

```
 8 4889 d6ff 7
f 0000 90e8 8424 8800 0
83c4 70c3 4889 bce9 d7ff 488b
4 2410 e89f 7fd9 0c24 4889 5c24 0
c24 1848 8d15 8b44 2420 488
1553 1661 0048 4889 1424 488b
8944 2418 e85f 4889 4c24 1048
0848 8b40 0848 8b44 2420 488b
2490 0000 0090 0000 0048 8984
4883 c470 c348 ff48 8b6c 2468
2490 898c
4883 2468
4883 2468
cccc ffcc
0048 0800
0600 0048 81ec 3b41 100f 8617
0000 488d ac24 4889 ac24 0801
0000 0000 0000 48c7 8424 3801
0000 488b 8424 4001 0000 0000
8c24 3001 0000 4889 0424 488b
88b 4424 2048 08e8 98d9 ffff
c 2468 488b 5424 2460 488b 4c24 1
1848 8974 2440 3048 8b5c 242
4 2430 0100 4885 db0f 4
c 8b0f 004b 0
```

DATA RESCUE PDX

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	-	

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	-	

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

The screenshot shows the homepage of the Open White House Data Catalog. At the top, there's a banner with the quote: "President Obama is committed to a more transparent federal government that offers better digital services to the American people." Below the banner, a quote by President Barack Obama is displayed: "I want us to ask ourselves every day, how are we using technology to make a real difference in people's lives." -President Barack Obama. The main content area features several cards: "Data Catalog" (27 Datasets), "President Obama's 2017 Budget" (with a photo of him and a woman), "The White House - Nominations & Appointments" (with a photo of two people), and "Developers" (with a photo of a person at a computer). There are also links for "Data Catalog" and "Developers".

The screenshot shows the search results page of the Open White House Data Catalog. The search term is "2014 Report to Congress on White House Staff". The results list four datasets:

- 2014 Report to Congress on White House Staff** (Dataset)
Since 1995, the White House has been required to deliver a report to Congress listing the title and salary of every White House Office employee. Consistent with President Obama's commitment to transparency...
More
Tags: white house staff, congress, government, salary, whitehouse
API Docs
- White House Visitor Records Requests** (Dataset)
A list of White House Visitor Record requests.
Tags: government, transparency, white house, visitor records
API Docs
- The White House - Nominations & Appointments (New)** (Dataset)
Tags: whitehouse
API Docs
- 2016 Report to Congress on White House Staff** (Dataset)
Since 1995, the White House has been required to deliver a report to Congress listing the title and salary of every White House Office employee. Consistent with President Obama's commitment to transparency...
More
White House
Tags: white house staff, congress, government, salary, whitehouse
API Docs

...as long as nothing changes

Wayback machine (Internet Archive)

Opendata.whitehouse.gov January 18, 2017

Links out to data!

https://open.whitehouse.gov

Check back soon for new data.

Data Catalog
Browse White House datasets and view types.

Join the team

The White House - Nominations & Appointments

White House Visitor Records Requests

Developers
Use public data from the White House public data to help build your next product.

Data on White House Staff

...something changed

Opendata.whitehouse.gov
March 2, 2017

Contains no links to data!

Everyone noticed this change

https://open.whitehouse.gov/browse

Categories

- Business
- Education
- Finance
- Government
- Health

View Types

- Calendars
- Charts
- Data Lens pages
- Datasets
- External Datasets
- Files and Documents
- Filtered Views
- Forms

0 Results

No Results



maxwell ogden
@denormalize

Following

Today Trump removed all open data (9GB) from the White House open.whitehouse.gov/browse but I grabbed it all Jan 20! Will distribute soon

A screenshot of a web browser displaying the 'Search & Browse | Open Data' page at <https://open.whitehouse.gov/browse>. The page shows a search bar and a 'Categories' dropdown menu with options like Business, Education, and Finance. Below the categories, there are two sections: '0 Results' and 'No Results'. The main content area is filled with a massive amount of JSON data, which appears as a single, long string of characters. At the bottom of the page, there are metrics for 'RETWEETS 8,895' and 'LIKES 12,907', along with a row of small user profile pictures.

11:12 AM - 14 Feb 2017

294 8.9K 13K



Maggie O. @mayflowergal1 · Feb 14

.@denormalize People like you are the true heroes of #TheResistance for upholding and maintaining the truth. Bless you 🙏❤️



Mark Headd @mheadd · Feb 14
@denormalize Max, you are a patriot.



Dan Ford @DanJFord · Feb 14
@mheadd @denormalize I concur. Thank you.

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 dataset backup

No way to assess changes over time

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

Server Error

404 - File or directory not found.

The resource you are looking for might have been removed,

Server Error

404 - File or directory not found.

The resource you are looking for might have been removed,

Server Error

404 - File or directory not found.

The resource you are looking for might have been removed,

Server Error

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



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



I don't want people
questioning
data,
so I'm here
to stand up
for it.

Berhan Taye



PGK @patrickgage · Feb 11

We have stories. Hear why we are saving data.

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





Tonight's Workshops

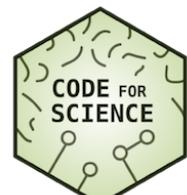
Humans will make metadata with Danielle and Robin (room)

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!



The
OHSU
Library

