

Ferrum Research: A Small Standalone Institutional Repository with OA Ingest and Server-Rendered Discovery

Jermaine Harris

Hutts Enterprises / Ferrus Labs

2026-09-20 · CC BY 4.0

Keywords: cs.DL, institutional repository, InvenioRDM, OAI-PMH, open access ingest

Source `git 6ef9c2d` · technical report

Abstract

Ferrum Research (<https://ferrumresearch.com>) is a MIT-licensed, InvenioRDM-shaped institutional repository implemented in Rust (Axum, SQLite/WAL, FTS5, content-addressed objects on Fil S3, ClamAV on file commit). It accepts only dataset and publication deposits. Open-access ingest uses arXiv, Unpaywall, and OpenAlex. Sci-Hub is forbidden. SSRN is a UX reference, not a scrape target.

A 2026-09-20 probe from an operator workstation found `/healthz 200` with `storage.backend=s3` and ClamAV ok. HTML routes `/`, `/search`, `/communities`, `/disciplines`, and OAI-PMH Identify also returned 200. Mean search TTFB for `q=arxiv` was 0.334 s ($n=10$) over the public Internet path. The first of those ten samples was 0.767 s; later samples sat near 0.22 s, which is WAN jitter plus whatever the edge had not yet cached, not an in-process FTS5 microbench.

The contribution is a documented small-IR subset and the refusals that keep the service operable on one VM, not a claim of Zenodo-scale hosting.

1. Introduction

Institutional repositories in the InvenioRDM tradition usually mean PostgreSQL, OpenSearch, Redis, RabbitMQ, and a Python web stack. That is the right shape for CERN-scale deposit. It is the wrong shape for a single VM that should stay restorable from SQLite backups and an S3 bucket.

Research question. What subset of InvenioRDM-like deposit (draft, malware scan, publish, communities, FTS5 search, OAI-PMH, OA ingest) is sufficient to run a public research portal on one VM, and what did we refuse?

Contributions:

- C1. Architecture (Figure 1) and crate list at `ferrum-rdm 7769685e5559fbf3db3068f211cfe7a9b59b7f70`.
- C2. Live probe table (Table 1) in `results/04/deployment-facts.json`.
- C3. Policy refusals: Sci-Hub, SSRN HTML scrape, WordPress or Next.js on `ferrumresearch.com`, GPL/Zenodo source in-tree.

Companion technical reports 1-3 in this series are intended deposits in the same IR.

2. System under test

Crates. ferrum-rdm-core, ferrum-rdm-db, ferrum-rdm-search, ferrum-rdm-storage, ferrum-rdm-jobs, ferrum-rdm-server, ferrum-ingest.

Runtime. Listen 127.0.0.1:3014 on host role utility-one; nginx TLS to https://ferrumresearch.com. SQLite WAL plus FTS5; durable jobs in SQLite; in-process cache; S3-compatible Fil One bucket for bytes; ClamAV 127.0.0.1:3310 on file commit.

Deposit API (InvenioRDM-shaped). POST /api/records draft with metadata.resource_type.id in {publication, dataset}; POST .../draft/files; PUT .../content; POST .../commit (ClamAV); POST .../draft/actions/publish. Auth: Authorization Bearer. Each publish is a new version id; older snapshots stay resolvable.

Discovery. Server-rendered /search (FTS5), /communities, /people/{slug}, /records/{id}, /oai2d, /disciplines. Anonymous home is search-first: hero query, latest publications, community cards, optional researcher spotlight.

Clean room. Independent MIT history; ignored .reference/ InvenioRDM docs; automated GPL/Zenodo marker check. Not affiliated with CERN or Zenodo.

Figure 1.

```
[browser] -> nginx TLS -> ferrum-rdm-server :3014
|
|- SQLite WAL (records, users, jobs)
|- FTS5 (search_fts)
|- Fil S3 (cas objects)
+- ClamAV (commit)

ferrum-ingest -> OA PDF (arXiv / Unpaywall) -> same API
```

3. Method

We read IMPLEMENTATION.md, deposit-types.md, ingest-policy.md, and HTTP contract tests, then probed production GET routes with curl from the operator box on 2026-09-20. Search timing: 10 GETs of /search?q=arxiv, mean TTFB. No secrets dumped. Optional ingest of a new arXiv id was not required for this qualitative paper. Probe times include the public Internet path to utility-one, not loopback.

4. Results

Table 1. Public probes, 2026-09-20. All HTTP 200. TTFB is WAN from the operator box. Source: results/04/deployment-facts.json.

Route	TTFB (s)	bytes	notes
/healthz	0.661	193	db, search, s3, jobs, clamav ok
/ping	0.470	2	OK
/	0.606	9099	anonymous portal HTML
/search	0.831	7868	FTS5 HTML
/search?q=machine+learning	0.465	6053	query page
/communities	0.664	5649	

/disciplines	0.894	6766	
/oai2d?verb=Identify	1.268	663	OAI-PMH
/api/vocabularies/resourceTypes	1.078	178	publication, dataset

Table 2. Search TTFB for q=arxiv, n=10, public HTTPS.

stat	seconds
mean	0.334
min	0.214
max (first sample)	0.767

Health body (abridged): status ok, storage backend s3 ok, clamav enabled ok, jobs pending 0, dead_letter 0.

OAI Identify returned repositoryName Ferrum Research. Resource-type vocabulary returned exactly two ids: dataset and publication. That is the deposit surface; there is no software or lesson resource type.

5. Discussion and threats

The sufficient subset on one VM is: two resource types, SQLite, FTS5, object store, ClamAV, OAI, OA ingest, server-rendered HTML. We refused a PHP CMS on the research hostname, a 69-discipline taxonomy as a database schema, email alerts as a go-live blocker, and any shadow library.

InvenioRDM in production typically means PostgreSQL, OpenSearch, Redis, RabbitMQ, and Python. Ferrum collapses that to SQLite and in-process jobs. That is an operational choice with obvious scale limits: single-node writer, FTS5 not a distributed index, object bytes on one Fil bucket.

WAN TTFB in Table 1 is not an in-DC microbench. /healthz at 0.661 s and /oai2d at 1.27 s are dominated by TLS plus Cloudflare plus the path to Hetzner, not by SQLite. The search series is the more useful kernel signal once the first sample is treated as cold.

Threats. (1) Probe TTFB is WAN. (2) Portal information architecture may change independently of this SHA. (3) We did not measure concurrent deposit throughput. (4) ClamAV “ok” is a sidecar ping, not a malware corpus evaluation. (5) OA ingest correctness (Unpaywall is_oa, publisher PDFs) is policy, not a precision/recall study.

6. Related work

InvenioRDM, Dataverse, EPrints, OAI-PMH, Unpaywall, OpenAlex, arXiv, SQLite FTS5. SSRN informed information architecture only. We do not scrape SSRN HTML and we do not fetch from Sci-Hub.

7. Ethics

OA ingest only. No Sci-Hub. No SSRN HTML scrape. Malware scanning on every committed file. This writing campaign did not flip FERRUM_RESEARCH_PORTAL_IS_LIVE and did not wipe /opt/ferrum-rdm/data.

8. Conclusion

A small IR can be a Rust binary, SQLite, an S3 bucket, and ClamAV if the operators accept two deposit types and OA-only ingest. The interesting part is the refusal list as much as the crate list.

Appendix A. Artifact

```
python3 harness/04-ferrum-ir/probe.py
```

Facts file: results/04/deployment-facts.md. The ZIP supplement includes JSON and the probe script.

References

- InvenioRDM. <https://inveniordm.docs.cern.ch/>
- The Dataverse Project. <https://dataverse.org/>
- EPrints. <https://www.eprints.org/>
- OAI-PMH. <https://www.openarchives.org/OAI/openarchivesprotocol.html>
- Unpaywall. <https://unpaywall.org/>
- OpenAlex. <https://openalex.org/>
- arXiv. <https://arxiv.org/>
- SQLite FTS5. <https://www.sqlite.org/fts5.html>
- SSRN. <https://www.ssrn.com/>
- ferrum-rdm (pinned). git SHA 7769685e5559fbf3db3068f211cfe7a9b59b7f70