The NIST Research Data Framework and Data Democratization

Julia Lane, NYU and RTI
And many colleagues

Overview

Context

Challenges

Opportunities

Practical Next Steps

RDAF

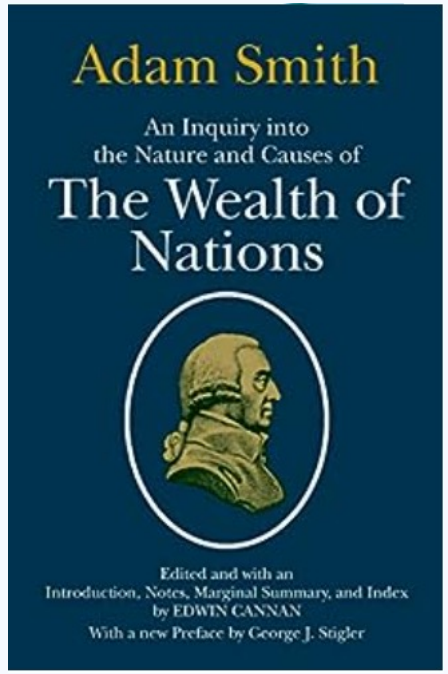
The NIST Research Data Framework (RDaF) is a multifaceted and customizable tool that aims to help shape the future of open data access and research data management (RDM). The RDaF will allow organizations and individual researchers to customize an RDM strategy.

Some thoughts

Research data is largely a public good

"It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest." ~ Adam Smith

“There are two units of academic currency: publications and grants” ~Dan Hamermesh



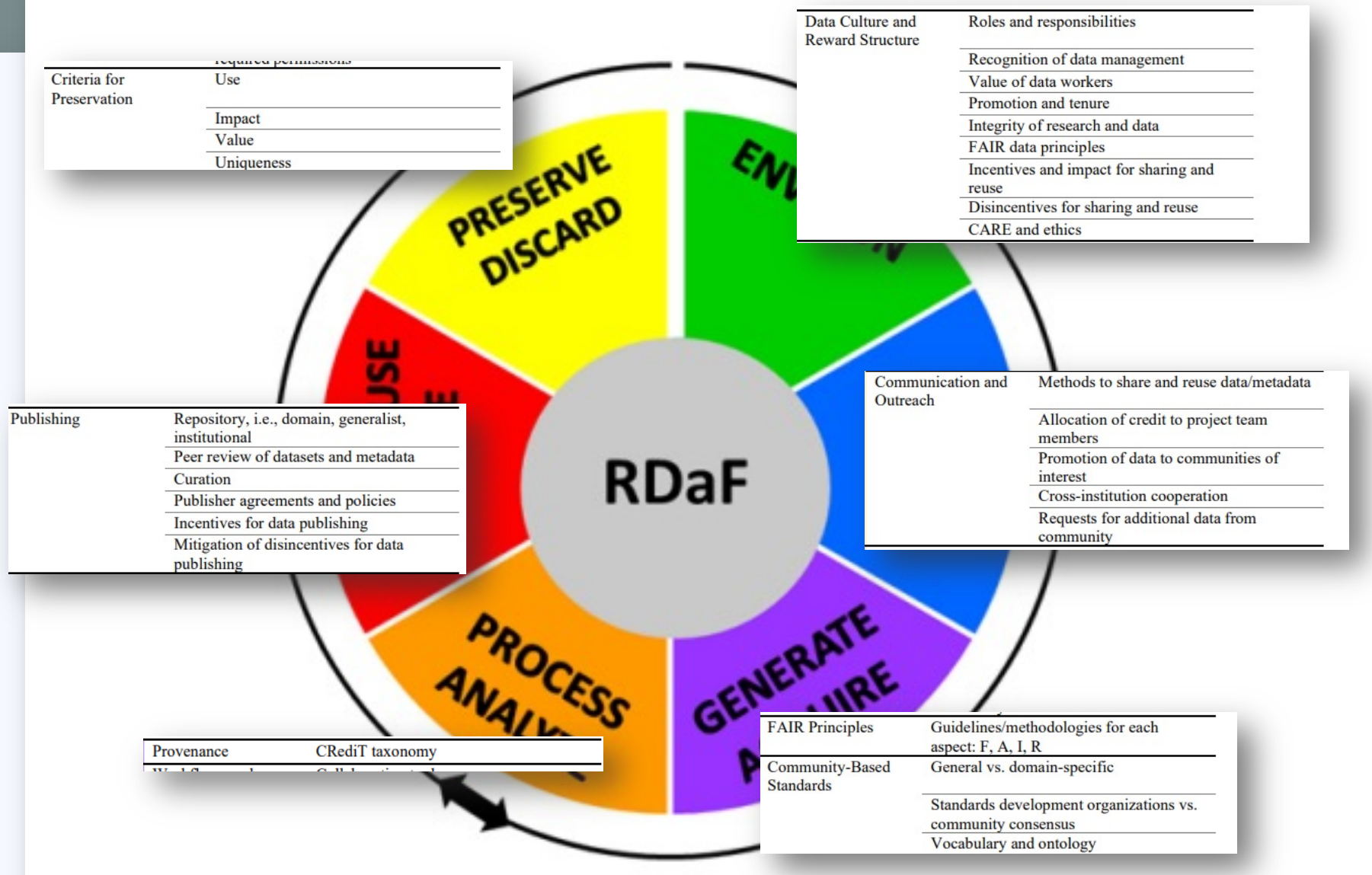


Fig. 2. Research Data Framework Lifecycle Stages

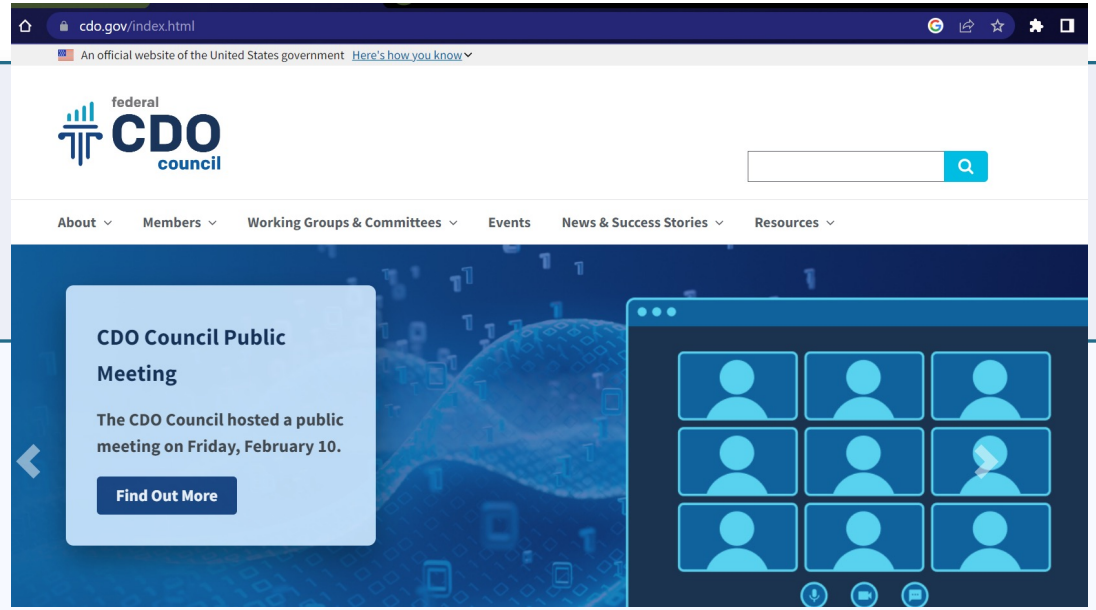
Incentives

Agencies

Researchers

Publishers

Institutions



1. Basic Usage Information

How much are agency datasets used in research and how has that usage changed over time?

How often is each one of an agency's identified dataset used in research and how has that usage changed over time?

2. The Agency's Portfolio

What topics are an agency's datasets being used to study and what publications are associated with each topic?

What topics is each one of an agency's identified dataset used to study in research and what publications are associated with each topic?

What other datasets are being used to study each topic?

3. Drilling Into the Details for Each Dataset

Who are the main authors using each agency's datasets? Who are the main authors using each specific dataset?

What are the publications associated with each author?

What institutions are the centers of use for each agency dataset and in what geographic locations are the institutions located?

SJVN: Of course, there's nothing new about answering programming questions. In a way, Stack Overflow is a direct descendant of Usenet group FAQs such as those on comp.lang.c, which began in the 1980s. Since then, there have been other efforts to answer developer questions. But, you are so much more successful than anyone else. How did you do it?

PC: It's all thanks to the brilliance of our founders, Joel Spolsky, and Jeff Atwood, who created fast, automatic social management tools in 2008. They also brought together a community, and that's where they were really brilliant.

Overview

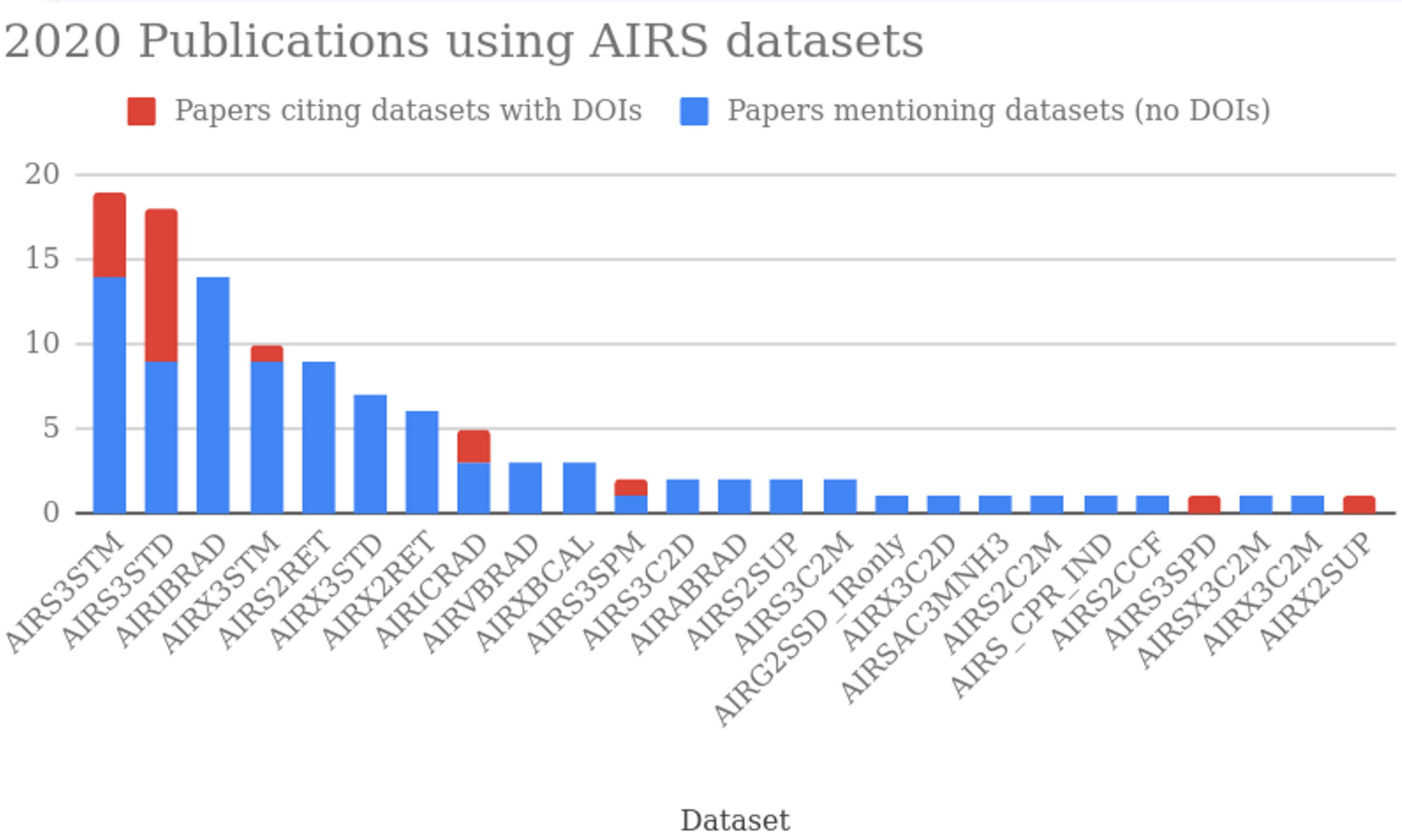
Context

Challenges

Opportunities

Practical Next Steps

How do you get there from here?



100 publications:

- 18 with dataset DOI citations
- 82 manually reviewed
- 10-15 minutes for paper review
- ~14 hours total review time

Or from here?

Research Policy 48 (2019) 1487–1492

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Federal funding of doctoral recipients: Wh

Wan-Ying Chang[a], Wei Cheng[b], Julia Lane[c,*], Bruce

[a] National Science Foundation, United States
[b] School of Business, East China University of Science and Technology, China
[c] New York University, United States
[d] Ohio State University, United States

ARTICLE INFO

JEL classification:
O30
O38
C8
C81

Keywords:
UMETRICS
Linked survey transaction data
Doctoral workforce
Survey of earned doctorates
Research impact

ABSTRACT

This technical note descri
describe the richness and
funding channels can be
university payroll and fin
important US survey data
funding and the doctoral
disciplines and by differen
incorporate more linkages

3. Filling data gaps

There are now new administrative data that can be combined with the Survey of Earned Doctorates to fill the gap. The STAR METRICS project, which was initiated by federal agencies in 2009 in response to the Roadmap findings, was intended to (National Science Board, 2015) provide policymakers with a better understanding of the process of research and (Romer, 1990) provide the research community with a common data infrastructure that connected research funding with research outcomes (Lane et al., 2015). Since it was impossible to collect and link data on all individuals supported by research funding from across federal agencies, the STAR METRICS approach drew the information directly from the research organizations themselves. The key information came from administrative grant records, which contain record level information on wage payments made from federal grants to all university personnel, including doctoral recipients.

The program evolved to be led by universities (and called UMETRICS). It became institutionalized at the Institute for Research on Innovation and Science (IRIS) at the University of Michigan (Lane et al., 2014). It also included information that permitted linkages to Census data, ProQuest dissertations, US Patents, PubMed, and public in-

Table 1
Survey sources of federal funding.

Source of Funding	SED[1]	SED-UMETRICS[2]	Federal SED-UMETRICS[3]
Research assistantship	6117	4006	3410
Fellowship, scholarship	5703	3036	2522
Teaching assistantship	4745	2613	2166
Grant	2534	1494	1239
Missing (did not respond)	2584	1084	852
Traineeship	2054	882	689
Spouse's, partner's, or family's earnings or savings	1712	663	501
Foreign (non-U.S.)	1568	541	399
Personal earnings during graduate school	338	270	231
Loans (from any source)	391	200	164
Personal savings	550	177	135
Employer reimbursement/ assistance	356	163	132
Other	375	117	81
Internship, clinical residency	680	341	268
Other assistantship	5	2	1

Responses to SED Question A5: Which of the following were sources of financial

Failure to align incentives has predictable results

Beginning in the Obama Administration, Agencies have been making datasets available for public use via Data. Gov. The Trump Administration augmented this by prioritizing data sets for AI R&D and those that support healthcare initiatives.

This has grown from a few datasets contributed by each Agency to today's status with over 300,000 datasets that are available in multiple formats, searchable, and tagged with industry protocols.

But just being available, does not mean that the data is "of value"

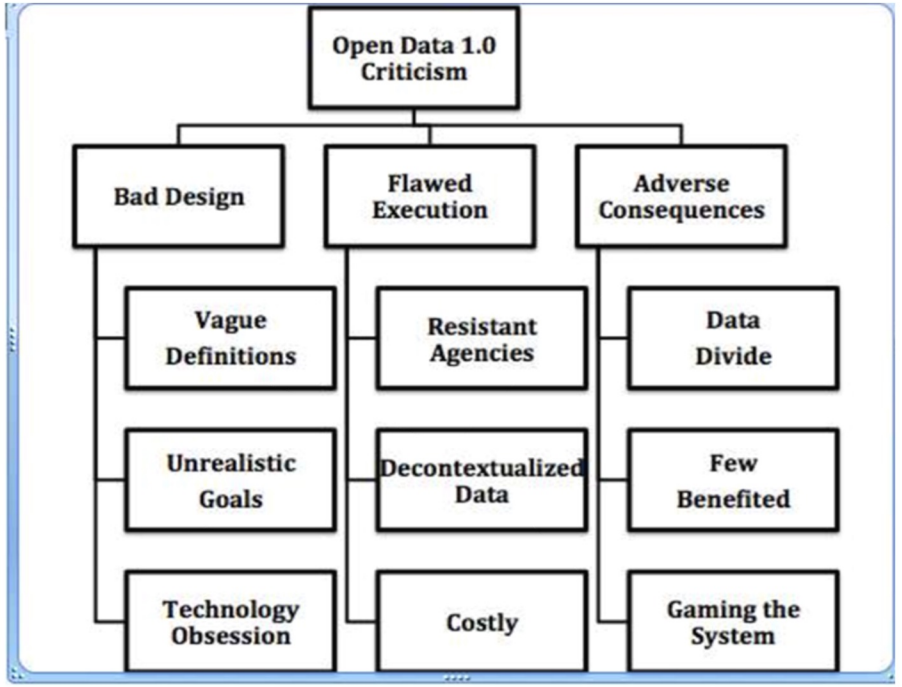


Figure 1: Open Data 1.0 Criticism

Peled, Alon. "Re-designing open data 2.0." *Conference for E-Democracy and Open Government*. 2013.

Opportunities: Evidence Act

Congress's intent for prioritizing evidence building is woven throughout the Evidence Act, including the following:

- The Evidence Act's Title I requires multi-year learning agendas, or evidence-building plans. In addition, Title I includes requirements for analyzing the capacity of federal agencies to engage in evidence-building activities. Agency "capacity assessments" assess agency capacity to support the development and use of evaluation.
- Title II, or the OPEN Government Data Act, establishes that the default for government data is open availability unless otherwise prohibited by law and requires agencies to publish data inventories. Notably, this requirement applies broadly to government data assets to support transparency and has implications and foundational benefits for evidence building across government.

ACDEB
Advisory Committee on Data for Evidence Building

- Title III, or the Confidential Information Protection and Statistical Efficiency Act of 2018 (CIPSEA), requires federal agencies to make data accessible to federal statistical agencies within a strong privacy framework and for those statistical agencies to ensure secure access to confidential data assets and to create a Standard Application Process (SAP).

Public Law 115–435 115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activi- ties improve Federal data management, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

SHORT TITLE.—This Act may be cited as the "Foundations for Evidence-Based Policymaking Act of 2018".

TABLE OF CONTENTS.—The table of contents for this Act is as follows:

Sec. 1. Short title; table of contents.

TITLE I—FEDERAL EVIDENCE–BUILDING ACTIVITIES

Sec. 101. Federal evidence-building activities.

TITLE II—OPEN GOVERNMENT DATA ACT

Sec. 201. Short title.

Sec. 202. OPEN Government data.

is Important
al Agencies?

2 – OPEN Government Data Act:

n with non-Government entities (including
rs, and the public for the purpose of
ata users value and use government data

using public data assets of the agency and
on by publishing on the website of the
asis (not less than annually), information
ssets by non-Government users

Assist the public in expanding the use of public data assets

Opportunities: NAIRR

research community.

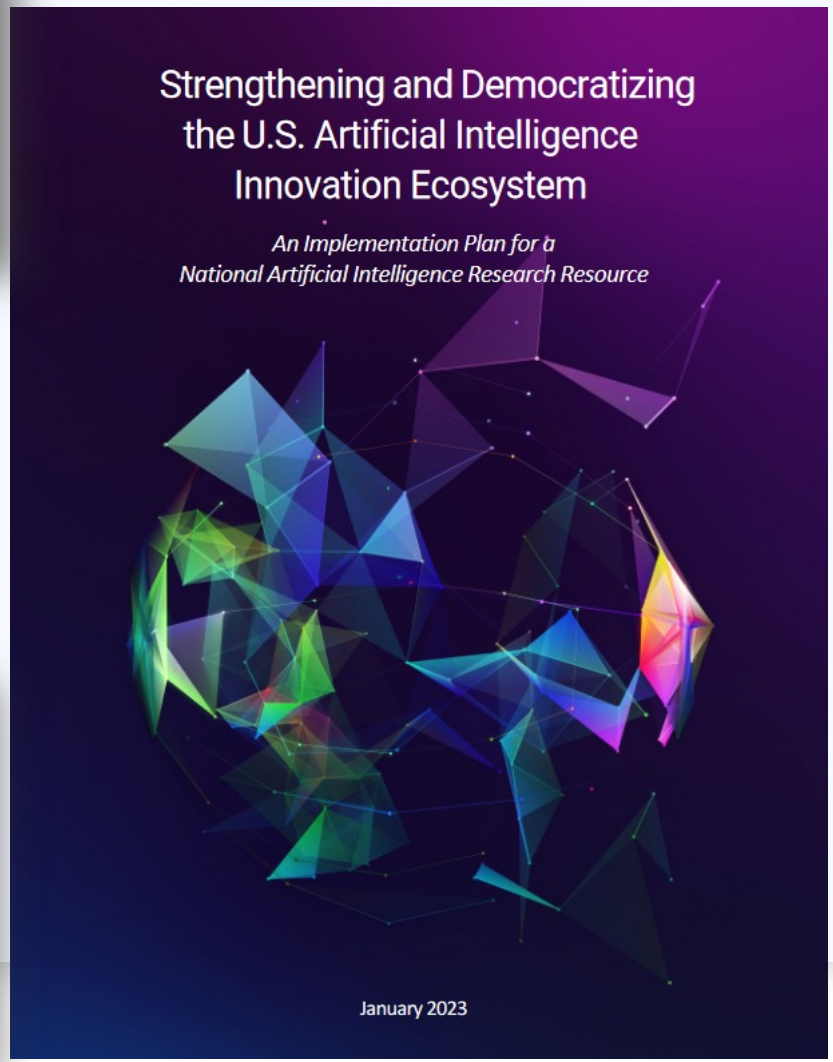
The NAIRR must be broadly accessible to a range of users and provide a platform that can be used for educational and community-building activities in order to lower the barriers to participation in the AI research ecosystem and increase the diversity of AI researchers. The NAIRR access portal and public website should provide catalogs and search and discovery tools to facilitate access to data, testbeds, and educational and training resources serving a range of experience levels.

Access Portal and User Interface

The Operating Entity is responsible for development of an NAIRR user portal that supports key user functionalities such as single sign-on, team allocations, data search and discovery, collaboration tools, resource discovery, job submission, consolidated accounting, spend alerts, information about data use, and cost-optimization of workflows. The portal will be one way to access NAIRR resources. Alternate access methods such as secure shell or scripting interfaces should also be made available for advanced users. The portal will allow users to select their AI applications, computational resources, and data sources from a curated catalog, and to launch and monitor jobs from a portal that provides a uniform, integrated view.

monitor jobs from a portal that provides a uniform, integrated view.

The portal should have built-in help functions and an integrated help desk ticketing system. The portal should maintain an up-to-date catalog of resource provider user documentation and training materials. Chat functions, meeting rooms, forums, and other functionality may be included to support collaboration and community building among students, researchers, resource providers, and other users. The portal should also enable data search and discovery and leverage automated technologies so that (1) metrics on data use can drive data acquisition and (2) diverse, community-driven data curation, linkage, and validation activities can be fostered. A user account would be required to manage computational allocations, monitor usage, submit jobs, and post to the community forum.



Opportunities: CHIPS and Science

National Center for Science and Engineering Statistics

About | Areas of Interest | Surveys & Analysis | Explore Data | Co

- **The National Secure Data Service Demonstration Project**
- Authorizing Legislation
- Oversight and Partnerships
- Privacy and Confidentiality
- The NSDS-D Project and America's Data Hub
- Demonstration Projects +

The National Secure Data Service Demonstration Project

The National Secure Data Service Demonstration (NSDS-D) project is required under the 2022 CHIPS and Science Act to inform a governmentwide effort on strengthening data linkage and data access infrastructure. This effort facilitates statistical activities in support of increased evidence building for the American public. The goal of the NSDS-D project is to inform efforts for developing a shared services model that would streamline and innovate data sharing and linking to enable decision-making at all levels of government and in all sectors.

Overview

Context

Challenges

Opportunities

Practical Next Steps

Researchers

Evidence Act

-Statute

-Recommendations

Common Goals

Combined Approach

Research Policy 48 (2019) 1487–1492

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Federal funding of doctoral recipients: Wh

Wan-Ying Chang[a], Wei Cheng[b], Julia Lane[c,*], Bruce

[a] National Science Foundation, United States
[b] School of Business, East China University of Science and Technology, China
[c] New York University, United States
[d] Ohio State University, United States

ARTICLE INFO

JEL classification:
O30
O38
C8
C81

Keywords:
UMETRICS
Linked survey transaction data
Doctoral workforce
Survey of earned doctorates
Research impact

ABSTRACT

This technical note discr
describe the richness and
funding channels can be
university payroll and fin
important US survey data
funding and the doctoral
disciplines and by differen
incorporate more linkages

3. Filling data gaps

There are now new administrative data that can be combined with the Survey of Earned Doctorates to fill the gap. The STAR METRICS project, which was initiated by federal agencies in 2009 in response to the Roadmap findings, was intended to (National Science Board, 2015) provide policymakers with a better understanding of the process of research and (Romer, 1990) provide the research community with a common data infrastructure that connected research funding with research outcomes (Lane et al., 2015). Since it was impossible to collect and link data on all individuals supported by research funding from across federal agencies, the STAR METRICS approach drew the information directly from the research organizations themselves. The key information came from administrative grant records, which contain record level information on wage payments made from federal grants to all university personnel, including doctoral recipients.

The program evolved to be led by universities (and called UMETRICS). It became institutionalized at the Institute for Research on Innovation and Science (IRIS) at the University of Michigan (Lane et al. 2014). It also included information that permitted linkages to Census data, ProQuest dissertations, US Patents, PubMed, and public in-

Table 1
Survey sources of federal funding.

Source of Funding	SED[1]	SED-UMETRICS[2]	Federal SED-UMETRICS[3]
Research assistantship	6117	4006	3410
Fellowship, scholarship	5703	3036	2522
Teaching assistantship	4745	2613	2166
Grant	2534	1494	1239
Missing (did not respond)	2584	1084	852
Traineeship	2054	882	689
Spouse's, partner's, or family's earnings or savings	1712	663	501
Foreign (non-U.S.)	1568	541	399
Personal earnings during graduate school	338	270	231
Loans (from any source)	391	200	164
Personal savings	550	177	135
Employer reimbursement/ assistance	356	163	132
Other	375	117	81
Internship, clinical residency	680	341	268
Other assistantship	5	2	1

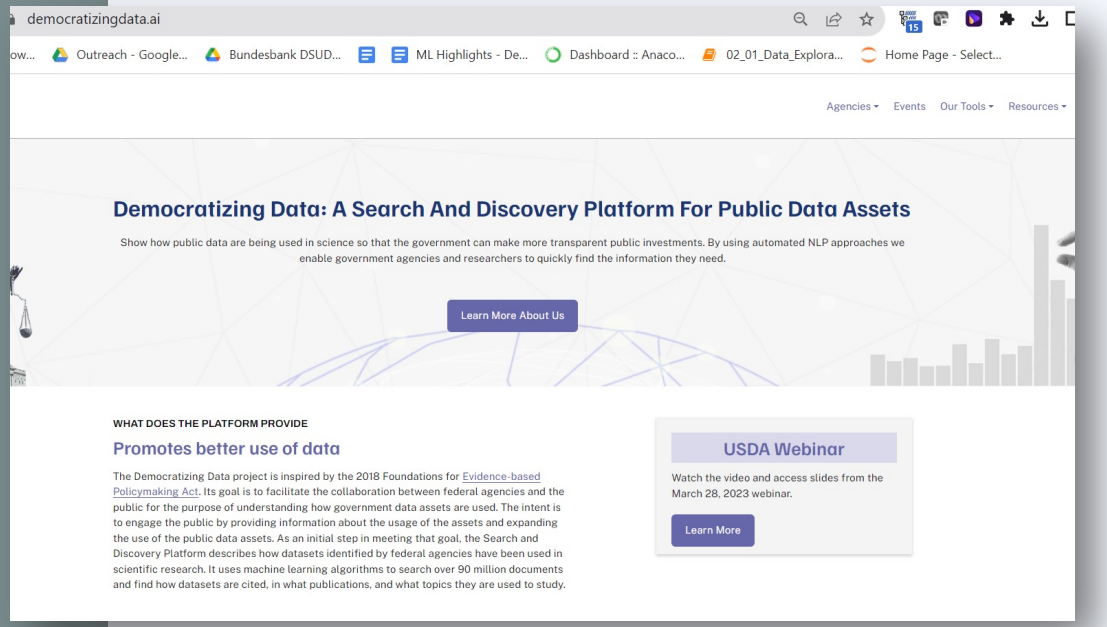
Responses to SED Question A5: Which of the following were sources of financial

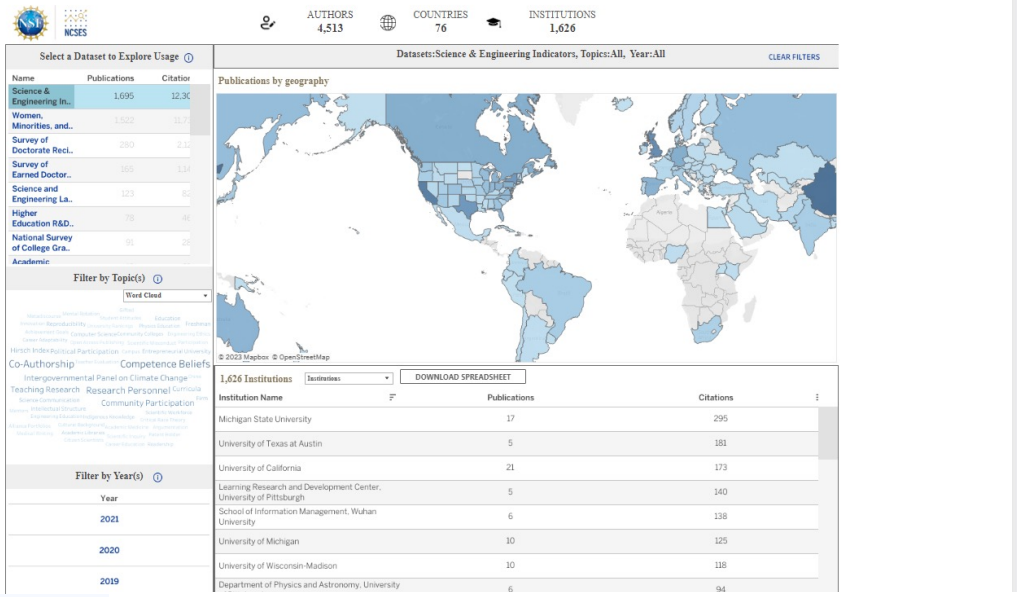
Dashboard

Jupyter Notebooks

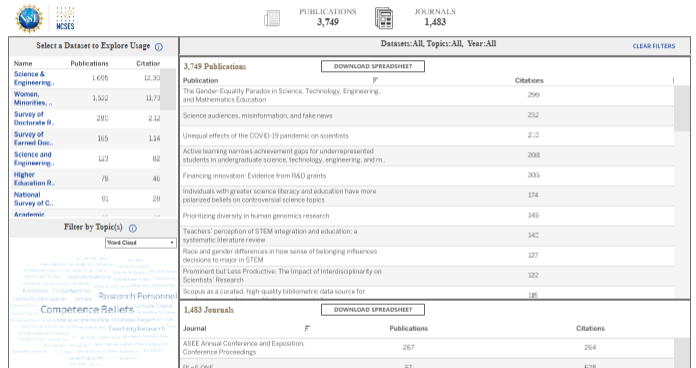
API

Agencies, Researchers, Institutions, and Publishers





NCSES Dashboard: Usage At Publication Level



NCSES Dashboard

https://democratizingdata.ai/agencies/ncses-landing/

67%

NCSES Dashboard

Explore how NCSES data assets are used in published research.

The goal of the Democratizing Data Initiative is to enable different communities to understand how government data assets are used.

One way to build this understanding is through usage dashboards. Each dashboard on the right draws from a database that describes how NCSES data assets are referenced in research publications.

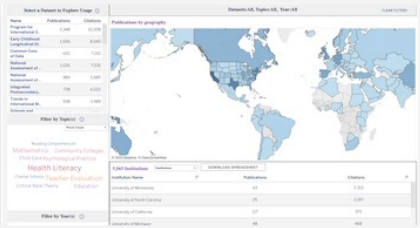
The database links mentions of NCSES data assets in research publications with the research topics of those publications, the publication authors, and their affiliated institutions.

Please note that this pilot project only reflects the information associated with the datasets requested by the agency for the project and is not intended to find all references to all datasets and data assets produced or supported by the agency.

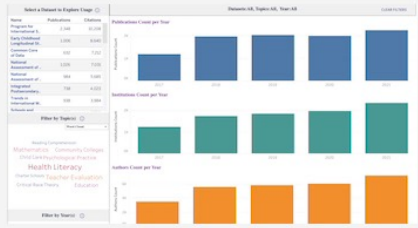
Further exploration:

- For more information about how the database is generated, go to our website (https://democratizingdata.ai) or to the user guide https://soda-umd.gitbook.io/userguide/
- For users interested in exploring the data further, check out the APIs at the link below.
- For our community of users interested in exploring data further, please

View usage at publication level



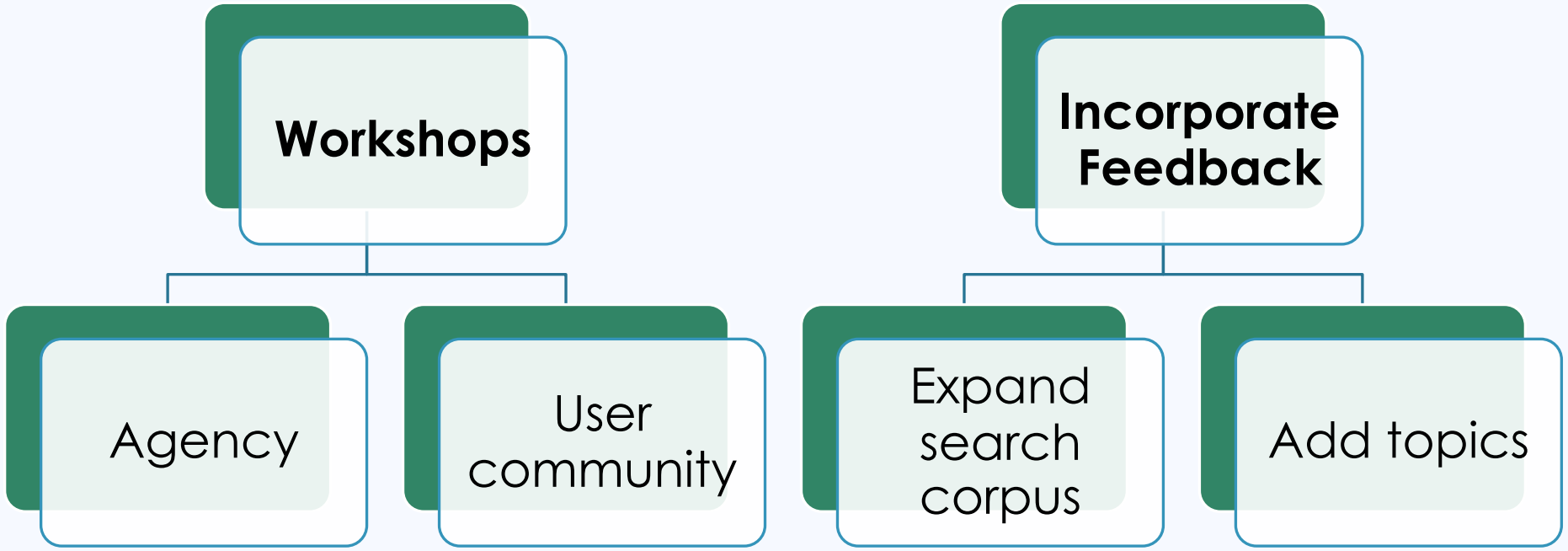
View usage by geography and institution



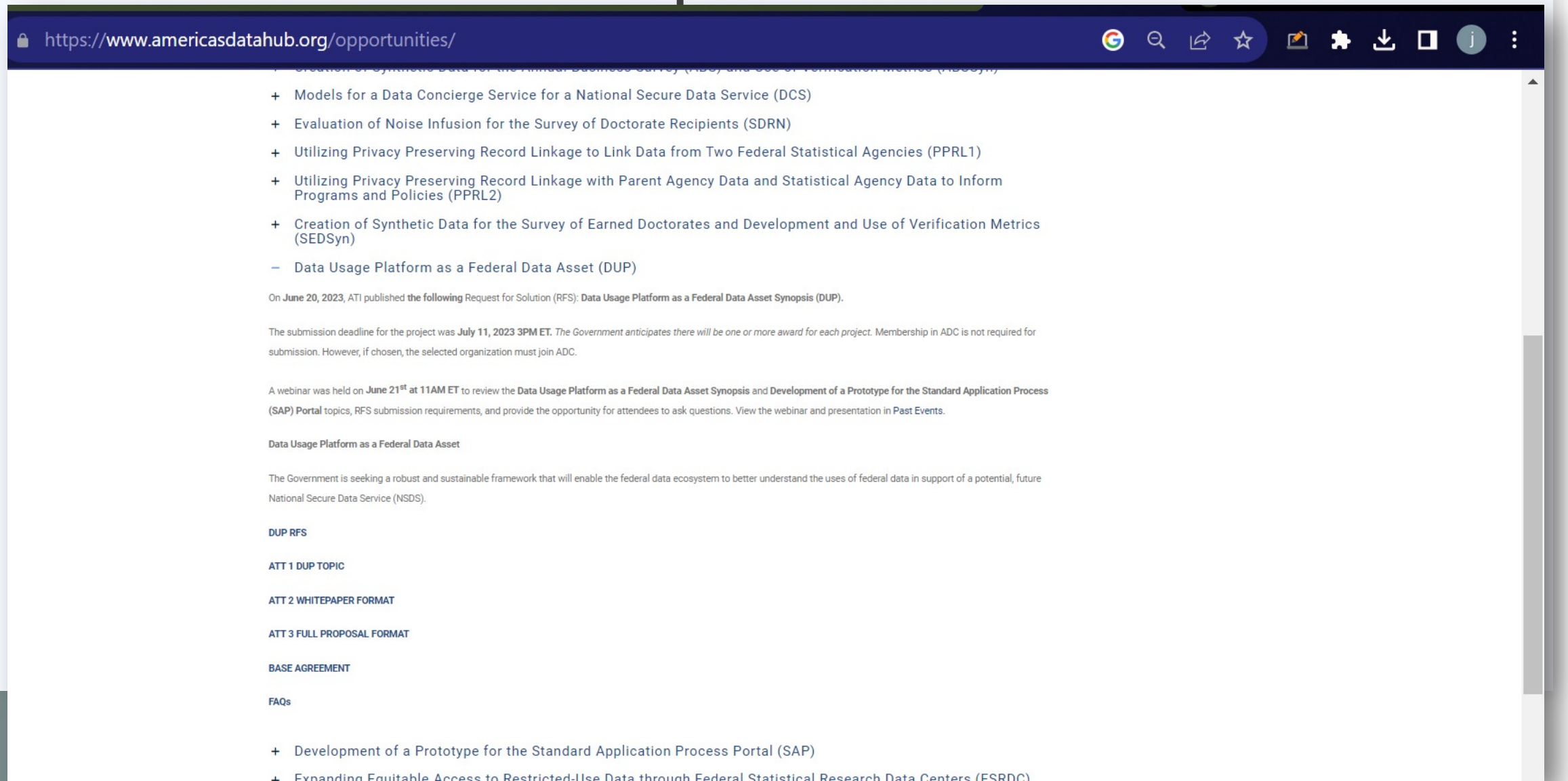
View usage change over time

A	B	C	D	E	F
agency	dataset_title	parent_alias_id	Alias ID	Alias Name	alias_type
NOAA	Sea Level Rise Data	2001	2001	Sea Level Rise Data	Parent
NOAA	Sea Level Rise Data	2001	2002	SLR	Acronym
NOAA	Sea Level Rise Data	2001	2003	Sea Level Rise Wetland Impacts and Mig	Alias
NOAA	Sea Level Rise Data	2001	2004	Sea Level Rise Viewer	Alias
NOAA	Global Forecast System	2005	2005	Global Forecast System	Parent
NOAA	Global Forecast System	2005	2006	GFS	Acronym
NOAA	Global Forecast System	2005	2007	GFS model	Acronym
NOAA	Global Forecast System	2005	2008	GDAS	Acronym
NOAA	Stock SMART	2009	2009	Stock SMART	Parent
NOAA	Stock SMART	2009	2010	Fish Stock Assessment	Alias
NOAA	Stock SMART	2009	2011	Fish Stock Status	Alias
NOAA	Stock SMART	2009	2012	Fish Stock Management, Assessments & Re	Alias
NOAA	Special Weather Statement	2013	2013	Special Weather Statement	Parent
NOAA	Special Weather Statement	2013	2014	Hazardous Weather Outlook	Alias
NOAA	Special Weather Statement	2013	2015	SAME event code SPS	Alias
NOAA	Special Weather Statement	2013	2016	SPS	Acronym
NOAA	Special Weather Statement	2013	2017	SPSs	Acronym
NOAA	Integrated Water Vapor Data	2018	2018	Integrated Water Vapor Data	Parent
NOAA	Integrated Water Vapor Data	2018	2019	Atmospheric River Data	Alias
NOAA	Integrated Water Vapor Data	2018	2020	IVV	Acronym

Design Incentives



Build an incentive based platform



Data Usage Platform as a Federal Data Asset – Objective

To research and develop a robust and sustainable framework that will enable the federal data ecosystem to better understand the uses of federal data in support of a potential, future National Secure Data Service (NSDS).

Related work includes the dashboard prototype efforts of the Democratizing Data Initiative, which promotes the use of federal data and assets for evidence building (https://democratizingdata.ai/).

This project will produce possibilities for a future, state-of-the- art, updatable publicly accessible platform that provides information on federal data usage as part of the National Secure Data Service demonstration project.

Data Usage Platform as a Federal Data Asset - Background

The Democratizing Data Initiative is a cross-agency, multidisciplinary effort to demonstrate the value of statistical data through aggregated usage statistics that are displayed on dashboards and for use in other tools.

- The Advisory Committee on Data for Evidence Building recommended the development and application of usage statistics to respond to the future needs of a National Secure Data Service (https://www.bea.gov/system/files/2022-10/acdeb-year-2-report.pdf).
- Building on prototype efforts, this project aims to explore alternative dashboard technologies, approaches, techniques, and methods to meet the sustainability and transparency needs of the federal statistical system and to inform the efforts of the National Secure Data Service Demonstration Project as required by the 2022 CHIPS and Science Act.

Data Usage Platform as a Federal Data Asset - Information Gaps

What is user feedback on the Democratizing Data pilot dashboards and tools? What information do users report is useful in the current dashboard and what additional information would be useful?

What open data software, data science, best practices, and other cutting-edge technologies can be used to generate aggregated usage data that meet the quality and sustainability requirements of federal agencies?

What additional types of federal data, features and functions of the dashboard or other interface are required to support evidence-building use cases for a wide variety of stakeholders, federal agencies, and consumers of federal statistics? Does this vary by sector, geographies, federal agencies, and user groups?

Do these usage statistics offer a vision for understanding the health and progress of the federal statistical system?

What are the resources required for flexibility, sustainability and scalability of the dashboard platform and related tools?

How can these usage data be leveraged in a profile to inform tiered data access needs, transparency priorities, and open data requirements for federal agencies?

How can these usage data be used as a shared service to inform public trust of official statistics, and engagement in a possible, future National Secure Data Service (NSDS)?

Data Usage Platform as a Federal Data Asset – Project Objectives

1. Identify sustainable, state of the art, updatable solutions, tools, or frameworks to ingest, aggregate and deploy usage statistics. Approaches may include but are not limited to data science techniques and tools such as machine learning and natural language processing.

2. Identify, interview, and document use cases across sectors and subgroups to ensure usability and accessibility of usage data for collaboration, evidence building, and other purposes.

3. Identify alternative data, formats, and tools (e.g. APIs) to present and complement usage data.

4. Develop, refine, and/or enhance dashboard interface, functions, and features to encourage use across various stakeholder communities.

Data Usage Platform as a Federal Data Asset – Project Objectives (continued)

5. Identify workflow processes and other mechanisms to ensure efficient, quality, timely and up-to-date usage data, and information.

6. Implement technologies, techniques, and processes to build a prototype dashboard platform and tools using publicly available federal data and information.

7. Identify approaches and best practices to implement data quality flags to communicate reliability and fitness for use of usage statistics.

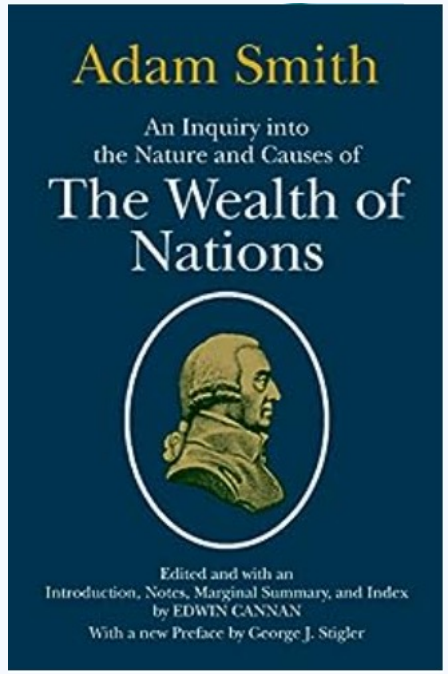
8. Building on findings and lessons learned, advise on the integration of the platform and tools within the federal data ecosystem and possible, future NSDS.

Some thoughts

Research data is largely a public good

"It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest." ~ Adam Smith

“There are ~~two~~ THREE units of academic currency: publications, datasets, and grants”
~Dan Hamermesh and RDAF



Questions?

Julia Lane

Julia.lane@nyu.edu

https://www.linkedin.com/in/julia-lane/