

Portland State University

PDXScholar

Online Northwest

Online Northwest 2018

Mar 30th, 9:00 AM - 10:00 AM

Keynote: A Vision for Decentralized Data Preservation Across a Network of Libraries and Trusted Institutions

Danielle Robinson

Code for Science & Society, daniellecrobison@gmail.com

Follow this and additional works at: <https://pdxscholar.library.pdx.edu/onlinenorthwest>

Let us know how access to this document benefits you.

Robinson, Danielle, "Keynote: A Vision for Decentralized Data Preservation Across a Network of Libraries and Trusted Institutions" (2018). *Online Northwest*. 21.

<https://pdxscholar.library.pdx.edu/onlinenorthwest/2018/presentations/21>

This Presentation is brought to you for free and open access. It has been accepted for inclusion in Online Northwest by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

A vision for decentralized data preservation across a network of libraries and trusted institutions

DANIELLE ROBINSON, PhD
Co-Executive Director at Code for Science & Society
@daniellecrobins

GORDON AND BETTY
MOORE
FOUNDATION



CS&S Code for
Science &
Society



People power change

CS&S



**Transparency is critical
to modern research**

CS&S

@daniellecrobins

Index of /pub/data

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 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	-	

**Data are accessed via existing
web infrastructure**

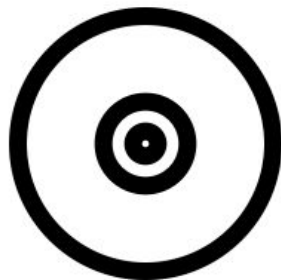
CS&S

@daniellecrobins



California Fish Data and Management Software

Download PISCES
Software



View Species Range
Maps



Download PISCES
Range Data



Reduction in sad directories

CS&S

@daniellecrobins

Sharing and depositing data:

Collaborative document sharing solutions include:

- Box
- Dropbox
- Google docs

Repositories for domain-specific data:

Many repositories are available for data sharing and publishing. Some examples are listed below:

- OHSU Digital Commons
- Figshare
- Dryad

Many data publishing options

CS&S

@daniellecrobins



**Siloed info, centralized
gate keepers control access**

CS&S

@daniellecrobins



What's next?

[Shiratski](#)

CS&S

[@daniellecrobins](#)

CS&S

Code for
Science &
Society



CS&S

@daniellecrobins



**Research &
Scholarship**



**Government
& Society**



**Journalism &
New Media**

1. Data live on the web
2. Assumptions around data preservation
3. Decentralize now!
4. Decentralized data preservation
5. Reimagine the web

1. Data live on the web

2. Assumptions around data preservation

3. Decentralize now!

4. Decentralized data preservation

5. Reimagine the web

Across domains, data live online

Early work of a writer

Government data

Newspaper archives

Your family photos

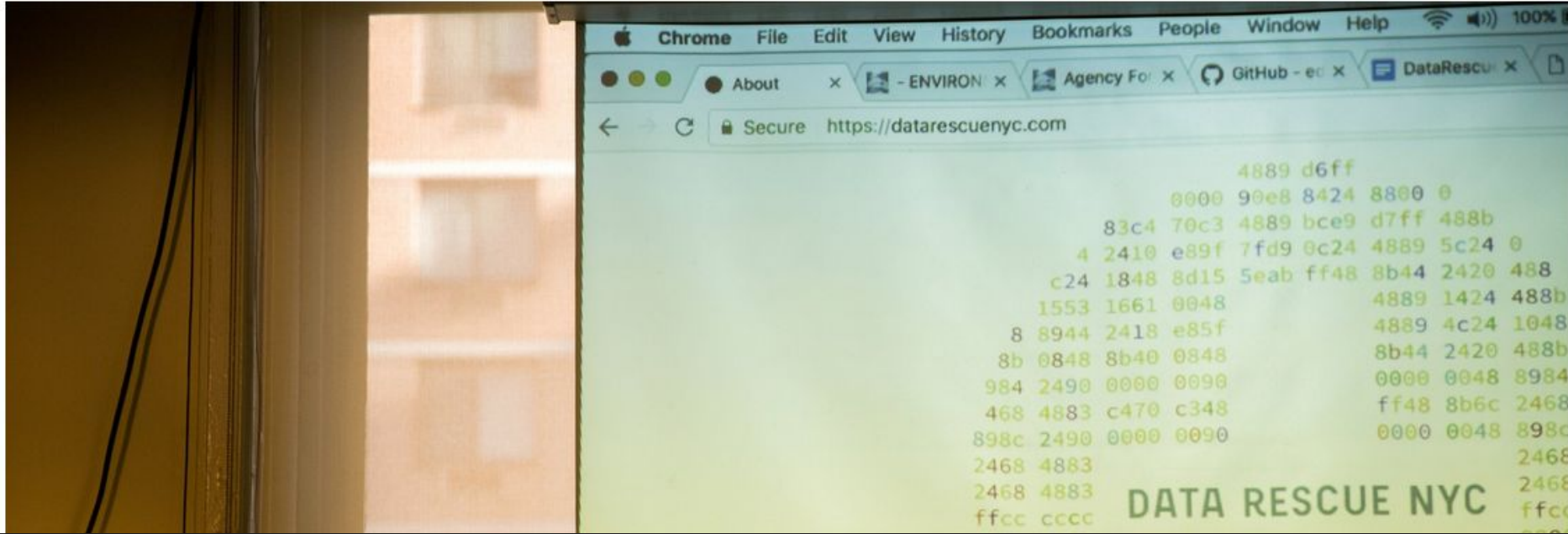
Scientific data

CS&S

@daniellecrobins

Activists Rush to Save Government Science Data — If They Can Find It

By AMY HARMON MARCH 6, 2017



Data Rescue 2016 - present

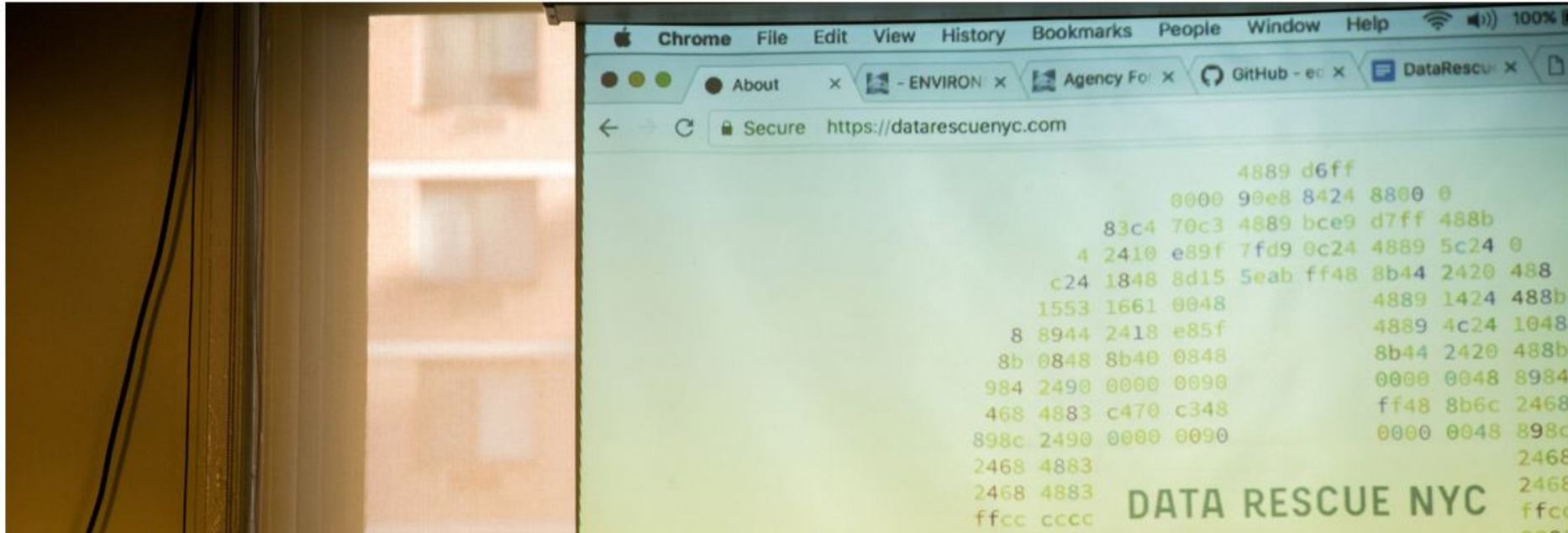
CS&S

<https://www.nytimes.com/2017/03/06/science/donald-trump-data-rescue-science.html>

@daniellecrobins

Activists Rush to Save Government Science Data — If They Can Find It

By AMY HARMON MARCH 6, 2017



Data Rescue 2016 - present

CS&S

Data Mirror - a copy of Data.gov

Data Mirror: Complementing Data Producers

by **John Chodacki** (Director, University of California Curation Center) <john.chodacki@ucop.edu>

Data Mirror is a collaborative project between the **University of California Curation Center (UC3)** and **Code for Science & Society (CSS)**, a non-profit organization committed to improving access to data for the public good. We are interested in preserving federal data because we know that the research produced, collected, or funded by the federal government are an integral part of the rich tapestry of the nation's cultural and scholarly record, and are critical resources for advancing scholarship, public policy, and governmental transparency and accountability. However, we in the library and preservation community often forget that the data producers within the federal government have comprehensive preservation strategies and workflows of their own. Although we are focused on helping solve problems, many times we unnecessarily create duplicative or parallel solutions that cut the federal research groups out of the conversation and can cause



additional issues down the road. The Data Mirror project (datamirror.org) is working to exemplify a different possible path forward.

Data Mirror is a complete, and routinely updated, copy of the main federal government research data portal, *data.gov*. Hosted by the UC3 at the **California Digital Library (CDL)**, Data Mirror points back to the “datasets of record” on federal agency websites for routine access. Why? Because those are the copies that are cared for and handled by the data producers themselves, and therefore, those copies should be referenced and used by researchers. However, should these access paths become interrupted or inaccessible, Data Mirror also includes pointers to

CDL-managed copies, as well as additional registered replicas hosted by other institutions. In this model, *data.gov* and the mandates that it works under remain the center of the workflow. Basically, Data Mirror works as a back-up of the

existing systems and offers redundancy to the *data.gov* metadata catalog and preservation services to its underlying datasets. Providing alternative search and retrieval opportunities helps to ensure that these important data remain available for study and use in perpetuity while keeping existing Federal workflows intact. Without building entirely new systems or processes, government research groups can continue to rely upon their existing workflows.

We have worked directly with the team at *data.gov* to ensure we are respecting their existing workflows. With the support of the wider library and preservation community, we would like to enhance the Data Mirror portal to include the ability for our communities to propose enriched metadata or the addition of new datasets through the portal, which would be communicated back to the agencies and *data.gov*. It is that round-tripping of federal data preservation (through existing channels!) that would truly build long-term collaboration between those producing government data and those focusing on the preservation of government data. 🐼

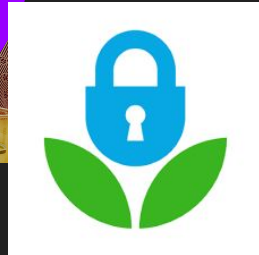
A screenshot of the Data Mirror website interface. The browser address bar shows 'datamirror.org/dataset'. The website header features the 'DATAMIRROR' logo with a stylized American flag. Below the logo, there are navigation links for 'DATA CATALOG', 'Home / Datasets', and 'Organizations'. A search bar is present with the text 'Search datasets...' and a magnifying glass icon. To the right of the search bar, there is an 'Order by:' dropdown menu set to 'Popular'. The main content area displays '152,192 datasets found'. Below this, there are three dataset entries, each with a 'Federal' label in a blue box and an 'HTML' link. The first entry is 'National Wildlife Chemical Effects Database' with a description: 'Department of Agriculture - Contains bioassay records and data for chemicals analyzed and evaluated for repellency, toxicity, reproductive inhibition, and immobilization.' The second entry is 'Clery Act Reports' with a description: 'Department of Education - The Jeanne Clery Disclosure of Campus Security Policy and Campus Crime Statistics Act is a federal statute requiring colleges and universities participating in...'. The third entry is '50th Percentile Rent Estimates' with a description: 'Department of Housing and Urban Development - Rent estimates at the 50th percentile (or median) are calculated for all Fair Market Rent areas. Fair Market Rents (FMRs) are primarily used to determine...'. On the left side of the search results, there is a 'by location' filter section with a 'Clear' button and a map of the United States.

CS&S

@daniellecrobins

“The internet is a terribly unstable way to keep information available”

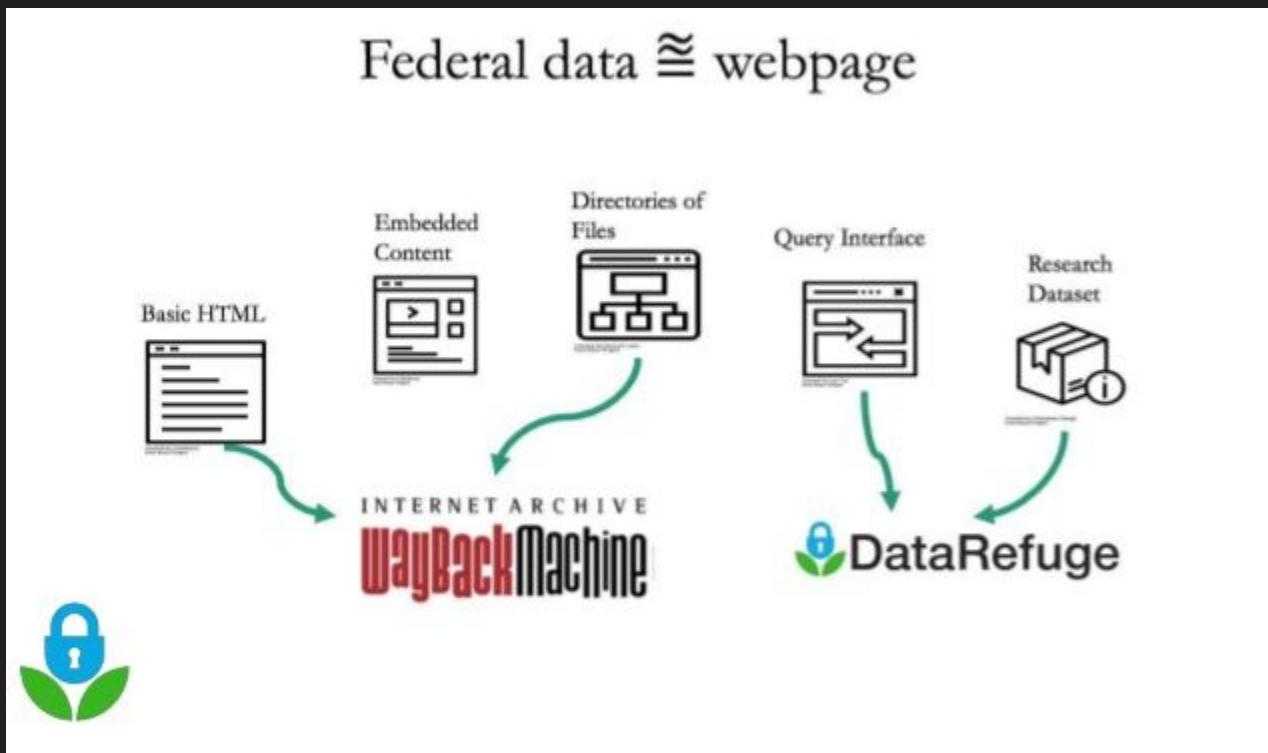
- Laurie Allen
Penn Libraries'
Assistant Director
for Digital Scholarship



CS&S

@daniellecrobins

Why are federal data \cong webpages?



Why are federal data \approx webpages?

To find an object online:

1. Discover the link
2. Link still works
3. Trust the info at the link

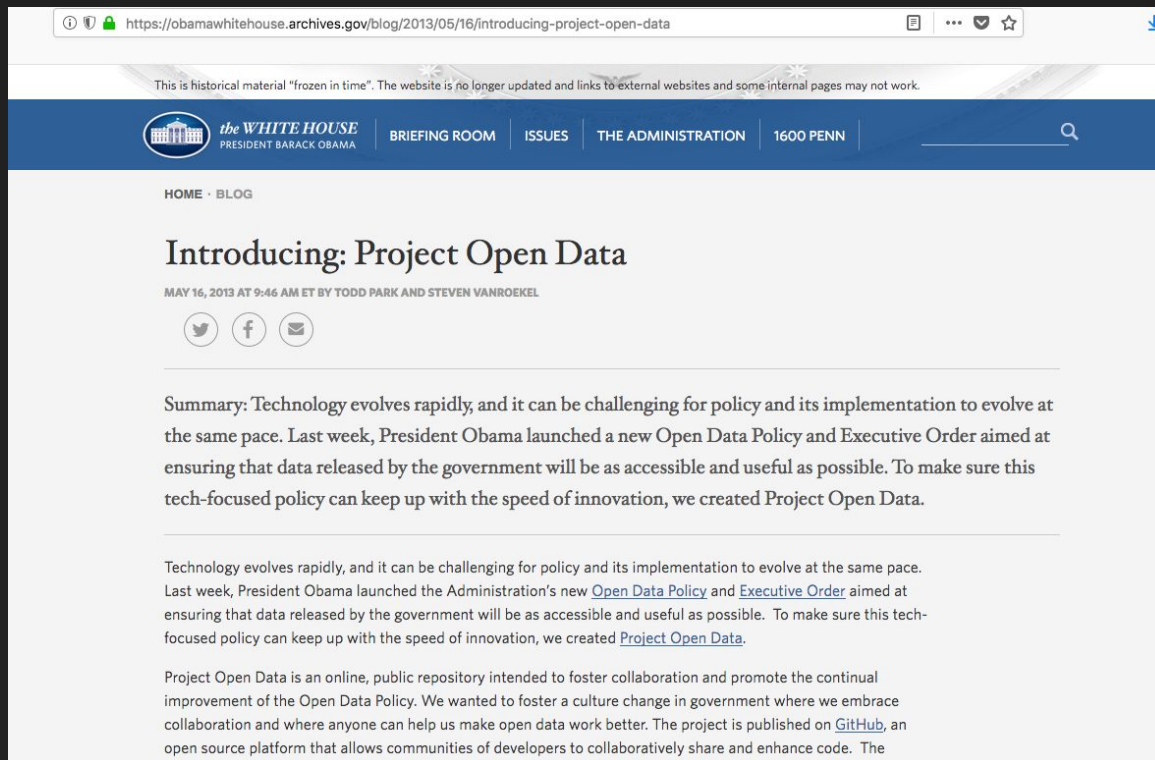
Why are federal data \approx webpages?

https://www.nasa.gov/mydataset/final_version/pubdata.csv

CS&S

@daniellecrobins

Data Rescue tested Project Open Data



The screenshot shows a web browser displaying a blog post from the White House Archives. The address bar shows the URL: <https://obamawhitehouse.archives.gov/blog/2013/05/16/introducing-project-open-data>. A warning message at the top states: "This is historical material 'frozen in time'. The website is no longer updated and links to external websites and some internal pages may not work." The navigation bar includes the White House logo, "the WHITE HOUSE PRESIDENT BARACK OBAMA", and links for "BRIEFING ROOM", "ISSUES", "THE ADMINISTRATION", and "1600 PENN". The main content area features the title "Introducing: Project Open Data" and the byline "MAY 16, 2013 AT 9:46 AM ET BY TODD PARK AND STEVEN VANROEKEL". Below the title are social media icons for Twitter, Facebook, and Email. The summary text reads: "Summary: Technology evolves rapidly, and it can be challenging for policy and its implementation to evolve at the same pace. Last week, President Obama launched a new Open Data Policy and Executive Order aimed at ensuring that data released by the government will be as accessible and useful as possible. To make sure this tech-focused policy can keep up with the speed of innovation, we created Project Open Data." The main body text repeats this summary and adds: "Technology evolves rapidly, and it can be challenging for policy and its implementation to evolve at the same pace. Last week, President Obama launched the Administration's new [Open Data Policy](#) and [Executive Order](#) aimed at ensuring that data released by the government will be as accessible and useful as possible. To make sure this tech-focused policy can keep up with the speed of innovation, we created [Project Open Data](#)." The final paragraph states: "Project Open Data is an online, public repository intended to foster collaboration and promote the continual improvement of the Open Data Policy. We wanted to foster a culture change in government where we embrace collaboration and where anyone can help us make open data work better. The project is published on [GitHub](#), an open source platform that allows communities of developers to collaboratively share and enhance code. The

CS&S

@daniellecrobins

Web as data archive :(

Project Open Data Dashboard Agencies Validator Converters ▾ Rubric Help ▾ About Sign in with MAX

Selected: Milestone 19 - May 31st 2018 ▾

	Last Crawl	Last Modified	Public Datasets	Valid Metadata	Programs	Bureaus	Public Datasets	Restricted Datasets	Non-public Datasets	Datasets with downloads	Total Download URLs	Working Download URLs	Correct Format	HTML Downloads	PDF Downloads
Department of Agriculture	25-Mar-2018 03:15:07 EDT	06-Mar-2018 17:01:22 EST	1328	100%	31	17	86.5%	13.3%	0.2%	91.6%	2645	15.3%	73.8%	63.7%	1.5%
National Aeronautics and Space Administration	25-Mar-2018 04:45:28 EDT	26-Feb-2015 18:39:23 EST	9128	100%	17	1	95.8%	0.0%	4.2%	100%	9128	9.5%	34.9%	99.1%	0.2%
National Science Foundation	25-Mar-2018 05:12:16 EDT		168	100%	2	1	95.8%	3.6%	0.6%	95.8%	178	21.3%	89.5%	21.1%	2.6%
Department of Energy	25-Mar-2018 06:35:46 EDT	20-Mar-2018 11:00:46 EDT	2771	100%	24	6	93.1%	0.2%	6.7%	93.4%	6583	87.8%	80.6%	4.7%	10.7%
Department of Health and Human	25-	25-	2043	92.4%	23	10	97.0%	0.0%	3.0%	75.6%	3549	60.5%	0.4%	9.8%	0.2%

CS&S

@daniellecrobins

Contents not found

Server Not Found	5.4% (489 of 9128)
Working links (HTTP 2xx)	9.5% (863 of 9128)
Broken links (HTTP 4xx)	0.9% (79 of 9128)
Error Links (HTTP 5xx)	0.0% (0 of 9128)
Redirected Links (HTTP 3xx)	84.3% (7697 of 9128)
Correct format	34.9% (301 of 863)
PDF for raw data	0.2% (2 of 863)
HTML for raw data	99.1% (855 of 863)
Bureaus Represented	1
Programs Represented	17
License Specified	0.0% (1 of 9128)
Detects with Redactions	0.0% (0 of 9128)

Contents not found

Server Not Found	5.4% (489 of 9128)
Working links (HTTP 2xx)	9.5% (863 of 9128)
Broken links (HTTP 4xx)	0.9% (79 of 9128)
Error Links (HTTP 5xx)	0.0% (0 of 9128)
Redirected Links (HTTP 3xx)	84.3% (7697 of 9128)
Correct format	34.9% (301 of 863)
PDF for raw data	0.2% (2 of 863)
HTML for raw data	99.1% (855 of 863)
Bureaus Represented	1
Programs Represented	17
License Specified	0.0% (1 of 9128)
Datasets with Redactions	0.0% (0 of 9128)

Link rot: When links fail

Content Drift: When referenced
content are changed

Link rot + content drift =
Reference rot

The Internet is broken

and we are using it
to access and distribute
all of human knowledge

-_(\ツ)_/-

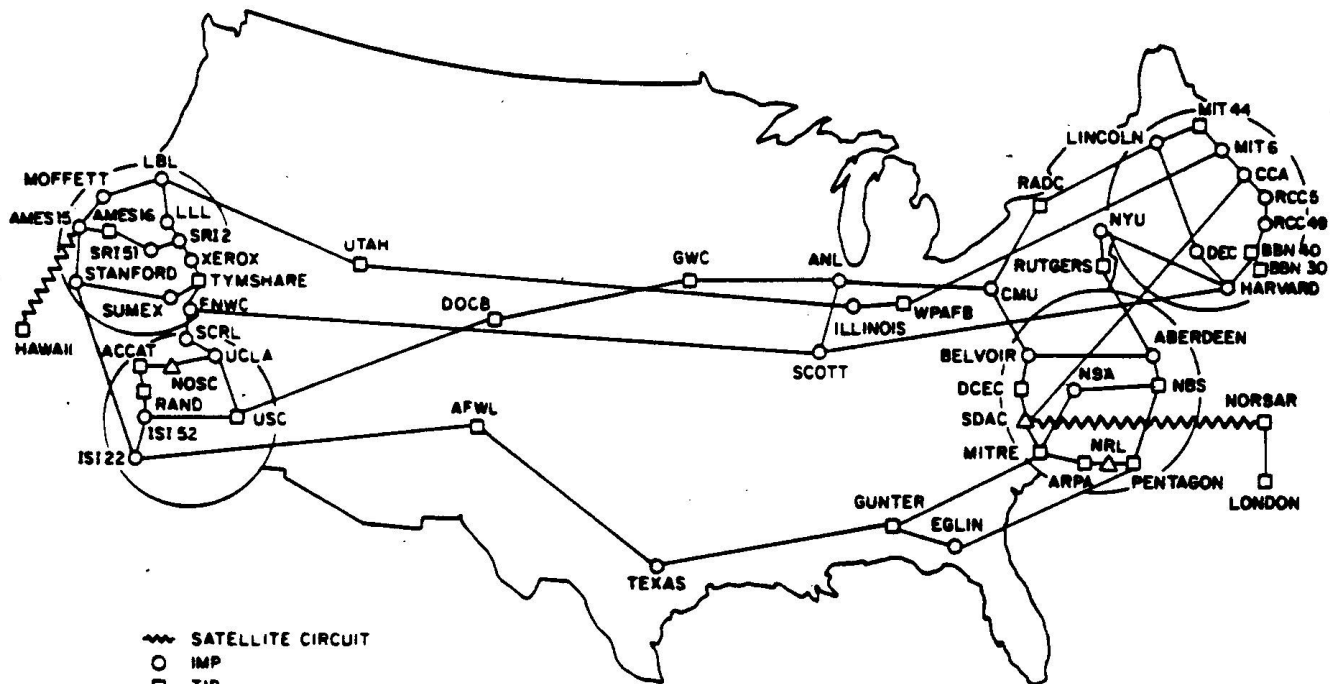
HOW DID WE EVEN



GET HERE

CS&S

@daniellecrobins

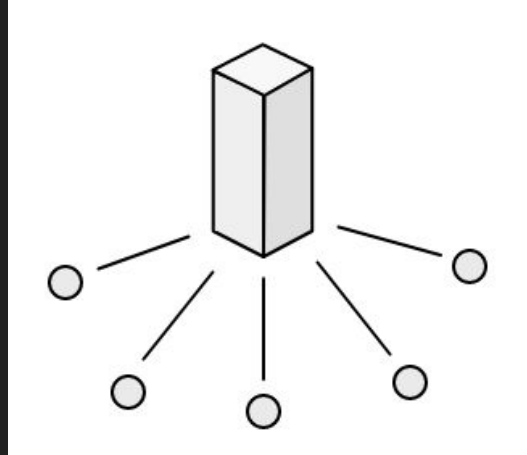


Distributed beginnings

CS&S

@daniellecrobins

Web centralization



Easier to manage and
monetize a silo

Web centralization

We trust the server to locate,
not change objects

Silos are the natural state

Data may be in multiple silos

Today's web relies upon

URLs to identify location of objects

Ability to change information
without changing location

Aggregating content for discovery

Today's web lacks

Persistent identifiers

Transparent change log

Links between silos



The web is being reimagined

[its all about Rock \(:](#)

CS&S

@daniellecrobins

1. IS IT THE TRUTH ?
2. IS IT FAIR TO ALL CONCERNED ?
3. WILL IT BUILD GOODWILL AND BETTER FRIENDSHIPS ?
4. WILL IT BE BENEFICIAL TO ALL CONCERNED ?

What's important to you?

CS&S

@daniellecrobins

1. Data on the web
- 2. Assumptions around data preservation**
3. Decentralize now!
4. Decentralized data preservation
5. Reimagine the web

Assumptions

Preservation initiated by creators

Centralized storage is stable

Preservation requires custody

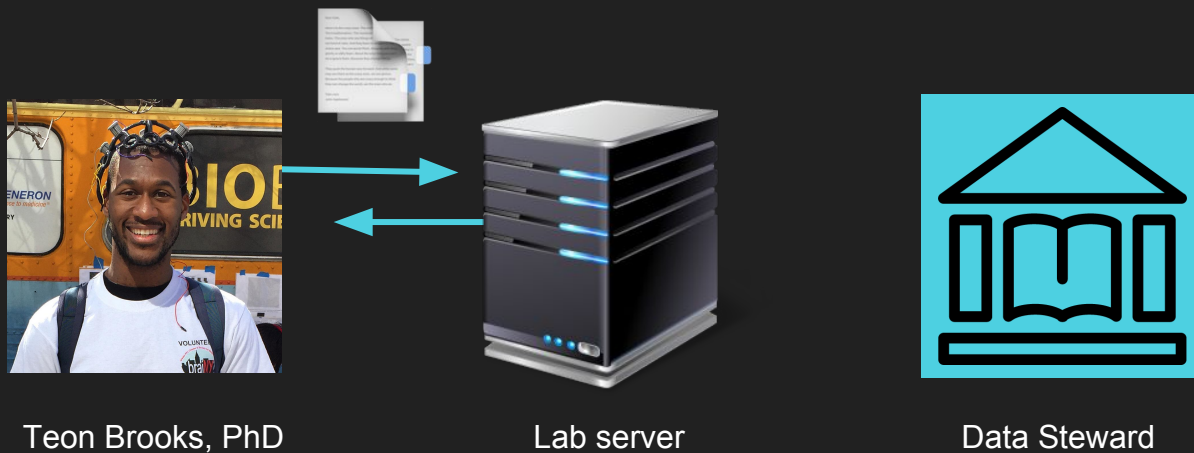


Preservation starts here

CS&S

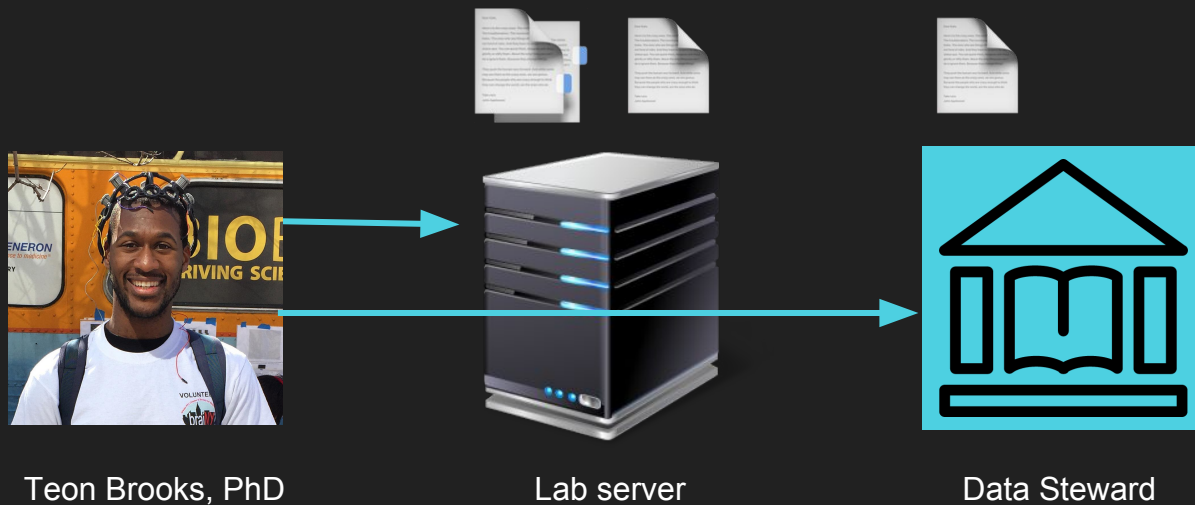
@daniellecrobins

Real data preservation workflow



Learning from researchers

Real data preservation workflow



Deposit of final version, maybe

and preserving
**“Sharing research data is not well
understood, incentivized,
or accessible”**

Daniella Lowenberg
Research Data Specialist
Product Manager of @uc3dash
California Digital Library

CS&S

@daniellecrobins



Centralized storage is stable



Is centralized storage stable?

[Timothy Vollmer](#)

CS&S

[@daniellecrobins](#)

Scholarship = Online Content Creation



-_(ツ)_/-

(sorry)

CS&S

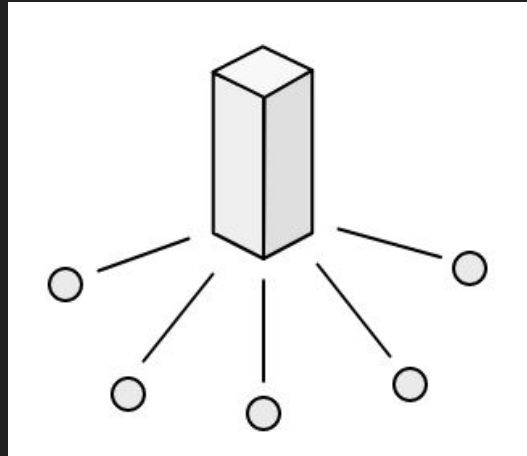
@daniellecrobins



Preservation requires custody

CS&S

Centralized model requires custody to provide access



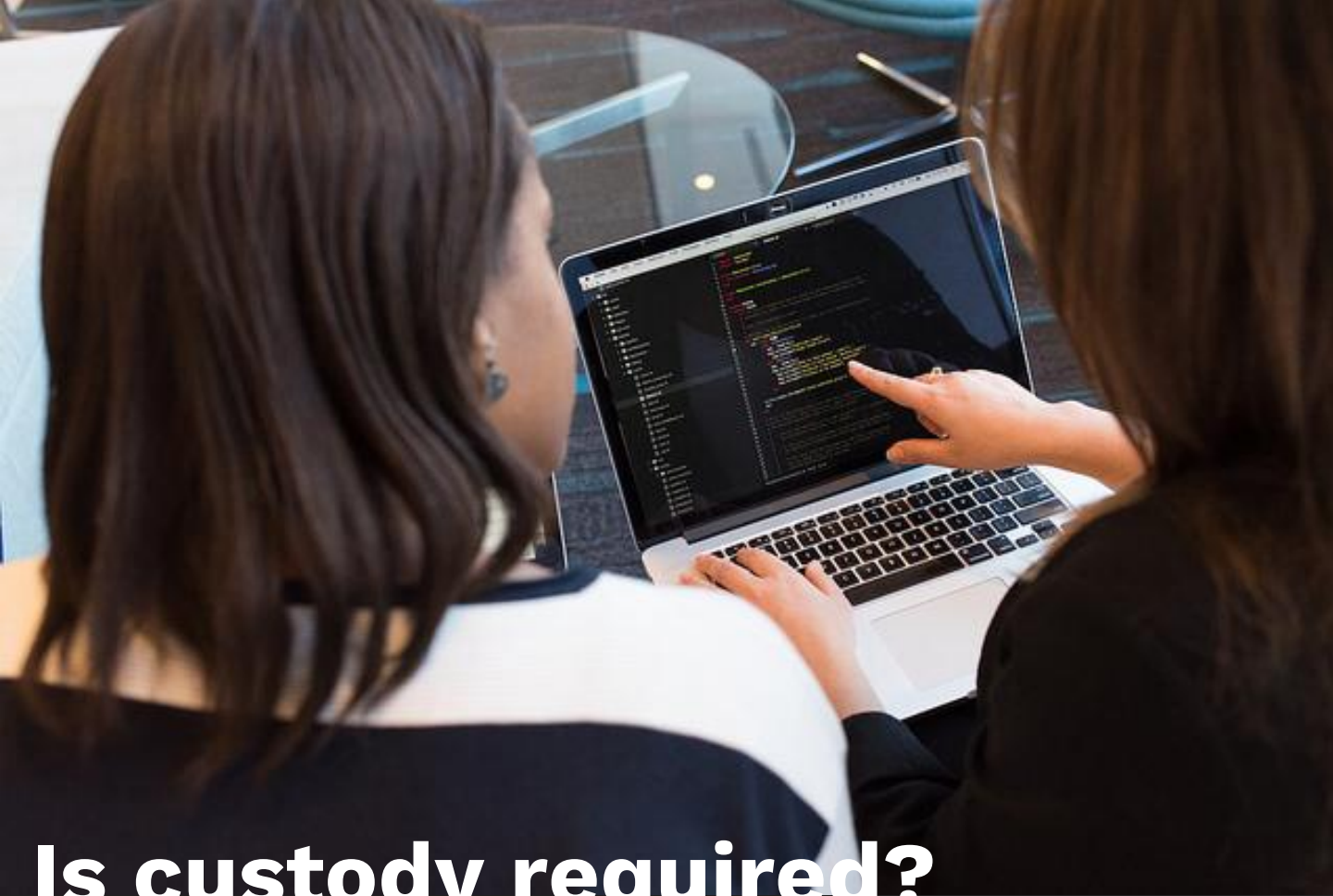


Web accessible objects

Via Agency

CS&S

@daniellecrobins



Is custody required?

[#WOCinTech Chat](#)

CS&S

[@daniellecrobins](#)

**“Preservation in place...
Bring preservation services
to the content”**

-Stephen Abrams
Preservation without Possession
California Digital Library

What do we need?

FAIR data standards:

Findable

Accessible

Interoperable

Reusable



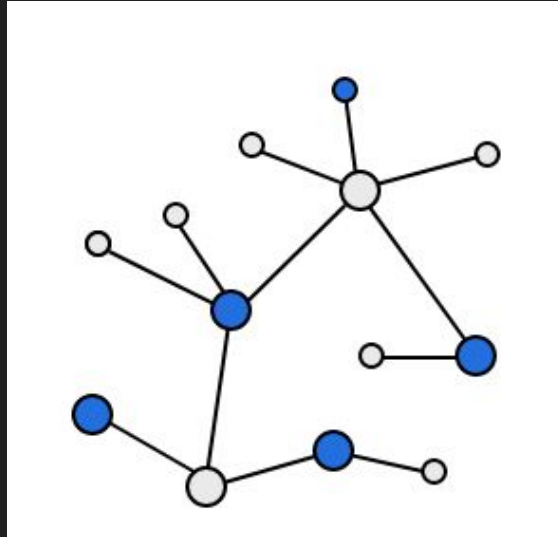
Leverage existing infrastructure

www.force11.org/group/fairgroup/fairprinciples

CS&S

@daniellecrobins

Link trusted institutions



Leverage researcher practices



Visions are nice!

[Peter Miller](#)

CS&S

[@daniellecrobins](#)



Now let's get real

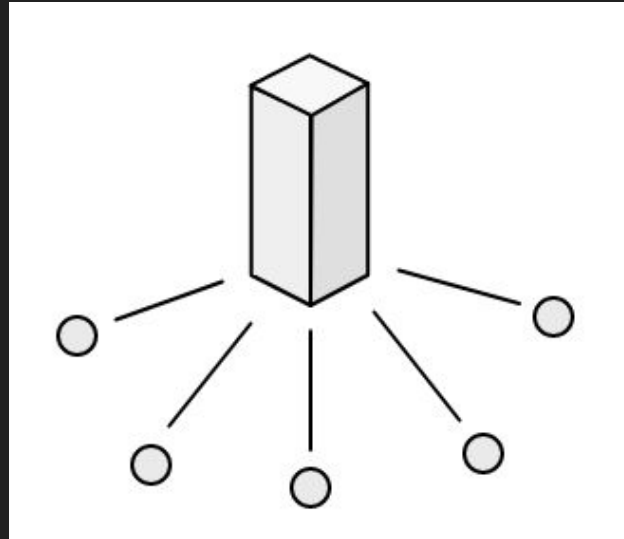
[vladeb](#)

CS&S

[@daniellecrobins](#)

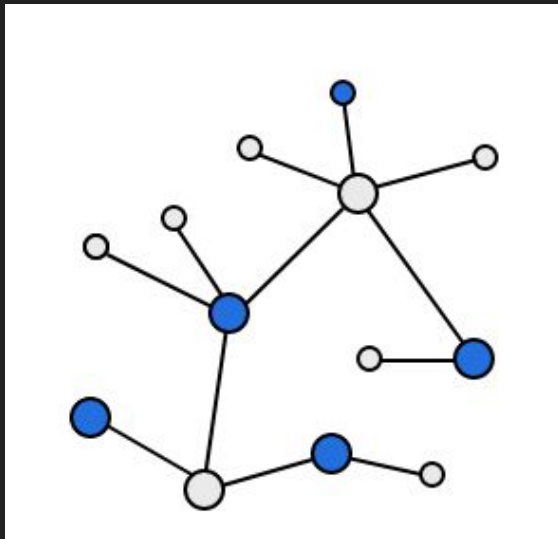
1. Data on the web
2. Assumptions around data preservation
- 3. Decentralize now!**
4. Decentralized data preservation
5. Reimagine the web

Centralized “hub and spoke” model



Data stored at central location,
accessed by independent users

Decentralized models



Data persistently identified,
networked ability to scale

Multiple decentralized approaches

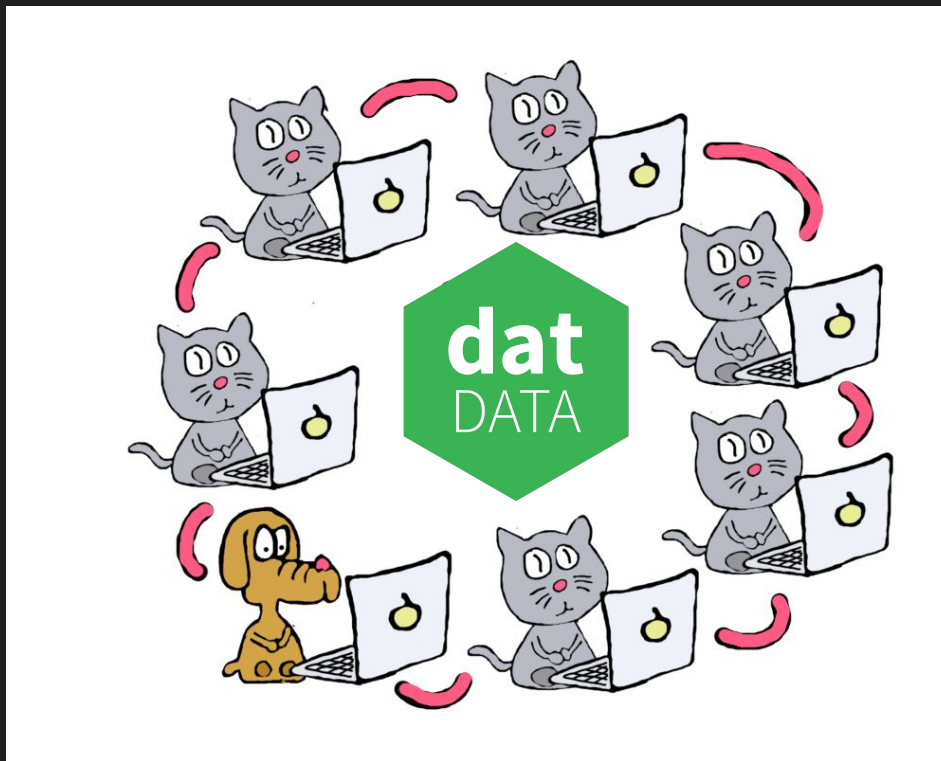


Blockchain



Peer-to-peer

Peer-to-peer public technology



The Dat Project



Nonprofit-backed, open source
peer-to-peer file sharing protocol

What's Dat?

Persistent identifiers

+

Network of peers



<https://github.com/datproject/docs/blob/master/papers/dat-paper.pdf>

CS&S

@daniellecrobins

Why Dat for scholarly data?

1. Automate preservation
2. Find data regardless of storage location
3. Spread burden of bandwidth, storage across network
4. Foundational links between silos



Data sharing tools for research
Digital Democracy indigenous rights
make the web with Beaker Browser
Enoki blogging platform
Archetype artist project space
ScienceFair publication library
and of course social media
... all built with Dat



CS&S

@daniellecrobins

For more on Dat



Details at the [white paper](#)
[try-dat.com](#)

CS&S

@daniellecrobins



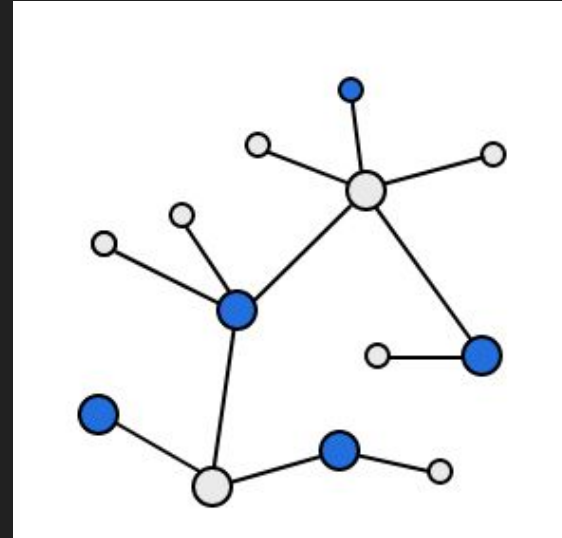
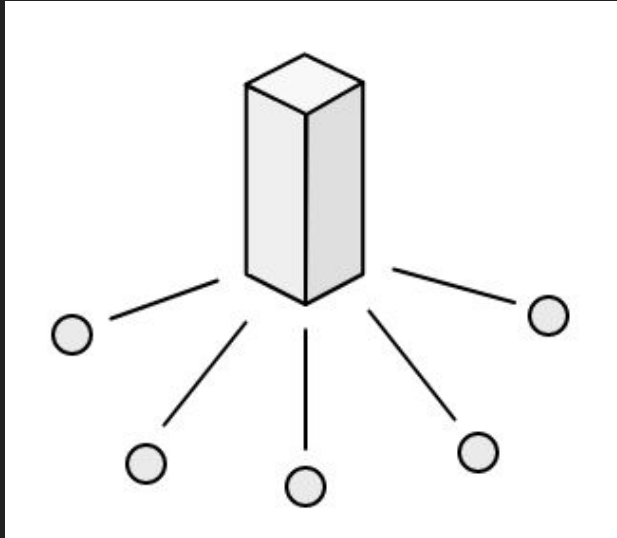
Reimagine data preservation

浪宏 葉

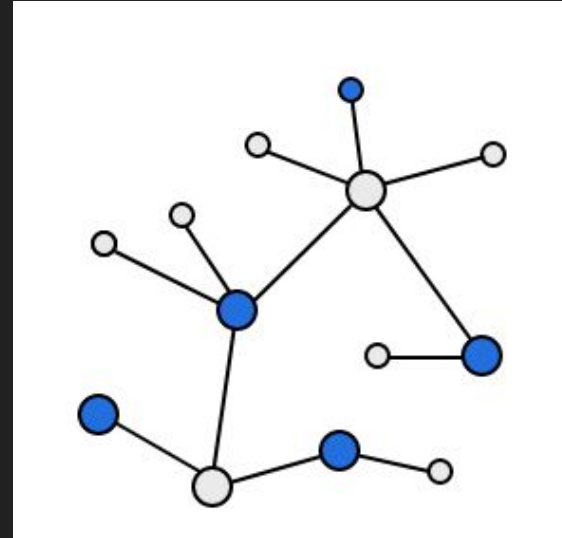
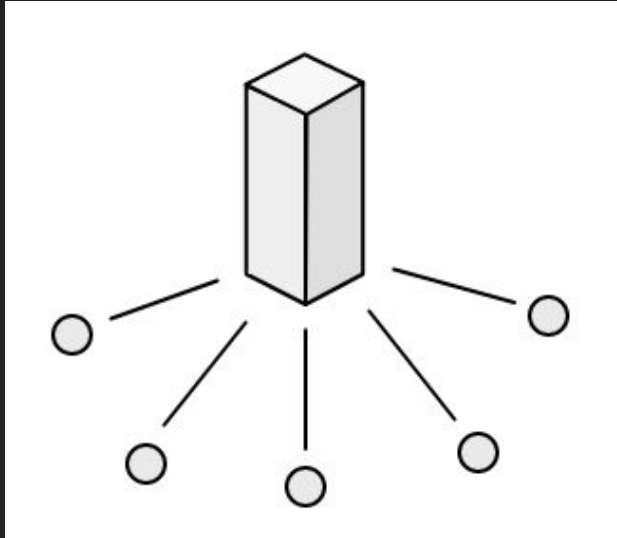
CS&S

@daniellecrobins

It's all about TRUST



... and I trust LIBRARIES



1. Data on the web
2. Assumptions around data preservation
3. Decentralize now!
- 4. Decentralized data preservation**
5. Reimagine the web



Let's build it!

[Eran Sandler](#)

CS&S

[@daniellecrobins](#)



Start with data creation

Dr. [Dannise V. Ruiz-Ramos](#) describes sea star genome annotation pipeline

CS&S

@daniellecrobins

Dat in the Lab lessons:

Leverage existing workflows

Automate data versioning, preservation

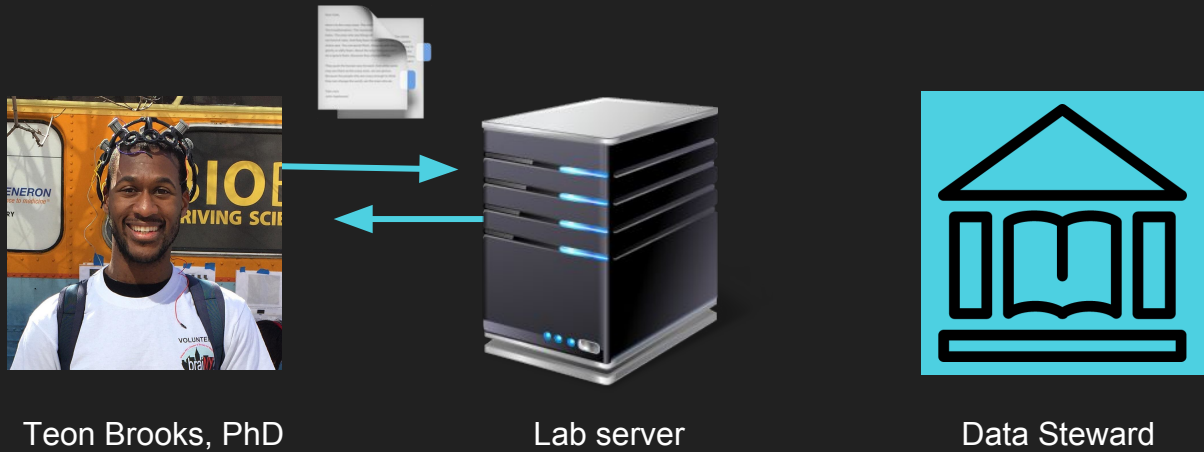
Link researchers to library

Then link libraries to each other

CS&S

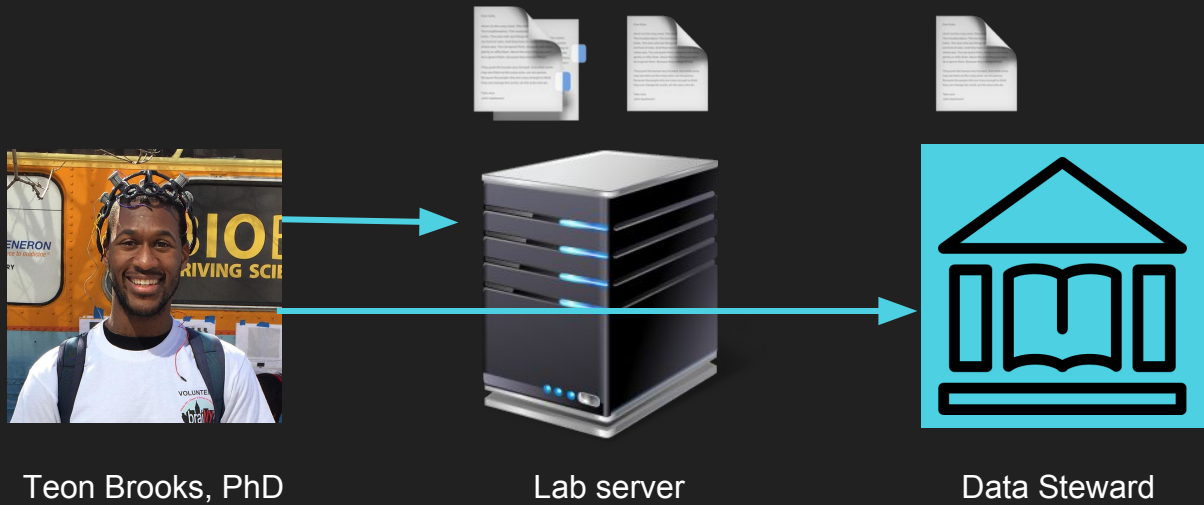
@daniellecrobins

Automating data preservation



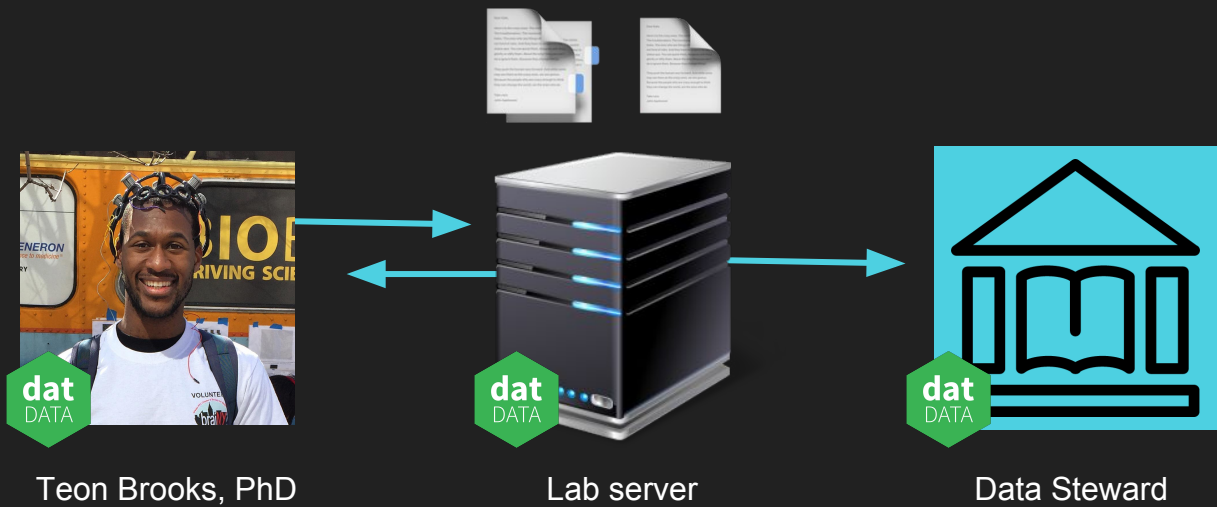
Current workflows don't preserve

Automating data preservation



Researcher may deposit a final copy

Automating data preservation

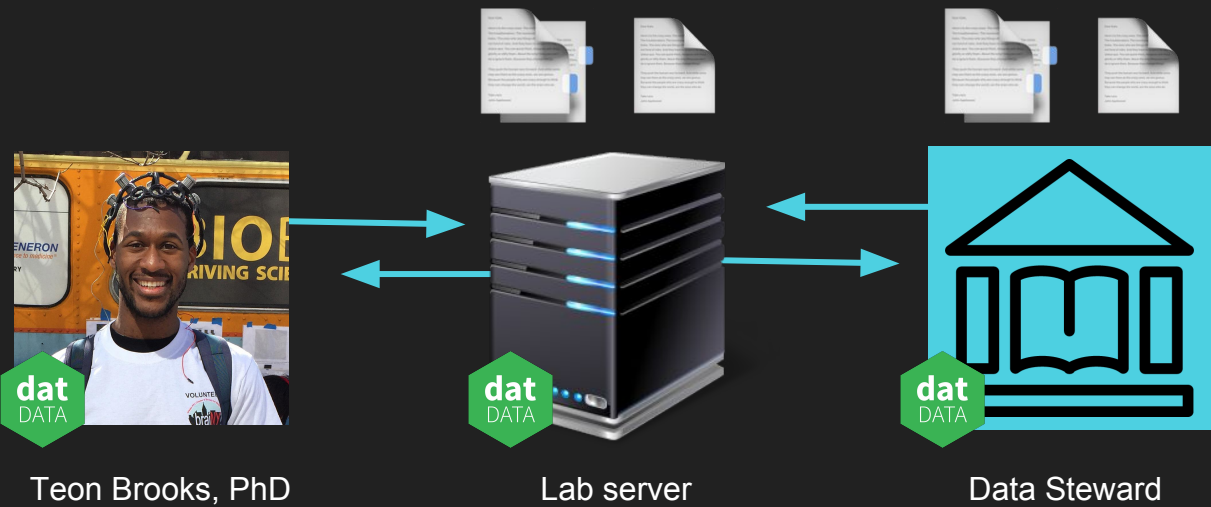


by leveraging researcher practices

CS&S

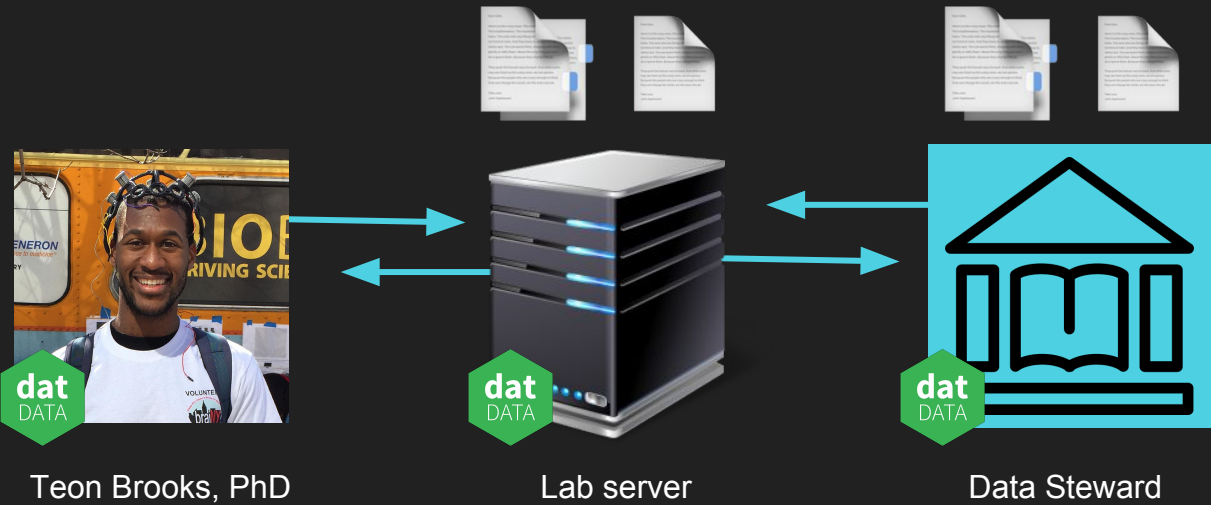
@daniellecrobins

Automating data preservation



by leveraging researcher practices

Automating data preservation



by leveraging researcher practices

Automating data preservation



dat
DATA

Teon Brooks, PhD



dat
DATA

Lab server

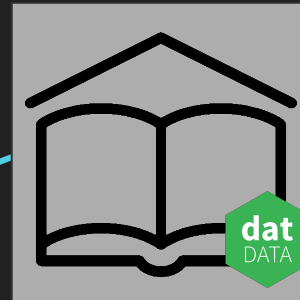


dat
DATA

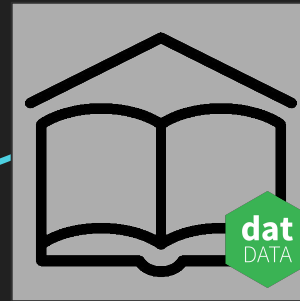
Data Steward

... and avoiding disaster

Link people via institutions



Link people via institutions



CS&S

@daniellecrobins

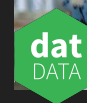
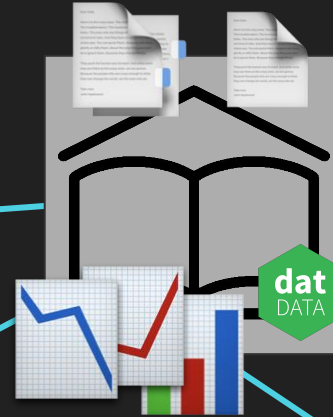
Link people via institutions



Data preserved in place



Data stewards pull from network



Every institution contributes

Storage, bandwidth

Metadata on their collection

Commitment to preserve their collection

To the network



CS&S

@daniellecrobins

Any user can access

Information on library collections

History of objects

Whole or partial data sets

from the network



CS&S

@daniellecrobins



How can we get there from here?

Eddy Josafat Hernández Vega

CS&S

@daniellecrobins



We cover more ground together

1. Data on the web
2. Decentralize now!
3. Decentralized data preservation
4. Assumptions around data preservation
- 5. Reimagine the web**

**“We embed values into our
technology whether we are
aware of it or not”**

- Stephen Whitmore (@noffle)

[Digital Democracy](#)

CS&S

@daniellecrobins

A hand is holding a white rectangular card in the center of the frame. The card has the text "What's important to you?" written on it in a bold, black, sans-serif font. The background is dark with many out-of-focus, colorful lights in shades of red, green, and orange, creating a bokeh effect. The lighting is soft, highlighting the hand and the card.

**What's
important to
you?**

Reimagine...

What scholarship is
(is it “just online content”?)

How scholarly resources identified

How institutions are connected

How preservation happens

Reimagine scholarly data preservation with us!

Apply to work with us as a Ford-Mozilla
Open Web Fellow

<https://foundation.mozilla.org/fellowships/apply/>
Applications close on April 20th (Friday at 5pm ET)

CS&S

@daniellecrobins

Thank you!

Questions?

Extra thanks to the Online Northwest
organizing committee!

DANIELLE ROBINSON, PhD
Co-Executive Director at Code for Science & Society
[@daniellecrobins](#)

GORDON AND BETTY
MOORE
FOUNDATION



CS&S Code for
Science &
Society

Citations & links

1. Dat project whitepaper: <https://github.com/datproject/docs/blob/master/papers/dat-paper.md>
2. Harmon A. Activists Rush to Save Government Science Data — If They Can Find It. The New York Times. March 6, 2017.
<https://www.nytimes.com/2017/03/06/science/donald-trump-data-rescue-science.html? r=0>
3. Preservation is not a Place: <http://www.ijdc.net/article/view/98/73>
4. https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/div-classtitlereferenc-e-rot-an-emerging-threat-to-transparency-in-political-sciencediv/54F56CFC2CBE05778130E40CABB2CC_C5
5. <https://www.nature.com/news/the-trouble-with-reference-rot-1.17465>
6. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0167475>
7. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253>
8. <https://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416>
9. Preservation without possession: Content-addressable identifiers for post-custodial preservation: https://figshare.com/authors/Stephen_Abrams/4788273
10. [Birds, Bees, and EZIDs: Where Do CDL's Persistent Identifiers Come fr](#)

CS&S

[@daniellecrobins](#)