Modernizing Data Management at Jefferson Lab

Document Management: DocDB ⟶ InvenioRDM

Anil Panta

2026

Clas12 Collaboration Meeting

STRIDE
Software Tools for Research, Infrastructure, Data, and Experiments

Jefferson Lab

DocDB

Jefferson Lab

History

- Created 2001 in BTeV @ Fermilab
- Basic Metadata Support.
- Event Management
- One instance per community

DocDB

Jefferson Lab

History

- Created 2001 in BTeV @ Fermilab
- Basic Metadata Support.
- Event Management
- One instance per community

System Limitation

- Legacy architecture limits extensibility
- Weak metadata model and search.
- No native REST API or automation layer.
- Poor integration with modern data platforms
- Limits to documents only.

DocDB

Jefferson Lab

History

- Created 2001 in BTeV @ Fermilab
- Basic Metadata Support.
- Event Management
- One instance per community

System Limitation

- Legacy architecture limits extensibility
- Weak metadata model and search.
- No native REST API or automation layer.
- Poor integration with modern data platforms
- Limits to documents only.

Governance gap

- Don’t adhere to standard metadata Schema
- Don’t adhere to FAIR principle
- More…

DocDB

Jefferson Lab

History

- Created 2001 in BTeV @ Fermilab
- Basic Metadata Support.
- Event Management
- One instance per community

System Limitation

- Legacy architecture limits extensibility
- Weak metadata model and search.
- No native REST API or automation layer.
- Poor integration with modern data platforms
- Limits to documents only.

Governance gap

- Don’t adhere to standard metadata Schema
- Don’t adhere to FAIR principle
- More…

Operation Challenges

- Separated public vs private
- Custom SSO AUTH to be maintained
- One instance one experiment…
- Migration is painful for large #instance

DocDB

Jefferson Lab

History

- Created 2001 in BTeV @ Fermilab
- Basic Metadata Support.
- Event Management
- One instance per community

System Limitation

- Legacy architecture limits extensibility
- Weak metadata model and search.
- No native REST API or automation layer.
- Poor integration with modern data platforms
- Limits to documents only.

Governance gap

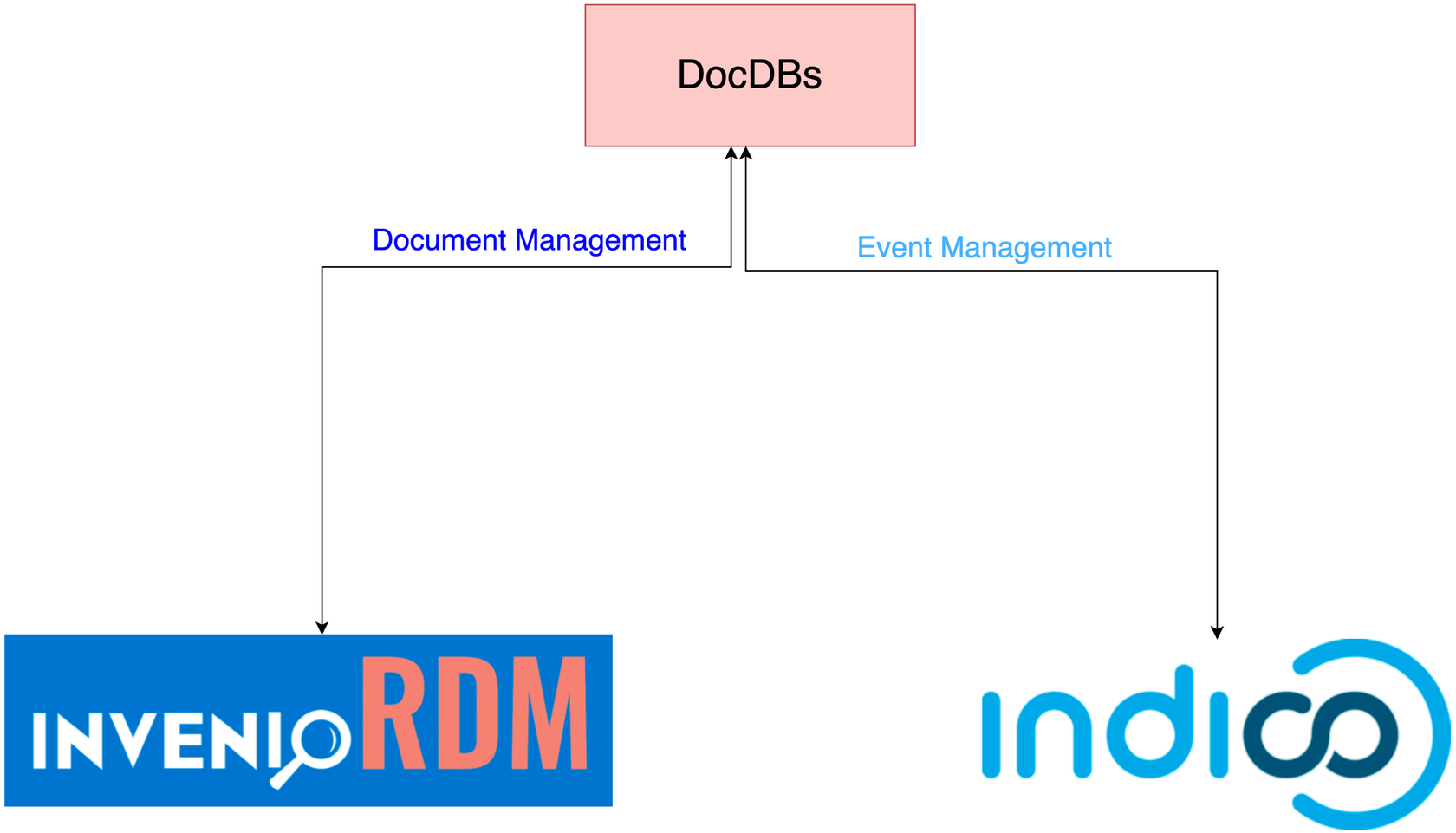
- Don’t adhere to standard metadata Schema
- Don’t adhere to FAIR principle
- More…

Operation Challenges

- Separated public vs private
- Custom SSO AUTH to be maintained
- One instance one experiment…
- Migration is painful for large #instance

“DocDB served the community for over two decades, but modern research collaboration now requires richer metadata, APIs, automation, and scalable discovery.”

Move Away from DocDBs: InvenioRDM and Indico



InvenioRDM

Open-Source research repository- Turn Key research data management platform.
Document, Image, Data, Software etc

- Developed at CERN in Collaboration with many research institution.
- Now has been deployed in many large to small data/document repository.
  CDS-RDM (CERN), Zenodo (CERN), Caltech, TU Graz/Wein, BNL, NYU, University of Hamburg etc)

DataCite Metadata Schema

International Standard

RestAPI

Programatic Access

Communities

Experiment, Halls, groups

Full Text Search

Programatic Access

Secure & Compliant

Public, Restricted
& Embargo

OAUTH

Federated Logins
Institutional Integration

InvenioRDM

Open-Source research repository- Turn Key research data management platform.
Document, Image, Data, Software etc

- Developed at CERN in Collaboration with many research institution.
- Now has been deployed as many large to small data repository.
  CDS-RDM (CERN), Zenodo (CERN), Caltech, TU Graz/Wein, BNL, NYU, University of Hamburg etc)

DataCite Metadata Schema

International Standard

RestAPI

Programatic Access

Communities

Experiment, Halls, groups

Full Text Search

Programatic Access

Secure & Compliant

Public, Restricted
& Embargo

OAUTH

Federated Logins
Institutional Integration

"Moving from a document archive to a research knowledge platform"

FAIR in InvenioRDM

Jefferson Lab

FAIR: Findable	InvenioRDM Features
F1: (meta)data are assigned a globally unique and persistent identifier	• A DOI is issued to every published record in InvenioRDM. • Supports use of additional unique IDs such as ORCiD, ROR
F2: data are described with rich metadata (defined by R1 below)	• InvenioRDM's metadata is compliant with DataCite's Metadata Schema minimum and recommended terms, with a few additional enrichments.
F3: metadata clearly and explicitly include the identifier of the data it describes	• The DOI is a top-level and a mandatory field in the metadata of each record.
F4: (meta)data are registered or indexed in a searchable resource	• Metadata of each record is indexed and searchable directly in InvenioRDM's search engine immediately after publishing. • Metadata of each record is sent to DataCite servers during DOI registration and indexed there.

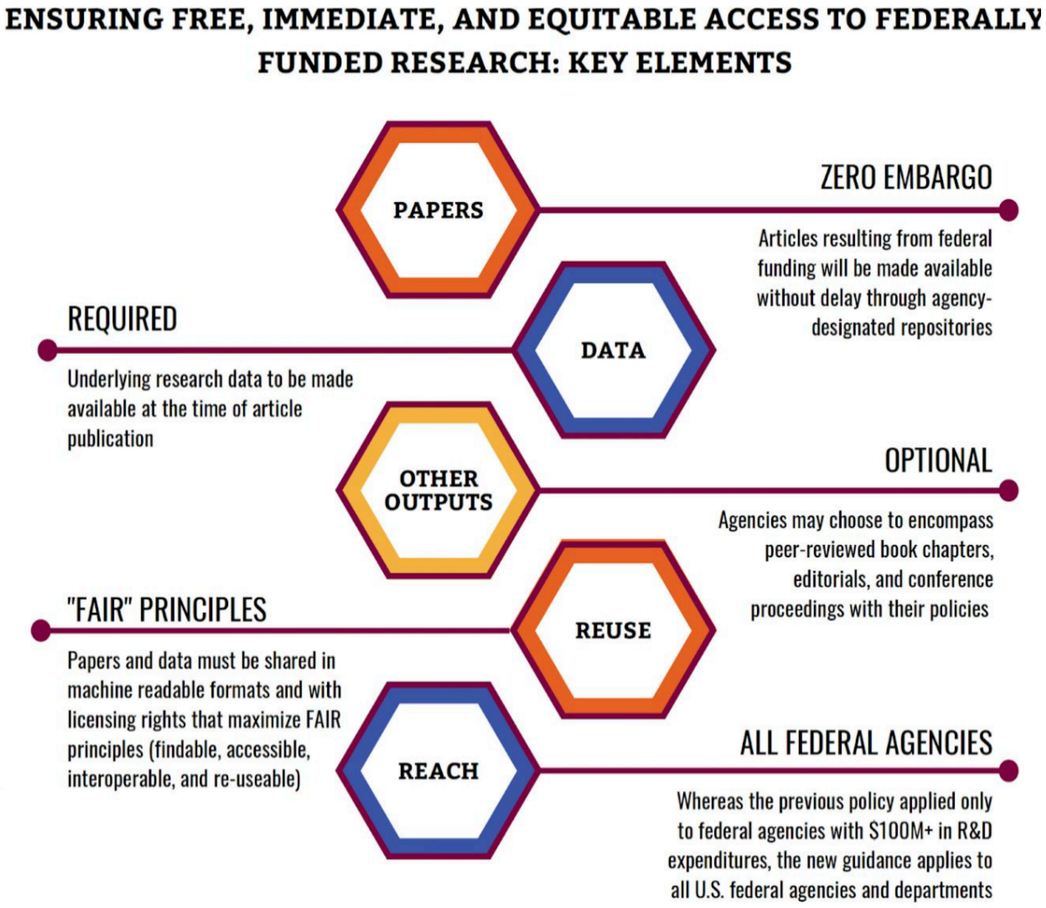
FAIR: Accessible	InvenioRDM Features
A1: (meta)data are retrievable by their identifier using a standardized communications protocol	• Metadata for records are harvestable using the OAI-PMH protocol by record identifier and the community name. • Metadata is also retrievable through the public REST API.
A1.1: the protocol is open, free, and universally implementable	• See point A1. OAI-PMH and REST are open, free, and universal protocols for information retrieval on the web.
A1.2: the protocol allows for an authentication and authorization procedure, where necessary	• Authentication via OAuth 2 / OpenID Connect or SAML • Record metadata can be public, restricted, or embargoed. • Metadata required by DataCite are publicly accessible.
A2: metadata are accessible, even when the data are no longer available	• Data and metadata can be retained for the lifetime of the repository (determined by policies at administering institution). • Metadata are stored in a database (separate from the data)

FAIR: Interoperable	InvenioRDM Features
I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	• InvenioRDM uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core, MARCXML, DataCite, DCAT-AP, CSL, BibTex.
I2: (meta)data use vocabularies that follow FAIR principles	• DataCite metadata schema • Can be configured to use additional vocabularies • For certain terms we refer to open, external vocabularies, e.g.: license (SPDX), funders (FundRef) and grants (OpenAIRE).
I3: (meta)data include qualified references to other (meta)data	• Each referenced external piece of metadata is qualified by a resolvable URL.

FAIR: Reusable	InvenioRDM Features
R1: (meta)data are richly described with a plurality of accurate and relevant attributes	• Each record contains a minimum of DataCite's mandatory terms, with optionally additional DataCite recommended terms and InvenioRDM's enrichments.
R1.1: (meta)data are released with a clear and accessible data usage license	• A variety of licences can be chosen from SPDX. • Data downloaded by the users is subject to the license specified in the metadata by the uploader.
R1.2: (meta)data are associated with detailed provenance	• All data and metadata uploaded is traceable to a registered InvenioRDM user. • Metadata can optionally describe the original authors and contributors of the published work.
R1.3: (meta)data meet domain-relevant community standards	• InvenioRDM is not a domain-specific repository, yet through compliance with DataCite's Metadata Schema, metadata meets one of the broadest cross-domain standards available.

OSTP Requirement Fulfillment by InvenioRDM

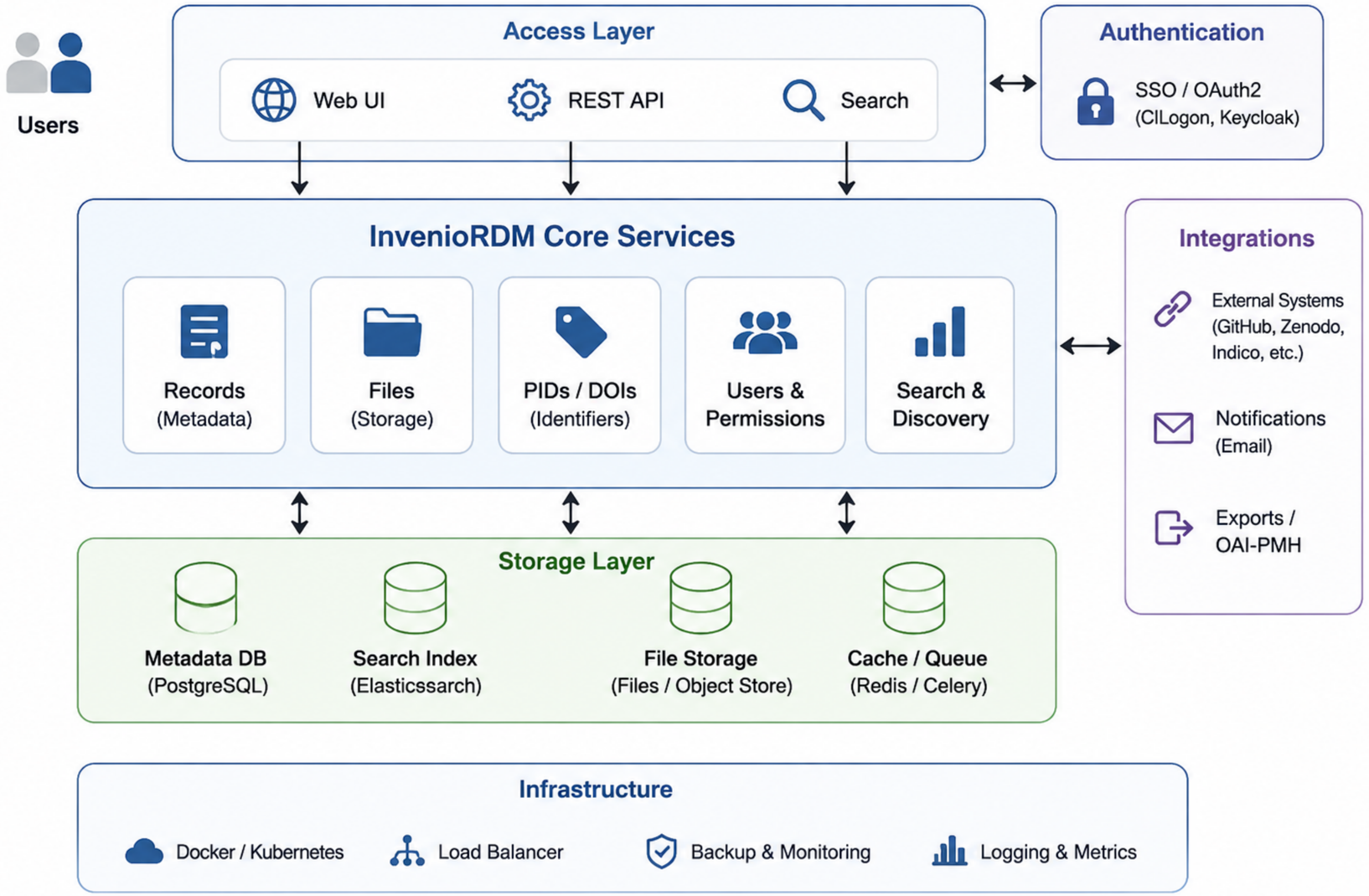
Not just a JLab policy requirement, but a mandate from **DOE** and **OSTP** policies that we must comply with.



01	Zero Embargo	• Deposit and link supporting information with automated embargo removal at a specified date
		• Unique identifiers support discovery:
		  ○ DOI assigned to deposits
		  ○ ORCiD integration for authors & contributors
		  ○ ROR integration for institutions
		• Ability to link funding sources

02	Underlying research data must be shared at the time of article publication	• Contextualization of records via secondary identifiers to publications, project sites, external data, and other associated research outputs • A wide variety of licensing options • Ability to archive software via GitHub integration within records, allowing data to be linked to code [coming soon]
03	Other research outputs such as book chapters, editorials, and grey literature may be required	• A variety of resource types are available • Contextualization of records via secondary identifiers to publications, project sites, external data, and other associated research outputs
04	Articles and data must adhere to FAIR Principles	• Deposited records check all FAIR principles (Findability, Accessibility, Interoperability, Reusability) • Export formats include JSON, GeoJSON, DataCite (JSON or XML), Dublin Core XML, DCAT-AP, CSL, MARCXML, and BibTex • A wide variety of licensing options • Citations in multiple formats provided for each record

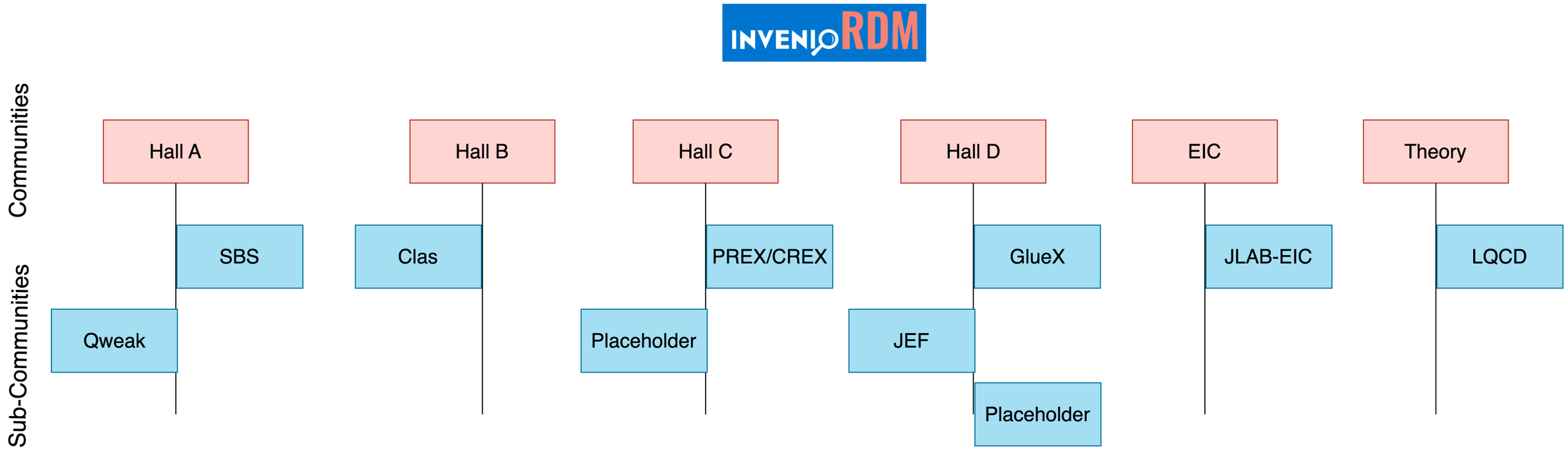
InvenioRDM Architecture.



Communities Organization

One instance Serving All JLab communities

Access control at Sub-Community level



Migration Objective

Goal: 100% record fidelity

Zero-loss transition from legacy DocDB to a unified, FAIR-compliant InvenioRDM platform.

Completeness

zero record loss
zero data loss

Integrity

Preserve exact
access control

Traceability

Retain docID
for back reference

Onboarding

Document

Smooth Transition

No Friction

No extra account

Same login must work

"A complete, traceable, zero-loss migration that unifies every record, preserves every permission, and transitions every user to a InvenioRDM platform."

Migration workflow

DocDB

Extract

- Query Records and Relationship.
- Correlate File Paths to records
- Create md file for meeting agendas

Transform

- Translate Metadata Schema -> DataCite Schema
- Translate Access Control
- Route to sub communities
- Retain back reference of docID and topics
- Archive meeting agendas.

Load

- Create Record Via REST API
- Upload and Attach Files to records.
- Assign PIDs and Index for Search

INVENIO RDM

Authentication Flow

CILogon is only Authentication provider.

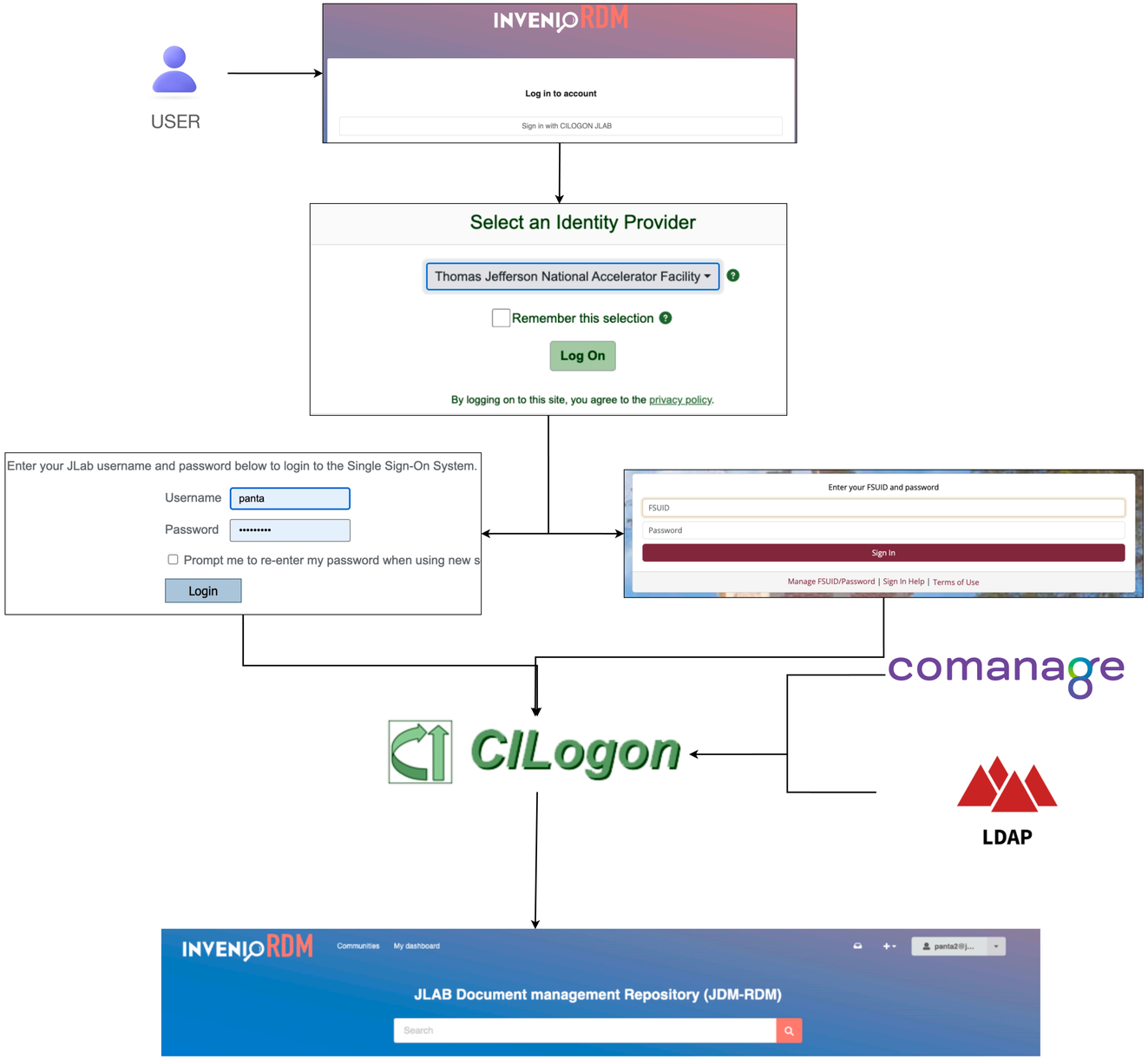
CILogon is Federated authentication thus supporting many institution login

Similar to code.jlab.org

Two path to login

1. For Experiment with JLAB only login
   - Must Select “TJNAF” in CILOGON IdP.
   - Pulls JLAB CUE/LDAP group to identify user

2. For Experiment with Federated login
   - Must create group in comanage registry.
   - User enroll to that group vis institutional credential.
   - Use can login via institutional account.



Authentication Flow

CILogon is only Authentication provider.

CILogon is Federated authentication thus supporting many institution login

Similar to code.jlab.org

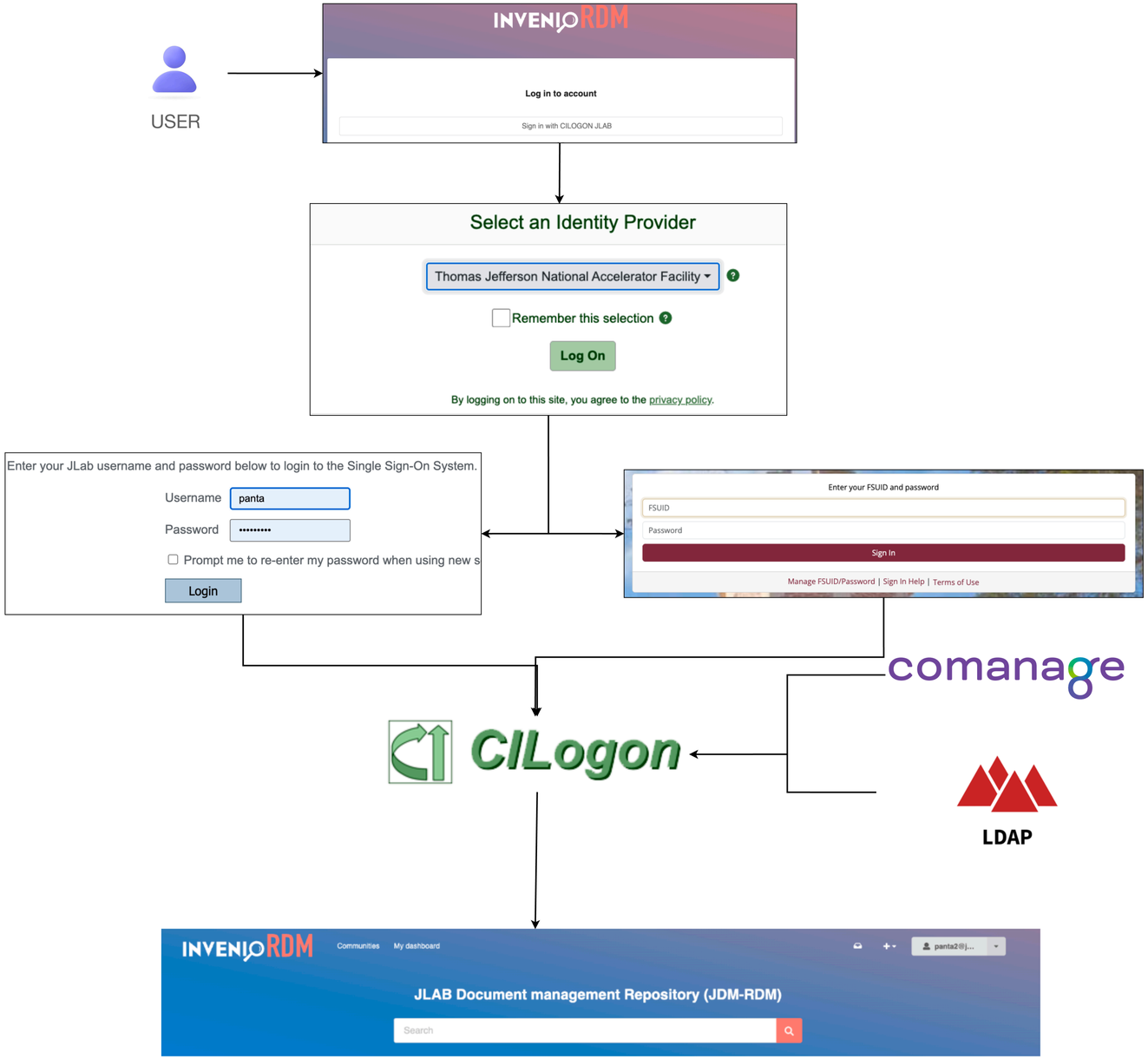
Two path to login

1. For Experiment with JLAB only login
   - Must Select “TJNAF” in CILOGON IdP.
   - Pulls JLAB CUE/LDAP group to identify user

2. For Experiment with Federated login
   - Must create group in comanage registry.
   - User enroll to that group vis institutional credential.
   - Use can login via institutional account.

What is comange?

- IAM platform for federated auth and org
- It has Collaborative groups and roles defined
- Links multiple accounts to single profile



Migration Status:

Migration complete for following

- PREX/CERX (Hall A)
- GlueX (Hall B)
- EIC-JLAB (EIC)

Authentication and Infrastructure

- OpenSearch central deployment in open shift
- Reroute docdb document link to InvenioRDM records

GlueX
Part of Hall D
https://www.gluex.org/ Thomas Jefferson National Accelerator Facility

Hall B
https://www.jlab.org/physics/hall-b Thomas Jefferson National Accelerator Facility

Hall D
https://www.jlab.org/physics/hall-d Thomas Jefferson National Accelerator Facility

EIC-JLAB
Part of EIC
https://www.jlab.org/eic Thomas Jefferson National Accelerator Facility

EIC
Project https://www.bnl.gov/eic/ Thomas Jefferson National Accelerator Facility and 1 more organizations

PREX
Part of Hall A
https://hallaweb.jlab.org/parity/prex/ Thomas Jefferson National Accelerator Facility

Hall A
https://www.jlab.org/physics/hall-a Thomas Jefferson National Accelerator Facility

LQCD
Part of Theory
Thomas Jefferson National Accelerator Facility

Theory
https://www.jlab.org/theory Thomas Jefferson National Accelerator Facility

Conclusion

JLab DocDBs

Legacy Archive (<2026/2027)

INVENIO RDM

Unified - FAIR- Future-Ready

“Preserving more than two decades of legacy. Building one platform for next decades”

Test Instance: inveniordm.jlab.org