Peer-Learning Session · Institutional Repositories & Open Academic Data

FAIR²

Making research data work for people and machines — and protecting how it is reused

What we are learning about AI-ready research data

Cristina E. Gonzalez Espinoza
AI Data Science Specialist · Frontiers Media SA
Product Owner, open FAIR² specification

Monday 7 September 2026 · Lausanne

Open is not the same as usable

Most institutional repositories already do FAIR well.

Persistent DOIs, ORCID identifiers, Schema.org on landing pages, signposting, curated deposits. That foundation is largely built.

But hand a deposited dataset to a machine and it still needs a human:

to work out which file matters, guess what the columns mean, read the paper to find the units, and judge whether this kind of reuse is even permitted.

That gap is where FAIR² sits.

✓ **What FAIR gave us**

A record that people can find, reach and cite — discoverability and basic interoperability, backed by persistent identifiers.

? **What is still missing**

A machine-readable schema. Types and units. Provenance. Documented limitations. And explicit, machine-checkable conditions for reuse — including AI training.

THE CONCEPT

FAIR² = FAIR + AIR + RAI

FAIR

THE 2016 BASELINE

Findable, Accessible, Interoperable, Reusable. Discovery and human reuse — but not machine-actionability, and nothing about responsible use.

AIR

AI-READY

Machine-actionable structure on open standards, primarily MLCommons Croissant. Schemas, types, units and statistics encoded, so an ML framework loads the dataset natively — no bespoke parser.

RAI

RESPONSIBLE AI

Human-in-the-loop curation, privacy by design, documented bias and limitations, and graded access — expressed with the Croissant RAI vocabulary.

BUILT ON STANDARDS THAT ALREADY EXIST — NOT A NEW SILO

- MLCommons Croissant
- Schema.org
- PROV-O
- QUDT (units)
- CRediT
- Croissant RAI

Open specification at fair2.ai, governed by a non-profit community association and vendor-neutral. Conformance is testable: metadata validated with SHACL, access is graded as identity tiers L0–L2, per column.

Dumping files, and building a package

What usually gets deposited

- A folder of files, named by whatever convention the lab used
- A README, if you are lucky
- The schema is implicit — in the file names, or in someone's head
- Units, methods and conditions live in the paper, or its supporting information
- A licence, and a link to the article

Reusable by: *a person, with time — and ideally the author's phone number.*

What ships in a FAIR² package

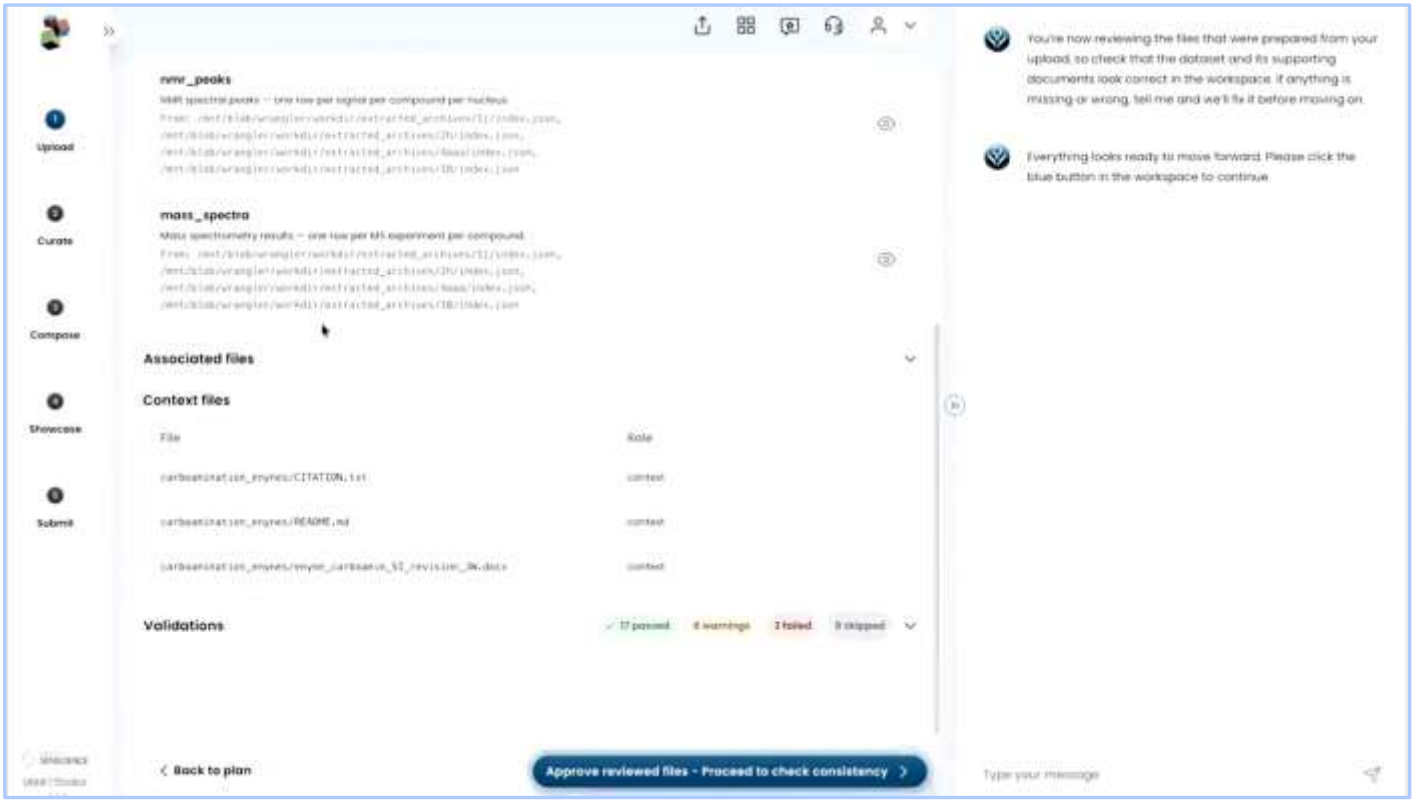
- Typed tables, every field described, with units attached
- fair2.json — a Croissant manifest a standard ML library can load
- Provenance: where each file came from and what was done to it
- Extended Methods, structured down to the individual step
- Machine-readable reuse conditions, including AI training
- An interactive portal, a notebook, and a DOI

Reusable by: *a person and a machine, without asking anyone.*

The files are the deposit. The package is the deposit plus everything that was in someone's head.

How a dataset becomes a package

The same catalysis dataset, going through the app. AI proposes at every step; a person approves before anything moves on.



FIVE STEPS IN THE APP

- **Upload —** files in, and a plan for how they are organised
- **Curate —** types, units and a data dictionary, proposed then approved
- **Compose —** the article and the Methods, drafted from what is there
- **Showcase —** the portal and its explorers built for this dataset
- **Submit —** journal, licence and rights confirmed by the author

Try the sandbox
app.sandbox.sen.science/auth

Nothing here is automatic. Every step ends with someone clicking Approve and Continue.

What FAIR² Data Management delivers

The specification is open and anyone can implement it. This is what Frontiers builds on top — one submission, seven components.

Data Article

Peer-reviewed and citable, in a Frontiers journal — narrative plus extended Methods.

Data Portal

An interactive explorer for that dataset. Nothing to download.

Data Package

Data, Croissant manifest, dictionary, units, provenance. SHACL-validated, own DOI.

Notebook

Loads via mlcroissant — AI-ready made executable.

Podcast

A short audio walkthrough of the dataset and how it was produced.

Social media kit

Ready-made posts and graphics, so the dataset gets seen.

AI data steward

Clara answers questions about the study from the metadata — how it was run, who contributed, what a field means.

See one live: `sen.science/doi/10.71728/senscience.4f2j-8h1k`

Atoll biodiversity, Indo-Pacific — 310 atolls · 90 variables · 4,215 species records from 677 sources · taxa linked to GBIF, BOLD and NCBI

Curation → **Certification** → **Publication** → **Preservation** → **Reuse**

EXAMPLE 1 · LIVE DEMO

A dataset that was already open

Catalysis chemistry from NCCR Catalysis — NMR, IR and mass-spectrometric characterisation of 69 synthesised compounds

BEFORE · ALREADY OPEN ON ZENODO

`zenodo.org/records/18851416`

- 72 files, in folders named after compound numbers — or, in the deposit's own words, "self-describing"
- Structures included as .mol files
- Conditions and equipment: "in the supporting information of the article"
- CC-BY-4.0, a DOI, and a link to the paper

AFTER · AS A FAIR² DATA PACKAGE

`sen.science/doi/10.xxxxx/senscience.x8ou-mo2o`

- Five typed tables — compounds, NMR spectra, NMR signals, IR peaks, MS spectra — explorable in the page
- Methods as addressable steps, down to the individual bench operation
- Croissant manifest, a Jupyter notebook, the analysis scripts
- Same licence, same authors, same science

Solé-Àvila, Brownsey, Senelle & Waser (EPFL) · article in Angewandte Chemie · deposit CC-BY-4.0, shared with me by the group's data officer

EXAMPLE 2 · A DIFFERENT STARTING POINT

When there is no deposit to start from

Neuroscience with EPFL and the Open Brain Institute — June to August 2026

What arrived

- Raw instrument files, in non-standard formats, straight off the rig
- Supporting material scattered across related publications
- Nine datasets: patch-clamp, multimodal single-cell, connectivity and its derived analysis, dendritic spines, synthesised morphologies, cell density series

What it took

- Transformation before description was even possible — and iterative, not one-shot
- Methods drafted from related papers, then corrected by the authors
- A purpose-built explorer per data type: spectra, spatial maps, individual spine locations
- Every article and portal back to the authors before submission

2
institutions

9
datasets

3
months

Published — one of the nine:
sen.science/doi/10.71728/senscience.bt07-sboc
244 morphologies · 641 recordings · 396 slices · SWC + NWB + JPEG, linked by cell ID
The rest is live work — and none of it started from a deposit we could point a tool at.

What it cost us, three times over

1

Heterogeneity is the real cost

Data never arrives standard, and transformation is iterative rather than one-shot: first runs failed, scripts were revised, subsets were tested before the full run — one run stopped because it filled the storage. Budget for iteration, not conversion.

2

Structuring data is a discovery process

The gaps only appear when you try to build a complete, validated record: files on disk with no matching acquisition record, sample metadata incomplete for part of a cohort, inclusion decisions that had never been written down. The specification finds the gaps; only the original team can close them.

3

Methods are the scarce resource

The last thing missing before review was almost never the data. Semantics and provenance live in people and publications, not in the files — and we ask for them at the worst possible moment, after the paper is out and the people have moved on.

Sharing as a requirement, or sharing as a plan

For most researchers, data sharing is now something that happens after the article is accepted. That is the expensive place to put it.

Sharing as compliance

a deposit obligation, met at the end

- Data prepared under deadline, after the science is finished
- Conventions reconstructed from memory
- Metadata written by whoever is still on the project
- Licence and access decided last, sometimes by no one

Sharing as design

a research-plan decision, made at the start

- Formats, vocabularies and identifiers named up front
- Metadata captured at acquisition, not reconstructed
- Methods written while the work is being done
- Access and reuse conditions agreed before collection

A data management plan is the cheapest place to make this happen. Five things worth committing to in writing:

File and format conventions

Vocabularies and units

Where sample metadata lives

Who writes Methods, and when

Access & AI-reuse conditions

WHERE THIS IS GOING

FAIR² Discovery

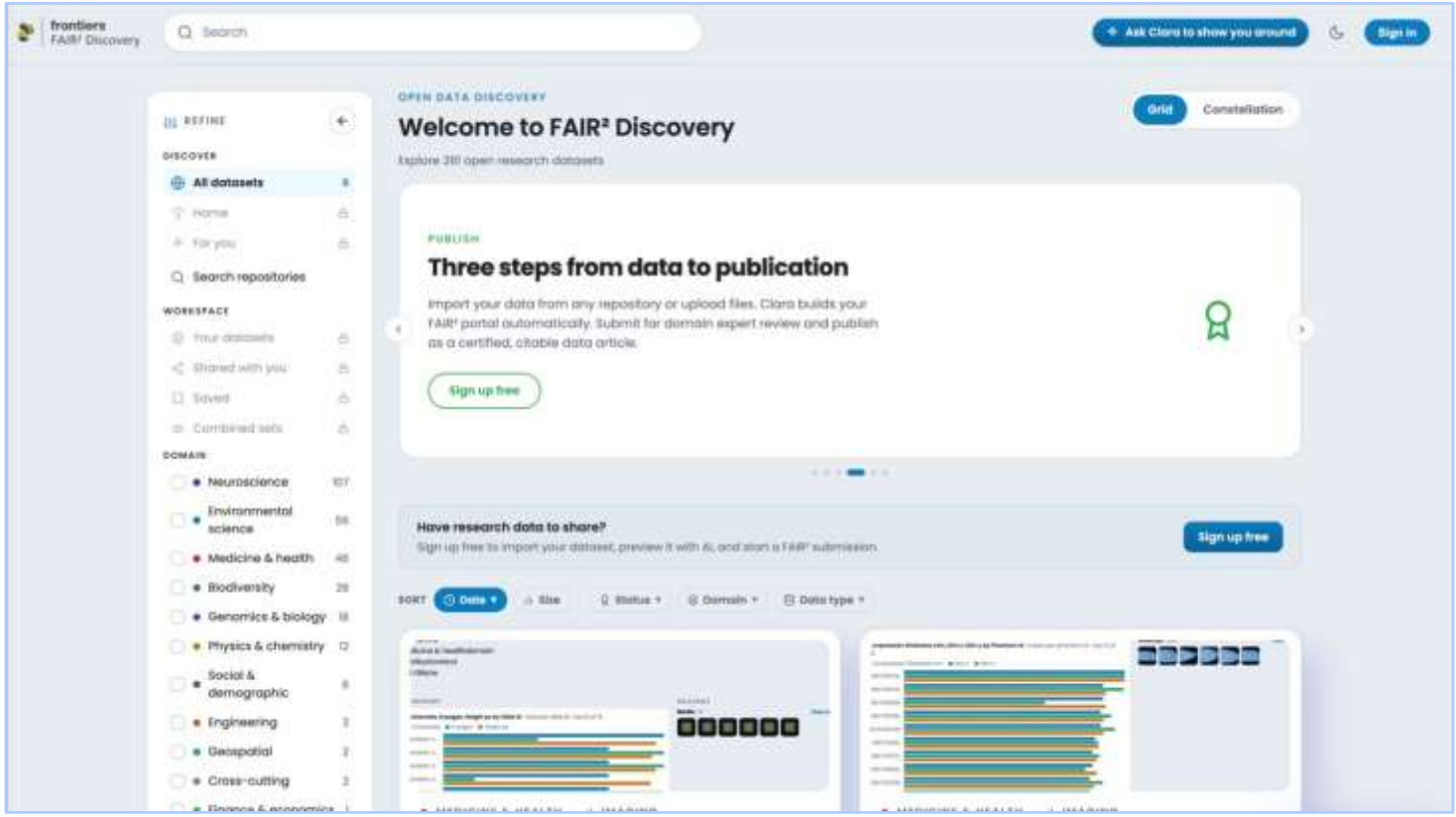
An early project: a discovery layer across repositories, where every dataset gets its own portal

THREE STEPS FROM DATA TO PUBLICATION

1. Import from any repository, or upload files
2. Clara drafts the FAIR² portal automatically
3. Domain-expert review, then a certified, citable data article

LIVE DEMO

fair2-discovery.dev.internal.senscience.cloud



How access is graded: three tiers

A data use agreement is a promise, not a mechanism. Every major research-data breach of the last decade involved people who had signed one.

L0 Open

- No authentication. Data is distributed unencrypted.
- The document still carries provenance, licence and citation — but enforcement is contractual.
- No audit trail of who used it.

L1 Registered

- Proved control of an identifier — a login — and agreed to the terms.
- No real-world identity check by default.
- Columns are encrypted; keys released per tier.

L2 Credentialed

- The steward composes the requirements, rather than accepting a fixed set.
- Default is a signed DUA; add institutional affiliation, ethics training, identity assurance.
- The highest-sensitivity columns live here.

One dataset can use all three, column by column. A clinical dataset might publish demographic summaries at L0, diagnosis at L1, and patient identifiers at L2 — and record partitions do the same job row by row, so a cohort can be open in part and credentialed in part. L0 assets are plaintext with no key at all; L1 and L2 columns each get their own.

The ladder is aligned with the GA4GH and PhysioNet models — not a scheme we invented.

Governance composed from ten primitives

The tier ladder is the first. These nine say **under what conditions** — combined per dataset, so a census table and a paediatric study need not share one setting.

+G Community governance

Named approvers must vote. For community-governed data the key is split among them — CARE and OCAP® as a prerequisite, not a policy statement.

+S Secure environment

Where processing may happen: open, a secure environment, or compute-to-data where only results leave.

+C Client certification

Only certified software receives keys — and a certificate can be revoked across the network.

+P Purpose binding

Declare why you need it, checked against allowed and prohibited lists.

+J Jurisdiction

Where data may be processed. A property of the environment, not the requester.

+Q Query governance

Query granularity per tier, cross-dataset linkage, and registration of derived data.

Constraint axes

AI training, commercial use, export and combination limits.

Embargo

A temporal gate, with named exemptions for the consortium that produced the data.

Record partitions

Which rows a rule applies to — a scoped view, so aggregates stay honest.

Honest about the edges: this controls decryption, not perception, and limits on AI training after decryption stay contractual — audit trails and revocation, not physics. What are we missing?

OVER TO YOU

Four questions I would like your experience on

1. Where does your repository lose the most time — at the moment of deposit, or reconstructing what happened months earlier?
2. Does your institution’s DMP process actually change what arrives at the repository? If not, what would have to change so that it did?
3. Who writes Methods documentation at your institution — and at what point in the project?
4. And on governance: what are we missing?

Cristina E. Gonzalez Espinoza
cristina.gonzalez@senscience.ai · Frontiers Media SA

The open specification
fair2.ai

FAIR² Data Management
frontiersin.org