



Sustainability Data Hub

Nathan Chasse, Shane Zarechian, Andrew Litcofsky, Devin Wilson
Computer Science, University of New Hampshire, Durham, NH 03824

Introduction

Collaborative sustainability research requires access to diverse datasets and collaborative analysis platforms. However, no centralized platform currently exists at UNH for researchers to easily share and discover datasets and analyses outside of their research group.

To address this, we designed and prototyped the Sustainability Data Hub (SDH) around the open-source metadata management platform DataHub. The SDH integrates *existing* databases and analysis tools, enabling researchers to collaborate without overhauling their workflows. It is the connective tissue turning isolated data environments into an ecosystem.

Project Goals

Our project was organized around the following goals:

- **Research** existing platforms and tools suitable for building the Sustainability Data Hub (SDH)
- **Design** a modular system architecture based on research and user needs
- **Build** a prototype using real datasets and tools that support interactivity for non-technical users
- **Clarify** the purpose and scope of the SDH within the UNH research ecosystem
- **Envision** a development roadmap outlining project vision and next steps for future SDH teams

We completed research and design, created a prototype, and wrote an initial version of the roadmap.

Implementation

The core of our system is DataHub, an open-source metadata management platform that connects existing databases, storage systems, and analysis tools. Rather than replacing tools researchers already use, DataHub provides the infrastructure to link them together.

For datasets not already hosted in a managed database, we propose a centrally managed solution where researchers can upload floating or underutilized datasets. We explored a platform called CKAN, a well-supported open-source data management system.

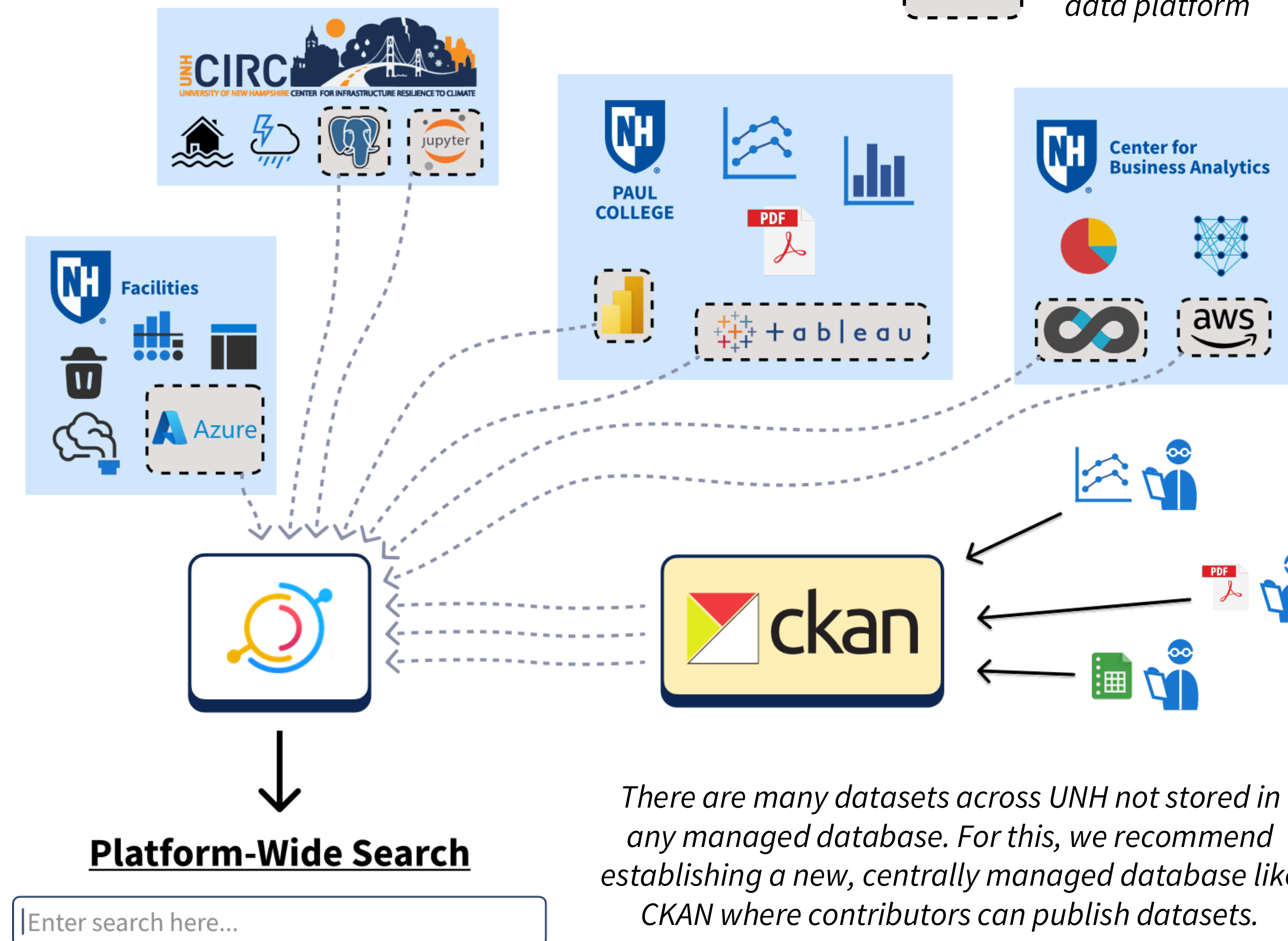
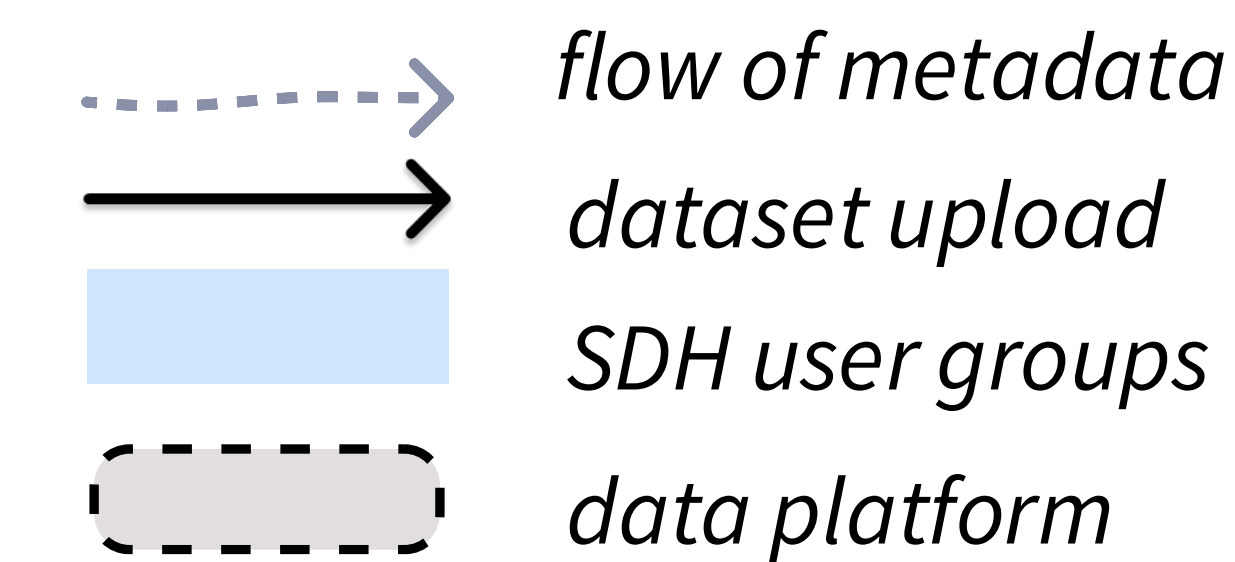
The goal of the SDH is not to centralize everything under one stack, but to support the diversity of tools in already in use at UNH by connecting them.

System Design

DataHub is the core of the SDH. To clarify: DataHub is the *tool* we used to build the SDH; they are separate concepts.

DataHub organizes metadata about datasets and analyses, such as storage locations, file types, contributors, and schemas, so that others can discover them.

LEGEND



Community Feedback and Testing

Early Feedback

We received encouraging anecdotal support for the SDH's purpose, but we did not systematically gather feedback. We strongly recommend future SDH leadership engage with researchers and stakeholders across UNH to understand their needs and align the system with institutional priorities.

Testing

Preliminary testing involving our advisors, Nagaraj Bukkapatnam and Shuili Du, was effective but limited in scope. We strongly recommend frequent testing and demonstrations during the next development phase.

Next Steps

Next steps for future development teams and leadership of the Sustainability Data Hub include:

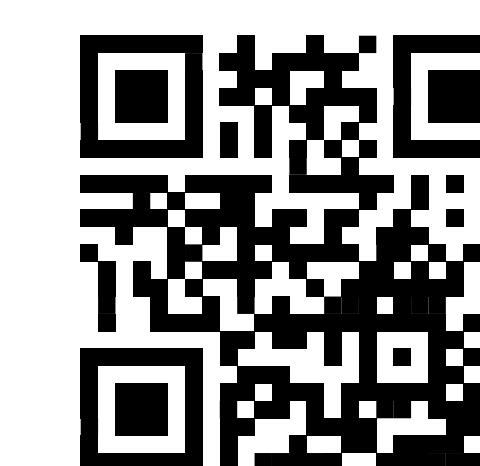
1. **Deploy** a production-ready dataset storage solution (e.g. CKAN) and start gathering datasets
2. **Integrate** additional data platforms according to user interest; conduct surveys to prioritize
3. **Collaborate with domain experts** to identify intuitive and useful dataset tagging schemes
4. **Pilot the SDH** to gather initial feedback, working college-level leadership to identify departments
5. **Secure administrative support** and funding for infrastructure hosting, platform maintenance, staffing needs, and communication

These steps are discussed at length in the roadmap.

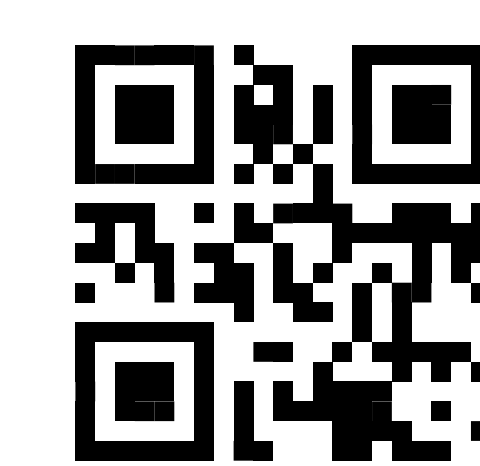
Acknowledgements

A special thank you to Shuili Du and Nagaraj Bukkapatnam for their patient support throughout the project, for enabling us to learn, and for offering valuable perspective. Thanks to Scott Kitterman for support with hosting in the CS department cluster.

References



datahubproject.io



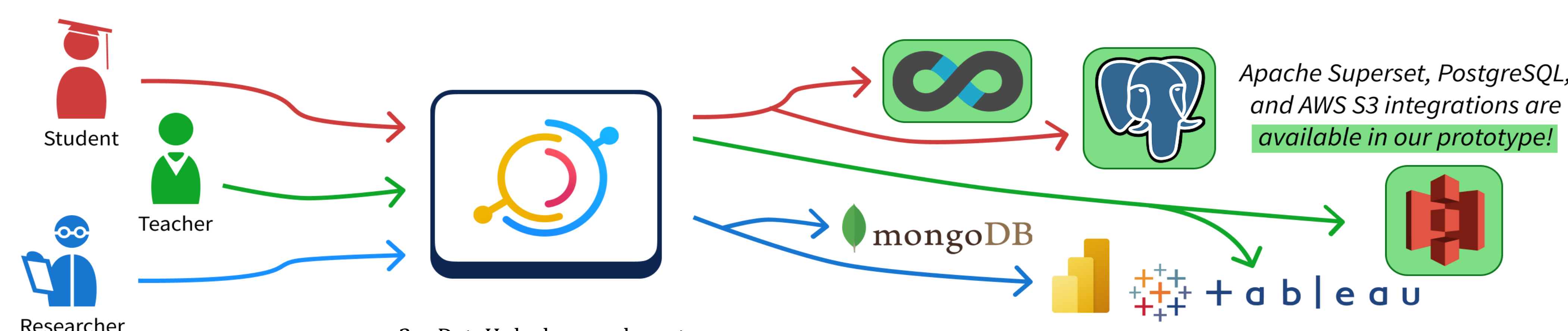
ckan.org



DataHub: The Platform That Makes Data Work for You

Use Cases

Users access DataHub to discover available datasets and analyses via an intuitive search interface. DataHub then directs them to the original platforms where the content is stored. As users interact with those platforms, relevant metadata is automatically sent back to DataHub.



NOTE: these users and platforms are just examples. Any user can access any configured platform with the appropriate permissions.