The MSD-LIVE Data and Computational Platform

MSD LIVE

Casey Burleyson, Carina Lansing, Zoë Guillen, Matthew Macduff, Devin McAllester, and Jon Weers

---

1. Facilitating Open Science with MSD-LIVE

Journals largely skipped to this end of the pyramid...

MSD-LIVE is about tackling these foundational elements of the pyramid...

- Policy — Make it required
- Incentives — Make it rewarding
- Communities — Make it normative
- User Experience — Make it easy
- Infrastructure — Make it possible

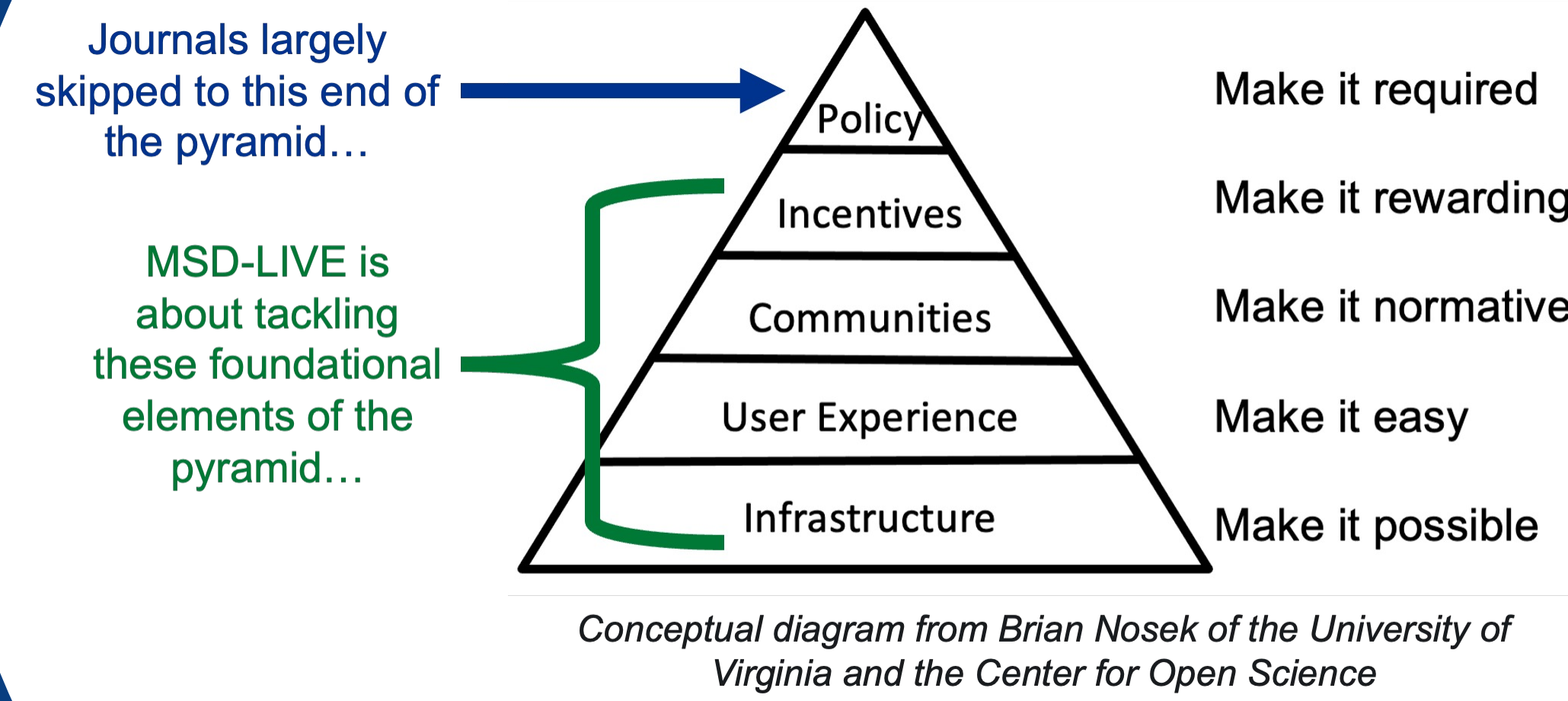*Conceptual diagram from Brian Nosek of the University of Virginia and the Center for Open Science*

MSD-LIVE is providing the foundational infrastructure to support open MSD science.
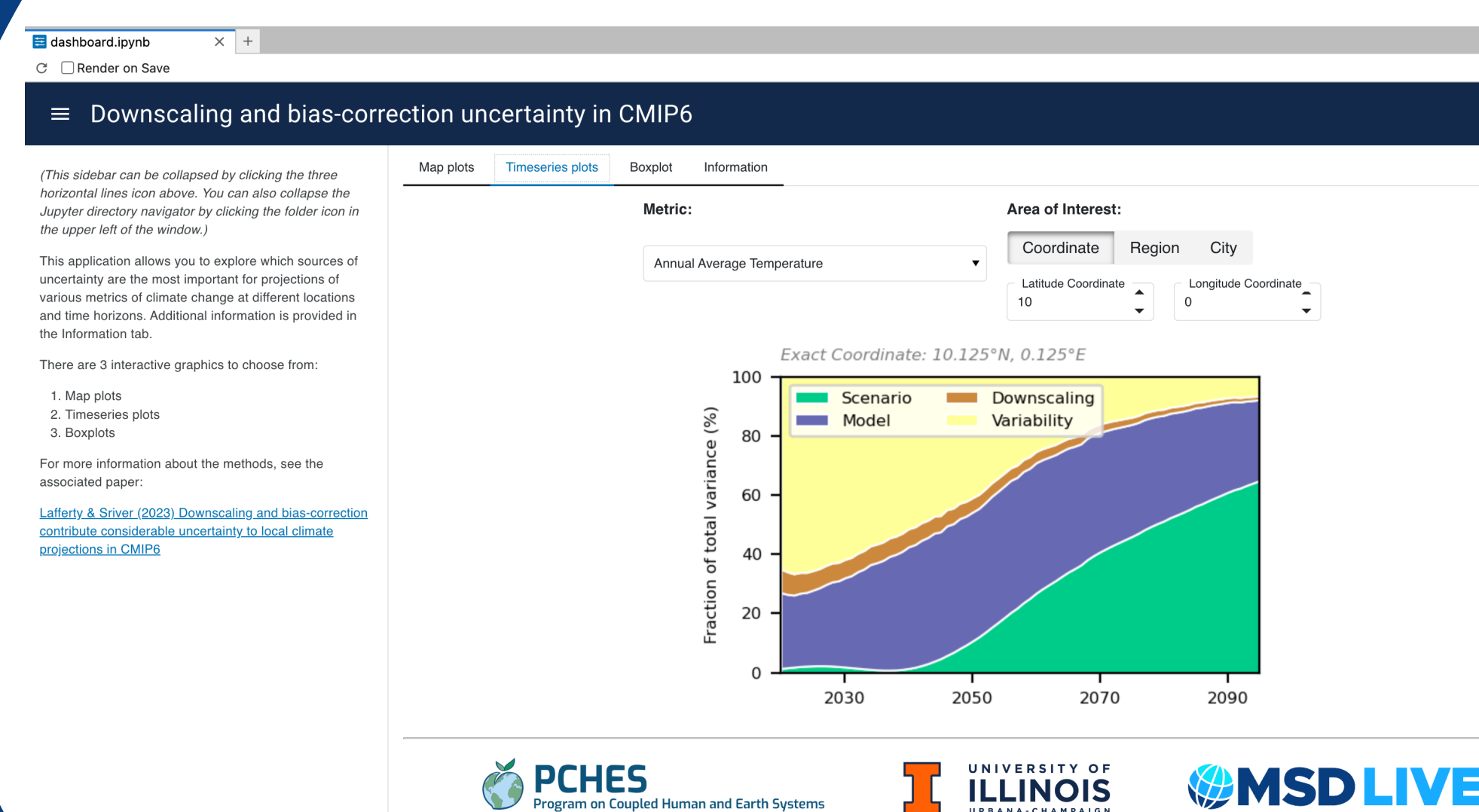
---

2. The Vision for MSD-LIVE

A cloud-based data management system and advanced computing platform that will enable researchers to document and archive their data, run their models and analysis tools, and share data, software, and multi-model workflows.

https://msdlive.org

---

3. Use Cases

Name	Description
Find Data	Quickly and easily find datasets produced by other users and projects
Archive Data	Permanently archive small (<250 MB), medium (250 MB - 50 GB), and large (50 GB to 20 TB) final-form datasets and generate data Digital Object Identifiers (DOIs) in order to meet journal requirements for data sharing
Version Control	Use an intuitive web-based user interface to document and share versioned datasets and associate data with the code used to produce it
Training	Train new team members on MSD projects to effectively manage data and code and capture the institutional knowledge of members that leave a project
Control Access	Create and manage teams that cross institutions in order to quickly and easily grant access to data and code without having to obtain multiple sets of institutional credentials
Share Data	Share working datasets across multiple institutions collaborating on a project in real-time
Analyze Data	Create, run, and share Jupyter notebooks to analyze or visualize datasets in MSD-LIVE
Multi-Model Workflows	Create, execute, document, and publish multi-model workflows where component models run on computational resources at different institutions
ML/AI	Virtually assemble data lakes and execute ML/AI algorithms on large pools of data that may be physically located in different places
Move Code to the Data	Easily deploy models to run on new computational resources and make it easy for users to bring their code to the data when transferring data is infeasible

MSD-LIVE is built to solve a series of ten community-inspired use cases for managing data and code.

---

4. Usage Statistics

- 9 projects
- 113 users from 15 states and 4 countries
- 39 published datasets
- 187+ Tb of data

---
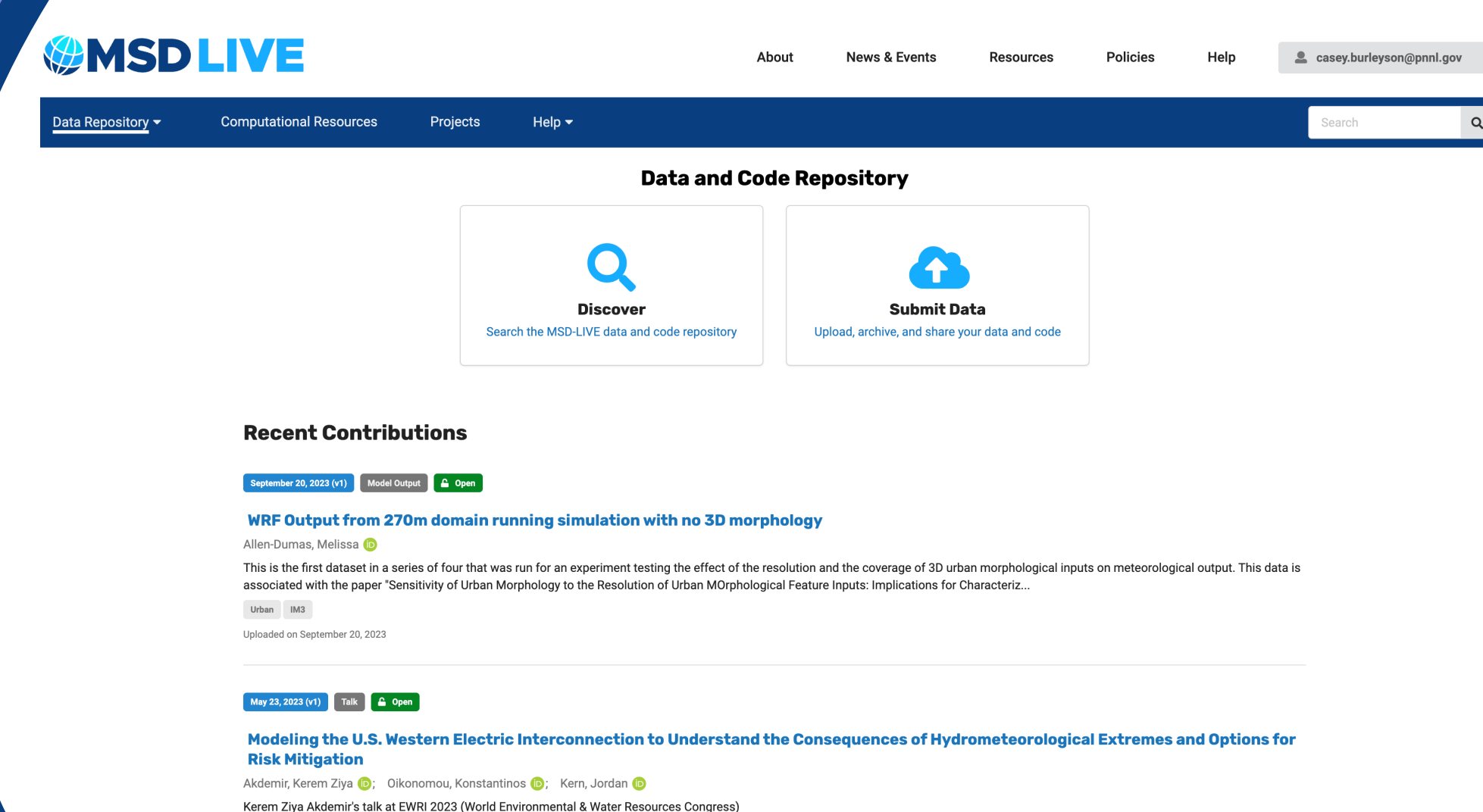
5. Data Repository

Our data repository allows users to share data (with DOIs) and find data produced by others.

---

6. Interactive Data Dashboards

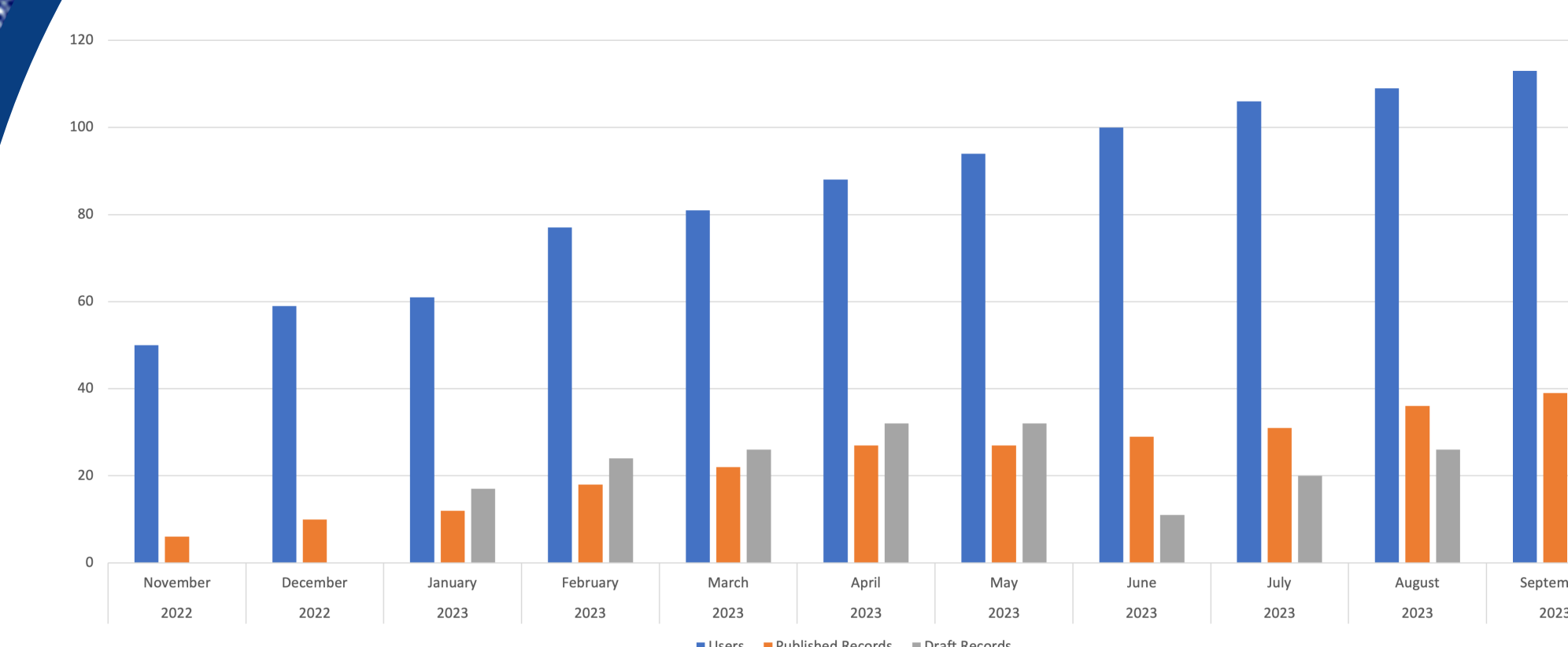One of our newest features allows users to build interactive dashboards to explore data.