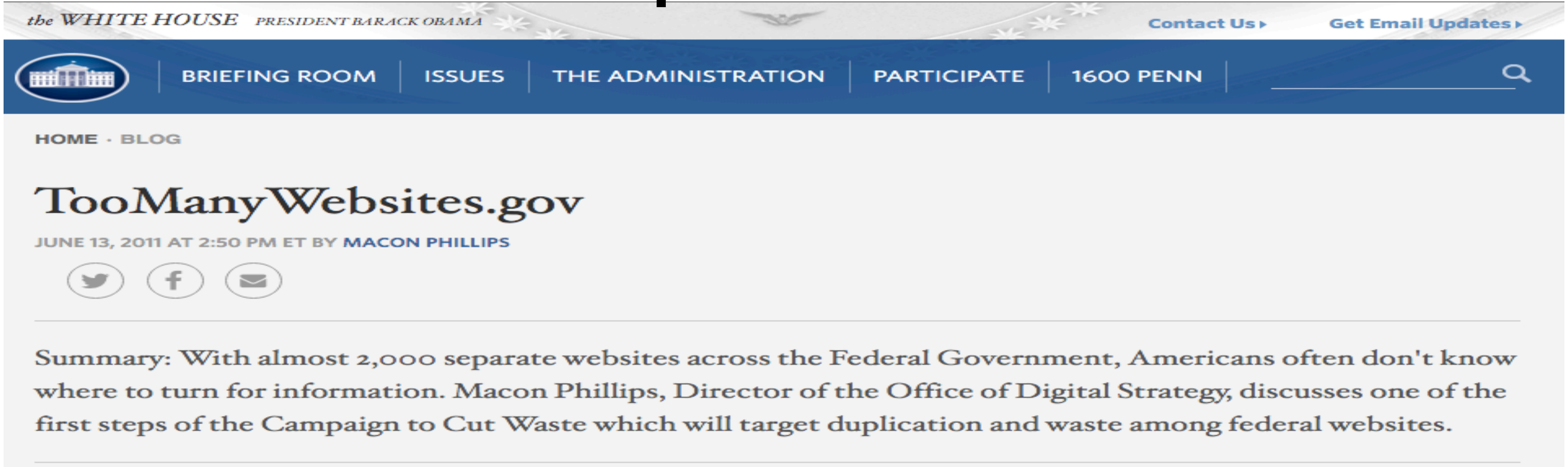
1. This computer,
2. This computer network,
3. All computers connected to this network, and
4. All devices and storage media attached to this network or to a computer on this network. This information system is provided for U.S. Government-authorized use only.

Unauthorized or improper use of this system may result in disciplinary action, and civil and criminal penalties.

By using this information system, you understand and consent to the following:

- You have no reasonable expectation of privacy regarding any communications transmitted through or data stored on this information system. At any time, the government may monitor, intercept, search and/or seize data transiting or stored on this information system.
- Any communications transmitted through or data stored on this information system may be disclosed or used for any U.S. Government-authorized purpose.

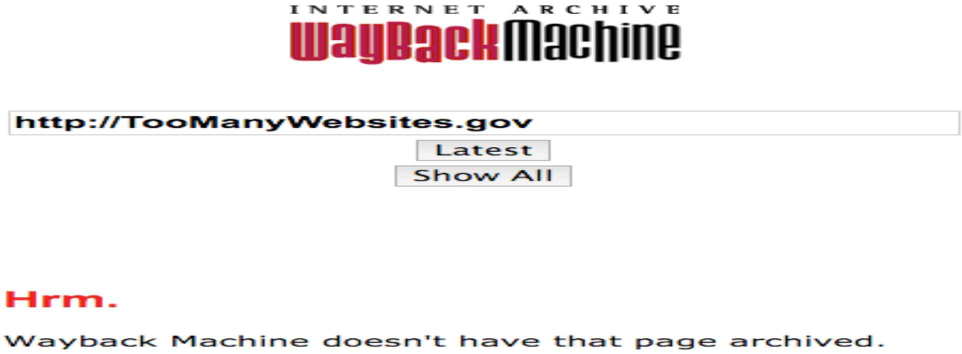
“web waste” & preservation mentalities



TooManyWebsites.gov

JUNE 13, 2011 AT 2:50 PM ET BY MACON PHILLIPS

Summary: With almost 2,000 separate websites across the Federal Government, Americans often don't know where to turn for information. Macon Phillips, Director of the Office of Digital Strategy, discusses one of the first steps of the Campaign to Cut Waste which will target duplication and waste among federal websites.



http://TooManyWebsites.gov

Hrm.

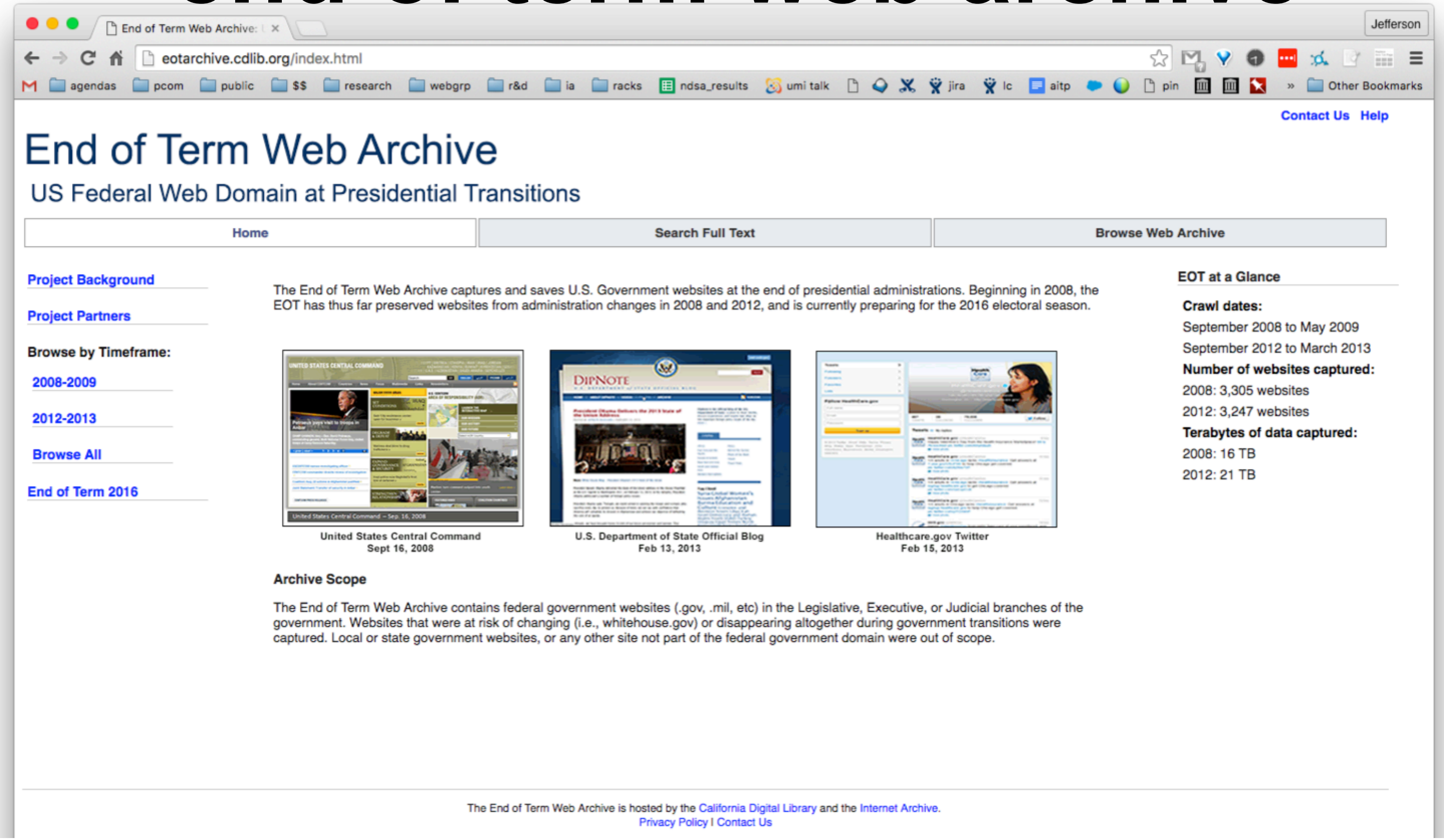
Wayback Machine doesn't have that page archived.

EOT Collections to Date

- 2008 – 16 TB, 160 million URLs/files
- 2012 – 35 TB, 194 million URLs/files
- 2016 – ~250 TB, >350 million URLs/files

Currently transferring data between partners

end of term web archive



Contact Us Help

End of Term Web Archive

US Federal Web Domain at Presidential Transitions

Home | Search Full Text | Browse Web Archive

Project Background

Project Partners

Browse by Timeframe:

- 2008-2009
- 2012-2013
- Browse All

End of Term 2016

The End of Term Web Archive captures and saves U.S. Government websites at the end of presidential administrations. Beginning in 2008, the EOT has thus far preserved websites from administration changes in 2008 and 2012, and is currently preparing for the 2016 electoral season.

United States Central Command
Sept 16, 2008

U.S. Department of State Official Blog
Feb 13, 2013

Healthcare.gov Twitter
Feb 15, 2013

Archive Scope

The End of Term Web Archive contains federal government websites (.gov, .mil, etc) in the Legislative, Executive, or Judicial branches of the government. Websites that were at risk of changing (i.e., whitehouse.gov) or disappearing altogether during government transitions were captured. Local or state government websites, or any other site not part of the federal government domain were out of scope.

EOT at a Glance

Crawl dates:
September 2008 to May 2009
September 2012 to March 2013

Number of websites captured:
2008: 3,305 websites
2012: 3,247 websites

Terabytes of data captured:
2008: 16 TB
2012: 21 TB

The End of Term Web Archive is hosted by the California Digital Library and the Internet Archive.
Privacy Policy | Contact Us

http://eotarchive.cdlib.org/