Searching…
for How Data Have Been Used:
Intuitive Labels for Data Search & Discovery

Emilda B. Rivers, NCSES Director

May 22, 2024

Democratizing Data: Discovering Data Use and Value for Research and Policy

NATIONAL CENTER FOR SCIENCE AND ENGINEERING STATISTICS
SOCIAL, BEHAVIORAL AND ECONOMIC SCIENCES DIRECTORATE
NATIONAL SCIENCE FOUNDATION

Today's Presentation

- NCSES & The Democratizing Data Pilot Initiative
- Searching & Labeling for Data Discovery
- Common Classification Frameworks to Generate Intuitive Data Labels
- Future Considerations

NCSES & The Democratizing Data Pilot Initiative

NCSES provides data for key insights into the U.S. science and engineering enterprise often in a global context

Science & Engineering Workforce

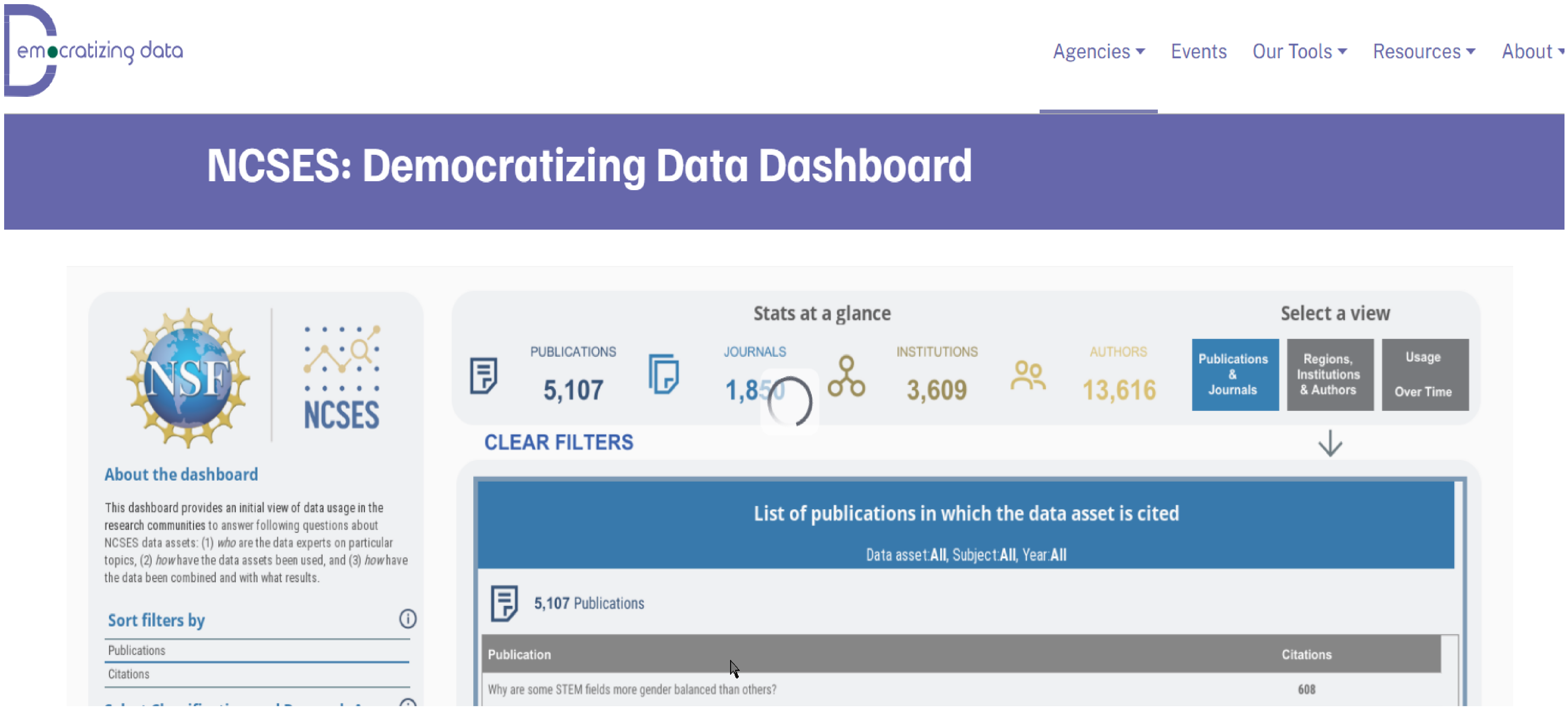
STEM Education

Innovation & Global Competitiveness

Research & Development

Government Funding for S&E

Higher Ed R&D



NCSES: Democratizing Data Dashboard

About the dashboard

This dashboard provides an initial view of data usage in the research communities to answer following questions about NCSES data assets: (1) *who* are the data experts on particular topics, (2) *how* have the data assets been used, and (3) *how* have the data been combined and with what results.

Sort filters by

- Publications
- Citations

List of publications in which the data asset is cited

Data asset:**All**, Subject:**All**, Year:**All**

5,107 Publications

Publication	Citations
Why are some STEM fields more gender balanced than others?	608

Robust metadata needed to understand data usage

The Evidence Act authorizes agencies to promote evidence building with federal data.

What is required to do this effectively?

- Produce statistics on product use
- Find and apply for use of restricted data assets (Standard Application Process)
- Identify information gaps for data producers and users using salient metadata (labels).

What is the metadata problem?

Metadata are essential but perennial challenge within federal data ecosystem

Agencies prioritize resources towards high-quality data often at expense of utility

No government-wide common standards or ontologies to categorize data

We need good metadata to support data usage statistics

Democratizing Data Pilot: NCSES Case Study

- Investigate the creation of intuitive labels that align with agency mission areas and emerging technologies

- Three approaches tested
  - All Science Journal Classification
  - Science-Metrix Classification
  - NCSES Taxonomy of Disciplines

- These approaches were insufficient in identifying NCSES data usage in emerging technologies and our mission areas

Searching & Labeling for Data Discovery

If only it was this easy…

Metadata are labels. Critical components to decision making.

- **Promote transparency, discoverability and governance**

Understand and store your ingredients

- **Facilitate data access and data use**

Retrieve your ingredients to cook your meal

- **Enable reproducibility**

Store the recipe for future use by you and others

Common Classification Frameworks to Generate Intuitive Data Labels

The Publisher Approach

Publishers and research analysts have developed journal- and publication-level taxonomies.

Create ontologies that serve a different function that label data sets to organize and grow a corpus of scientific literature.

How do you map metadata on authors and subjects to agency missions and research areas?

The Federal Statistical Agency Approach

Research classification mappings are typically agency specific.

Rule-based to create classifications through manual or machine-learning methods.

No commonly accepted federal standard for mapping data sets to either agency missions or research fields

Community-Driven Approach

- Classify (label) data sets by how they are used by researchers and their use of common terms to describe the work that they are doing
    - Published works that use federal data can be analyzed and grouped to signal information about a “topic”
    - Create classification systems based on these topics
    - Use a people-based framework based on evolving terminologies and topics across fields over time.

A Labeling Approach for the Search & Discovery Platform

Machine Learning to Create Labels: An approach

- Conceptualize technology areas as research fields and emphasize points of similarity

- Classify data sets into different mission areas or research fields, using the terms
  - by the publications in which they are found and
  - by the authors who do research in the specific areas

- Provides a solid basis for designing, training, validating, and explaining a classification model to generate labels

Artificial Intelligence as an Example Topic

Include terms that are closely related to AI and people working in AI

The advantage is that most common terms that AI researchers use can change rapidly while the people doing the work are more consistent

Future Considerations

Creating useful metadata for a Democratizing Data platform requires consistent attention

The people and areas of research will change over time. Taxonomies need to be **flexible**.

There is **no ‘one-size-fits-all**’ approach to developing data set labels.

Emerging research/technologies push us to consider a search approach that **uses people and terms** to label fields instead of an approach based solely on terms.

What is a future agenda?

- Use natural language processing, semantic analysis, and machine learning for a variety of use cases
- Develop different sets of metadata topic labels for different needs

Involve communities at all stages of the classification process

https://ncses.nsf.gov

@NCSESgov