Data Rescue

Why

Who

What's next

Danielle Robinson @daniellecrobins

Mozilla Science Fellow, OHSU Neuroscience Grad Program

mozilla Science Lab

Avoiding data disaster

"The internet is a terribly unstable way to keep information available"

– Laurie Allen

Assistant Director for Digital Scholarship
PENN Libraries
DataRefuge

Why do we need Data Rescue?

Piecemeal data infrastructure

Lack of standards

Functional as long as nothing changes

Regime change

Index of /pub/data

Name	Last modified	Size	Description
Parent Directory		-	
#rw-check	08-Jun-2012 14:09	0	
15min_precip-3260/	17-Mar-2016 10:02	-	
96300w60	16-Apr-1997 15:24	5.5M	
109020/	07-May-2015 14:20	-	
ASOS_Station_Photos/	03-Apr-2014 15:00	-	
EngineeringWeatherData_CDROM/	21-Mar-2014 07:12	-	
Impact/	21-May-2015 12:41	-	
Videoclip_50years_MCSS.wmv	08-Sep-2014 06:43	388M	
access.del/	21-Mar-2014 12:38	-	
aewc-v1/	16-Sep-2014 15:17	-	
airsea/	19-Aug-2014 14:58	-	
annualreports/	20-Jun-2014 09:18	-	
anomalie/	06-Sep-2012 06:28	-	
anomalies/	27-May-2014 14:27	-	
asos-fivemin/	03-Feb-2017 07:00	-	
asos-onemin/	03-Feb-2017 07:23	-	
blizzard/	27-Jun-2014 12:56	-	
ccd-data/	26-Sep-2016 14:26	-	
cdmp/	02-Jul-2014 12:12	-	
cdo/	20-Jun-2016 10:10	-	
cirs/	18-Aug-2014 09:29	-	
climgrid/	19-Apr-2016 10:05	-	
cmb/	09-May-2016 11:45	-	

https://www1.ncdc.noaa.gov/pub/data/

Piecemeal infrastructure

Every data generating entity hosts data in its unique way

Users know where datasets live

Links from agency websites direct

Index of /pub/data

Name	Last modified	Size	Description
Parent Directory		-	
#rw-check	08-Jun-2012 14:09	0	
15min_precip-3260/	17-Mar-2016 10:02	-	
96300w60	16-Apr-1997 15:24	5.5M	
109020/	07-May-2015 14:20	-	
ASOS_Station_Photos/	03-Apr-2014 15:00	-	
EngineeringWeatherData_CDROM/	21-Mar-2014 07:12	-	
Impact/	21-May-2015 12:41	-	
Videoclip_50years_MCSS.wmv	08-Sep-2014 06:43	388M	
access.del/	21-Mar-2014 12:38	-	
aewc-v1/	16-Sep-2014 15:17	-	
airsea/	19-Aug-2014 14:58	-	
annualreports/	20-Jun-2014 09:18	-	
anomalie/	06-Sep-2012 06:28	-	
anomalies/	27-May-2014 14:27	-	
asos-fivemin/	03-Feb-2017 07:00	-	
asos-onemin/	03-Feb-2017 07:23	-	
blizzard/	27-Jun-2014 12:56	-	
ccd-data/	26-Sep-2016 14:26	-	
cdmp/	02-Jul-2014 12:12	-	
cdo/	20-Jun-2016 10:10	-	
cirs/	18-Aug-2014 09:29	-	
climgrid/	19-Apr-2016 10:05	-	
cmb/	09-May-2016 11:45	-	

https://www1.ncdc.noaa.gov/pub/data/

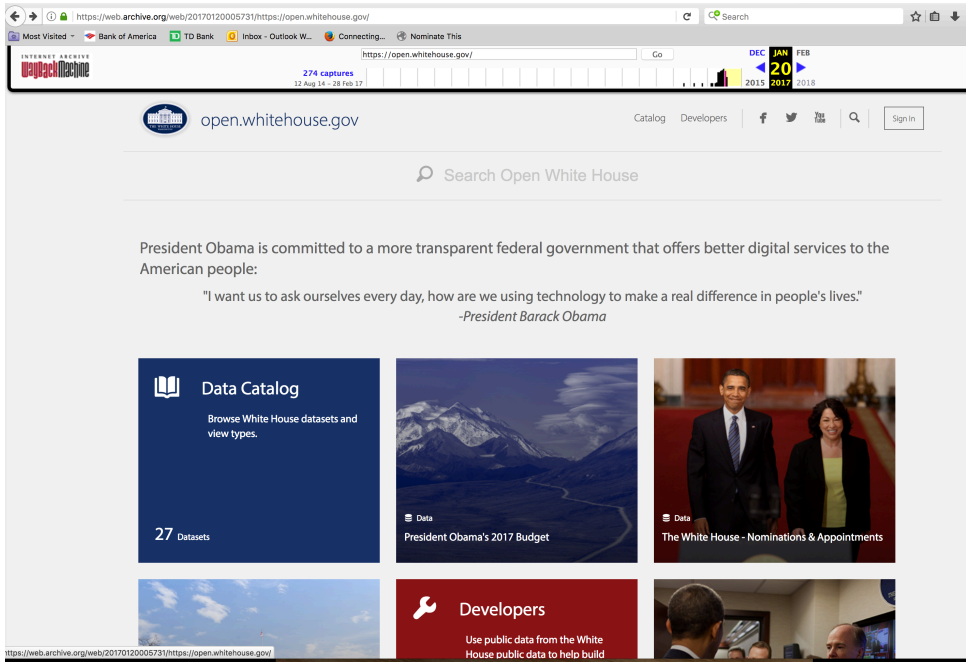
Lack of standards

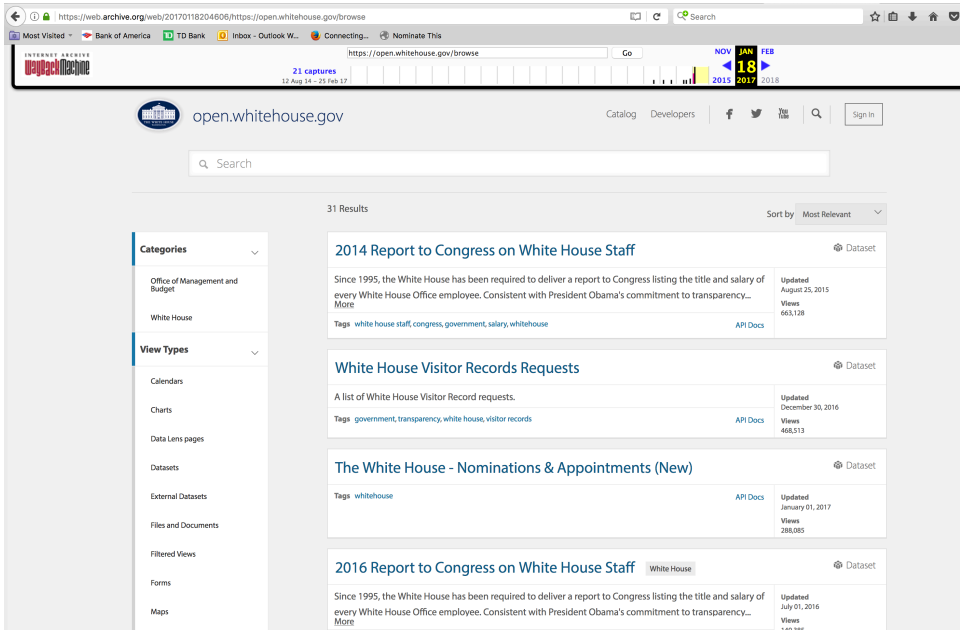
No machine readable metadata anywhere in this file structure

No agency pages link to this site

Only users know what/where

It's functional as long as nothing changes





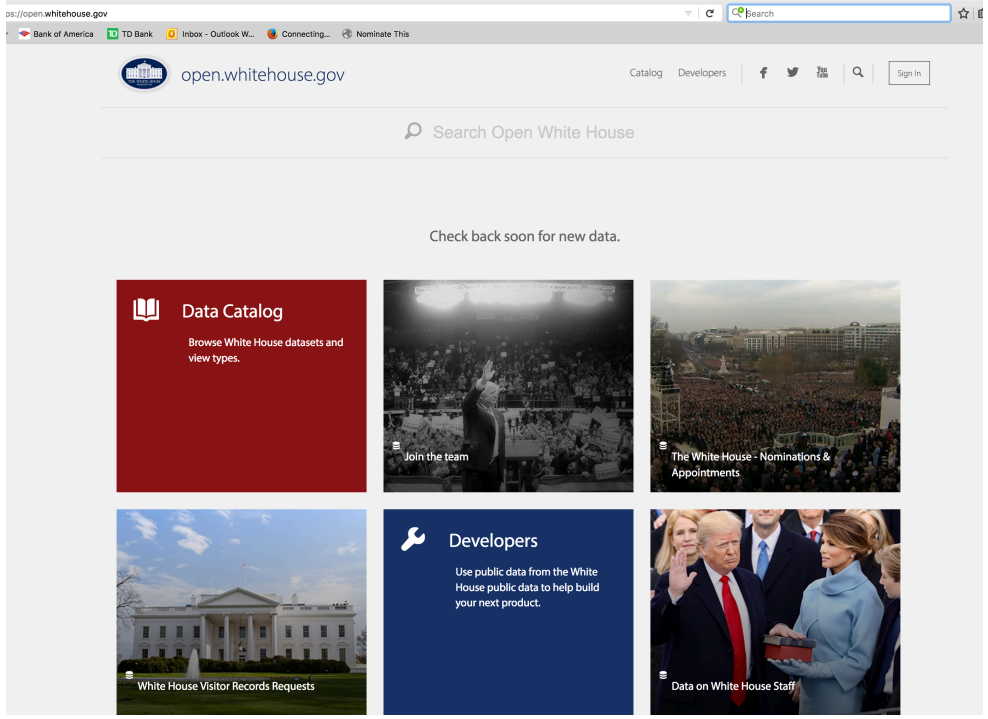
…as long as nothing changes

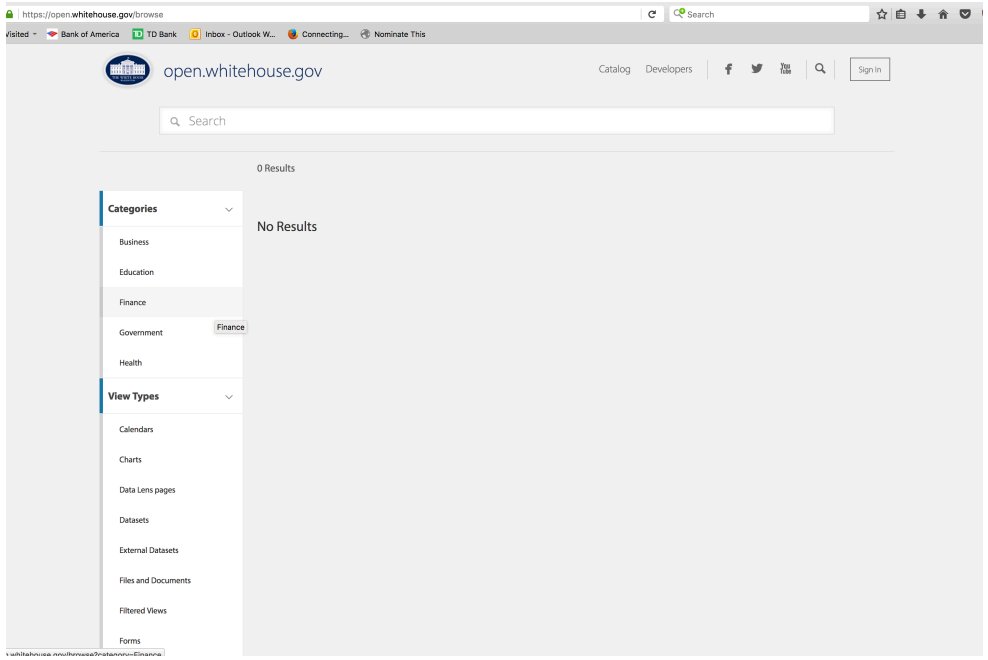
Wayback machine (Internet Archive)

Opendata.whitehouse.gov

January 18, 2017

Links out to data!





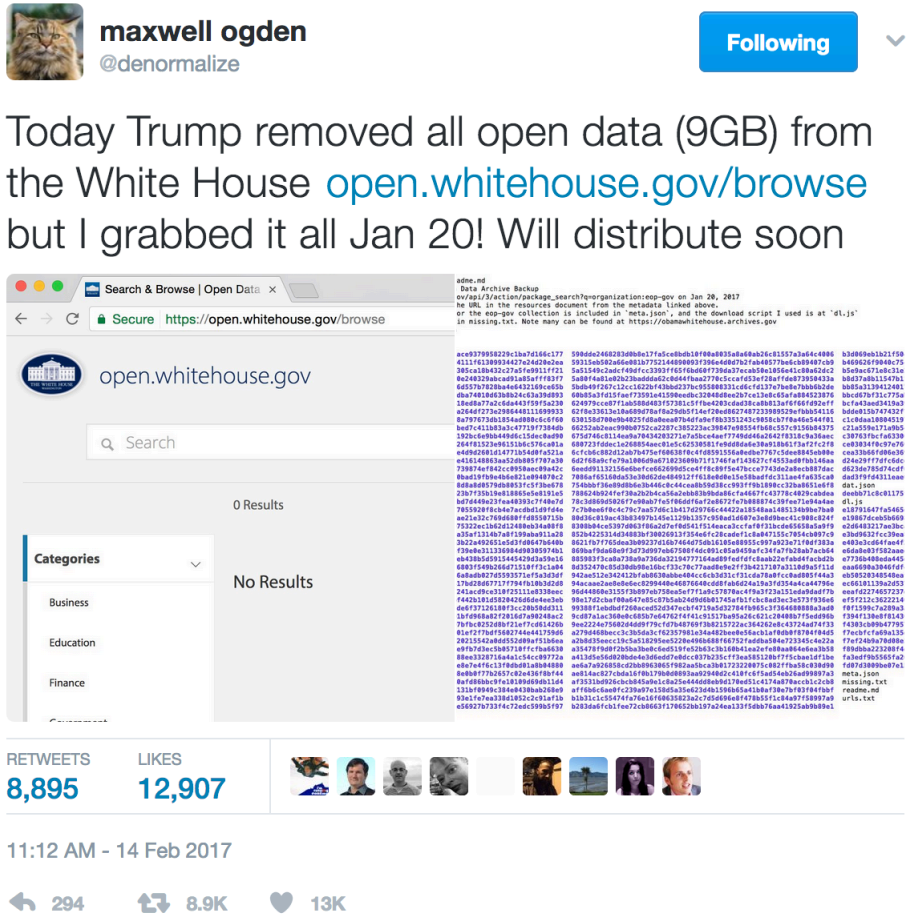
…something changed

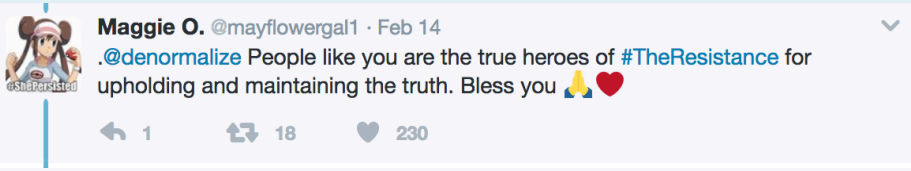
Opendata.whitehouse.gov

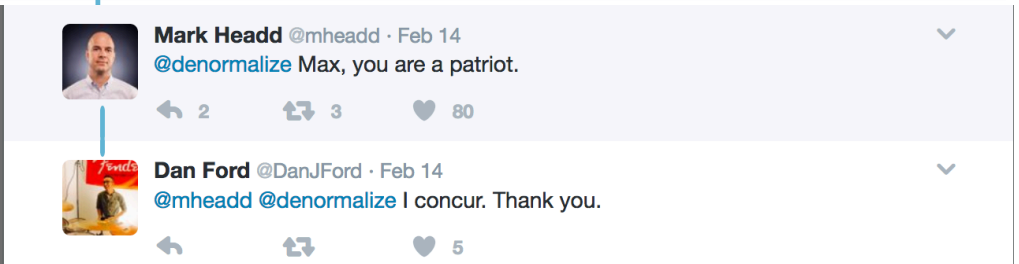
March 2, 2017

Contains no links to data!

Everyone noticed this change







Max got called a hero and a patriot

Everyone noticed this change!

Relatively small changes to websites can affect access to data

That's poor infrastructure!

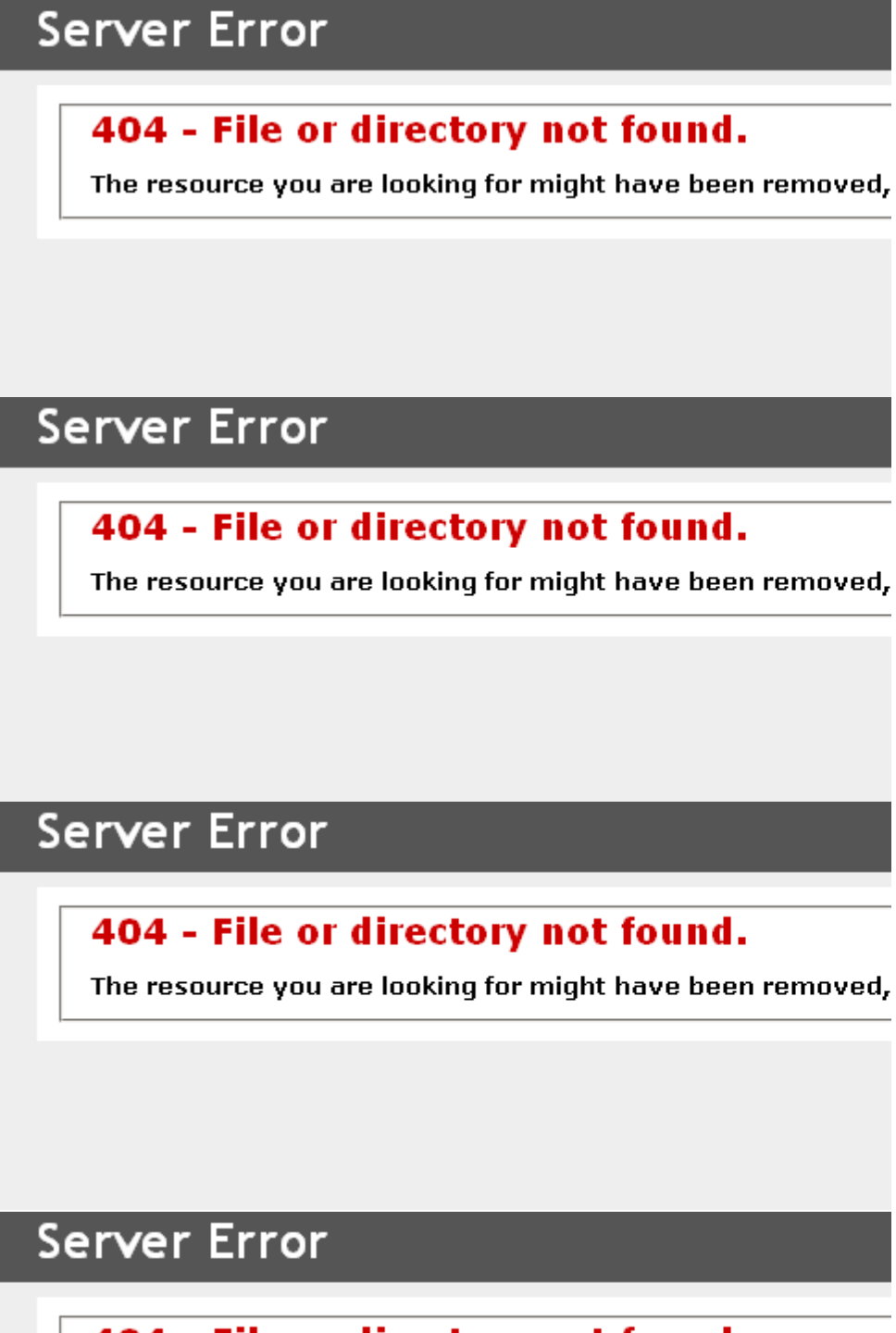
Piecemeal infrastructure + lack of standards =

No complete list of data, dataset locations, data volume

No guarantee of datset backup

No way to assess changes over time

No way to ask questions of the national or global volume of data



Avoiding data disaster

"The internet is a terribly unstable way to keep information available"

– Laurie Allen

Assistant Director for Digital Scholarship
PENN Libraries
DataRefuge

Need wide adoption of metadata standards … and documentation/mirroring of data now

Flickr user: drosen7900

The Data Rescue movement

Data.gov

Internet Archive & Archive Team

Libraries across the country!

- California Digital Library, PENN, and many more

Documentation, nomination, download via distributed events across the country

- DataRefuge @ PENN Libraries
- Climate Mirror
- Azimuth
- Environmental Data & Governance Initiative (EDGI)

Unprecedented public interest in scientific data archiving

PGK @patrickgage · Feb 11
We have stories. Hear why we are saving data.

@DataRescueSFBay @btayeg #datarefuge #resist pic.twitter.com/TzZfdSFhyt

5 3

Tonight's Workshops

Humans will make metadata with Danielle and Robin (room)

Scrapers (written by humans) will make metadata with Max, Steven, and Ted (room)

Open Hack all day tomorrow! Lunch Provided!

Scott Chamberlain of rOpenSci workshop, 10am Saturday

Open Data Scraping/Acquisition with R
R tools/techniques for exploring open data, scraping/downloading data, and more!

CODE FOR SCIENCE

ROpenSci

mozilla Science Lab

The OHSU Library