From Instruments to Reusable Data: an AI Layer for Experimental Data Capture and Reuse

Phase 1 proposal - AI and Collective Intelligence

Problem

Experiments often generate far more raw information than is used in the final analysis or publication. Data are usually selected for the original scientific question, while unused measurements, acquisition metadata, intermediate outputs and experimental context may remain local, poorly documented, or never become searchable. This limits reproducibility and wastes potentially valuable data precisely when biology increasingly needs large, heterogeneous datasets for AI and virtual-cell models.

Proposal

Develop a lightweight **AI data layer integrated with experimental instruments and laboratory workstations**. Its role would be to preserve as much reusable raw output as possible, automatically contextualize it, and make later reuse simple.

1. Capture at source. A local agent detects new files or data streams and records raw outputs together with instrument model, software version, acquisition settings, timestamps, calibration, sample identifiers and protocol information. AI maps these to standard vocabularies and asks the researcher only for missing or ambiguous metadata.

2. Package and publish. The system creates a FAIR, machine-readable package containing provenance, quality-control information, checksums, licence and access restrictions, then submits or links it to existing repositories rather than replacing them. Sensitive human data remain under controlled access; raw-data preservation respects consent and storage constraints.

3. Retrieve and contextualize. A researcher describes a new scientific task in natural language. The AI searches across repositories, labels and harmonizes relevant datasets, ranks them by relevance and quality, and warns about context differences such as species, cell type, assay, instrument, protocol, batch, calibration or environmental conditions.

Biological pilot

A realistic first implementation could focus on **cellular microscopy**, where instruments generate rich raw images and acquisition metadata that are often reduced to selected figures or processed measurements. The prototype could connect one microscope/workstation to an OMERO-style local store, automatically generate REMBI-compatible metadata, and prepare deposition to the BioImage Archive/BioStudies. A retrieval module would then test whether public image datasets can be assembled for a new task while exposing acquisition and sample differences. The same architecture could later extend to sequencing, flow cytometry, proteomics and perturbation data.

Relation to existing approaches

Existing approach	What it already does	Missing element addressed here
SRA / GEO / BioImage Archive / BioStudies	Public repositories for raw or study-level biological data.	Mostly deposition endpoints; no general instrument-to-reuse AI layer across modalities.
OMERO	Imports, manages, annotates and shares microscopy data.	Not a cross-repository system for contextual discovery, ranking and compatibility warnings.
FAIR Data Station / LEAF; d-post	FAIR metadata tooling and increasingly automated device-side capture.	Closest predecessors; this proposal adds AI semantic contextualization, publication routing and reuse-oriented retrieval across repositories.

Why it matters for collective intelligence

The main contribution is not another database, but a **continuous instrument-to-community loop**: **capture -> contextualize -> publish -> retrieve -> compare -> reuse**. Reducing the human cost of sharing would expose data that are currently invisible, while preserving provenance and uncertainty would let AI systems combine datasets without pretending that experiments performed under different conditions are equivalent. A first prototype can be small, but the architecture naturally scales toward community-built datasets for virtual-cell modelling.

Closest references: FAIR principles - Wilkinson et al., *Scientific Data* 3, 160018 (2016), doi:10.1038/sdata.2016.18. BioImage Archive / BioStudies (EMBL-EBI); OMERO (Open Microscopy Environment); FAIR Data Station / LEAF (FAIR by Design); Fitz et al., *d-post: Device-Side Data Capture for FAIR Laboratories* (2026), doi:10.52825/ocp.v9i.3317.