FAIR from Creation to Curation

Technologies, Lab Upgrades and HPC Needs

Stefano Cozzini

RIT Director · Area Science Park

EOSC Symposium 2026 · Florence, 15 October 2026 · Sala Verde

Setting the scene: research infrastructures at Area

AREA SCIENCE PARK · TWO RESEARCH INFRASTRUCTURES

RI LIFE SCIENCES

Life Science

Study of pathogens with a multi-omic approach: genomics, epigenomics, microbiology, structural, cellular and molecular biology, biophysics, microscopy

PRP@CERIC · Pathogen Readiness Platform for the CERIC-ERIC upgrade

INGenIO · next-generation digital infrastructure for the study of rare diseases

Area Lab participating Labs: Genomics and Epigenomics (Trieste) · Multiomics (Salerno) · Virology Unit (Trieste)

NFFA-DI · AREA IS PARTNER

Material and Nano Science

research infrastructure for **nano-sciences** and **nanotechnologies** with applications

NFFA-DI · Nano Foundries and Fine Analysis Digital Infrastructure: upgrade of NFFA for nanoscience and nanotechnology, with applications in materials

Area Lab participating : Electron microscopy Lab.

Digital research infrastructure at Area

AREA SCIENCE PARK · ONE DIGITAL ECOSYSTEM FOR BOTH INFRASTRUCTURES

RI LIFE SCIENCES

Life Science

PRP@CERIC · INGenIO

Area lab participating Labs: Genomics and Epigenomics (Trieste) · Multiomics (Salerno) · Virology Unit (Trieste)

NFFA-DI · AREA IS PARTNER

Material and Nano Science

NFFA-DI · Nano Foundries and Fine Analysis

Area lab participating: Electron Microscopy Lab

ORFEO digital ecosystem

storage, HPC/AI resources and FAIR repositories, transversal to both infrastructures

Data management is part of the research infrastructure, **not an add-on** to it.

Our data ecosystem: interoperability

HETEROGENEOUS DATA, ONE FAIR LAYER, ONE DIGITAL ECOSYSTEM

AREA SCIENCE PARK

Two roads to FAIR data

THE QUESTION

the view across a whole research infrastructure.

ROAD 1 · CURATION AFTER THE FACT

Make it FAIR later

- Metadata rebuilt from files, notes, memory
- Provenance reconstructed, often partially
- Curators do the work, long after the experiment

ROAD 2 · FAIR BY DESIGN

Make it FAIR at creation

- Metadata captured at acquisition, automatically
- Standard formats straight from the pipeline
- Provenance recorded as the sample moves

What each road costs, and what it buys

SIDE BY SIDE

	FAIR later	FAIR by design
Metadata	Rebuilt later from files and memory	Captured at acquisition, automatically
Completeness	What survived; gaps filled by ML or hand	Set by the schema; gaps visible at once
Provenance	Reconstructed, often partially	Recorded as the sample moves
Effort	Curators, per dataset, repeated	Engineers, per instrument, once
Cost curve	Low up front, grows with the backlog	High up front, low per dataset
Main risk	Information lost for good	Rigid pipelines, weak lab buy-in

Two roads to FAIR data: our experience

WHAT FOLLOWS · EXAMPLES FROM BOTH ROADS, AND THE LESSONS WE LEARNED

ROAD 1 · FAIRIFICATION AFTER THE FACT

Two examples, two outcomes

SUCCESSFUL

STM image archive · CNR-IOM

Metadata extracted and enriched with ML, provenance with W3C PROV, archive explorable online

NEGATIVE

Genomics data archive · LAGE

~150 TB produced 2014–2020: well stored but poorly annotated, still not FAIR

ROAD 2 · FAIR BY DESIGN

From blueprint to practice

NFFA-DI digital infrastructure · 11 operative units, OFED data lake, single entry point

OFED and its services · buckets, NOMAD Oasis, NeXus metadata generator

Bottom-up interoperability · two CNR labs, one common data layer

HPC in the lab · distributed basecalling at LAGE, no manual transfer

Then the lessons learned: **what each road costs, what it buys, and what we would do today**

AREA SCIENCE PARK

ROAD 1 · FAIRIFICATION AFTER THE FACT

The good one: FAIRifying an STM image archive

ROAD 1 · THE POSITIVE EXAMPLE · NFFA AND CNR-IOM

1

Extract

Instrument metadata pulled from each image file into a structured database

2

Enrich

Human annotation, machine learning and filtering to recover surface structure and composition

3

Trace

Provenance described with the W3C PROV standard

4

Expose

STM Explorer web service on Trieste Advanced Data services (TriDAS)

It worked, but what the instrument never recorded had to be **guessed, inferred or annotated by hand**.

PUBLISHED

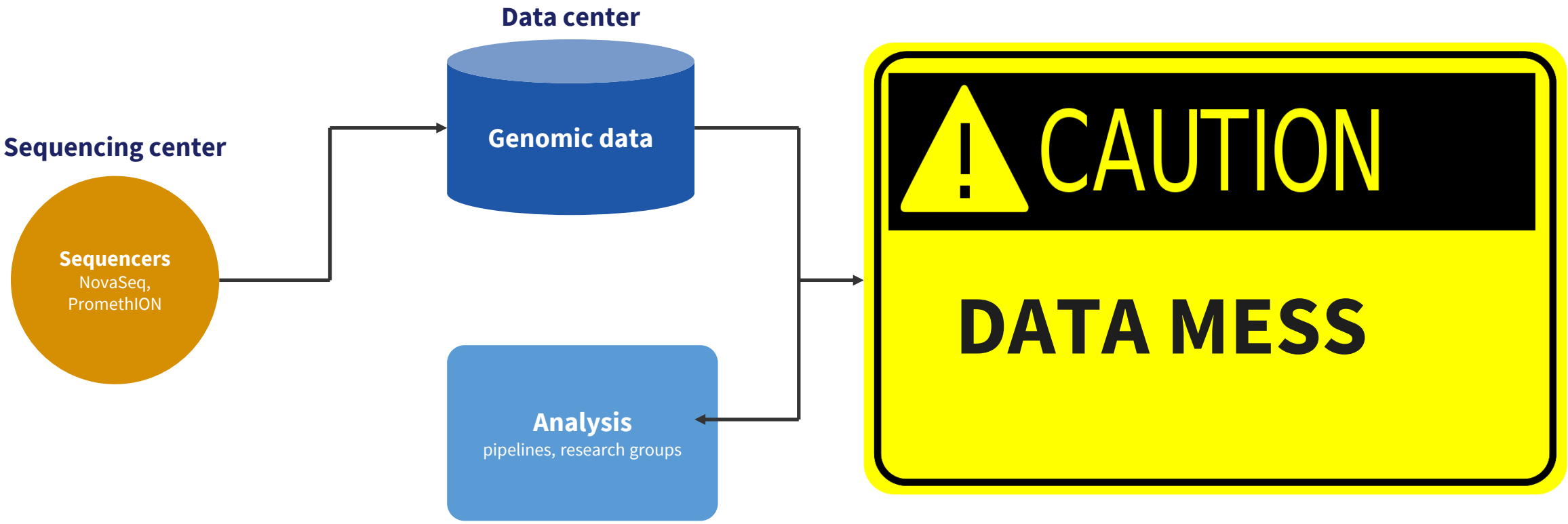
Towards the FAIRification of Scanning Tunneling Microscopy Images

T. Rodani, E. Osmenaj, A. Cazzaniga, M. Panighel, C. Africh, S. Cozzini · *Data Intelligence* 5(1):27–42, 2023 · doi:10.1162/dint_a_00164

AREA SCIENCE PARK

The bad one: genomic data workflow

ROAD 1 · NEGATIVE EXAMPLE · GENOMICS AT LAGE

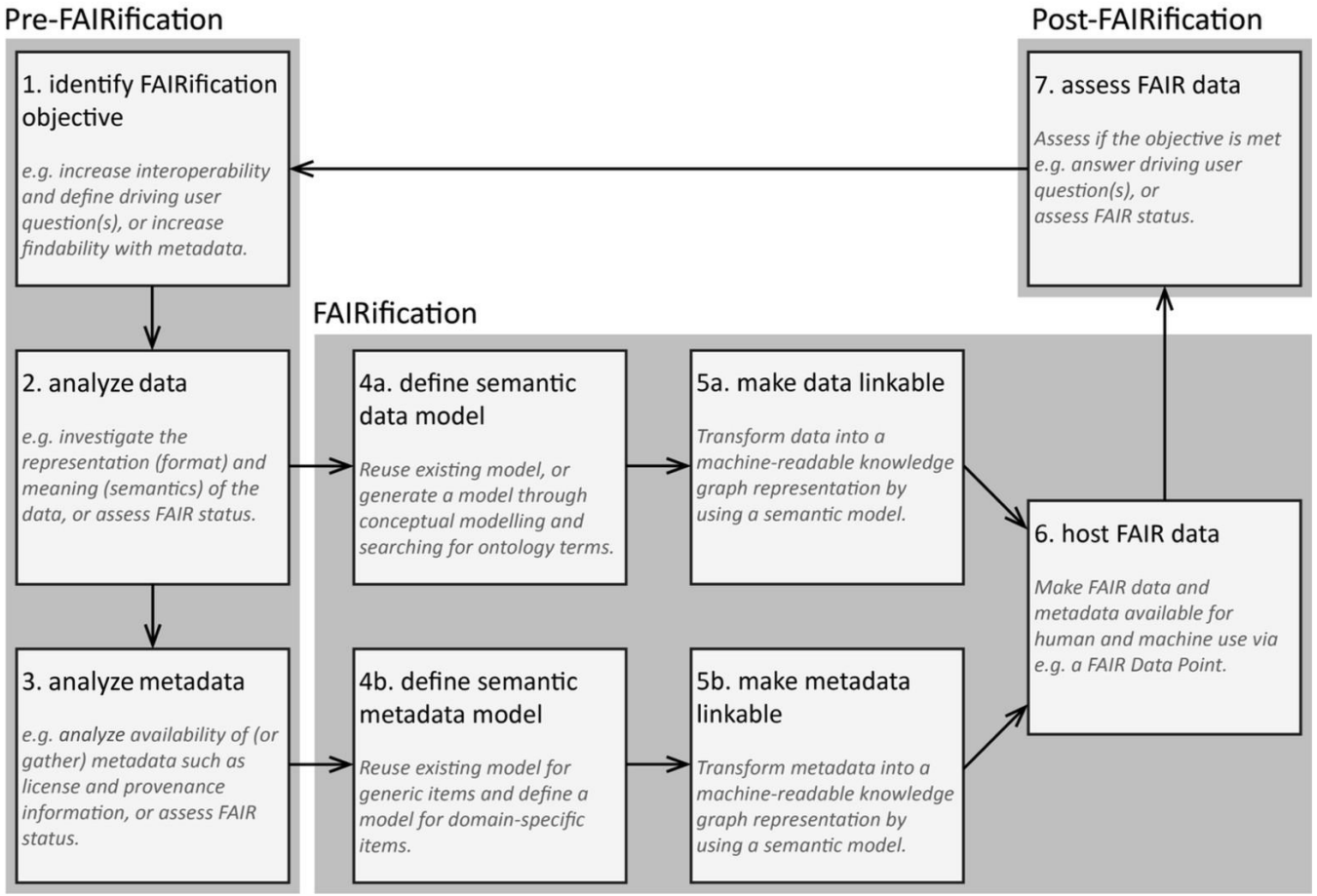


The data moved, the context did not: **stored safely, poorly annotated**

AREA SCIENCE PARK

Let us use a FAIRification workflow...

ROAD 1 · NEGATIVE EXAMPLE · THE TEXTBOOK APPROACH



From: A Generic Workflow for the Data FAIRification Process, *Data Intelligence* 2(1-2):56–65, 2020, doi:10.1162/dint_a_00028

AREA SCIENCE PARK

Our final workflow

ROAD 1 · NEGATIVE EXAMPLE · WHAT FAIRIFICATION AFTER THE FACT BECAME IN PRACTICE

fairification.sh

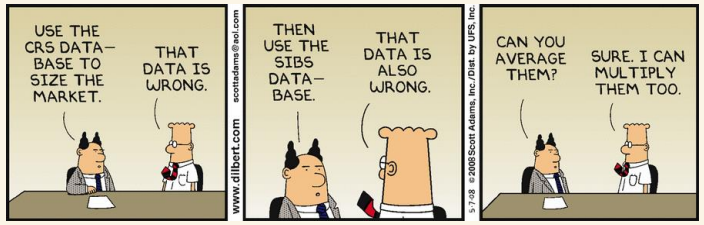
```
until (tired == true) do
    discuss_with_users       # days and days
    wait_for_user_feedback   # months and months
    discuss_again            # days and days
end

rm -rf /old-data/*
```

AREA SCIENCE PARK

Lessons learned

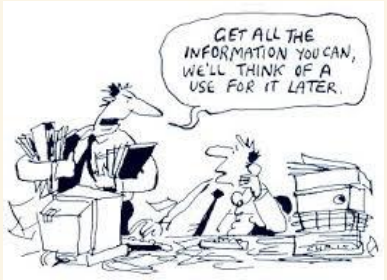
ROAD 1 · WHAT WENT WRONG

#1 Lack of agreement



- Scientists and data curators disagree on how to treat data
- Metadata registered by hand, only at publication time
- No clear, common (meta)data models

#2 Data wrangling



- Different formats, often proprietary
- No standardization
- Different information and units
- Incomplete, incincompatible data

#3 Inactivity of scientists

- Old "handmade" programs, pen and copybook
- Data intellectual property
- Data shared on hard disks and USB pens
- "Metadata registration is a waste of time" (cit.)

AREA SCIENCE PARK

ROAD 2 · FAIR BY DESIGN

AREA SCIENCE PARK

NFFA-DI: the research digital infrastructure

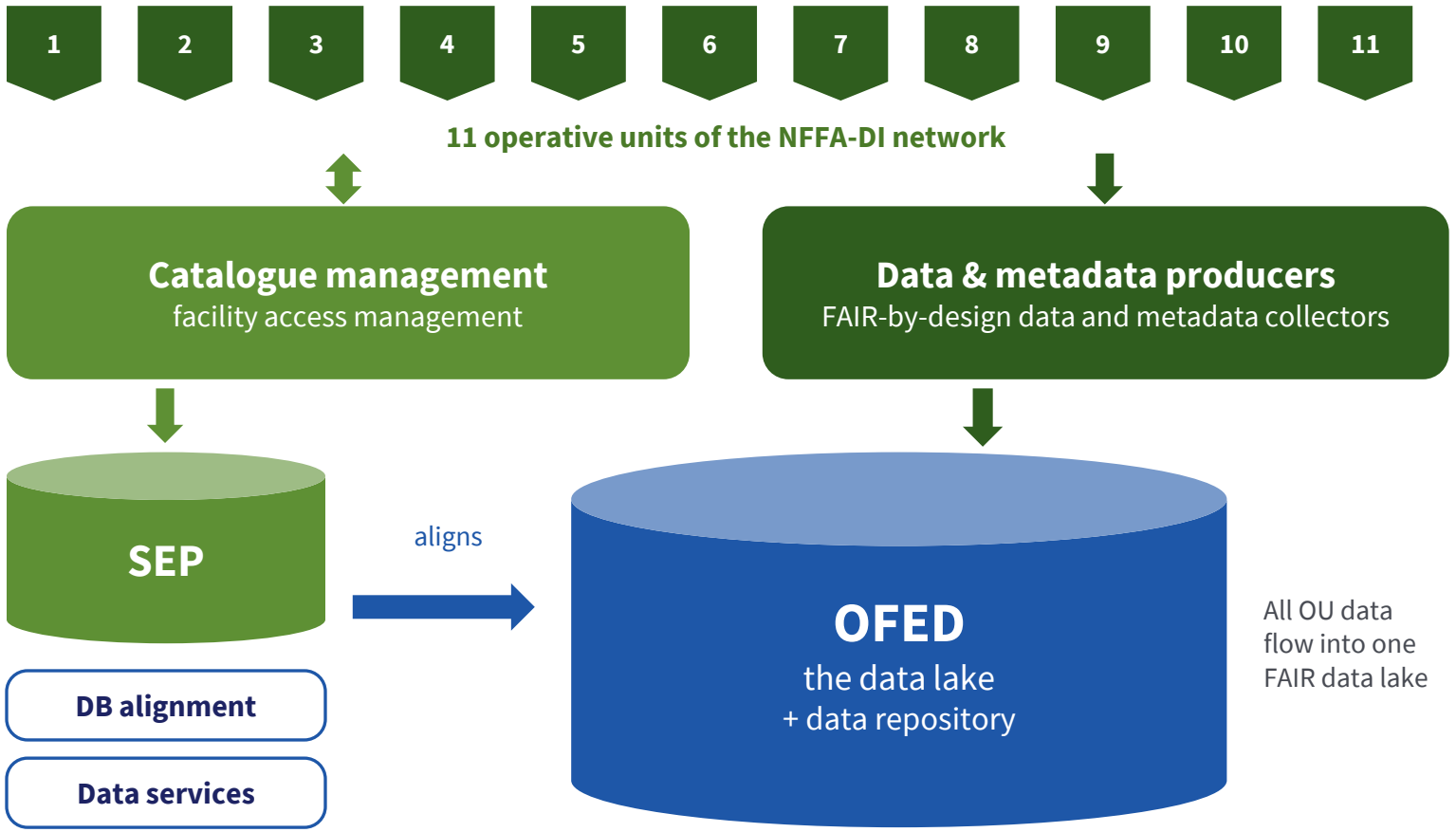
NANO FOUNDRIES AND FINE ANALYSIS · DIGITAL INFRASTRUCTURE (PNRR)

nffa–di

1 **The network**
11 operative units: labs and facilities producing data : new instruments fait-by-design ready.

2 **OFED**
the data lake, with the data repository

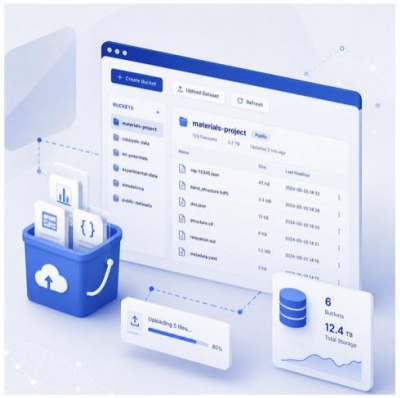
3 **SEP**
single entry point: coordinator and manager of access



AREA SCIENCE PARK

OFED and its services

NFFA-DI · DATA SERVICES ON TOP OF OFED



Buckets Explorer

A user-friendly interface that allows researchers to dynamically create and manage buckets, upload and list datasets.

Get access

Nomad Oasis

A platform developed by FAIRmat that maximizes material science data and metadata visualization, integrated by customized plugins and integrates the NORTH environment to support Jupyter Notebook for real-time data science analysis.

Get access

FAIR NeXus Metadata Generator

An AI-driven service designed to assist researchers in automating the transformation of heterogeneous datasets into the standardized NeXus format.

Get access

XCD

An interactive environment used by theoretical materials science researchers for constructing, visualizing, and manipulating atomic-scale models and structures.

Get access

AREA SCIENCE PARK

Bottom-up interoperability: two labs, one data layer

NFFA-DI EXAMPLE · CNR-IFN TRENTO AND CNR-ISMN BOLOGNA

Same techniques, different instruments

Two laboratories in different institutes, each with its own data production workflow

Common semantics

Agreed metadata, even though the community has no shared standard yet (Material Registry, NanoFabNet, ISO TS 80004)

Common syntax

One data format, although the original workflows produced different ones

One plugin

A single upload plugin for the shared data platform

Result: data produced at CNR-IFN and CNR-ISMN are **now fully interoperable**

AREA SCIENCE PARK

HPC infrastructures for data

A PERSONAL THREAD · FROM HPC INFRASTRUCTURE TO DIGITAL ECOSYSTEM

Then: infrastructure

HPC as infrastructure
clusters, grid, interconnects, queues, more flops

Now: a digital ecosystem

Hardware
HPC, cloud, storage: ORFEO

Software and DATA
codes, workflows, FAIR data

Brainware
people, training, data stewardship

Data is a fundamental part of the scientific process: not the only one, not necessarily the most important, but no longer optional.

AREA SCIENCE PARK

HPC in the lab: live basecalling at LAGE

HPC NEEDS · FROM THE SEQUENCER TO GPU NODES, AS DATA ARE PRODUCED

1 Why basecalling is hard

- Turning a noisy electrical signal into nucleobases
- From statistical methods to deep learning
- Precision vs speed : High Accuracy vs Super accuracy
- A very compute-intensive task: Sequencer equipped with GPU

2 The live basecalling gap

- Goal: basecall as fast as we sequence
- With its own computing units, our lab reaches only:

58%
of the minimum HAC speed

10%
of the minimum SUP speed

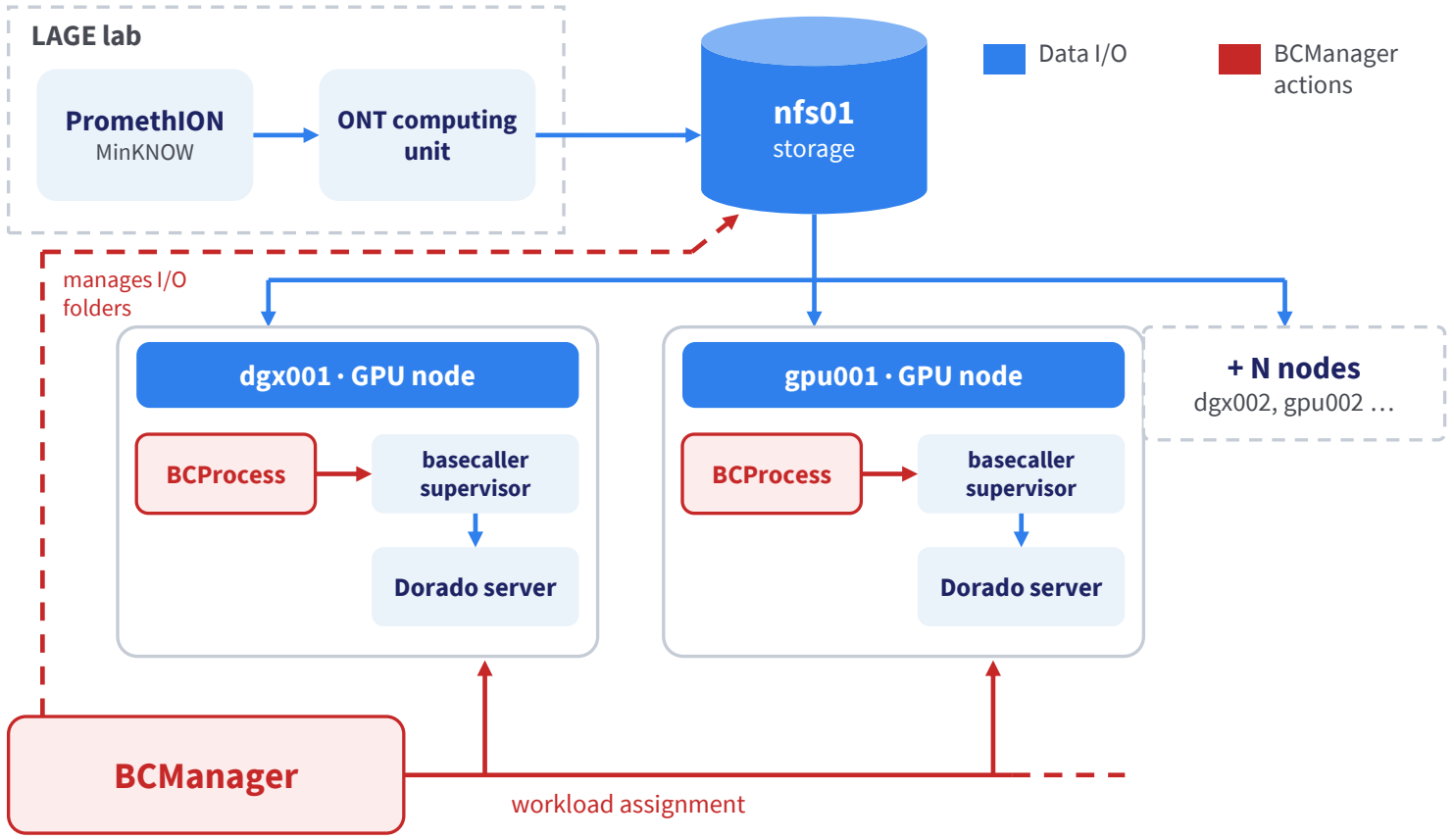
Still true with the latest ONT computing node (4×A100 GPUs)

Basecalling cannot keep up in the lab: **it has to move to HPC, as data are produced**

AREA SCIENCE PARK

HPC in the lab: distributed basecalling at LAGE

HPC NEEDS · FROM THE SEQUENCER TO GPU NODES, AUTOMATICALLY



3 Distributed basecalling on the fly

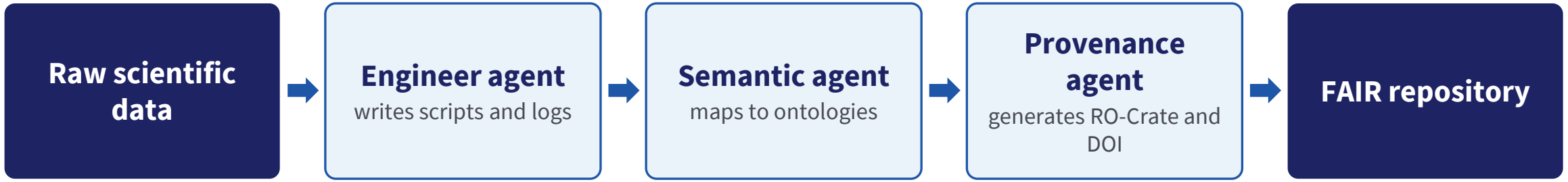
- Data flow on ORFEO
- basecaller software run in parallel on N nodes (dgx, gpu)

Data reach HPC as they are produced: **no manual transfer**

AREA SCIENCE PARK

Is agentic AI the solution?

AUTOMATING FAIRIFICATION WITH GOAL-DRIVEN AGENTS (E.G. LANGGRAPH, CREWAI)



What the evidence shows

Real progress in data preparation, workflow orchestration, semantic retrieval and multimodal synthesis

Persistent weak spots

Incomplete lineage, unstable dependencies, weak cross-modal semantics, inconsistent evaluation, miscalibrated trust

Agents produce plausible analyses faster than they can prove them complete and reproducible: **keep humans in the loop**

AREA SCIENCE PARK

Shah et al., From Raw Scientific Data to Trusted Discovery: Agentic AI for FAIR and Reproducible Data Science, Sch J Eng Tech 14(9):526–550, 2026 (systematic review of 218 studies)

Four things we would tell a lab starting today

LESSONS LEARNED

Budget the data path with the instrument

Controllers, converters and schemas belong in the procurement, as in NFFA-DI and PRP@CERIC.

Treat HPC as part of the lab

Ingest and storage tiers are design choices, not afterthoughts.

Curation does not go away

Legacy archives and external datasets still need road 1, as the STM archive and SPN01 show.

Brainware before hardware

Data curators and engineers inside the lab; we train them through the MDMC master and BOL.

MDMC: Master in Data Management and Curation

BRAINWARE · TRAINING THE PEOPLE OF THE ECOSYSTEM

"Shaping Data. Shaping the Future."

AREA SCIENCE PARK

AREA SCIENCE PARK

Thank you for attention!

Area Science Park

Padriciano 99, Trieste-Italy

www.areasciencepark.it