Introduction to FOXDEN

CHESS Scientific Computing Team

Valentin Kuznetsov, Keara Soloway, Werner Sun, Rolf Verberg

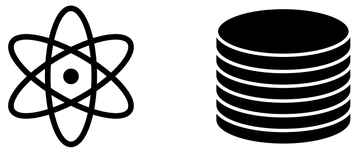
Cornell University

X-Cite workshop

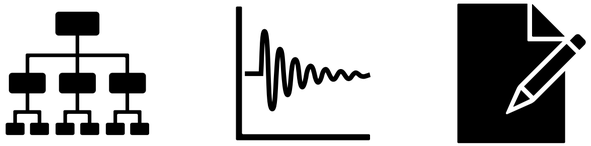
June 17, 2026

Scientific workflow(s)

Data Collection



Data Processing



Publication

FOXDEN

FAIR

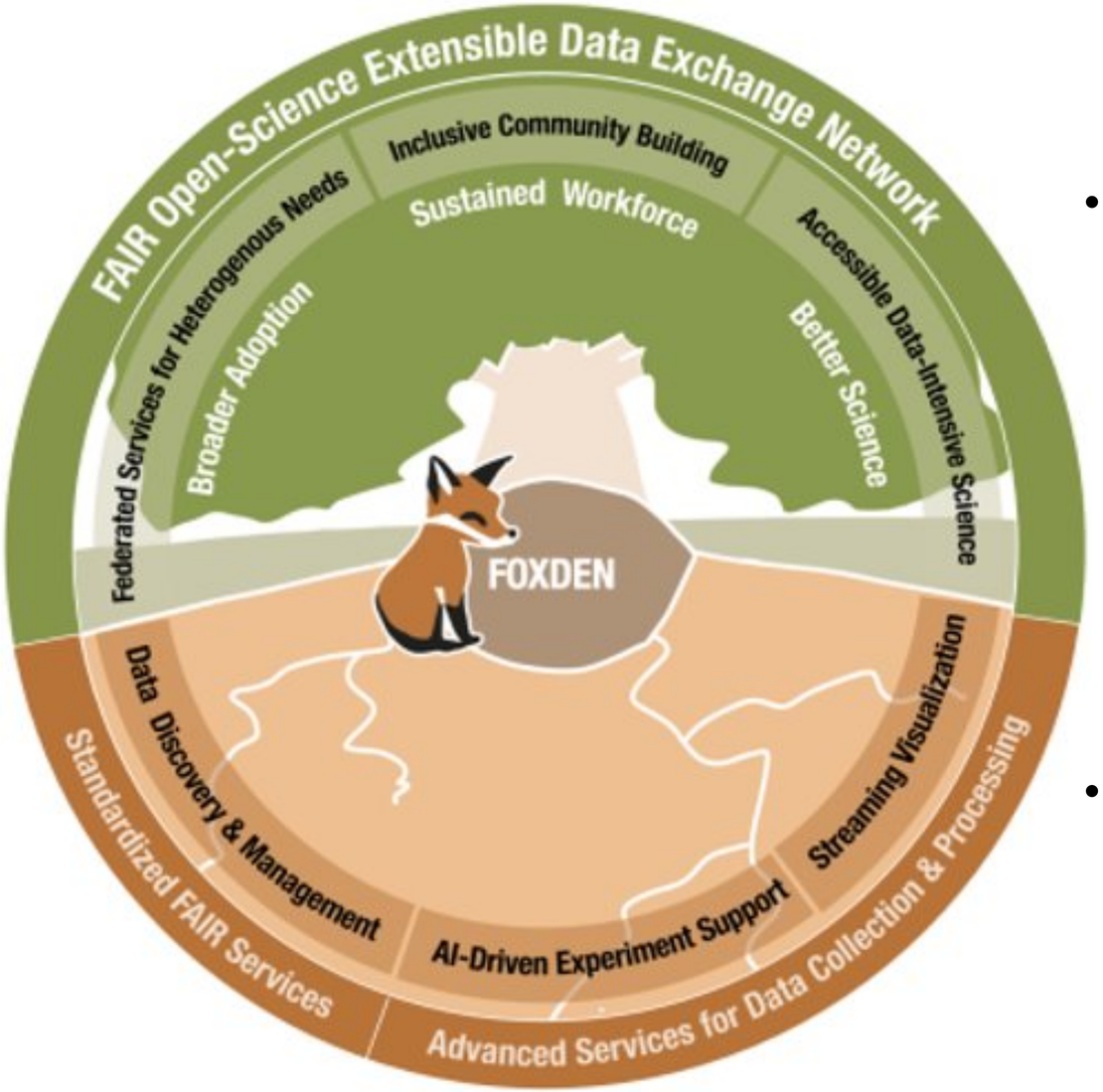
Open-Science

e**X**tensible

Data

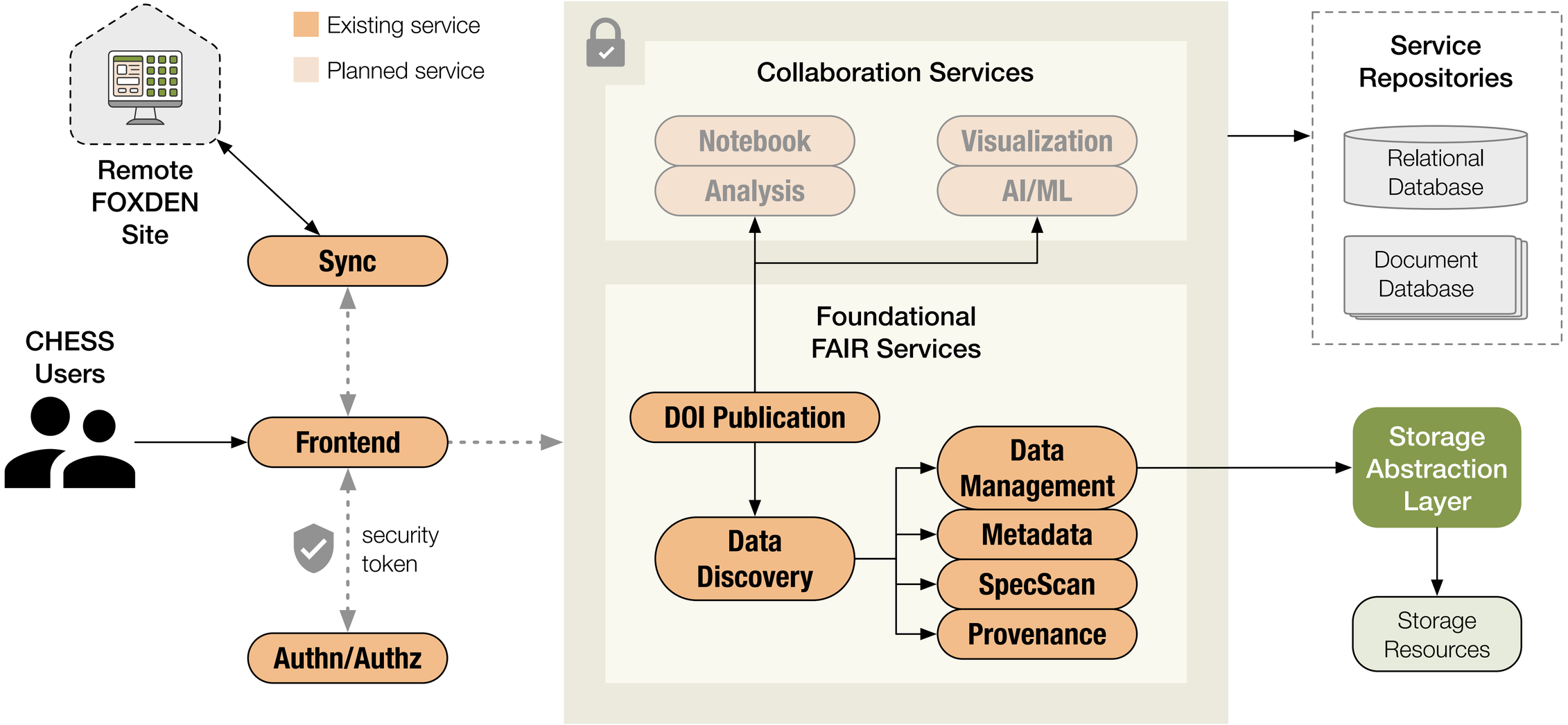
Exchange

Network

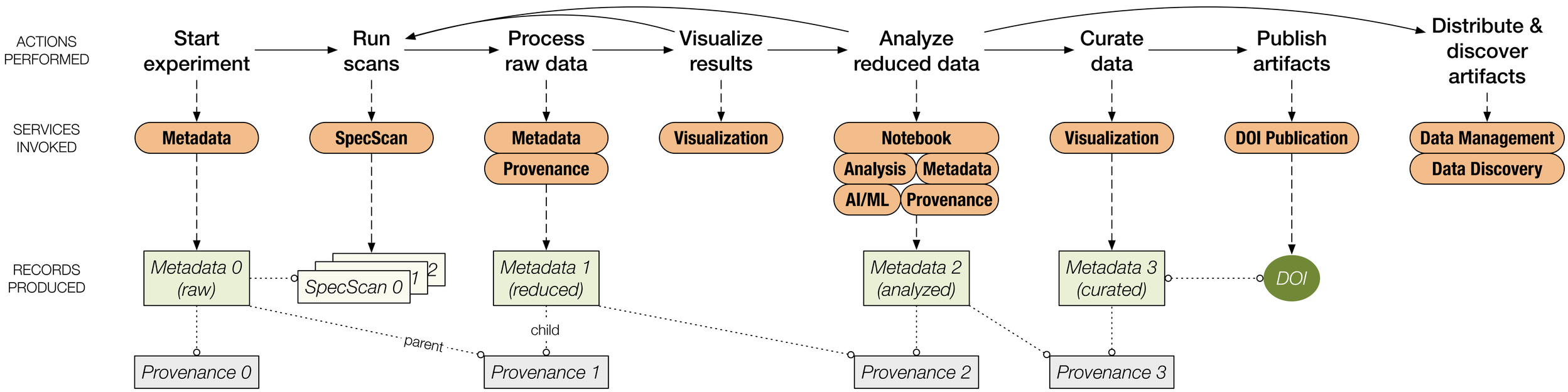


- A set of data services to:
  - Augments *existing* systems and workflows with capabilities for managing research artifacts
  - Service = web application with backend repository and command line API
  - Research artifacts = raw/reduced/ analyzed data, metadata, provenance, code, visualizations, AI/ML models
- An encapsulation of data science expertise and best practices for domain scientists

FOXDEN Services

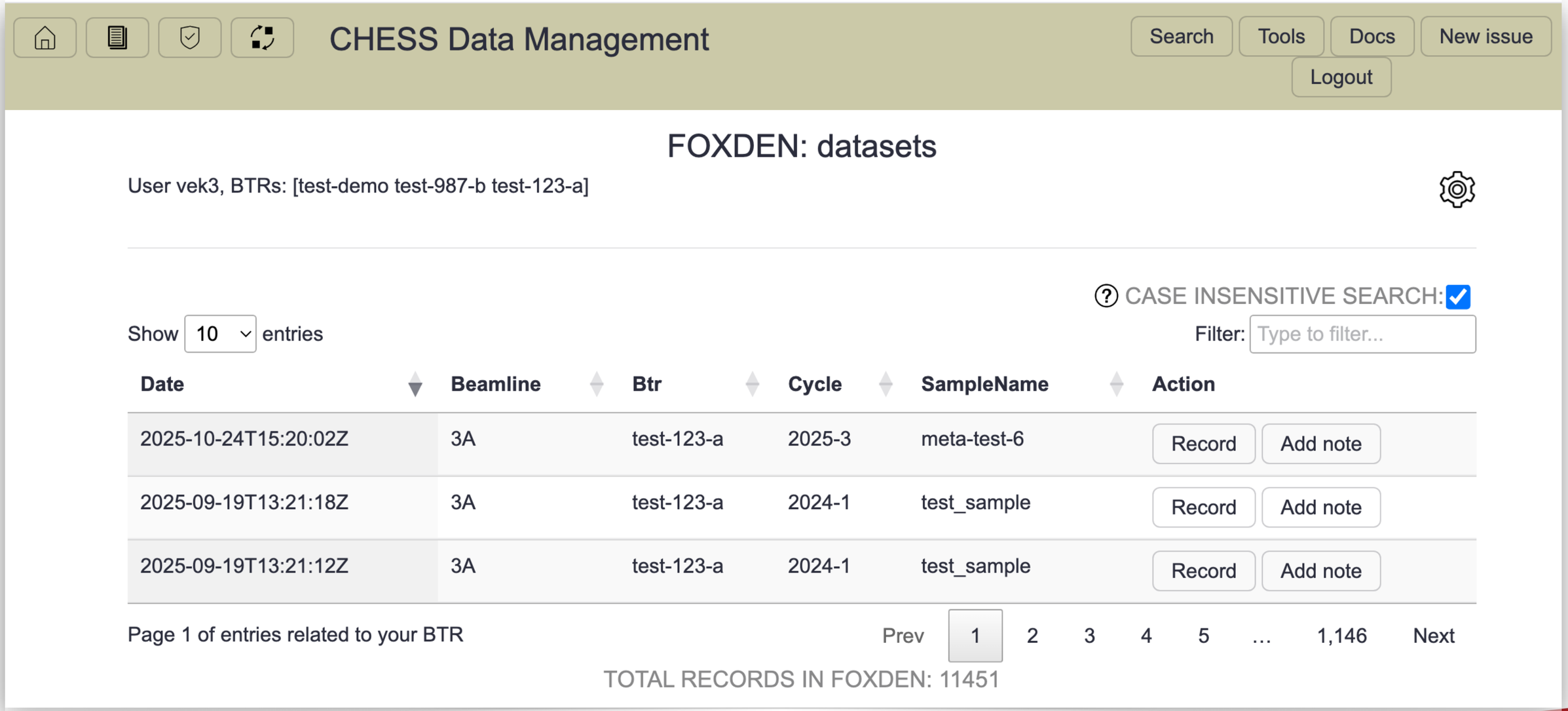


FOXDEN & Typical Research Workflow

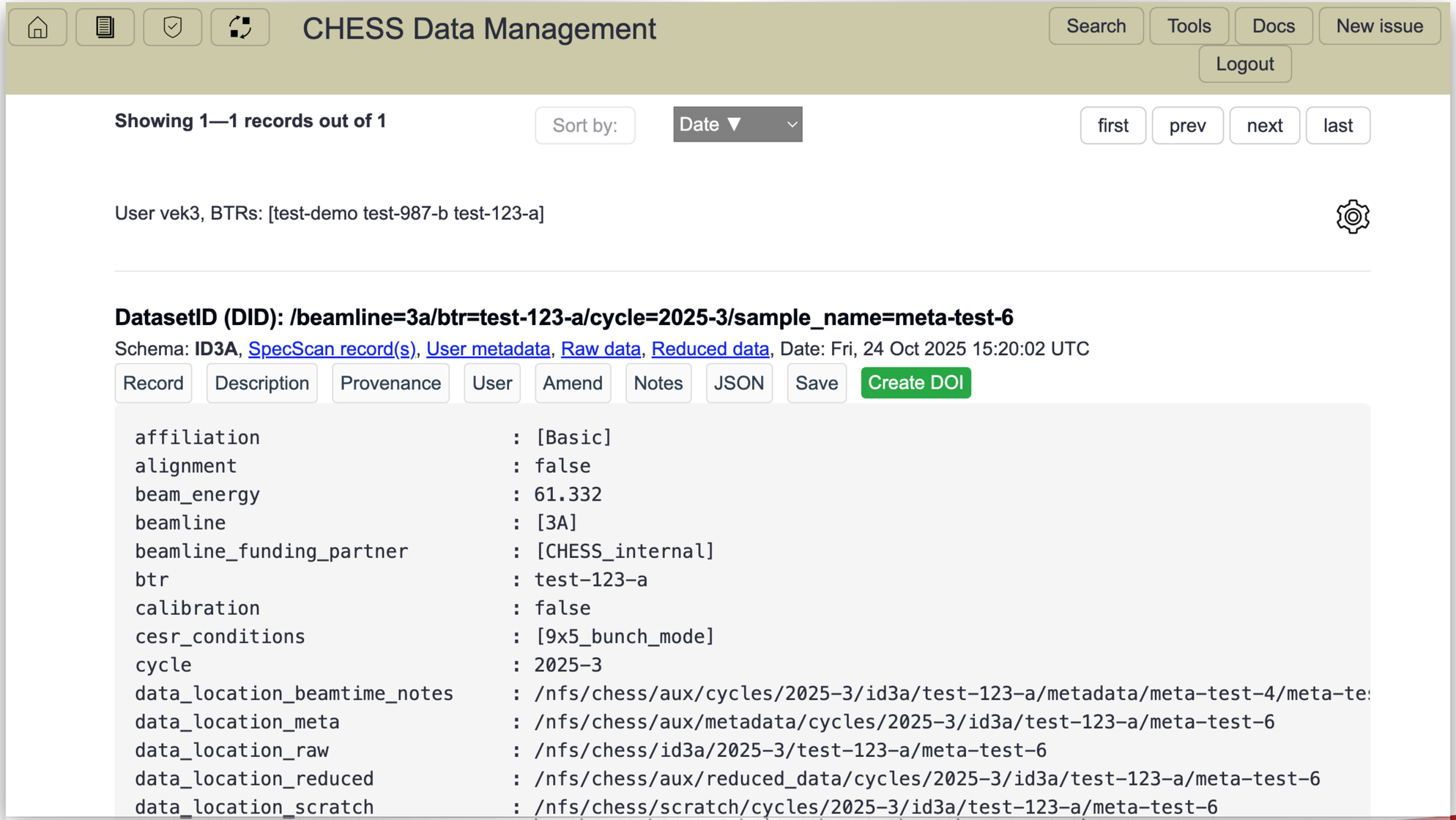


FOXDEN services shadow the research workflow and help organize its metadata, capture provenance, and easy publication process

FOXDEN: Web UI



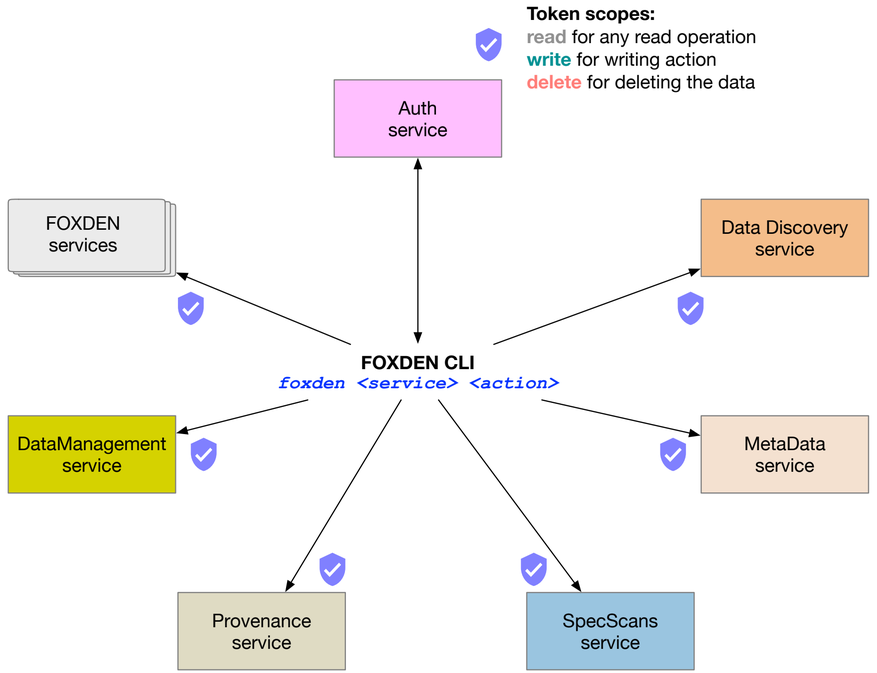
FOXDEN: Record



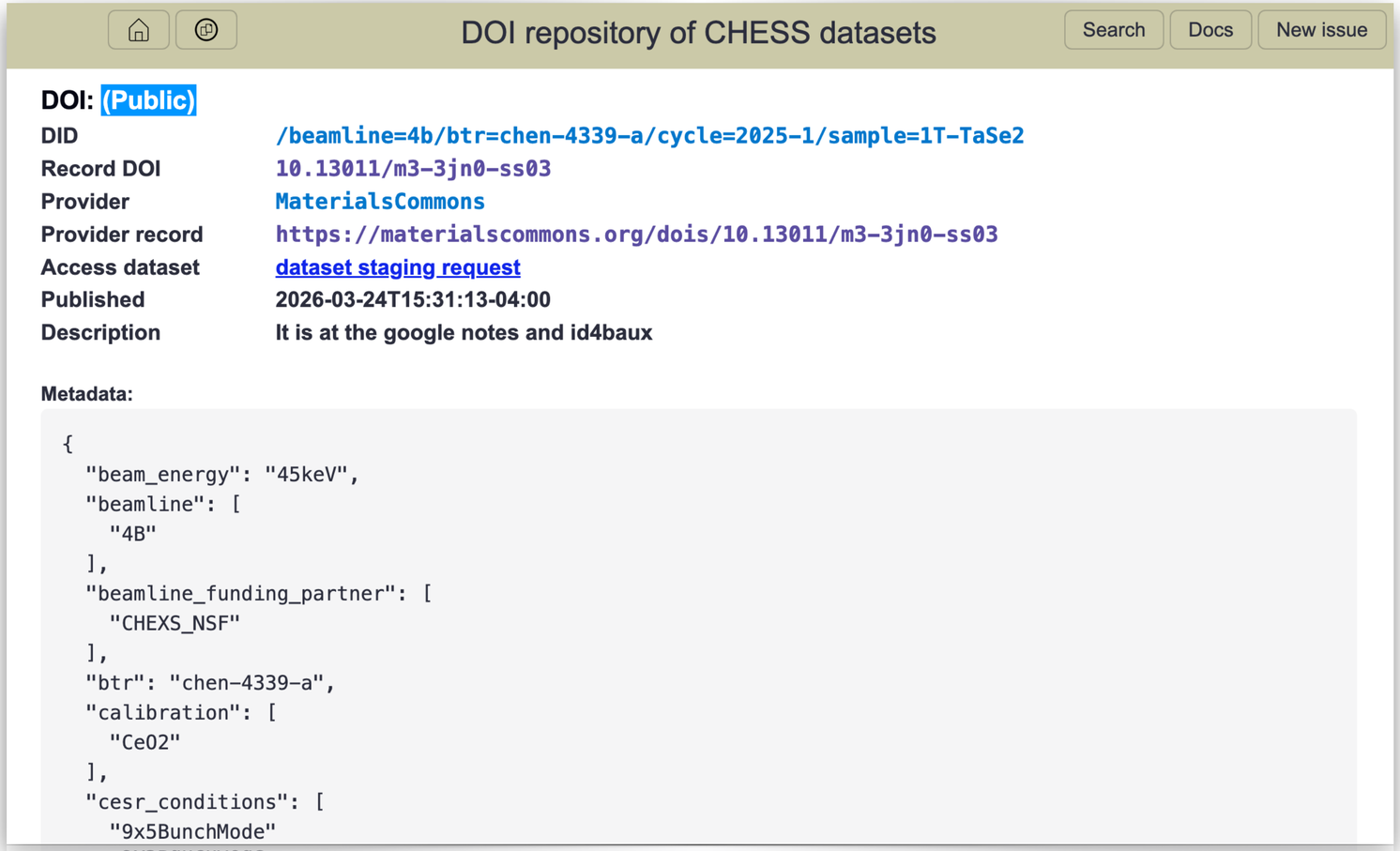
FOXDEN: Command line tool

```
foxden <command> <action> [options]
foxden meta add metadata.json
foxden prov add provenance.json
foxden search pi:name
foxden view <did>
foxden prov ls datasets
foxden ml predict …

# every command has its own help
foxden <command> –help

# get token with relevant scope
foxden token create write

# add new meta-data record
foxden meta add metadata.json

# add new provenance record
foxden prov add provenance.json
```

```
# find meta-data
foxden search {}
---
DID      : abc
Schema   : ID3A
Cycle    : 2023-3
Beamline: [3A]
BTR      : 3731-b
Sample   : <nil>

# view meta-data
foxden view abc
---
### MetaData records:
{"BTR": “3731-b”, …}
### Provenance dataset records:
[{…“dataset”:”/a/b/c”,”meta_id”:”abc”…}]
### Provenance files records:
[{…”file”:"/a/b/c/file1.png", …}]
```

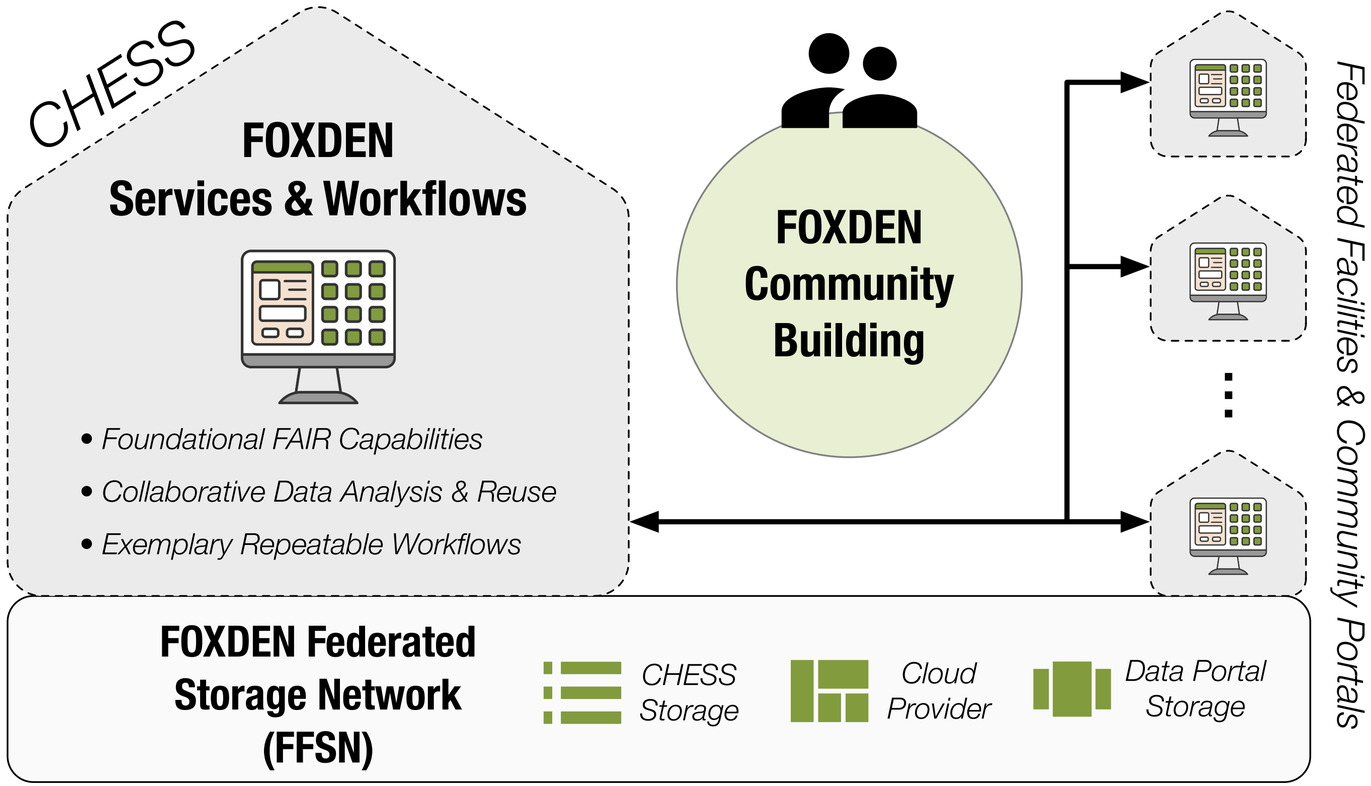


FOXDEN: dataset DOI



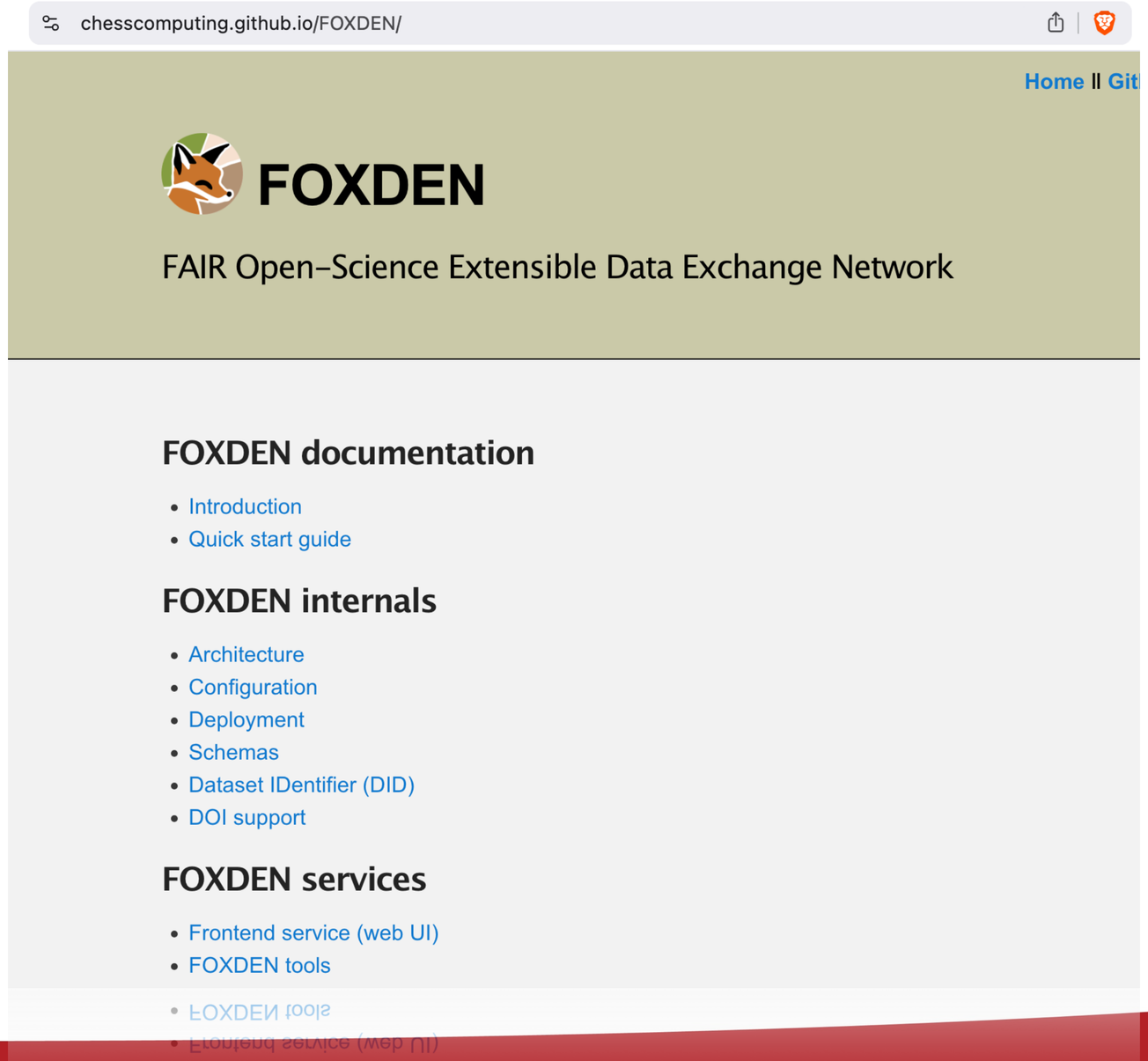
- FOXDEN provides integrated DOI publication service
- Works with multiple DOI providers (DataCite, MaterialsCommons, Zenodo)
  - support Draft and Public modes of DOI
- The actual data files are handled separately and can support different storage beckends, e.g. works with S3, Ceph, etc.
- Ideally all CHESS datasets should acquire DOI identifier to insurer their linkage at provenance layer

Future Vision: Federated Model

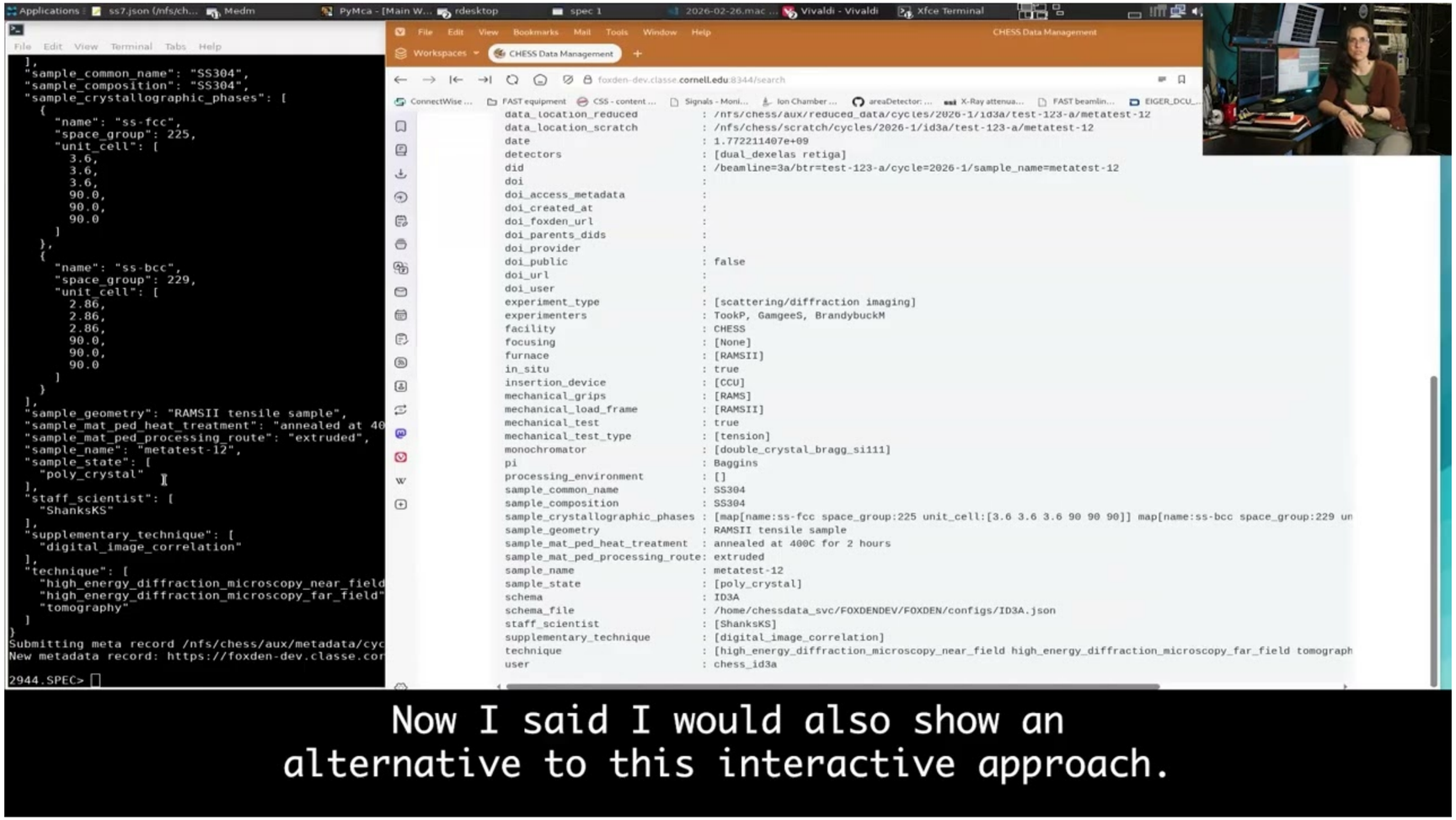


- **FOXDEN is domain-agnostic**
  - Services can be deployed at multiple facilities
  - Repositories periodically synchronized
  - Improves findability of facility data
- **Federated data storage**
  - FFSN: network of interoperable resources
  - Members contribute storage to a common pool
  - Facilitates dissemination, distributed computing
- **Streamline data sharing**
  - Hub for community building & data reuse by other users beyond original researchers
  - Use case: cross-facility datasets
    - Uniform interfaces to heterogeneous datasets, standardizes multi-modal analysis

FOXDEN: documentation



FOXDEN: overview



FOXDEN DEMO

https://foxden-demo.classe.cornell.edu:18344/

BACKUP SLIDE(S)

FOXDEN internals

- FOXDEN is designed to work with arbitrary schemas
  - but each schema file should define data-type, optionality of attribute, etc.
  - CHESS beamlines provide their own schemas
  - schema can be nested, DOI schema can be embedded into all CHESS beamlines schema, etc.
- The Dataset IDentifier (DID) is defined as a path of key-value pairs, e.g. /key1=value2/key2=value2/…
- DID are used to connect all FOXDEN services, like Provenance, Spec Scans, Data Management, Publication and refer to identify our datasets and define parent-child relationships

FOXDEN Services

- **Frontend:**
  - represent data in different views; provides uniform web UI and access to all FOXDEN services
- **Authz:**
  - handles LDAP, OAuth, OIDC providers and manage user's tokens
- **Metadata:**
  - experiment metadata are stored according to beamline schemas; user's metadata can be in free (unstructured) form
- **Provenance:**
  - provenance information is stored in RDBMS to allow easily scaling with data growth
  - we capture provenance information at different levels: OS, Environment, Scripts, Files, etc.
- **DataDiscovery**
  - discover and aggregate results across multiple FOXDEN services using common Query Language (QL)
- **DOI publication**
  - full integration with DataCite, MaterialsCommons and Zenodo DOI providers
- **AI/ML, Data Management, Notebook, Sync services**
  - supplementary services for specific tasks (still evolving)

FOXDEN Architecture

- **Services**
  - FOXDEN is composed by loosely coupled services via HTTP protocol/APIs
  - Each service is designed to take care of single task
    - currently we operate with 10 individual services
    - choice of service backend is driven by its use-case and scalability, e.g. Metadata are stored into document-oriented database (e.g. MongoDB) while Provenance data are backed by RDBMS (multiple backends are supported)
  - Data are interconnected via common Dataset IDentifier (DID)
- **Web frontend**
  - Users mostly interact with FOXDEN via web frontend
  - Data exploration, publication and management
  - Describes raw and reduced datasets
- **FOXDEN CLI tool**
  - One tool to answer all your questions about FOXDEN
  - Authentication, Metadata, SpecScans, Provenance management, DOI publications, synchronization, etc.
- **Technical details**
  - Code is written in Go language and (mostly) compiled into static executables
  - Zero setup/dependency requirements, OS agnostic, available on all hardware architectures (x86, ARM, Power8, RISC)
  - Build-in concurrency, minimal hardware requirements, JSON for data exchange and Yaml for configuration
  - Easily integrated with any software stack (via HTTP protocol)