

Building Toward iRODS at CRG

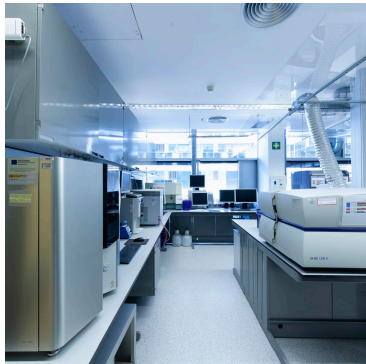
A pragmatic, incremental journey

Emyr James — Scientific IT, CRG

iRODS UGM 2026

About CRG

- Centre for Genomic Regulation — Barcelona; basic research in genomics & biomedicine
- International institute · Severo Ochoa Centre of Excellence
- Four research programmes: Computational Biology and Health Genomics · Quantitative Cell Biology · Genome Biology · Systems and Synthetic Biology
- Seven Core Technology Units (3dcores.crg.es): Genomics · Bioinformatics · Flow Cytometry · Proteomics · Microscopy (ALMU) · Protein Technologies · Tissue Engineering



The team — Scientific IT (SIT)

- We run HPC, storage and scientific-data infrastructure for CRG
- Methodology: infrastructure-as-code — reproducible, version-controlled
- We build in-house solutions where they fit better than off-the-shelf
- A team of 8, with specialties across:
 - storage systems · virtualization · DevOps · tape archive
 - web development · software development · HPC compute

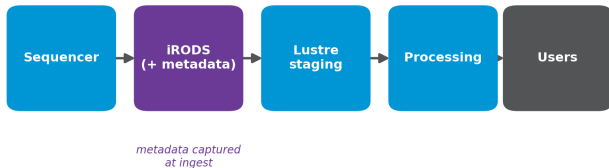


Storage — and why iRODS fits

- We already have plenty of storage: Isilon (NAS), Lustre (scratch), tape archive (LTO-8 / LTO-9)
- But storage $\neq$ data management
- The real questions: where does data live, who can see it, how is it described, how do we keep provenance?
- iRODS gives us catalogue + metadata + access control in one layer, on top of existing storage
 - we don't need iRODS to manage the bytes — Isilon/Lustre already do that well

Initial exposure: Wellcome Sanger Institute

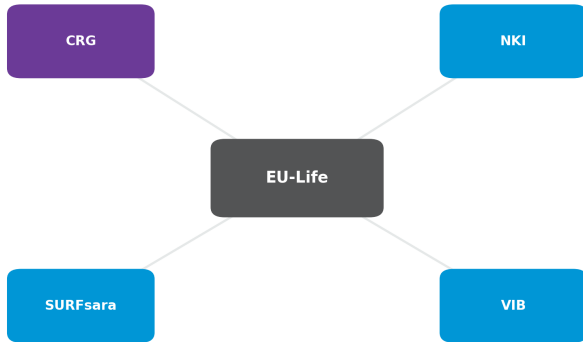
iRODS as bedrock at scale (Sanger)



- Sequencer → iRODS (+metadata) → Lustre → users
- Metadata captured at ingest
- Storage and compute separated
- My background: I saw it work at scale here

EU-Life: the idea spreads

EU-Life: lots of iRODS interest



- EU-Life meetings — institutes share what they're doing
- Lots of iRODS interest: NKI, SURFsara, VIB
- You don't have to figure this out alone

Analysis paralysis

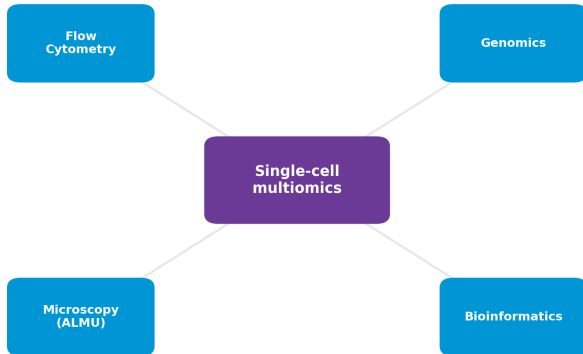
- The trap: trying to imagine the whole final state and 'get it right' before starting
- Endless architecture diagrams, no working system
- Months can disappear here

The catalyst: an RDM Lead

- A dedicated Research Data Management Lead joined (Jan 2025)
- A focal point whose job is to push this through
- Organisational readiness matters as much as technical readiness

Multiomics — a concrete problem

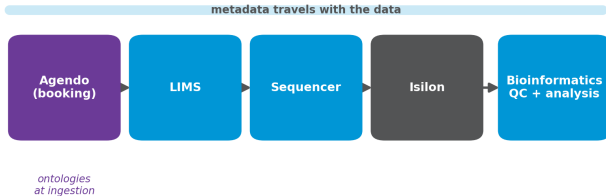
Many units, one sample — provenance + metadata at every step



- New instruments combine single-cell work across units
- Flow Cytometry · Genomics · Microscopy (ALMU) · Bioinformatics
- Data must flow between them while keeping provenance + metadata at every step

Agendo, the LIMS, and the data flow

From request to result



- Agendo — booking platform; metadata captured at the request
- Reused through the LIMS — not re-entered
- Pushed to iRODS in a community format:
- ENA ERC000011 (ELIXIR-Belgium)

A mission, and collaborators

- We picked a real mission: the multiomics drive
- Buy-in from the Genomics and Bioinformatics Core Technology Units
- A concrete problem with real users beats an abstract platform

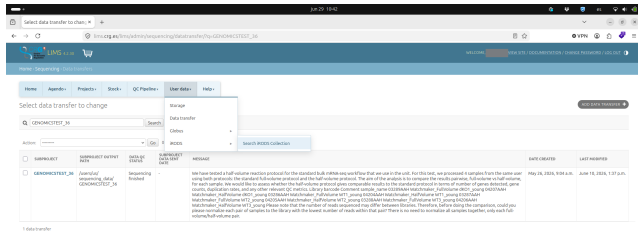
The reframe: MVP

- Minimal Viable Product — the smallest thing that delivers real value
- Start, learn, iterate — let the system grow with us

The minimal system we stood up

- iCAT in a VM + a single resource server mounting an Isilon folder
- No replication or storage features for now — Isilon handles all that
- Authentication — the hard but fundamental hurdle
 - once a cluster user could iput / iget, we had a foundation

An iRODS menu in the LIMS



- A small iRODS menu, added to the LIMS
- Search and open a collection
- (the push itself is on the next screen)



One screen: metadata + access control

Search IRODS Collection

Q GENOMICTEST_36 Search

IRODS Collection

SUBPROJECT	BIOS PATH	EXISTS	SIZE
GENOMICTEST_36	/CRG/home/qc_user/crg.es/us/GENOMICTEST_36	●	8.8 GB

Metadata

ATTRIBUTE	VALUE	UNITS	REMOVE
organization	Center for Genomic Regulation (CRG)	None	✖
unit	Genomics Unit	None	✖
instrument_brand	Illumina	None	✖
project	GENOMICTEST	None	✖
pi_email		None	✖
instrument	NextSeq2000, iSeq100	None	✖
subproject_name	GENOMICTEST_36	None	✖
application	Stranded mRNA-Seq	None	✖
organism	Mus musculus	None	✖
flowcell	222JL2NX, BWB90401-1509	None	✖
collection_size_bytes	9453472666	auto	✖

Add metadata

Attribute Value Units (optional) Add

Permissions

USERNAME	ZONE	ACCESS LEVEL	REMOVE
	CRG	read_object	✖
	CRG	read_object	✖
	CRG	read_object	✖
	CRG	read_object	✖
	CRG	own	

Add permission

Type to search LDAP users... read Add

Actions

Re-generate metadata

Delete IRODS collection

- Shows the Agendo metadata as AVUs; operator can add more
- Sets the ACLs (permissions) the data needs
- Operator verifies, then pushes via python-irodsclient
- Note: the AVU 'Units' column is mostly empty — remember this

Then we iterate — where next?

- Bioinformatics closed the loop with a Nextflow pipeline:
 - stage out to Lustre scratch → analyse → write derived data back to iRODS with more metadata
- Next: use the rules engine to do the staging and trigger pipelines
- Working with ALMU to integrate OMERO into the workflow:
 - OMERO captures all the microscopy metadata automatically
 - connecting OMERO to iRODS is a future aim
- Deliver QC'd data to external customers via Globus landing folders
- The approach now: small SOPs — or even parts of SOPs — some iRODS-native, some not yet

How we work now: weekly wins

- A core team of 6 now meets every week
- We discuss and commit to small, achievable wins for the week ahead
- Agile / scrum in spirit — a weekly cadence rather than daily
- Constant visible progress keeps motivation high
- A world away from the analysis-paralysis stage

The metadata question: are AVUs enough?

- iRODS metadata = AVU triples: Attribute / Value / Unit
- You just saw it: real AVUs, and the Unit column almost always empty
- RDF parallel — same intent, but NOT RDF: no URI identity, no graph, no inference; relational ICAT
- library_strategy = RNA-Seq is just a string, not a formal ontology term
- Not just us — the community says the same:
 - KU Leuven (UGM'23): AVUs fake hierarchy 'via namespacing and a custom use of units'
 - a UGM 2026 proposal (Groningen, 'iRODS Labels') tackles the same gap — one to watch

Making AVUs semantic: ontology binding

A statement that links three things — **subject** → **predicate** → **object**.

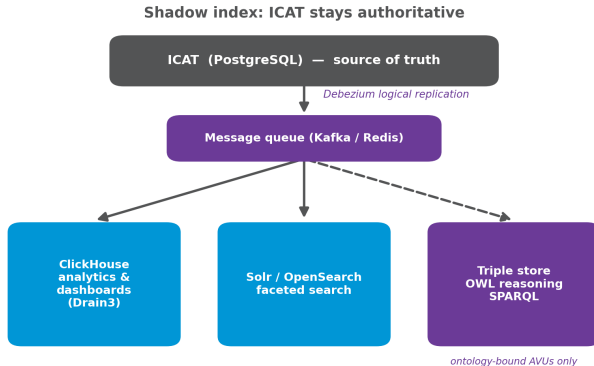
For example: *the sky* → *has colour* → *blue*.

An AVU is the same shape: the iRODS object is the **subject**, the **attribute** is the predicate, the **value** is the object.

	Attribute (predicate)	Value (object)	Unit (namespace)
Today (strings)	library_strategy	RNA-Seq	(empty)
Ontology-bound	library_strategy	EF0:0002770 (single-cell RNA seq)	https://www.ebi.ac.uk/efo/

Replace that text with an ontology term — EF0:0002770 — and software can tell it's a *kind of* sequencing assay; the unused **Unit** field names the ontology.

Where metadata could go



- Ontology binding: value = an ontology term; unit = its namespace
- ICAT = source of truth; sync changes out to query stores
- ClickHouse (analytics) · Solr (search) · triple store (SPARQL)
- Sync via Debezium CDC off the PostgreSQL WAL

Fitting in: the official iRODS indexing plugin

- iRODS already ships `irods_capability_indexing` — a shadow-index framework
 - indexes AVU metadata + full text; Elasticsearch only today
 - backend is a pluggable rule-engine plugin: register policies + speak HTTP/JSON
 - syncs via rule-engine delay rules (PEP-triggered) — not real-time CDC
- Two ways we could extend it:
 - A — backend plugin: add ClickHouse / Solr / triple-store backends (upstreamable)
 - B — Debezium CDC off the ICAT WAL: captures everything, incl. out-of-band changes
- Real question for the room / Consortium: where should this work live?

The bigger picture: core first, then everyone

- Build the data workflow for the Core Technology Units first
- Once that's smooth, research groups should slot into it naturally
- Add user-friendly interfaces — we're looking at Yoda and Mango
 - both are iRODS-based portals; they ride on the same metadata layer
- The interface can evolve — the iRODS core stays solid

And the bigger goal: FAIR data

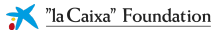
- Getting data + metadata into iRODS moves us much closer to the FAIR principles:
 - Findable — rich, queryable metadata (and, ahead, ontology-bound)
 - Accessible — controlled access via ACLs and clear protocols
 - Interoperable — structured, standards-aligned metadata (ENA-ready)
 - Reusable — provenance and context captured at every step
- Not a checkbox — a direction the architecture naturally supports

Acknowledgements

BOARD OF TRUSTEES



Barcelona Institute of
Science and Technology



MEMBERS OF



Thankyou

We're learning the system as we go — it's growing organically with us

See you Thursday at the troubleshooting session

Emyr James · emyr.james@crg.eu