DIA 2026 · SESSION #535 · PHILADELPHIA

Is your data AI-ready?

How would you even know?

Unlock AI's Full Potential: The Power of FAIR Data

Tuesday, June 16, 2026 · 1:45 – 3:00 PM ET · Moderator: Anuj Uppal

Panelists: Ted Slater · Kaveen Hiniduma · Kathleen Rand

Pharma's AI investment is enormous. What fraction of that spend goes to the data underneath it?

THE PANEL

Three layers. One conversation.

01

THEORY

Ted Slater

EPAM · Co-author, FAIR Principles (2016)

What FAIR was — and where it is today.

02

MEASUREMENT

Kaveen Hiniduma

Ohio State · Lead Author, AIDRIN

The dimension nobody scores.

03

REALITY

Kathleen Rand

Amgen · Operations Data Strategy

Five agreements before the data.

01

What was FAIR meant to do — and what's actually happened in the ten years since?

Ted Slater

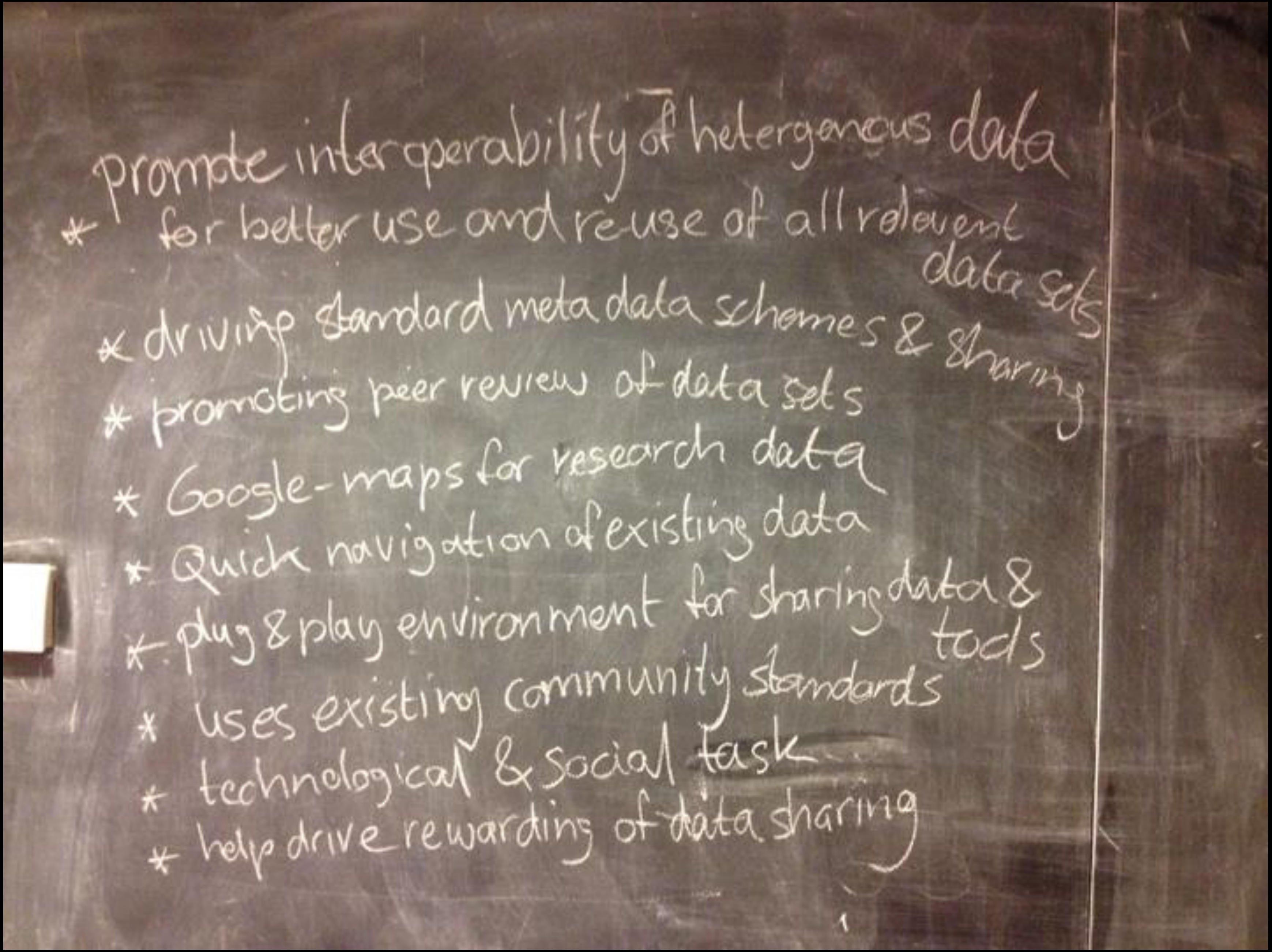
EPAM · Global Head, Knowledge Engineering & Data Advisory

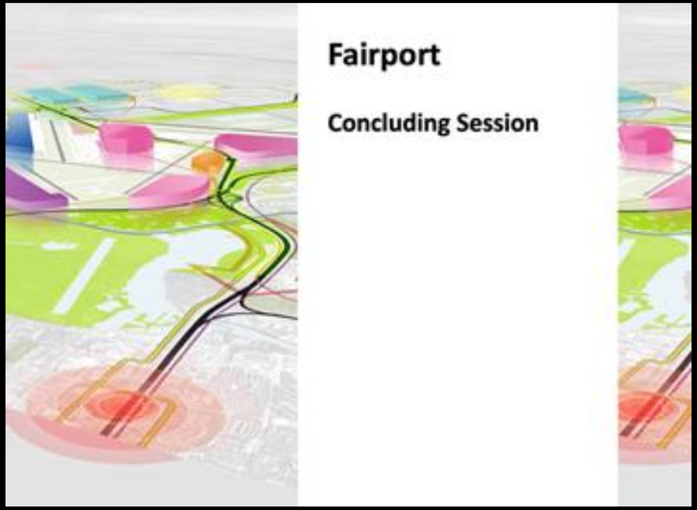
Co-author, FAIR Data Principles (Wilkinson et al., 2016)

FAIR Data is more important today than ever.

Why?

It turns heterogeneous, siloed data into machine-actionable, reusable assets that accelerate discovery and enable trustworthy AI.





Slides from the FAIRPORT meeting, Leiden, January 2014

Box 2 - 2016

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Box 2 - 2026

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

No drift in FAIR Principles.

The foundation remains solid.

"But 10 years on, we're still struggling. Why?"

Critical Reflection

A QUOTE

"There is an urgent need to improve the infrastructure supporting the reuse of scholarly data."

— Wilkinson, M., Dumontier, M., Aalbersberg, I. et al.

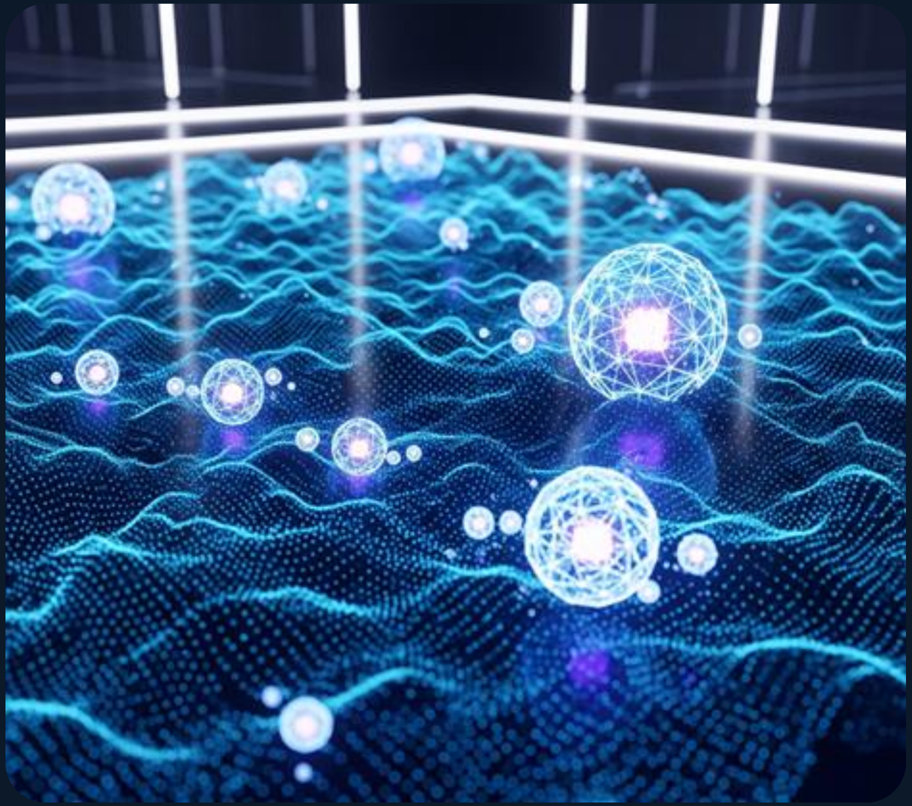
The FAIR Guiding Principles for scientific data management and stewardship.
Sci Data 3, 160018 (2016).

https://doi.org/10.1038/sdata.2016.18

But is **FAIR Data** an *infrastructure* problem?

FAIR Theater

"All my data is in this data lake, so it's **findable.**"



But no persistent, globally unique identifiers or semantics

FAIR Theater

"My data platform has a **semantic layer.**"

No GUPRIs, no semantics, no machine actionability

"The biggest obstacle to **FAIR Data** *today is shallow understanding of the Principles."*

AI cannot save you from this.

“What is the **ROI of roads?**”

How many FAIR Principles are there?

Do not edit
How to change the design

Does your organization have the competency to implement FAIR effectively and systematically?

The Slido app must be installed on every computer you're presenting from

02

If FAIR isn't enough, what is — and how do you measure it?

Kaveen Hiniduma
Ohio State University · PhD Student
Lead Author, AIDRIN (AI Data Readiness Inspector)

What if your **cleanest data** is your **least AI-ready** data?

Kaveen Hiniduma
PhD Student, Ohio State University
Lead author, AIDRIN: AI Data Readiness Inspector — SSDBM 2024

Three questions. Three different answers.

DATA QUALITY

"Is this data correct?"

No duplicates.
Valid formats.
Reasonable values.

MEASURED BY
DQ tools, governance.

FAIR COMPLIANCE

"Can a machine find, access, and interpret this data?"

Findable.
Accessible.
Interoperable.
Reusable.

MEASURED BY
Metadata standards, FAIR assessors.

AI READINESS

"Can a model learn something true and generalizable from this data?"

Balanced classes.
Independent features.
Representative sample.

MEASURED BY
Almost no one.

The six vital pillars of AI readiness.

01 Data quality

Is the data accurate, complete, and consistent enough to trust the model's output?

Model fails because the sensor data was not calibrated, leading the AI to learn patterns from equipment errors rather than actual signals.

02 Fairness and bias

Does the data contain historical biases or under-representations that will lead to unfair outcomes?

Model achieves high accuracy by simply guessing the majority group every time.

03 Understandability and usability

Is the data documented and standardized according to FAIR principles?

Absence of standardized ontologies and metadata leads to incorrect feature interpretation.

04 Structure and organization

Is the data formatted and scaled appropriately to support the chosen architecture?

Causing the model to overfit due to the lack of training data

05 Data governance

Is the data HIPAA (Health Insurance Portability and Accountability Act) compliant?

Poorly governed AI training data can be vulnerable to re-identification by adversaries

06 AI task-specific value

Are features secretly proxies for the target?

Model “predicts” survival from a column recorded after death.

Six dimensions. All measurable. Together they capture what FAIR, and data quality alone cannot.

Framework adapted from Hiniduma, Byna, Bez. “Data Readiness for AI: A 360° Survey.” ACM Computing Surveys, 2025.

Four warnings AIDRIN caught.

TCGA-LUAD · 1,000 patients · Lung adenocarcinoma
✓ Valid formats ✓ No duplicates ✓ Ranges plausible ✓ FAIR compliant
Data quality: PASSED.

AIDRIN flagged four issues before a single model was trained.

56%

COMPLETENESS

overall

Overall. Worst column: 22%.

0.27

CLASS IMBALANCE

dead / alive

Meaningful skew. Biases any survival model.

87%

REPRESENTATIVENESS

Caucasian

Caucasian. 65% female. Who is this model for?

0.15

DISCRIMINATION

race · TSD

Race shows a potential bias/discrimination signal requiring review.

Every one of these was measured automatically

TCGA-LUAD findings from Hiniduma, Byna et al., AIDRIN, SSDBM, 2024. DOI: 10.1145/3676288/3676296

What this means for pharma AI.

01

Your data quality program has a blind spot.

ALCOA+, GxP validation, completeness audits — all necessary. None of them measures whether a model can learn from this data.

02

Statistical readiness can be measured today.

The methods are published, measurable, and ready to pilot. This is not a roadmap item. It is a deployment decision.

03

Skipping it is not free.

Models trained on unaudited data fail review, embed unfair patterns, or don't generalize — wasting the model investment and everything upstream of it.

The methods are also evolving — toward customizable and agentic assessment that further reduces the burden on data teams.

AI readiness is not data quality plus FAIR.

It is a third, measurable dimension —

and we now know how to measure it.

github.com/**idtlab/AIDRIN**
aidrin.readthedocs.io
byna.1@osu.edu, hiniduma.1@osu.edu

Knowing what to measure is one problem.

Implementing it inside a global pharmaceutical organization — with legacy systems, entrenched SQL culture, and competing priorities — is an entirely different problem.

Kathleen, that's your world.

Do not edit
How to change the design

Where does statistical-integrity checking sit in your organization — BEFORE a model is trained?

The Slido app must be installed on every computer you're presenting from

03

Theory and measurement aren't enough. What has to happen inside a global pharma to make either land?

Kathleen Rand

Amgen · Operations Data Strategy

11 years in connected-data and knowledge-graph programs

Five Agreements Before the Data

Ted gave you the principles.

Kaveen gave you the measurement.

Neither lands without these five agreements in place first.

Most organizations skip them. That is why FAIR programs stall.

AGREEMENT #1 — PERSONA

Same System. Two Verdicts.

Agree what "ready" means — for these users, on this dataset.

SOLUTION ARCHITECT

5 / 5

"The schema is documented.
The IDs are stable.
It passes every test we defined."

"Will die on this hill."

SAME DATASET

QUALITY RECORDS

DATA SCIENTIST

0 / 5

"I can't find it.
I can't join it.
I can't tell what it means."

"Has stopped asking."

FAIR scoring without a shared, persona-aware definition of "ready" is theater.

Same Data. Two Different Conversations.

Agree what counts as objective proof — before the political review.

BEFORE

Binary self-assessment form

F1. Data assigned a persistent identifier?	NO
F2. Data described with rich metadata?	NO
A1. Data retrievable via standard protocol?	NO
I1. Uses formal, accessible language?	NO
R1. Released with clear usage license?	NO

"Everything's red. Months of work — nothing moves."

Progress is invisible. Effort stops.

AFTER

Algorithmic assessment of the actual data

41 / 100

composite structural health
Quality Record ontology —
measured, not declared.

Graph reachability	100%
Annotation coverage	100%
SHACL consistency issues	34
Controlled vocab gaps	16

Specific issues. Ranked by impact. Action plan attached.

Objective ground. Measurable progress.

Binary scoring doesn't just miss progress — it punishes the teams making it.

Born FAIR, or Pay 100x to Fix It Later

Agree where you invest the dollar — at the source, the warehouse, or the model.

1x *cost to fix*

SOURCE SYSTEMS

Veeva · SAP · QMS · LIMS · MES · ERP

→

10x *cost to fix*

WAREHOUSE / LAKE

Pipelines, cleanup scripts, curated tables

→

100x *cost to fix*

MODELS & CONSUMPTION

AI, analytics, data products

But the source systems are designed to resist this.

No normalized tables. No persistent identifiers. No shared masters.
Vendor lock-in is a feature, not a bug.

THE TENSION

You can't rip out your QMS. You can't rip out your ERP.
So you build the layer that absorbs the dysfunction.

Fixing FAIR at the consumption layer is where the screaming is loudest — and where every dollar counts as 100.

Who Has the Agency?

Agree who has the political authority to push fixes upstream.

	What they bring	What they lack	What FAIR work needs from them
EXECUTIVES *Air coverage*	Immediate value recognition · Air coverage	Time. Attention. Daily presence.	**Sustained protection of the program past the first political review.**
MIDDLE MANAGEMENT *Orchestrators*	Orchestration · Cross-team reach	Semantic skills. Vendor leverage.	**Political agency to push fixes back to source-system owners and vendors.**
DATA SCIENTISTS *Builders*	SQL fluency · Engineering capacity	Knowledge engineering. Org agency.	**Last-mile data product work — once upstream fixes land.**

THE MISSING ROLE

Someone with semantic skills AND political agency. Most pharma org charts don't have this seat. At Amgen we're building it now — the **data domain lead**, accountable for FAIR maturity in one data domain. Defined and assigned this month. Ask us in a year whether it worked.

Knowledge engineering is rarely taught as a core data-science competency. Neither is internal political agency.

Meeting Users Where They Are

Agree how the data will actually get used — before you build the museum.

Meet them in SQL.
Teach no one graph.

System-Aligned

- Quality Systems
- ERP
- Lab Systems

Semantic Layer

Concepts converge
Meaning is shared

Consumption

- SPARQL
- SQL
- Stardog
- Databricks

Build on the graph. Consume however you want.

Change one variable at a time. 2D to 3D.

You don't have to teach them graph. You have to give them what they're already asking for: answers, fast.

If You Remember Three Things

Five agreements before the data. Three things if you only remember three.

ALL 5

1

Stop waiting for “FAIR enough.”

There is no magic threshold. Pick the dimension that's hurting you most. Fix it. Measure it. Move on.

2

Born FAIR beats fixing downstream — but you'll have to fight for it.

The 1-10-100 math is real. Vendors won't help. Architects may not either. Born-FAIR happens only when someone senior protects the time to do it right at the source.

3

Three experts with the right tooling beat fifty generalists without it.

Throwing bodies at FAIR produces churn. A small expert team with algorithmic tooling produces measurable, defensible, executive-legible progress.

> *Anuj — over to you to pull this together.*

Do not edit
How to change the design

Of Kathleen's five agreements that have to be in place before the data work even starts, which is your biggest gap?

The Slido app must be installed on every computer you're presenting from

PUTTING IT TOGETHER

The AI-ready data stack

01	FAIR	Can the machine find, access, and interpret the data?
02	AI readiness	Can a model learn something true and generalizable from it?
03	Operating model	Can the organization fix what the measurement exposes?

If any one layer is missing, the model inherits the failure.

What to do in the next 90 days

Pick the use case. Measure the data. Own the gaps.

90 DAYS

1

Days 1–30: Pick the use case

Choose one AI use case where data risk matters: patient safety, trial design, signal detection, regulatory documentation, clinical operations.

Output: one AI use case, one target dataset, one risk hypothesis.

2

Days 31–60: Measure the data

Assess FAIR maturity, data quality, and AI readiness separately. Do not collapse them into one green/yellow/red score.

Output: separate FAIR, data-quality, and AI-readiness scorecards.

3

Days 61–90: Assign ownership and remediate

Name the data-domain owner, rank the gaps by model risk, and fund the first remediation sprint.

Output: named data-domain owner, ranked remediation backlog, funded first sprint.

Three layers. One plan.

RESOURCES · TRY IT YOURSELF

Two ways to measure your data

Full disclosure — DATA Compass is ours. Either way, the move is the same: stop guessing, start measuring.

F-UJI

FREE · OPEN SOURCE

Automated FAIR self-check, one dataset at a time.

f-uji.net · github.com/pangaea-data-publisher/fuji

GOOD FOR

- ✓ Free, instant, no install — paste a DOI or URL
- ✓ Standards-based (FAIRsFAIR metrics)
- ✓ A fast first FAIR check on published datasets

WHERE IT STOPS

- — Needs a PID or URL — data must already be published
- — Reads metadata, not your data — no quality checks
- — One dataset; no org, governance, or trend view

DATA Compass

ENTERPRISE

Beyond FAIR — the full AI-readiness picture.

- ✓ No data leaves your environment
- ✓ Data quality + statistical health
- ✓ Semantic / ontology analysis
- ✓ ALCOA+ & regulatory compliance
- ✓ ROI & business-value estimates
- ✓ Bridge AI — synthesis + prioritized actions

compassfordata.com

Scan for a demo

FAIR is the foundation. Compass measures the three layers on top.

Should FAIR evidence be required in regulatory AI/ML submissions?

A Yes. No FAIR evidence, no submission.

B Yes — but the field can't define “FAIR enough” yet.

C No. You can make data AI-ready without FAIR — it's just inefficient.

CLOSING

Is your data AI-ready?

Now you can prove it.

Every gap has a name, a measurement, and a fix.

Thank you.