04 ferrum research ir

git a4aed0a | 2026-09-20 | Jermaine Harris, Hutts Enterprises / Ferrus Labs

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

Jermaine Harris

Hutts Enterprises / Ferrus Labs

date: "2026-09-20"

license: "CC BY 4.0"

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

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). 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.

Keywords

cs.DL; institutional repository; OAI-PMH; open access; SQLite FTS5

1. Introduction

RQ. 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/Next.js on ferrumresearch.com, GPL/Zenodo source in-tree.

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 + 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`.

Discovery. Server-rendered `/search` (FTS5), `/communities`, `/people/{slug}`, `/records/{id}`, `/oai2d`, `/disciplines`.

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.

4. Results

Table 1. Public probes, 2026-09-20. All HTTP 200.

Route	TTFB (s)	notes
`/healthz`	0.661	JSON: db, search, storage s3, jobs, clamav ok
`/ping`	0.470	`OK`
`/`	(200)	anonymous portal HTML
`/search`	(200)	FTS5 HTML
`/search?q=machine+learning`	(200)	
`/communities`	(200)	
`/disciplines`	(200)	
`/oai2d?verb=Identify`	(200)	OAI-PMH
`/api/vocabularies/resourcetypes`	(200)	publication, dataset

Search TTFB mean for `q=arxiv`: **0.334 s** (n=10). These times include the public Internet path to utility-one, not loopback.

Health body (abridged): `{"status":"ok","storage":{"backend":"s3","status":"ok"},"clamav":{"enabled":true,"status":"ok"}}`.

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).

Threats. Probe TTFB is WAN, not an in-DC microbench. Portal IA may change independently of this paper. We did not measure concurrent deposit throughput.

6. Related work

InvenioRDM [inveniordm], Dataverse [dataverse], EPrints [eprints], OAI-PMH [oai-pmh], Unpaywall [unpaywall], OpenAlex [openalex], arXiv [arxiv], SQLite FTS5 [sqlite-fts5]. SSRN [ssrn] informed information architecture only.

7. Ethics

OA ingest only. No Sci-Hub. No SSRN HTML scrape. Malware scanning on every committed file. Companion technical reports 1–3 are intended deposits in this same IR.

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.

9. Artifact appendix

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

Facts file: `results/04/deployment-facts.md`.

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/