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
dat DATA
UC3 UC Curation Center
CS&S Code for Science & Society

People power change

CS&S

@daniellecrobins

daniellecrobinson.com/mini-wow-pdx

Transparency is critical to modern research

CS&S

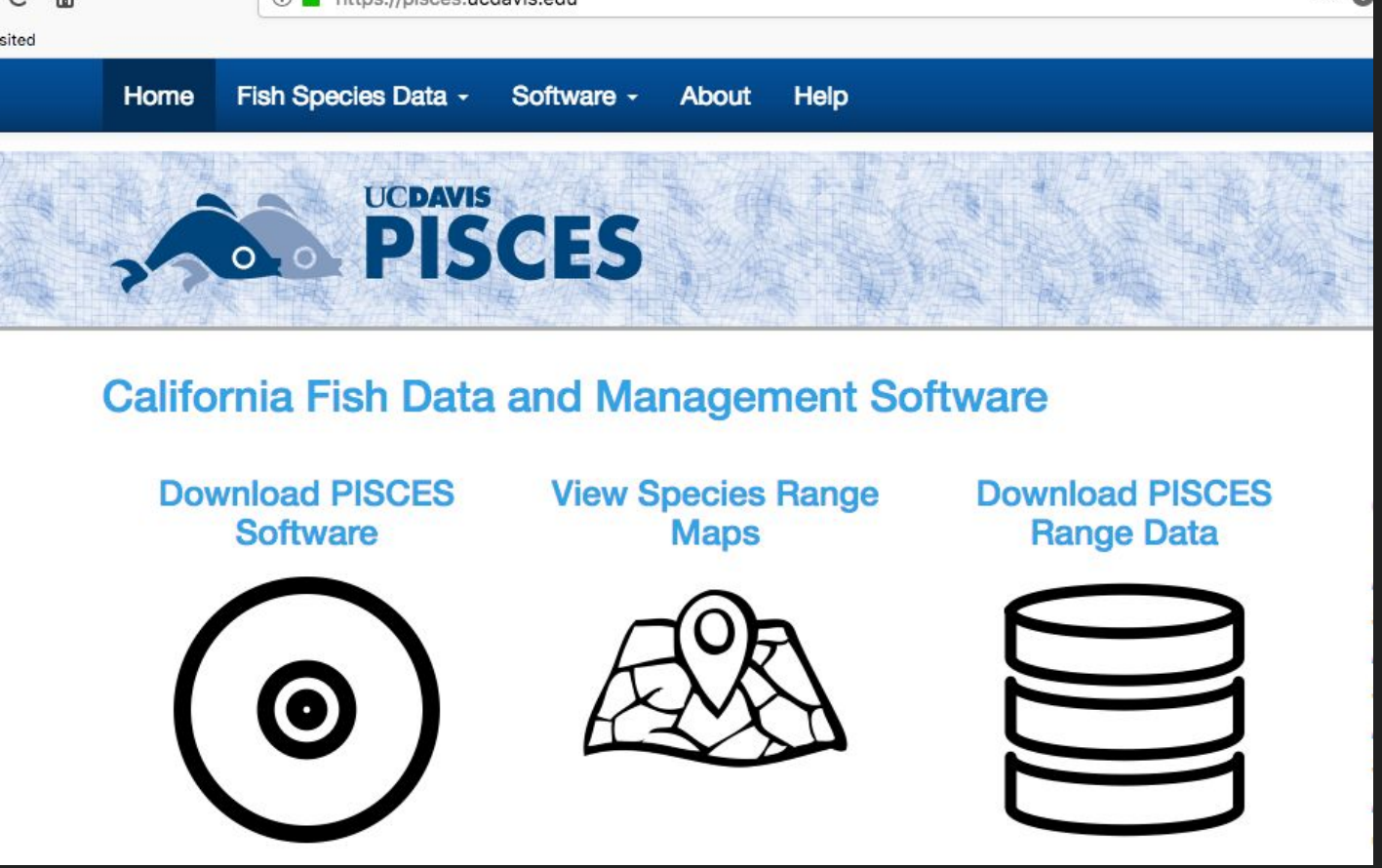
@daniellecrobins

Data are accessed via existing web infrastructure

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

CS&S

@daniellecrobins



Reduction in sad directories

https://pisces.ucdavis.edu/

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 be

- OHSU Digital Commons
- Figshare
- Dryad

Many data publishing options

https://www.ohsu.edu/xd/education/library/data/share-and-archive/index.cfm

CS&S

@daniellecrobins

Siloed info, centralized gate keepers control access

Doc Searls

CS&S

@daniellecrobins

What’s next?

Shiratski

CS&S

@daniellecrobins

CS&S

Code for Science & Society

Research & Scholarship

Government & Society

Journalism & New Media

CS&S

https://codeforscience.org/

@daniellecrobins

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

CS&S

@daniellecrobins

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

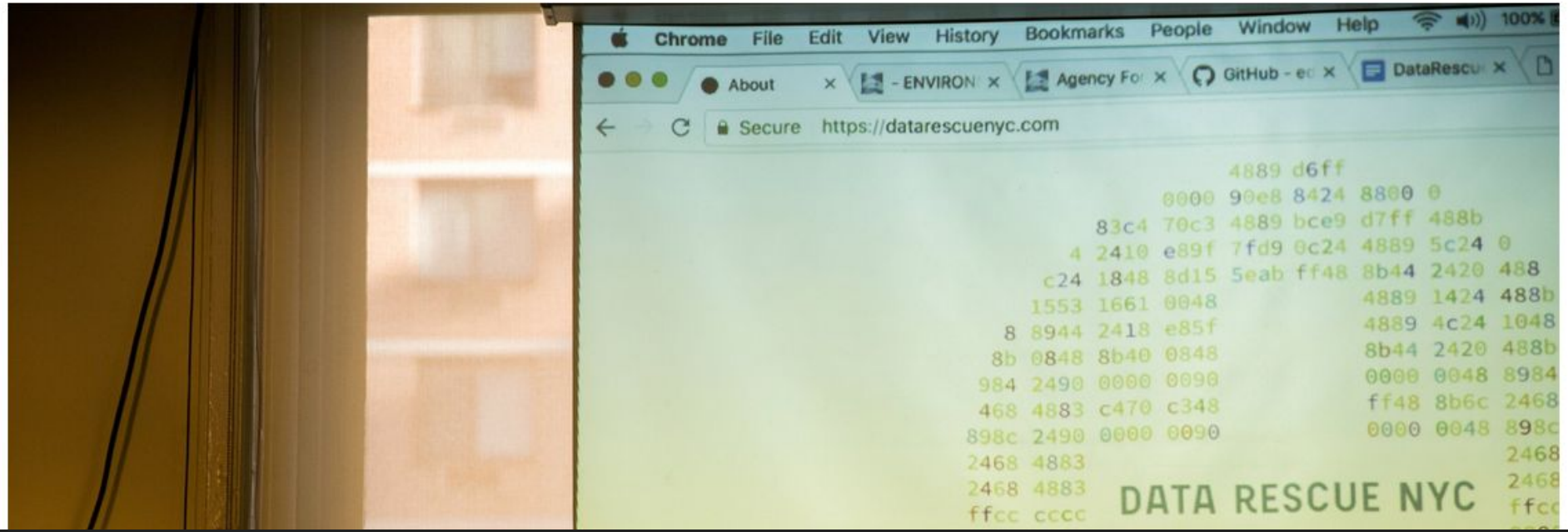
Scientific data

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

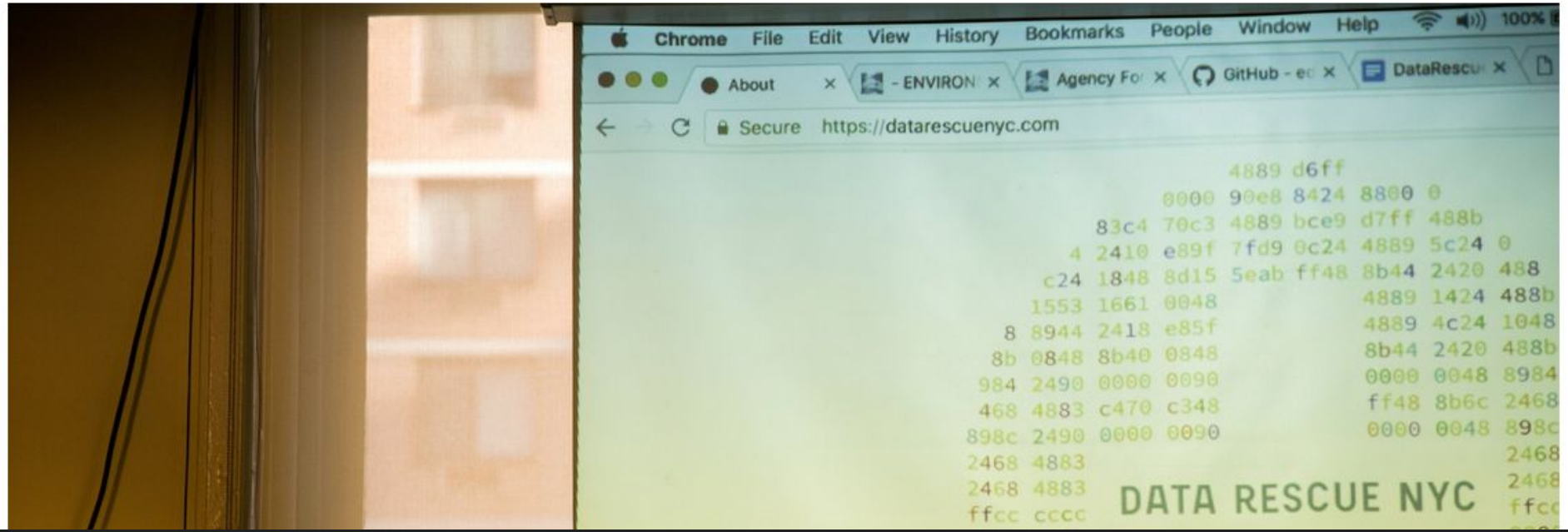
https://www.nytimes.com/2017/03/06/science/donald-trump-data-rescue-science.html

CS&S

@daniellecrobins

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

By AMY HARMON MARCH 6, 2017



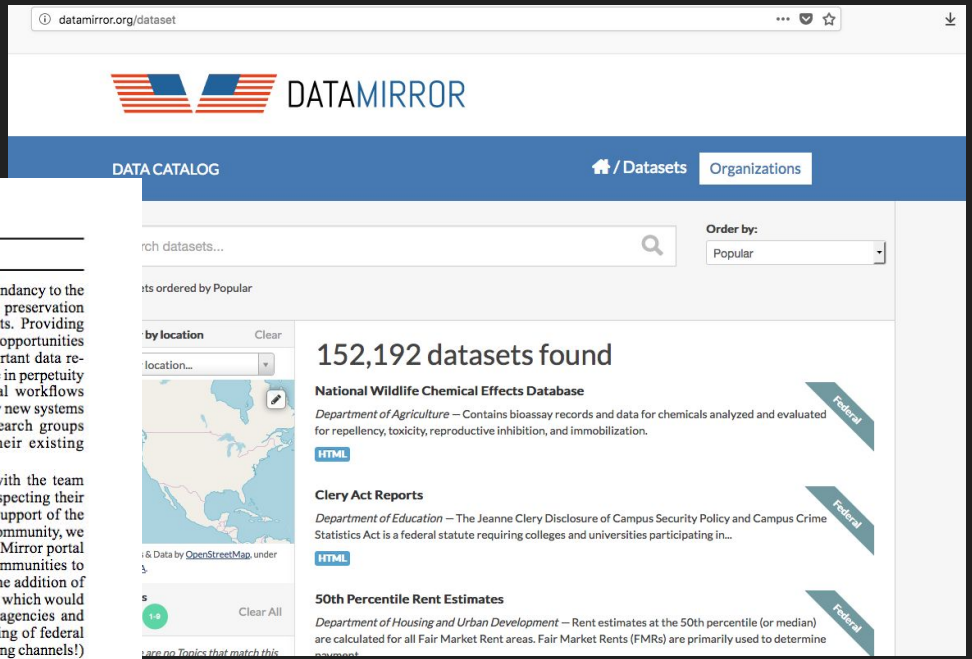
Data Rescue 2016 – present

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

CS&S

@daniellecrobins

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.

https://escholarship.org/uc/item/2n1715ff

CS&S

@daniellecrobins

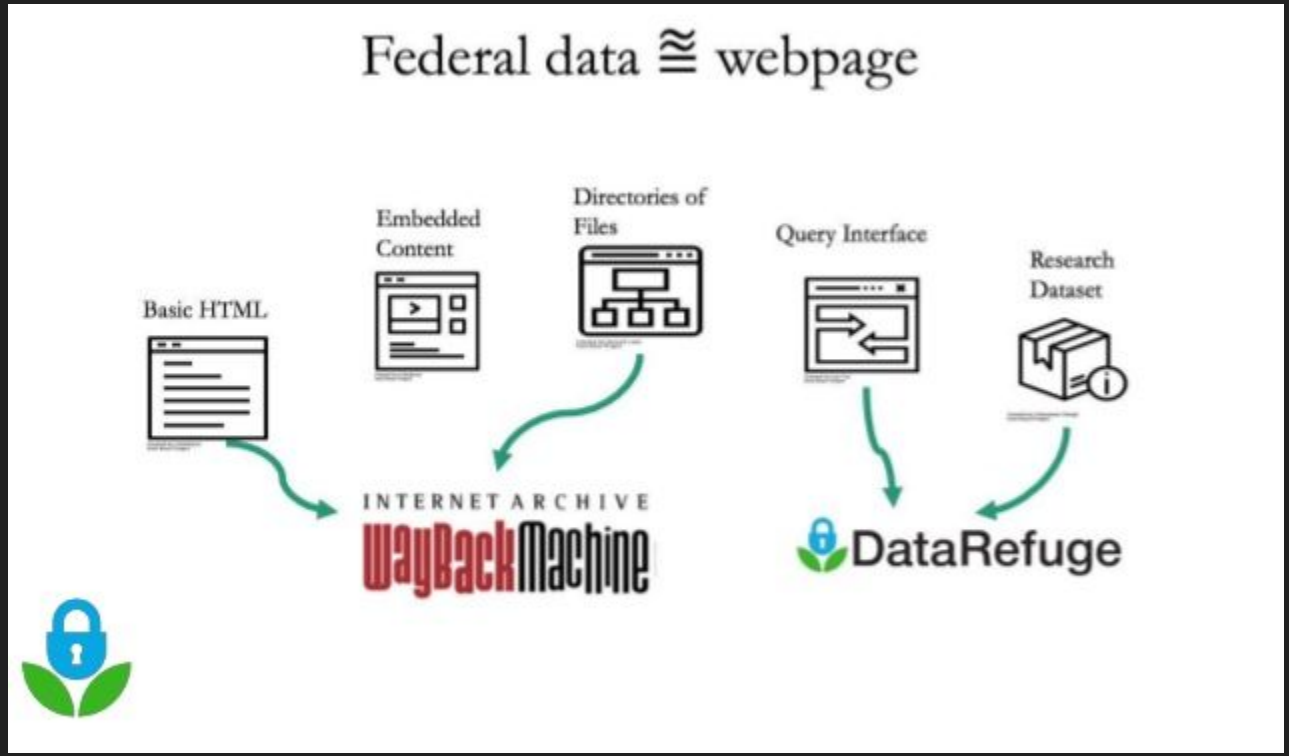
“The internet is a terribly unstable way to keep information available”

- Laurie Allen
  Penn Libraries'
  Assistant Director
  for Digital Scholarship

CS&S

@daniellecrobins

Why are federal data ≅ webpages?



https://www.slideshare.net/shefw/save-the-data-the-role-of-librarians-in-datarescue-collaborations

CS&S

@daniellecrobins

Why are federal data ≅ webpages?

To find an object online:

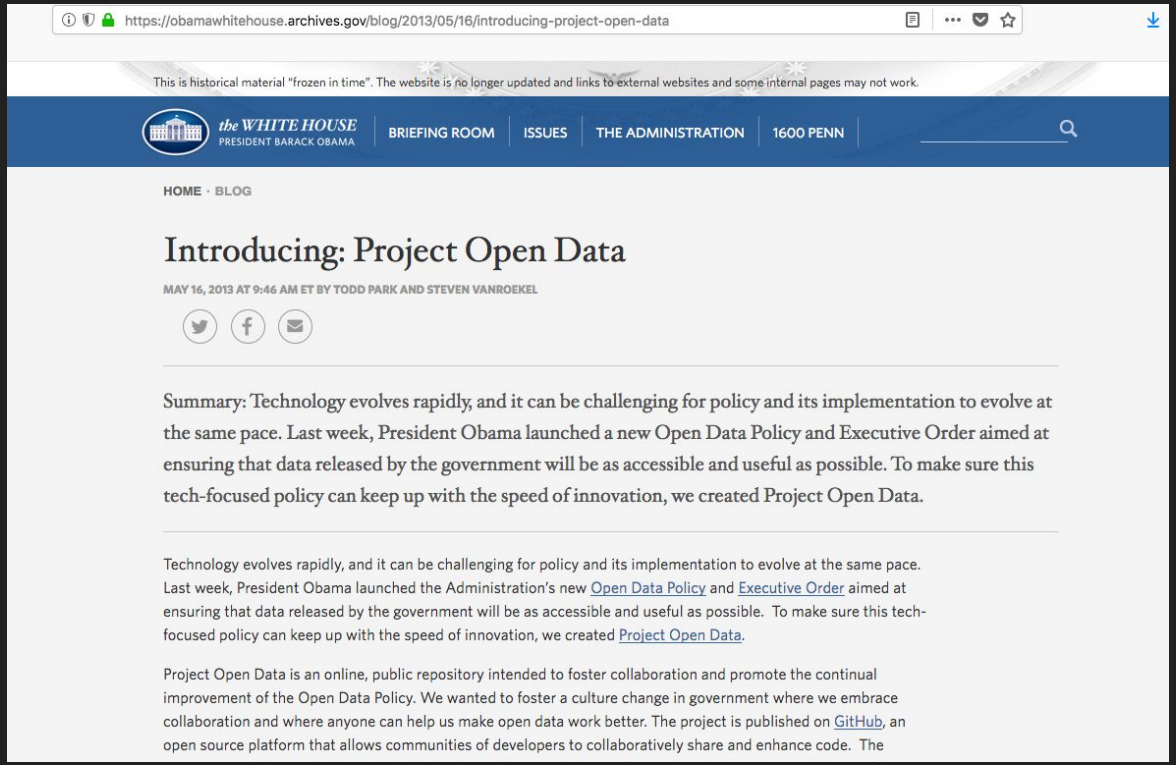
1. Discover the link
2. Link still works
3. Trust the info at the link

CS&S
@daniellecrobins

Why are federal data ≅ webpages?

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

Data Rescue tested Project Open Data

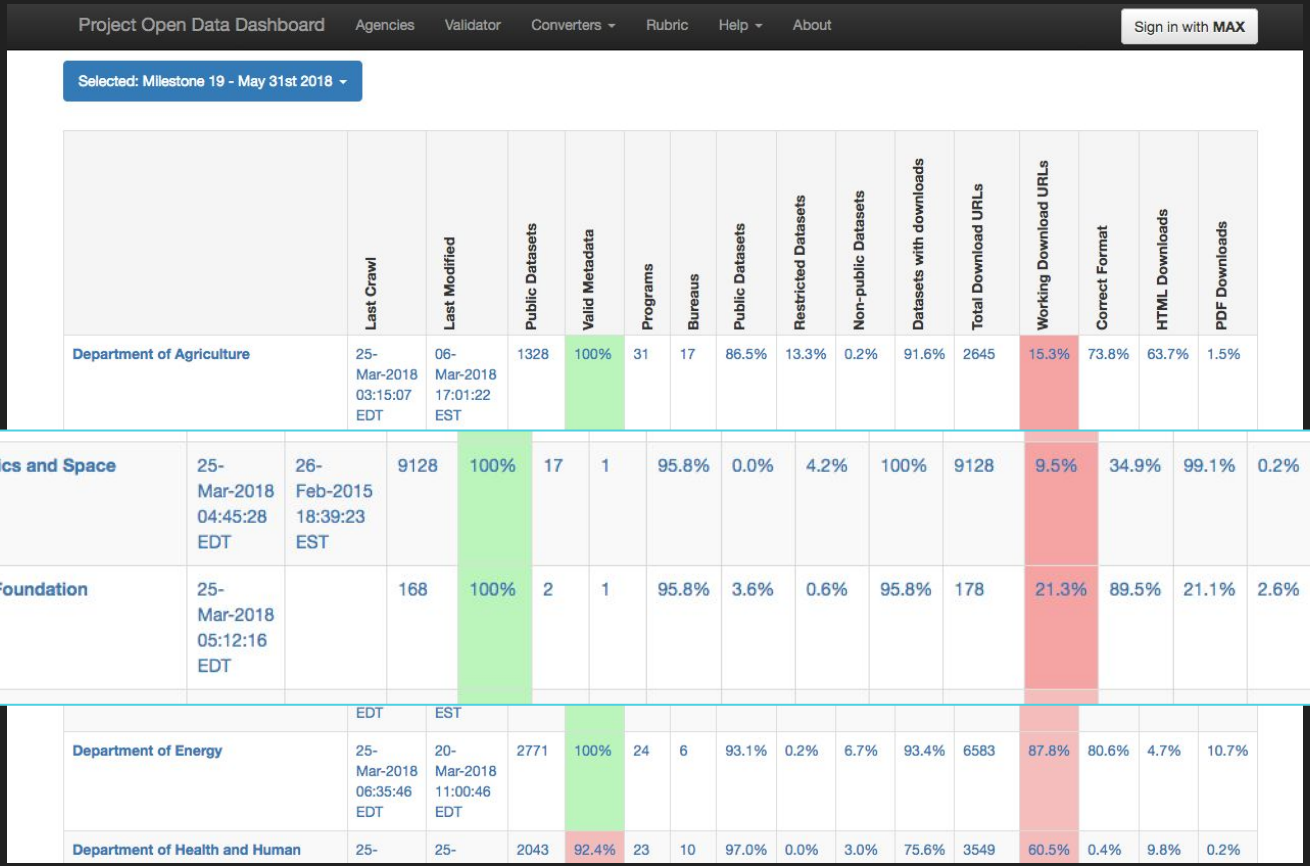


https://obamawhitehouse.archives.gov/blog/2013/05/16/introducing-project-open-data

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%
	EDT	EST													
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%

https://labs.data.gov/dashboard/offices/2018-05-31/qa

CS&S

@daniellecrobins

Contents not found

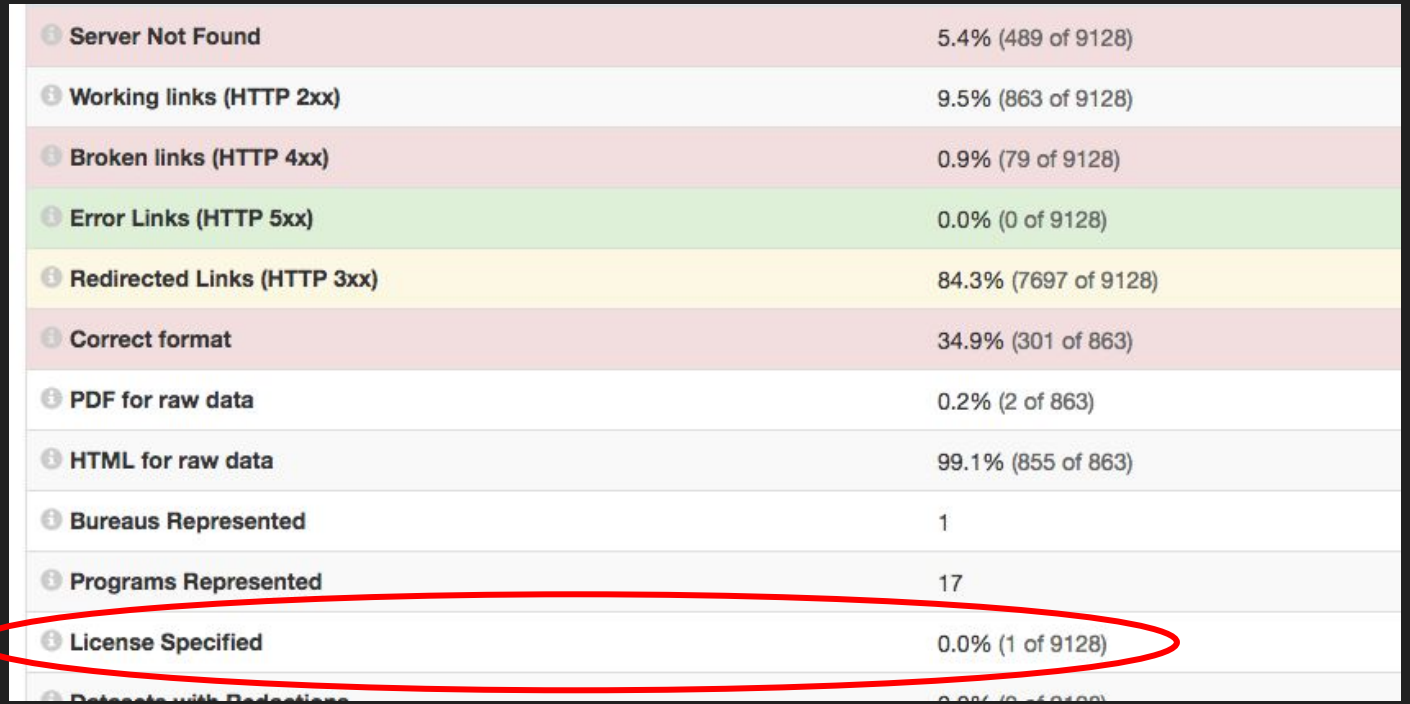
Server Not Found	5.4% (489 of 9128)
Working links (HTTP 2xx)	9.5% (863 of 9128)
Broken links (HTTP 4xx)	0.9% (79 of 9128)
Error Links (HTTP 5xx)	0.0% (0 of 9128)
Redirected Links (HTTP 3xx)	84.3% (7697 of 9128)
Correct format	34.9% (301 of 863)
PDF for raw data	0.2% (2 of 863)
HTML for raw data	99.1%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

https://labs.data.gov/dashboard/offices/detail/49476/2018-05-31

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)

https://labs.data.gov/dashboard/offices/detail/49476/2018-05-31

CS&S

@daniellecrobins

Link rot: When links fail

Content Drift: When referenced content are changed

Link rot + content drift =
Reference rot

M. Klein, several papers and talks, links at end

The Internet is broken

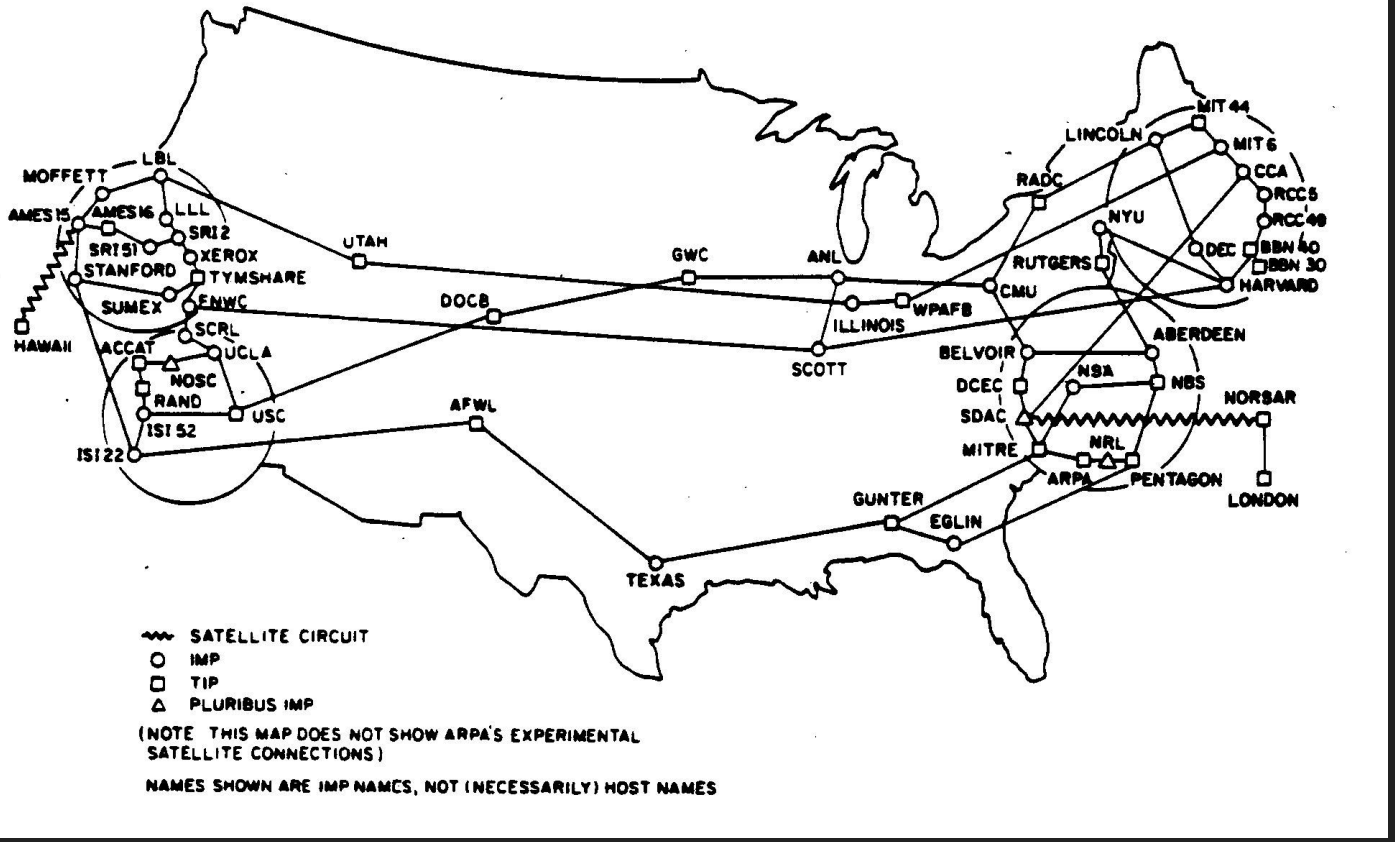
and we are using it
to access and distribute
all of human knowledge

¯_(ツ)_/¯

https://imgflip.com/memegenerator/Picard-Wtf

CS&S

@daniellecrobins



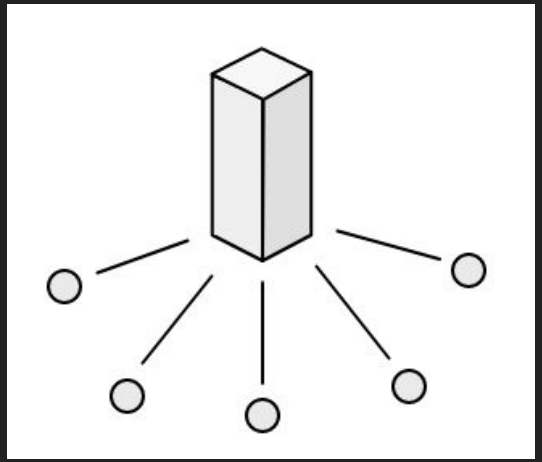
Distributed beginnings

http://som.csudh.edu/fac/lpress/history/arpamaps/

CS&S

@daniellecrobins

Web centralization



Easier to manage and monetize a silo

Image courtesy of Beaker Browser

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

What's important to you?

romana klee

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

CS&S

@daniellecrobins

Preservation starts here

CS&S

@daniellecrobins

Real data preservation workflow

Teon Brooks, PhD

Lab server

Data Steward

Learning from researchers

iconarchive.com

CS&S

@daniellecrobins

Real data preservation workflow

Teon Brooks, PhD

Lab server

Data Steward

Deposit of final version, maybe

iconarchive.com

CS&S

@daniellecrobins

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

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

https://medium.com/@UC3CDL/we-are-talking-loudly-and-no-one-is-listening-a108248693f7 / csv screenshot from https://peerj.com/preprints/2588/

CS&S

@daniellecrobins

Centralized storage is stable

Geir Friestad

CS&S

@daniellecrobins

Is centralized storage stable?

Timothy Vollmer

CS&S

@daniellecrobins

Scholarship =
Online Content Creation

¯_(ツ)_/¯

(sorry)

CS&S

@daniellecrobins

Preservation requires custody

seagen

CS&S

@daniellecrobins

Centralized model requires custody to provide access

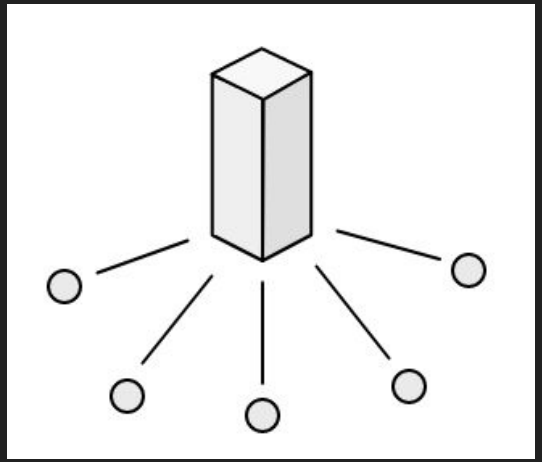


Image courtesy of Beaker Browser

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

https://figshare.com/articles/Preservation_without_possession_Content-addressable_identifiers_for_post-custodial_preservation/5844369

CS&S
@daniellecrobins

What do we need?

FAIR data standards:

Findable
Accessible
Interoperable
Reusable

www.force11.org/group/fairgroup/fairprinciples

CS&S
@daniellecrobins

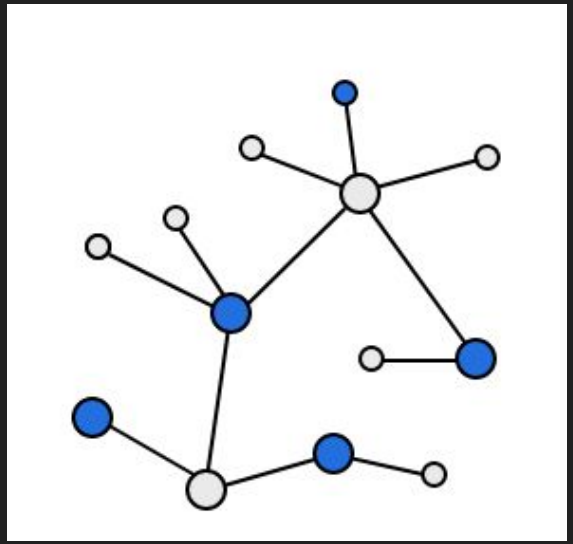
Leverage existing infrastructure

www.force11.org/group/fairgroup/fairprinciples

CS&S

@daniellecrobins

Link trusted institutions



Leverage researcher practices

Image courtesy of Beaker Browser

CS&S

@daniellecrobins

Visions are nice!

Peter Miller

CS&S

@daniellecrobins

Now let’s get real

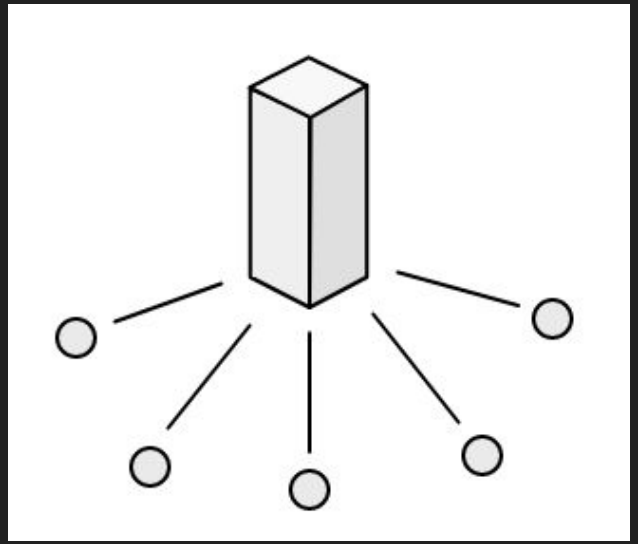
vladeb

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



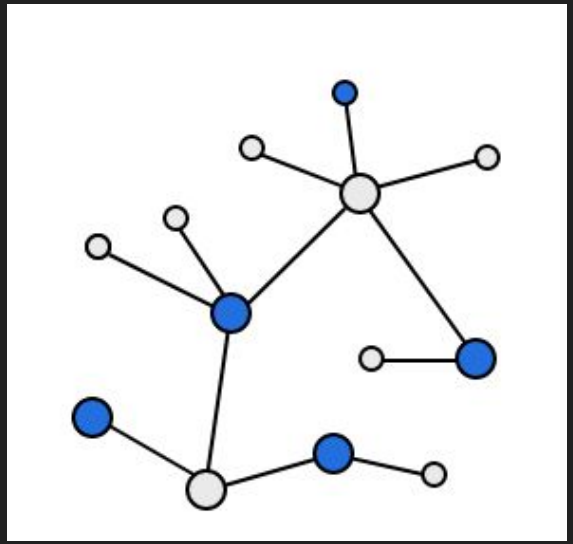
Data stored at central location, accessed by independent users

Image courtesy of Beaker Browser

CS&S

@daniellecrobins

Decentralized models



Data persistently identified,
networked ability to scale

Image courtesy of Beaker Browser

CS&S

@daniellecrobins

Multiple decentralized approaches

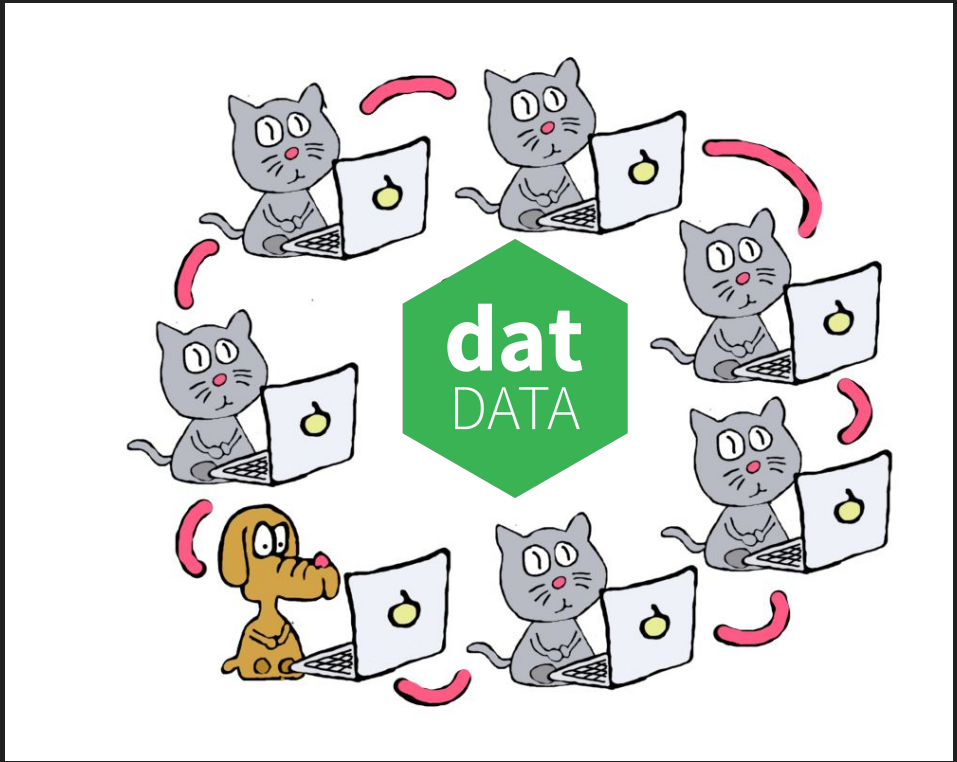
Blockchain

Peer-to-peer

BTC Keychain / Danilo / http://www.ala.org/tools/future/trends/blockchain / https://gist.github.com/mafintosh/bd9e6d350ebf02441c9707c5f799d05b

CS&S
@daniellecrobins

Peer-to-peer public technology



https://github.com/mafintosh/bws-2017

CS&S

@daniellecrobins

The Dat Project

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

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

CS&S

@daniellecrobins

What's Dat?

Persistent identifiers

+

Network of peers

dat
DATA

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

dat DATA

CS&S

@daniellecrobins

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

dat DATA

CS&S

@daniellecrobins

For more on Dat

Details at the white paper
try-dat.com

CS&S
@daniellecrobins

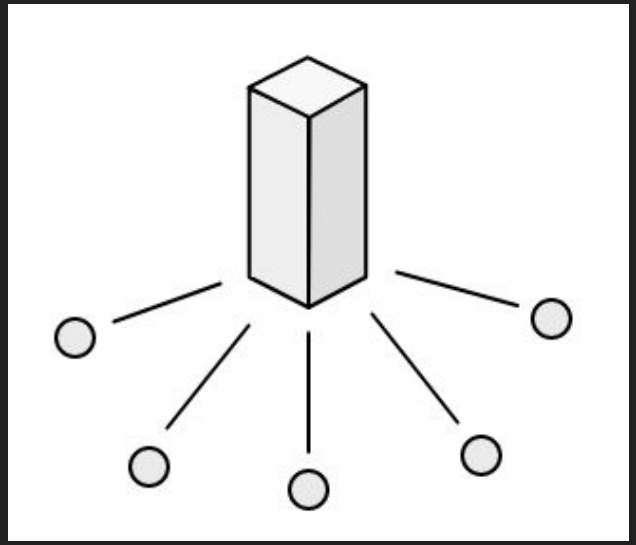
Reimagine data preservation

浪宏 葉

CS&S

@daniellecrobins

It’s all about TRUST



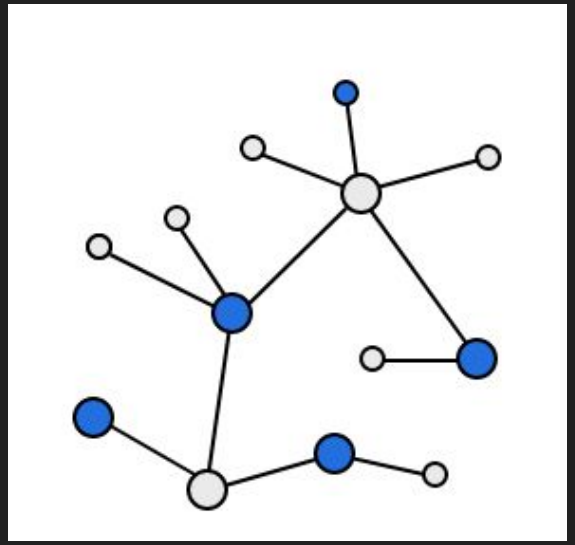
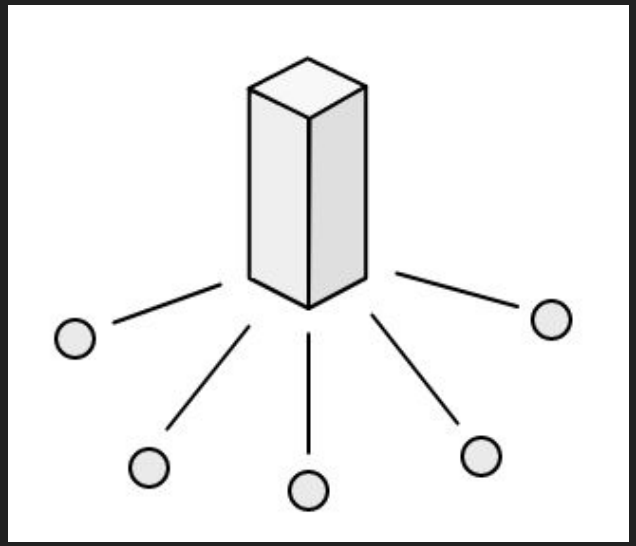


Image courtesy of Beaker Browser

CS&S

@daniellecrobins

… and I trust LIBRARIES



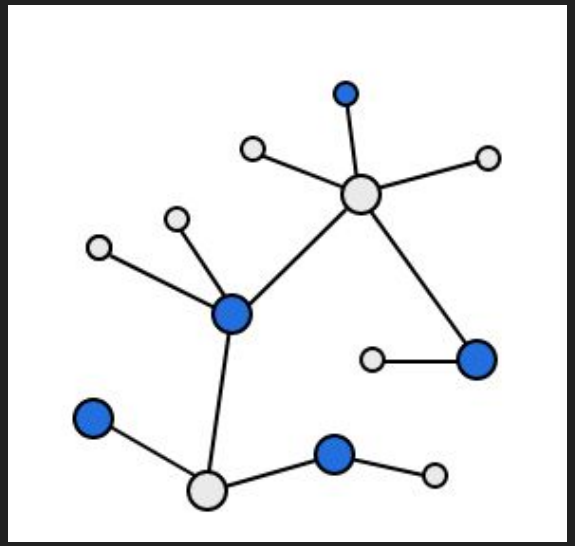


Image courtesy of Beaker Browser

CS&S

@daniellecrobins

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

CS&S

@daniellecrobins

Let’s build it!

Eran Sandler

CS&S

@daniellecrobins

GORDON AND BETTY MOORE FOUNDATION

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

https://blog.datproject.org/tag/science/

CS&S

@daniellecrobins

Automating data preservation

Teon Brooks, PhD

Lab server

Data Steward

Current workflows don’t preserve

iconarchive.com

CS&S

@daniellecrobins

Automating data preservation

Teon Brooks, PhD

Lab server

Data Steward

Researcher may deposit a final copy

iconarchive.com

CS&S

@daniellecrobins

Automating data preservation

Teon Brooks, PhD | Lab server | Data Steward

by leveraging researcher practices

iconarchive.com

CS&S

@daniellecrobins

Automating data preservation

by leveraging researcher practices

iconarchive.com

CS&S

@daniellecrobins

Automating data preservation

by leveraging researcher practices

iconarchive.com

CS&S

@daniellecrobins

Automating data preservation

Teon Brooks, PhD

Lab server

Data Steward

... and avoiding disaster

iconarchive.com

CS&S

@daniellecrobins

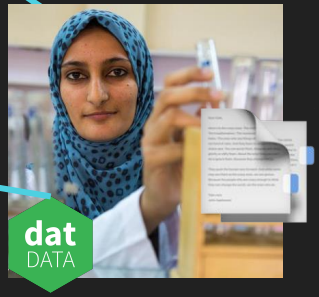
Link people via institutions

iconarchive.com

CS&S

@daniellecrobins

Link people via institutions

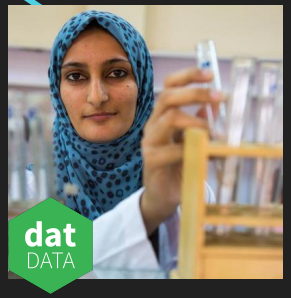


iconarchive.com

CS&S

@daniellecrobins

Link people via institutions

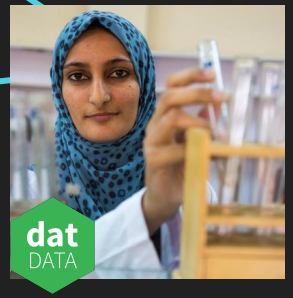


iconarchive.com

CS&S

@daniellecrobins

Data preserved in place

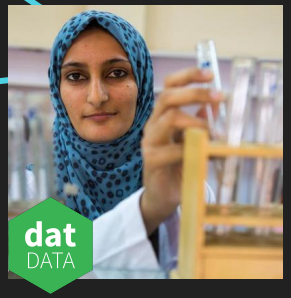


iconarchive.com

CS&S

@daniellecrobins

Data stewards pull from network



iconarchive.com

CS&S

@daniellecrobins

Every institution contributes

Storage, bandwidth

Metadata on their collection

Commitment to preserve their collection

To the network

dat DATA

CS&S

@daniellecrobins

Any user can access

Information on library collections

History of objects

Whole or partial data sets

from the network

dat DATA

CS&S

@daniellecrobins

How can we get there from here?

CS&S

Eddy Josafat Hernández Vega

@daniellecrobins

GORDON AND BETTY MOORE FOUNDATION

We cover more ground together

UC Davis Center for Watershed Sciences (CWS)

CS&S

@daniellecrobins

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

CS&S

@daniellecrobins

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

- Stephen Whitmore (@noffle)

Digital Democracy

www.liveoncelivewild.com

CS&S

@daniellecrobins

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
dat DATA
UC3 UC Curation Center
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-classtitlereference-rot-an-emerging-threat-to-transparency-in-political-sciencediv/54F56CFC2CBE05778130E40CABB2CCC5
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

CS&S

@daniellecrobins