Status and Trajectory of Supporting Tools and Systems

Julia Lane, Paco Nathan and Sophie Rand

The Coleridge Initiative

al Science

BIPARTISAN POLICY CENTER

Foundations for Evidence-Based Policymaking Act of 2018

The bipartisan Foundations for Evidence-Based Policymaking Act of 2018 builds off the work of the U.S. Commission on Evidence-Based Policymaking to strengthen data privacy protections, improve secure access to data, and enhance the federal government's capacity for producing and using evidence.

Strengthens Privacy Protections

Maintains Strong Confidentiality Protections for Sensitive Data. Reauthorizes the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), an existing law that gives the American public strong privacy safeguards and legal protections for appropriate uses of confidential data.

Institutes Processes to Assess Data Risks. Strengthens efforts to protect confidentiality while making data accessible for evidence building and transparent to the public by requiring comprehensive risk assessments for certain publicly released data.

Makes Administrative Records Available for Ev Building. Under a strong set of confidentiality pro encourages that government data can and should generate evidence about policies and programs, u otherwise restricted by law.

Creates a Common Portal for Researcher Appl Access Restricted Data. Reduces burden on res applying to access government data by establishi application system for qualified individuals to acc confidential data for approved projects.

Facilitates Continuous Feedback about Data C Promotes the use of data for evidence building by government advisory committee to review existing and availability of data.

Enhances Government's Evidence

those datasets.

Learn more at bipartisanpolicy.org/evidence

nd Engineering

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

DATA SCIENCE FOR UNDERGRADUATES

Agency Actions

	Foundations for Evidence-Based Policymaking Act of 2018 and Associated OMB Guidance	Executive Order on Maintaining American Leadership in Artificial Intelligence	Improving Implementation of the Information Quality Act (M-19-15)
1. Identify Data Needs to Answer	✓		✓
vernance Body	✓		✓
astructure Maturity			✓
rease Staff Data Skills	✓		
for Agency	✓	✓	✓
entories	✓		

ving tricky to fund and host.

Elena Semenova 9:09 PM
HI DOC data gurus! Do you know what the following indicates in reality? A person admitted first time in >= 2008 year with no previous incarcerations for lower offence class (1-3) being in jail for a few days but has sentence and custody dates goes back >=10 years. Does it mean that he/she was hiding from law enforcement all those years? How does custody date could go back like that in such situations? Is it just a bad data?

Vivek Ananda 11:27 PM
It mostly is bad data please email me the doc number so we can verify in the system

Experience

a. Biologists used the myGrid and Taverna workbench tools to conduct bioinformatics experiments and found that "manually, the processes undertaken by the workflows developed here could take at least 2 days, while the workflows achieve the same output in approximately an hour" (Stevens et al., 2004).
b. A group at the Otto-von-Guericke-Universität Magdeburg working on Diffusion Tensor Imagining imaging for functional MRI (fMRI) used the Grid Workflow Execution System (GWES), "a grid implementation of the algorithm with slice based parallelization," which, when applied to their image processing tasks, reduced "the processing down to 10% compared to a local cluster and 20% compared to sequential processing on the grid" (Solomonides, 2009).
c. Another group, also in neuroimaging, used the SHIWA Platform to create "integrated workflow systems [that] have been shown to reduce experiment times from weeks to minutes" (Terstvanszky et al., 2014).

XSEDE | USER PORTAL
Extreme Science and Engineering Discovery Environment
Search XSEDE...
SIGN IN
MY XSEDE | RESOURCES | DOCUMENTATION | ALLOCATIONS | TRAINING | USER FORUMS | HELP | ABOUT
Get Started | Manage Data | User Guides | Community Codes | News | Usage Policy | Knowledge Base | MFA

Top of page
t is XSEDE? ›
EDE New User Webinar
lore XSEDE Resources
your Campus Champion
ate a User Portal Account ›
age Your XSEDE Profile ›
an Allocation ›
n to your Allocated Resources ›
sfer your Files ›
puting Environment ›

Getting Started with XSEDE

Last update: October 31, 2019

What is XSEDE?

The Extreme Science and Engineering Discovery Environment (XSEDE) is a virtual collaboration funded by the National Science Foundation that facilitates free, customized access to advanced digital resources, consulting, training, and mentorship. XSEDE helps the nation's most creative minds discover breakthroughs and solutions for some of the world's greatest scientific challenges.

XSEDE's virtual cyberinfrastructure allows scientists to interactively share computing resources, data, and expertise. XSEDE resources may be broadly categorized as follows: High Performance Computing, High Throughput Computing, Visualization, Storage, and Data Services. Many resources provide overlapping functionality across categories.

Take a look at this brief introductory video "What is XSEDE?".

XSEDE What Is XSEDE?

NeuroImaging Tools & Resources Collaboratory (NITRC) Project: This resource for neuroinformatics software and data enables researchers to share software and data. Researchers using this resource saw a reduction in costs of 25%, and a reduction in time to process neuroimaging data by 85%. Users stated that having "access to several projects data in one central place.... saved us several weeks, and helped [them] select the best tools and conduct research of better quality."

SOC ARXIV SocArXiv Papers

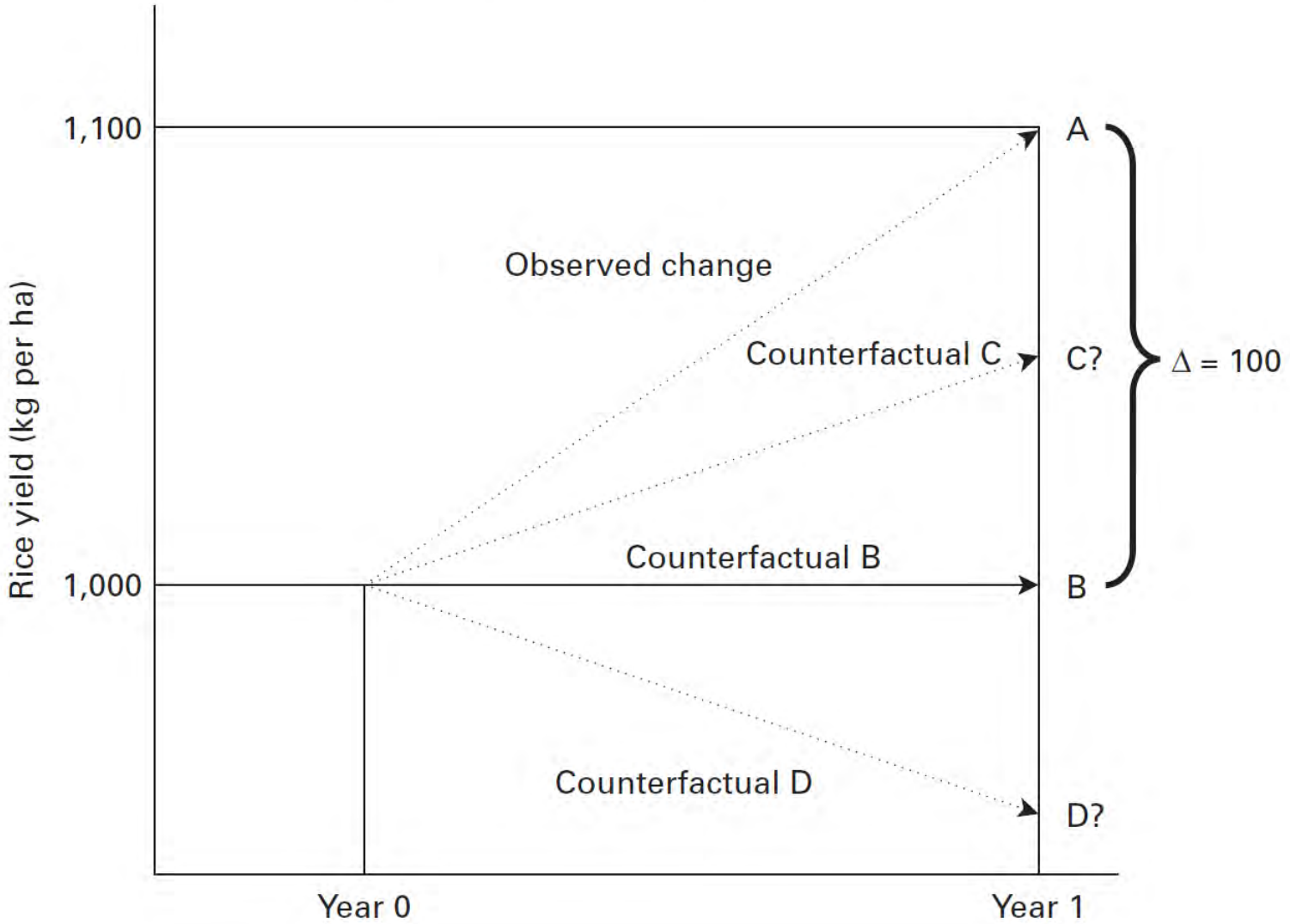
Enhancing and accelerating social science via automation: Challenges and opportunities

AUTHORS
Tal Yarkoni, Dean Eckles, James Heathers, Margaret Levenstein, Paul Smaldino, Julia Lane

SUBMITTED ON
January 24, 2019

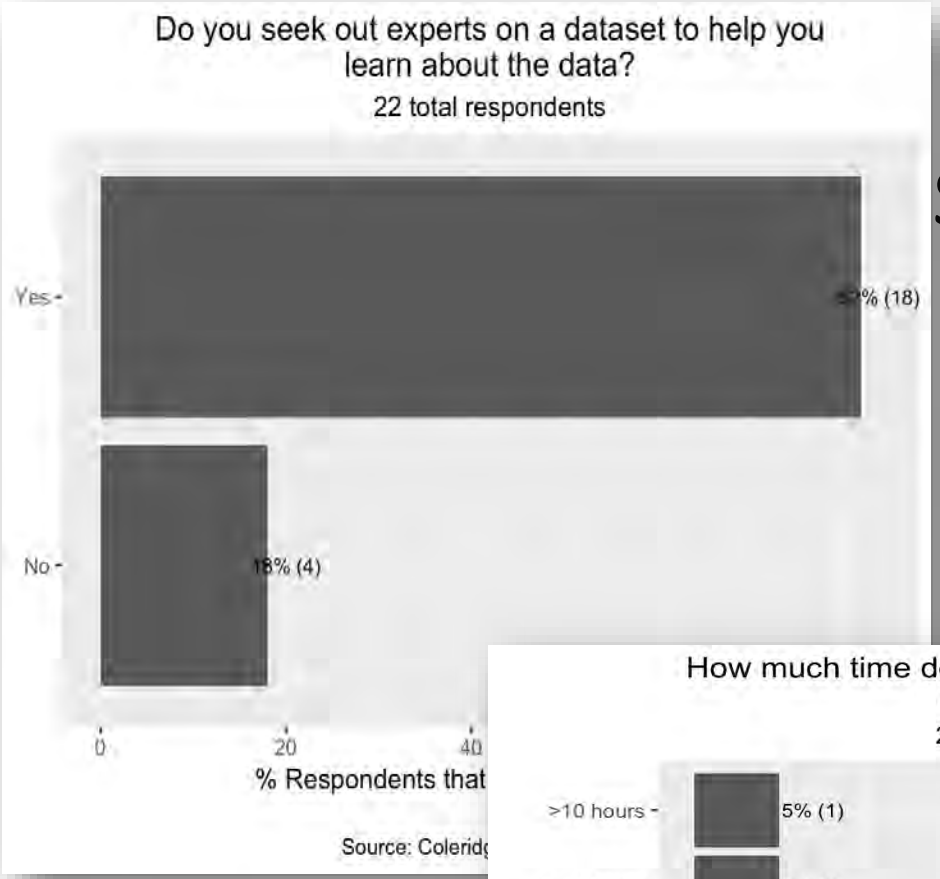
LAST EDITED
January 25, 2019

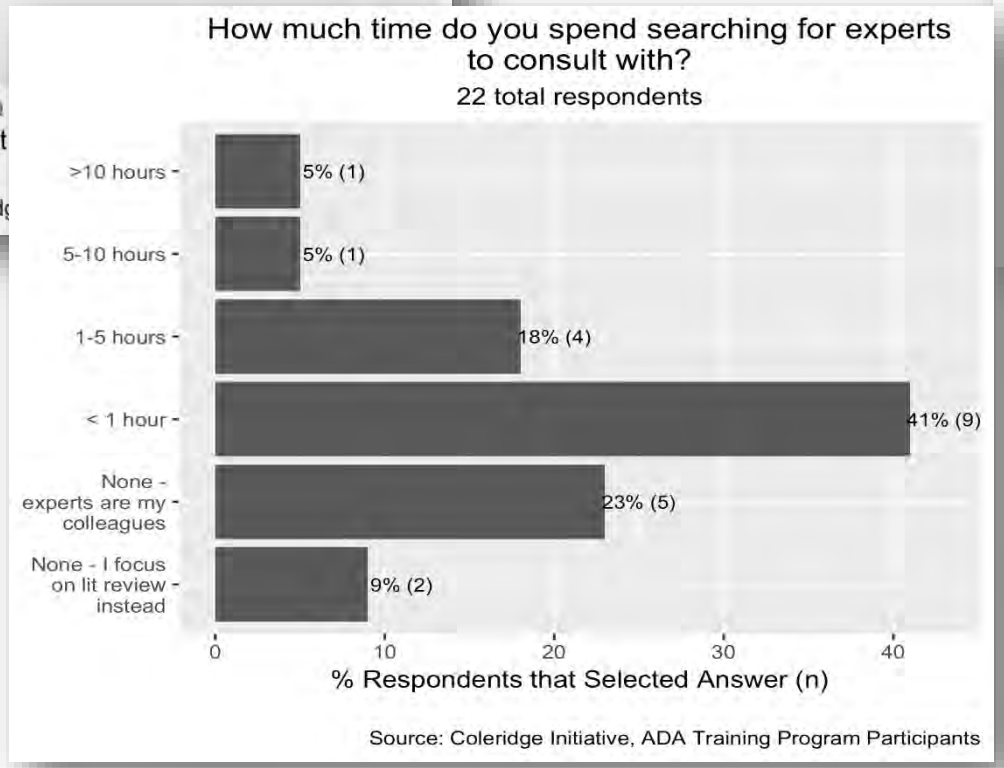
Figure 3.3 Before-and-After Estimates of a Microfinance Program

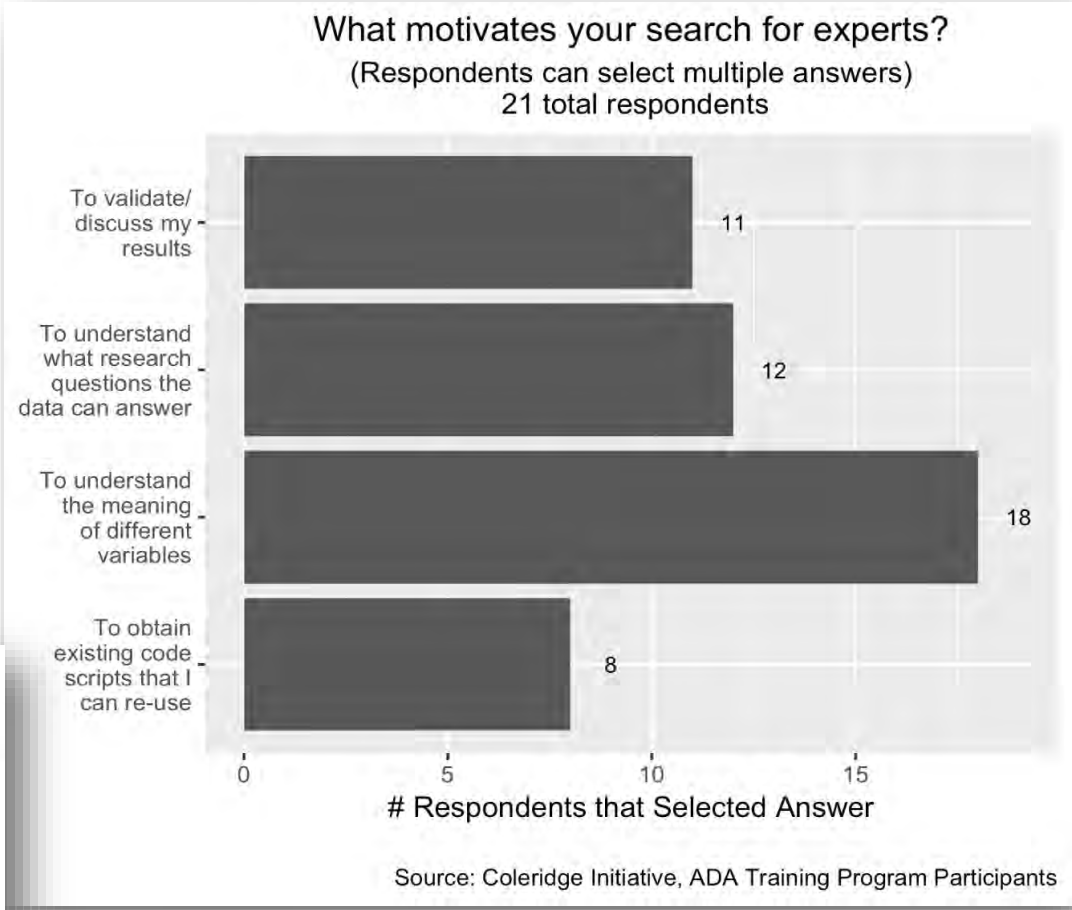


Note: Δ = Change in rice yield (kg); ha = hectares; kg = kilograms.

state?


Do you seek out experts on a dataset to help you learn about the data?
22 total respondents
Yes
No
18% (4)
0
20
40
% Respondents that
Source: Coleridg


How much time do you spend searching for experts to consult with?
22 total respondents
>10 hours
5% (1)
5-10 hours
5% (1)
1-5 hours
18% (4)
< 1 hour
41% (9)
None - experts are my colleagues
23% (5)
None - I focus on lit review instead
9% (2)
0
10
20
30
40
% Respondents that Selected Answer (n)
Source: Coleridge Initiative, ADA Training Program Participants

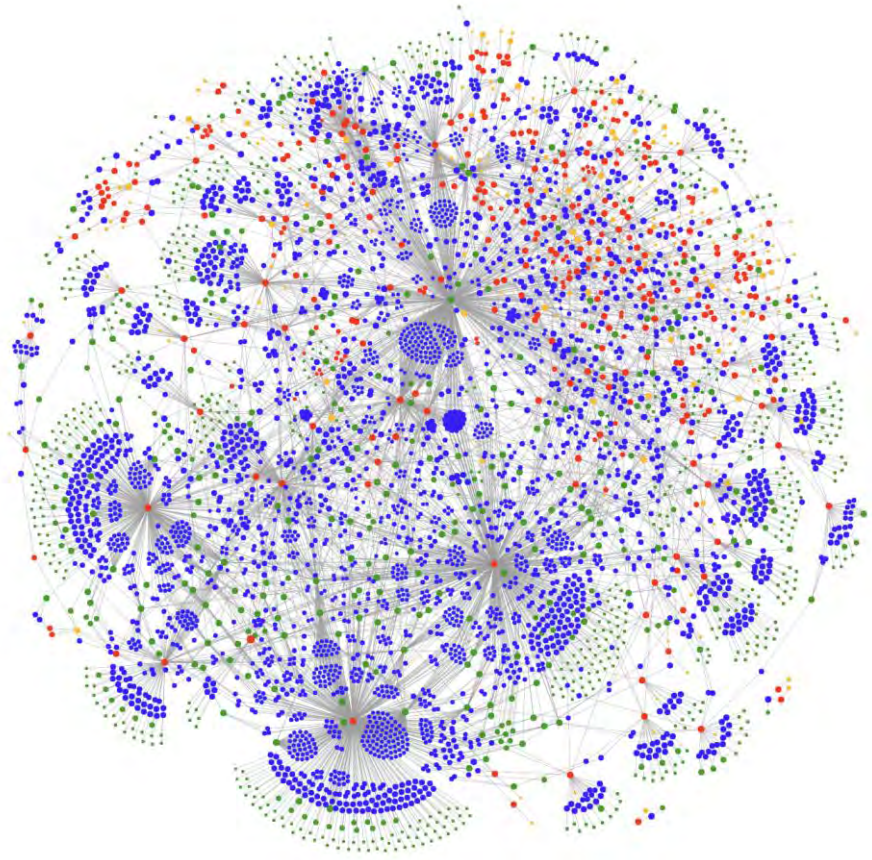

What motivates your search for experts?
(Respondents can select multiple answers)
21 total respondents
To validate/ discuss my results
11
To understand what research questions the data can answer
12
To understand the meaning of different variables
18
To obtain existing code scripts that I can re-use
8
0
5
10
15
Respondents that Selected Answer
Source: Coleridge Initiative, ADA Training Program Participants

Rich Context

The goal of the Rich Context project is to create a new platform that enables empirical analysts to search for and discover datasets.

The **challenge** that empirical researchers face is that, for a given dataset, it is difficult to find out **who** has worked with the data before, **what methods and code** were used, and **what results were produced**.

Leaderboard competition
github.com/Coleridge-Initiative/rclc



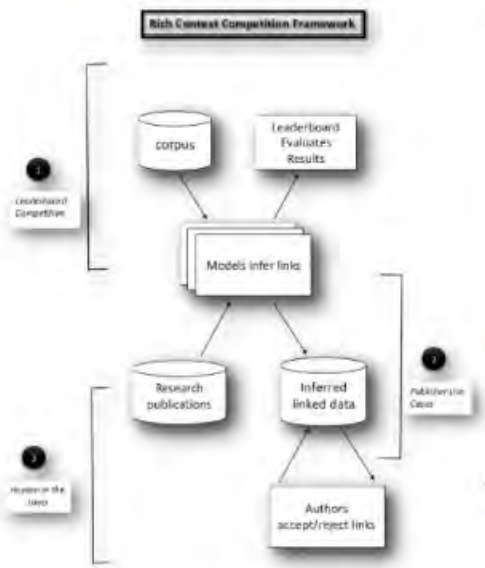
See also:

“Human-in-the-loop AI for scholarly infrastructure”

“New initiative to help with discovery of dataset use in scholarly work”, Christian Zimmerman

1. A Python library **richcontext-scholapi** which provides API integrations for federating exchange across multiple scholarly infrastructure providers. Additional APIs are cons

2. A **Knowledge Graph** of known links between datasets, research publications, resea sciences.



3. A **leaderboard competition** i models to automatically detec publications.

   Model development relies on a research publications and the The corpus continues to expa which are validated by domain

   Teams use a subset of the *cor* models. Using another subset and improve their models. Ou performance; scores are upda model versions. Read more at Competition Wiki.

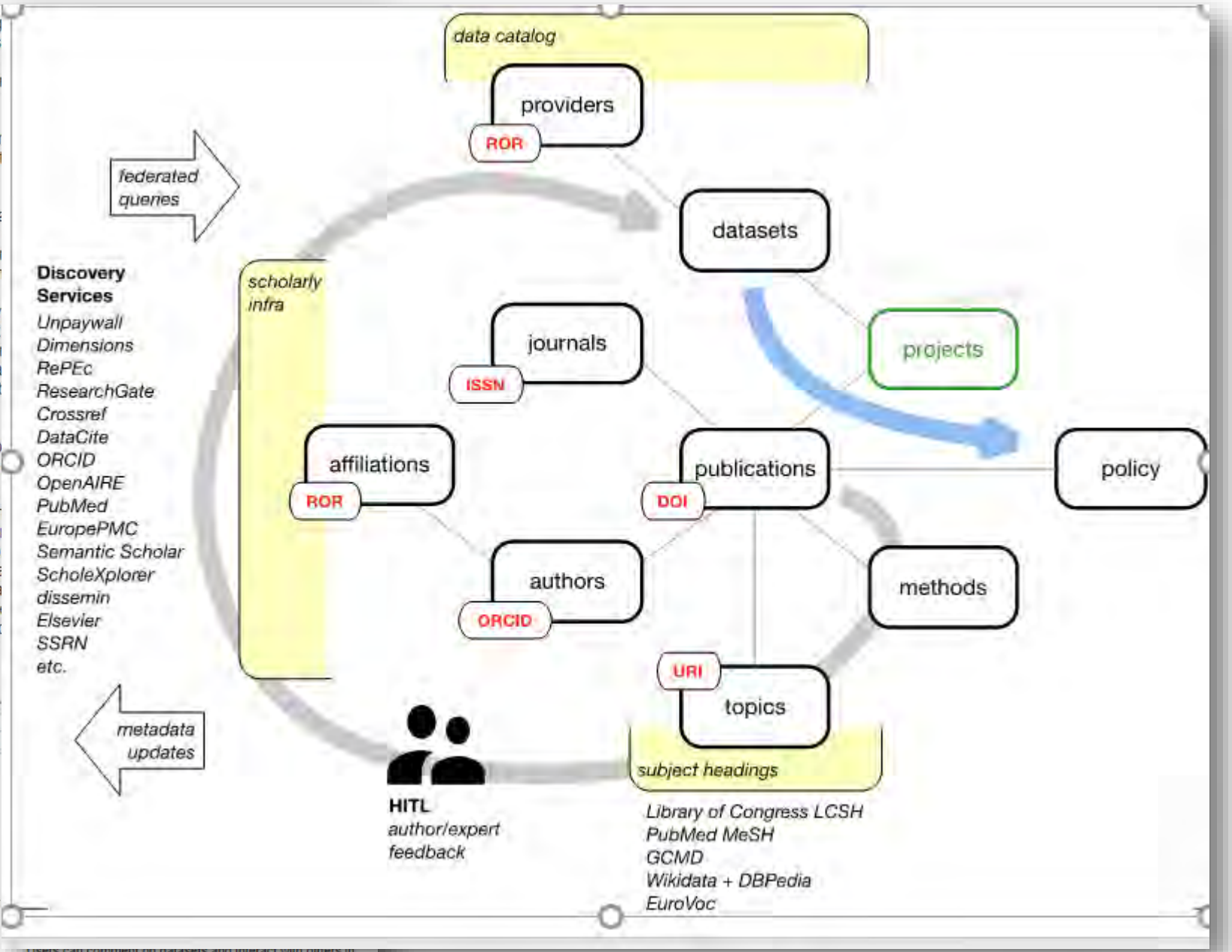
   Teams are also contributing to extraction. The competition is

4. A set of Jupyter extensions wh Features like the **Data Regist Commenting** enable users to browse metadata on their data collaborate in real-time on sha goal is to enrich the knowledg comments, collaboration and

and developing a recommendation engine for datasets and code.

Registry (left) and Metadata Browser (right) in terLab. *Project Jupyter*

Demo - Linked Data generated by user appears in Jupyter Data Registry. *Project Jupyter*



Result

Rich Context in Secure Facility

Managing, Monitoring, and Measuring the way datasets are used in research projects.

MANAGING

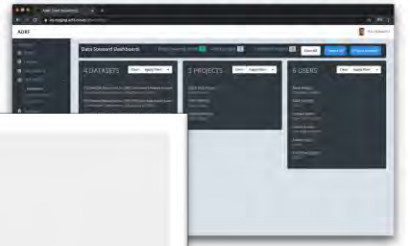
Research Project Request Workflow

Researchers can request access to datasets via the project request workflow. The appropriate Data Stewards are notified when their datasets are requested. They can then review and approve access.

MONITORING

Projects and Dataset Access

The Data Stewardship Dashboard gives Data Stewards a high-level overview as to which of their datasets researchers have access to as a part of their projects.

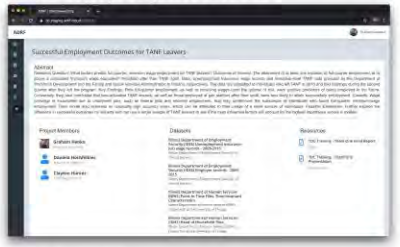


ll also provide high-level metrics on which
d by whom so that Data Providers can get
e most value.

MEASURING

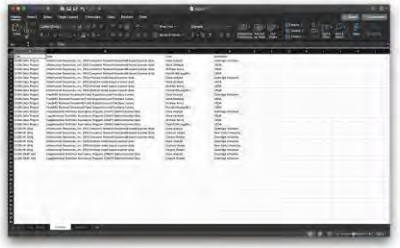
Research Papers and Presentations

Once a research project has finished, data products such as papers or presentations can be uploaded and associated with the datasets used, providing context and knowledge sharing with future researchers.



Export Metrics

Dataset usage information can be exported from the application so that Data Stewards can generate reports to share amongst key stakeholders at their agency.



https://coleridgeinitiative.org

Rich Context as an open platform

rc.coleridgeinitiative.org/?radius=2&entity=IRI%20Infoscan

IRI Infoscan radius: 2

datasets:
- Food Security Survey Module
- IRI Infoscan
- IRI Consumer Network

providers:
- IRI

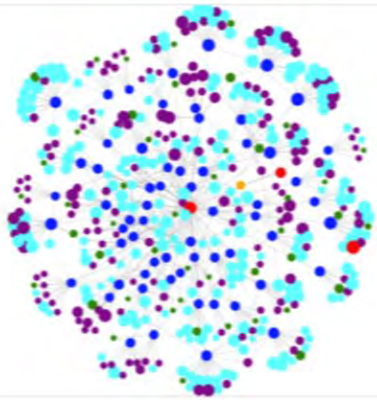
publications:
- WIC and the Retail Price of Infant Formula
- Spatial and Temporal Variation in the Value of the Women, In...
- Tobacco Marketing at SNAP- and WIC-Authorized Retail Food St...
- Household food security: Perceptions, behavior and nutrition...

authors:
- Gundersen, Craig
- Laska, Melissa Nelson
- Powell, Lisa M.
- Yen, Steven T.
- Smallwood, David M.
- Wu, Ping-Chao
- Zhen, Cheng

journals:
- Food Policy
- Nutr. Rev.
- J Health Econ
- Am J Agric Econ
- Prev Med Rep
- Health Econ
- SSRN Electronic Journal

topics:
- children
- food insecurity
- women
- wic
- food
- food security
- special supplemental nutrition program

Network Links



©2020, Coleridge Initiative • Feedback • Research • Configure • OpenAPI • ML Competition

Contact and more information

- Website: https://coleridgeinitiative.org/
- White paper: https://tinyurl.com/rcwp2019
- Book: http://34.82.145.119:8080/
- Email
  - dataanalytics@coleridgeinitiative.org
  - julia.lane@nyu.edu
- GitHub organization: https://github.com/Coleridge-Initiative