Next-Generation Metadata Management Infrastructure

for Intelligent and Efficient Scientific Data Sharing and Discovery at National-Scale

Jie Ren (William & Mary)

Lipeng Wan (Georgia State University)

8/4/2026

NSF Award #2609536

Rich in data, poor in discovery

AI-era science runs on data: every model, digital twin, and cross-domain analysis starts with finding the right data

domain + data / AI scientists working together

THE DATA ALREADY EXISTS

Billions of files

across labs, HPC centers, campuses & clouds

…but stored like this:

```
slide_20240513_0134.tif
run021_000123.mid
output_final_v3.nc
```

Hard to find data efficiently

SCIENCE NEEDS TO FIND IT

domain scientists | **data / AI scientists** | **AI agents**

"Plasma turbulence simulations under high magnetic fields"

"Microscopy images showing early-stage tissue degradation"

Vast scientific data goes unused — not because it's inaccessible, but because it's undiscoverable.

Why finding data fails: three root causes

Invisible by default

Sharing requires owner action; discovery runs on personal networks.

IN PRACTICE

Public data already exists at Lab A. Finding it means already knowing where it lives, or who made it.

FIXED BY

Federated metadata publication + proactive recommendations

Metadata without meaning

Files carry names, timestamps, and project IDs, not scientific concepts.

IN PRACTICE

output_final_v3.nc: what was measured, under what conditions, and when?

FIXED BY

Semantic schema: AI-oriented fields + concept-level descriptions

Closed to AI

Agents can't explore or recommend data that carries no semantics.

IN PRACTICE

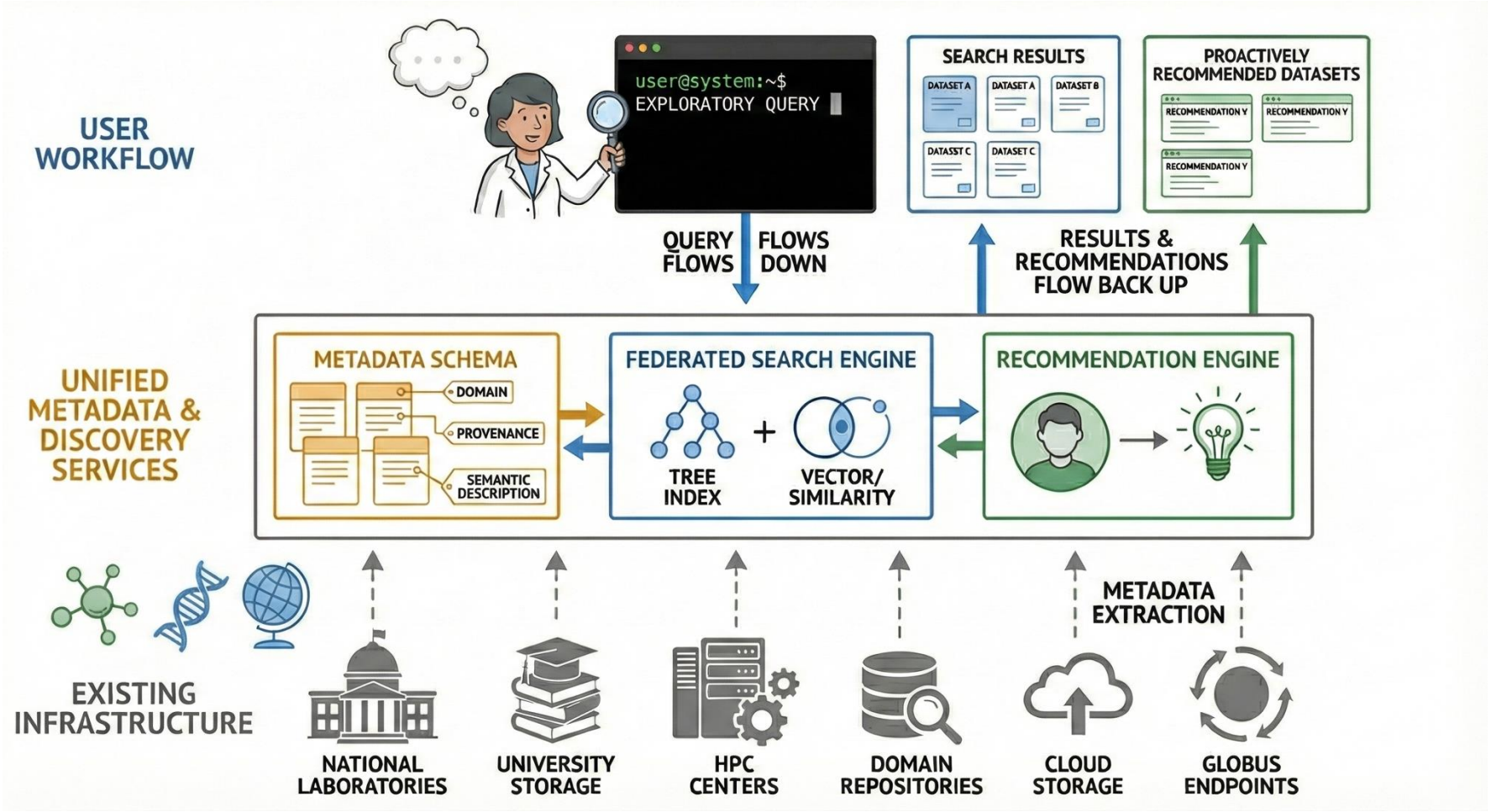
AI agents cannot ask repositories what they hold.

FIXED BY

AI-readable fields + hybrid filter & semantic search

Three birds with one stone: a shared semantic metadata layer.

Vision: a semantic discovery layer over existing infrastructure



Making existing data discoverable by people and by AI, with no new platform and no data movement.

This project: a two-year plan toward our vision

O1

CHARACTERIZE

Cross-domain discovery requirements — what researchers actually need.

O2

DESIGN

A community-informed, AI-ready metadata schema for existing scientific data.

O3

PLAN

A scalable architecture for operating discovery services at national scale.

Category I / II full proposal

ready to build, backed by evidence

Five activities, one pipeline: listen → design → deliver

LISTEN

- **A1** National survey
- **A2** Workshops
- **A3** Stakeholder consultations

DESIGN

- **A4** AI-ready metadata schema workgroup
- **A5** Scalable architecture workgroup

DELIVER

- Landscape Assessment
- Architecture White Paper
- Development Roadmap

Year 1: community and technical assessment

A1

National survey

~2,000

completed responses

Sampled

- observational & simulation science
- experimental facilities
- AI / ML workflows

Method

- survey firm designs and fields it
- cognitive interviews first
- follow-up campaigns

Metadata requirements

A2

Community workshops

2

workshops

Who comes

climate · materials · fusion · biomedical imaging · AI for science · CI experts

Why it matters

Cross-domain conversations

Metadata gap analysis

A3

Stakeholder consultations

3

types of consultation

Who we ask

- Software: Globus & database vendors
- Hardware: memory, GPU, FPGA
- Data facilities

What we assess

APIs, auth and ingestion · performance and cost at scale · operations and security

Technical landscape

Identify what the community needs, and what the technology allows.

Year 2: two working groups turn findings into blueprints

A4 Metadata schema

6 months · monthly · 6–8 experts from repositories, domain science, and AI

1. Which AI-oriented fields, and what values?
2. What makes a description useful outside its domain?
3. How much annotation can be automated?

PRODUCES

Schema specification, description guidelines, tooling recommendations

A5 Architecture

6 months · monthly · experts from Globus, PNNL, Jefferson Lab, and SK Hynix

1. How do sites stay consistent without global coordination?
2. How is the index partitioned and searched at scale?
3. How does it integrate with Globus, and who sustains it?

PRODUCES

Architecture plan: components, interfaces, deployment path, performance targets

Q&A

Thank you

We welcome your input on the survey and the workshops, and we welcome new collaborations.

CONTACT

Jie Ren PI · William & Mary
jren03@wm.edu

Lipeng Wan Co-PI · Georgia State University
lwan@gsu.edu

Supported by the National Science Foundation · IDSS · Award #2609536