### Title

Advancing FAIR data towards comparable, organized, predictive AI-ready data for community validation




### Journal

COMMUNICATIONS BIOLOGY, 9(1)

### ISSN

2399-3642



### Authors

Wood-Charlson, Elisha M

Akerstrom, Wolmar N

Anderson, Lindsey

et al.



### Publication Date

2026-07-18

### DOI

10.1038/s42003-026-10694-y

communications biology

A Nature Portfolio journal

Perspective

https://doi.org/10.1038/s42003-026-10694-y

# Advancing FAIR data towards comparable, organized, predictive AI-ready data for community validation



Elisha M. Wood-Charlson[1], Wolmar N. Åkerström[2], Lindsey Anderson[3], Mikayla A. Borton[4], Stephen K. Burley[5], Neil Byers[6], Ishwar Chandramouliswaran[7], Sylvain V. Costes[8], Paramvir S. Dehal[1], Mitchel Doktycz[9], Emiley Eloe-Fadrosh[6], Christopher Henry[10], Kjiersten Fagnan[6], Oliver Fiehn[11], Mahantesh Halappanavar[3], Bonnie Hurwitz[12], Marcin P. Joachimiak[1], Sean Jungbluth[13], Julia Koblitz[14], Gazi Mahmud[1], Lee Ann McCue[3], Thomas O. Metz[3], Nigel Mouncey[6], Christopher J. Mungall[1], Tiffanie M. Nelson[15], Valerie Skye[6], Amanda M. Saravia-Butler[16], V. Ratna Saripalli[3], Susannah G. Tringe[1], Tim Van Den Bossche[17,18] & Adam P. Arkin[1] ✉

The interrogation of data across biological and environmental systems has become increasingly complex. Fortunately, communities are adopting the FAIR (Findable, Accessible, Interoperable, Reusable) data principles for individual datasets, and continue to develop domain-specific, machine-actionable standards. However, integrating FAIR data for meta-analysis across data resources is still challenging. Understanding how disparate datasets are organized remains a manual, time-consuming process. Updating FAIR databases to reflect changes in knowledge is slow, allowing stale annotations and incorrect relationships to propagate, amplified by Artificial Intelligence (AI) systems that harvest data. Building on FAIR, we argue that data should be iteratively updated and improved. FAIR + COPE (Comparable, Organized, Predictive, Engaged) takes FAIR data and makes it Comparable, rapidly Organized (applying / updating standards) for Predictive models, which can be validated and improved by an Engaged community. We provide examples of FAIR + COPE resources and science use cases that highlight the importance of FAIR + COPE in scientific research.

Scientific research requires a deep understanding of data to formulate and test predictions. As research questions become increasingly complex, integrating diverse data types and sources becomes necessary. For example, early microbial DNA sequencing efforts acknowledged the feat of publishing a single genome[1]. Now, undergraduate students regularly generate microbial genome publications[2–6]. Understanding biological systems, however, requires the integration of numerous, diverse data types that range in data frequency (discrete samples vs continuous sensor data), source (people, places, instruments, repositories, samples)[7–9], and scale (space and time). Compiling complex datasets comes with challenges. The first - finding and accessing relevant data - has been made easier as data generators and repositories adopt the FAIR data principles (Findable, Accessible, Interoperable, and Reusable[10,11]). New curation tools and standards resources that support semantic integrations have emerged, as multidisciplinary research requires the combination of FAIR file-level standards with FAIR domain-specific, community-driven data standards, which also increases the AI-readiness of data[12].

However, challenges still exist. As described by Huttenhower et al[13]., the reuse and synergy in individual microbiome studies has many technicalities – assessing data quality, tracking data provenance, handling recalculation and versions of data products, and display and utilization of multiple, related analysis outputs. If we go beyond the FAIR principles, as they apply to static datasets, we still need to understand the complexity that underpins integration of many diverse, rapidly updating datasets. In brief, datasets must be made Comparable and Organized, per current standards and database versions, before they are ready for up-to-date, large-scale Predictive modeling, and then tested and validated by an Engaged community (Fig. 1). This must be an iterative process – as new knowledge is

A full list of affiliations appears at the end of the paper. ✉e-mail: aparkin@lbl.gov

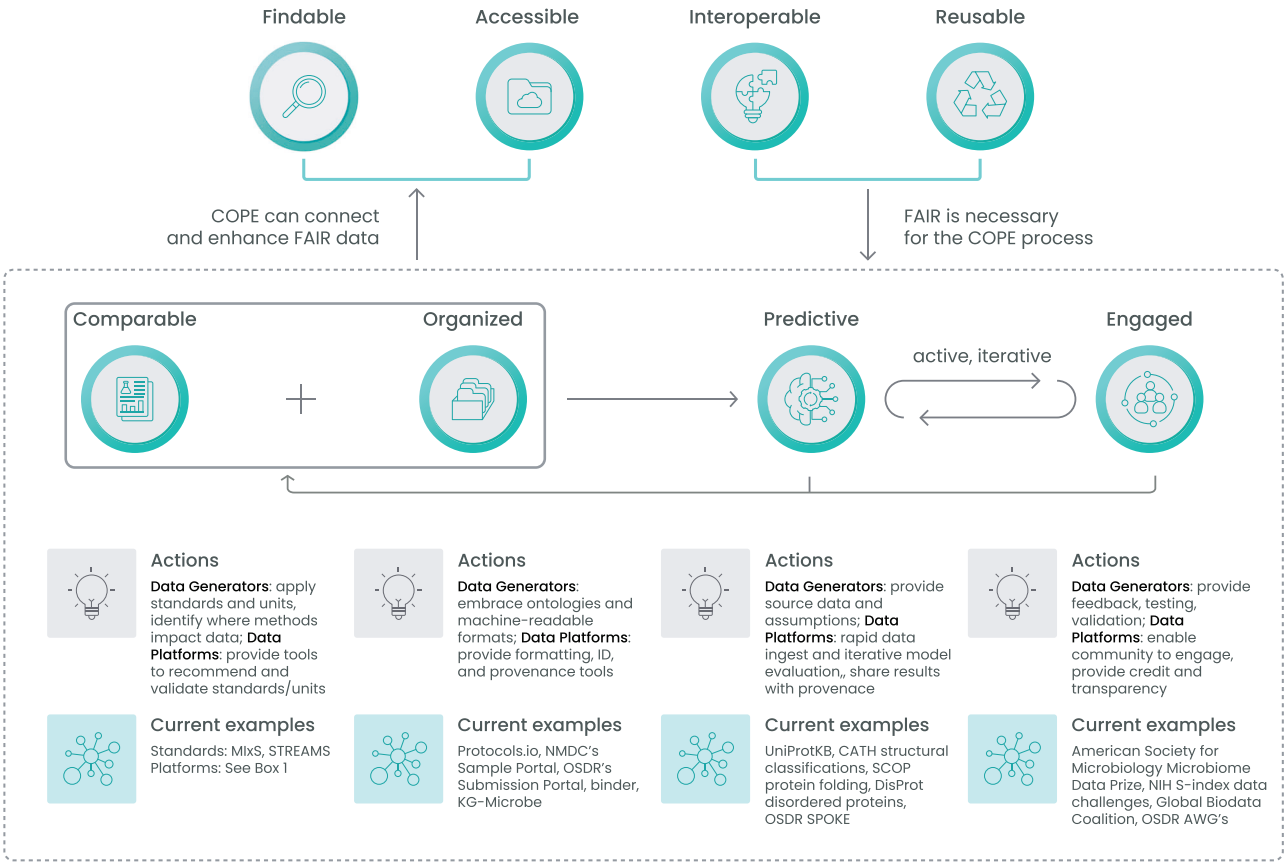


**Fig. 1 | FAIR + COPE improvements to data enables iterative improvements to predictions that can be iteratively validated by the community.** The FAIR data principles are necessary, but being able to make data Comparable and Organized for Predictive biology that can be validated by an Engaged community (FAIR + COPE) is an iterative process. Many resources support various aspects of FAIR + COPE, but the scientific community needs to invest in expanding and connecting those resources to fully enable the vision it embodies. Action and examples are provided to demonstrate feasibility.

generated through testing and validation, updates to existing Comparable and Organized data need to be made and Predictive models need to be regenerated and re-evaluated. COPE (comparable, organized, predictive, and engaged) goes beyond being a set of principles; COPE acknowledges the active, iterative scientific process required to create and maintain the predictive inferences that act as the foundation of tools, parameters, and data products used for Artificial Intelligence (AI)-driven modeling and long-term data reuse.

## The COPE process for data integration goes beyond the FAIR principles

The FAIR principles are intentionally high-level, providing a foundation for data discoverability and access. In practice, however, FAIR compliance is not universal and does not guarantee that datasets from different sources can be directly integrated for domain-specific predictive modeling or AI-driven analyses. While FAIR enables the linking of data products by persistent identifiers, FAIR + COPE further extends that linking to additional domain-relevant metadata (e.g., sampling conditions, experimental methods, units, etc.[14]) for the evaluation of comparability and establishing organizational relationships. Currently, the COPE process is often manual, focused on a particular research question, and done without full transparency or reproducibility. Even if the data are FAIR, integration of domain-specific data across studies still requires harmonization. The data must be converted to comparable units and methodological parameters, and organized with consistent ontological labels and standardized terms. The resulting predictive analyses are often done on local machines, sometimes using older database or tool versions, and testing and validation is performed by close collaborators / co-authors already engaged in the project. This siloed approach also highlights how implicit predictive inference, such as gene annotation, normalization, and labeling, is often not well documented and difficult to update, therefore making data difficult to reuse.

COPE, as a mostly manual process, is time consuming, scope-limited, and hard to reproduce. To accelerate FAIR + COPE, application and assessment of units, parameters, standards, tools, and database updates must be automated, documented, and available for validation. Also, predictive assertions are not limited to analysis outputs of comparable, organized data; it also includes the necessary updates to labels or models that underpin what makes data comparable and organized. FAIR + COPE does not specify which predictive models should be used, only that the labels and units be made explicit, versioned, and retain a link to the original FAIR data they were derived from. Predictive requires Comparable data to define valid inputs, Organized structures to capture and propagate inferred relationships, and Engaged communities to iteratively evaluate, validate, and revise predictions as knowledge advances. This is what enables data from different origins to be confidently integrated, updated, and made ready for modeling and AI-workflows. Examples of automation using large language models (LLMs) already exist: leveraging FAIR-aligned workflows and tools to apply standardized labels to data (comparable) and rapidly identify contextually-related research outputs (organized), when labeled by standard identifiers[15,16]. By embedding domain-aware AI tools, FAIR can be transformed into FAIR + COPE, a more automated process that reduces effort while maximizing synergisms across diverse data sources and data origins to enable novel exploration and *Predictive* analysis. By leveraging comparable, organized data, metadata, and provenance, reproducible *Predictive* analyses can generate models of biological systems with accompanying estimates of uncertainty, or reveal integration issues and inconsistent analysis. These predictive model outputs are ready for human-in-the-loop testing and validation by the community, and the retained link to the source data,

metadata, and provenance minimizes the data management effort required for data release or publication. Agent-assisted tools that support the initial steps in the FAIR + COPE processes will be essential for questions that extend beyond a single domain or institution; reach across time, space, and biological scales; and aim to integrate different data modalities. They will also accelerate the generation of AI-ready data.

## COPE is operationalized by integrating community engagement

FAIR has been embraced by the life sciences community[17]. However, once a FAIR dataset is released, it is rarely updated to incorporate new biological knowledge. FAIR + COPE embeds engagement into each step of Comparability, Organization, and Predictive analysis – from ontology curation and the formation of new standards[18], to workflow sharing or open model validation – creating an iterative loop where data and models are continually improved (Fig. 1). Engagement in FAIR + COPE is operationalized through defined feedback structures between data producers, consumers, and the communities establishing standards. Examples of these feedback loops already exist: Data competitions, like the Critical Assessment of Metagenome Interpretation[19], provide insight and user research opportunities, and community-led standards development and resource cataloging efforts[12,18,20–24] that ensure data analysis and standards evolve alongside needs of the research. New resources continue to make discovery easier. For example, the Multi-Omics Metadata Standards Integration Working Group (MOMSI) surveyed the landscape of standards and generated a set of 250 standards, universal and omics specific, released as a collection (https://fairsharing.org/5742)[12], with user exploration available at rda-momsi.github.io/Dashboard. Another example of deep community engagement was the development of the proposed nomenclatural code for uncultivated microbes, SeqCode[25,26], and the widely utilized Genome Taxonomy Database (GTDB[27]; https://gtdb.ecogenomic.org), which provides consistently rank normalized genome-based taxonomy for prokaryotic genomes.

Engagement has always been a key component of the scientific process, typically at the beginning (proposal review) and the end of a study (publication review). FAIR + COPE elevates engagement as an essential component throughout the data lifecycle. The challenge will be to grow the existing culture of structured participation (i.e., review panels and journal reviews) into a more iterative, interactive culture, where creating and sharing comparable and organized data, evaluating diverse data sources for cross-study integration, and testing, validation, and updates of predictions from those data, is the norm. This is going to be even more critical in the age of AI, as human-in-the-loop validation and testing will be necessary to test AI-generated predictions. FAIR + COPE, as a process, enables the community to participate in the exploration of novel engagement strategies that are iterative, inclusive, measurable, and accountable.

## Ever increasing data complexity has already launched FAIR + COPE efforts

Biological science has a long, successful history of data platforms and standards evolving with the needs of the community. In the 1980s, the International Nucleotide Sequence Database Collaboration (INSDC) was formalized to "capture and present the increasing volumes of sequence and annotation that arose from the emerging application of sequencing techniques"[28] with a continual evolution of repository products from each member: National Center for Biotechnology Information (NCBI)[29], European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI)[30], and DNA Data Bank of Japan (DDBJ)[31]. Globally coordinated efforts for mass spectrometry-based proteomics (ProteomeXchange[32]) and metabolomics (Global Natural Products Social Molecular Networking[33]) have expanded support into multi-omics data. Since then, the growth in volume and complexity of data, the need for greater contextual framing (e.g., environmental, conditional), and the rapid technological advancements have been truly remarkable. The concept of a data platform has evolved from a static record archive towards more knowledge base platforms, where dynamic, machine-actionable FAIR data records contain details about its origin and links to sample(s)[34] (important for the Nagoya Protocol and CARE principles[35]), sampling and laboratory protocols, related data products, the analysis software used for quality control or processing, and, finally, resulting scientific publications (Fig. 2). FAIR + COPE amplifies these integrated research products – expanding from static datasets to more interactive, cross-project modalities that support search, exploration, analysis, visualization, and publishing. Box 1 contains several representative FAIR + COPE data resources, many of which are affiliated with the authors of this manuscript. This list is not meant to be exhaustive, and we encourage the community to engage these projects as they continue to improve.

## Adoption of FAIR + COPE increases the value of scientific data

Adoption of FAIR + COPE can be incremental, and individual actions can improve the process of reliably integrating FAIR data over time. Here, we illustrate how we might transform a FAIR data meta-analysis into a FAIR + COPE process that supports microbial genotype-phenotype prediction, a process that can be updated to predict growth phenotypes of newly sequenced genomes or engineer strain optimization for target biomolecule chemical production. Full details regarding consideration involved in FAIR + COPE are discussed in the Supplemental Material.

**Data integrated from multiple FAIR sources**. For example: Microbial genomes with predicted genes and functional annotations (available from INSDC), growth × environment phenotype measurements (available for some genomes at BacDive[36]), gene essentiality data from RB-TnSeq experiments (available for a subset of genomes the Fitness Browser[37]), predicted homology groupings and genome similarity clusters (available at UniRef[38]), and taxonomic assignments via GTDB[27].

**Applying the COPE process to FAIR data**. Comparable: Normalize growth measurements to common units (Units Ontology[39] or Unified Code for Units of Measure) and conditions (Ontology of Microbial Phenotypes[40], updated at Microbial Ecophysiological Trait and Phenotype Ontology - bioportal.bioontology.org/ontologies/METPO). Map gene functions to updated database release and shared identifiers. Standardize taxonomic labels to the current database release version. Note: mappings are also a prediction. Tool / version should be recorded.

Organized. Link genomes to genes to homology groups to functions. Link phenotype measurements to genomes to environmental conditions. Link gene essentiality calls to genes to phenotypes. Note: Record annotation pipelines with tools / versions.

Predictive. Gene calls, functional annotations, homology groupings, and taxonomy assignments are predictive assertions, linked to evidence and methods (tool / version). Metabolic models or ML tools apply phenotypes as predictions to genomes lacking measurements. Discrepancies between predicted and measured phenotypes are identified, highlighting where models have possible limitations.

Engaged. Community evaluation during, for example, a data competition identifies a misprediction and traces it to a misannotated gene family in the databases. A targeted experiment provides empirical evidence that supports revising the functional annotation, triggering an update to that annotation across the data portals.

However, these COPE process stages are not fully automated and remain fragmented (see examples below), and the platforms for engagement are limited in scope, awareness, and adoption. As a diverse set of co-authors, from many organizations and with a global representation, we aim to help realize a FAIR + COPE future. Success will be measured by the 1) increase of rapid, iterative application of the FAIR + COPE process, 2) more examples of community evaluation and exploration appearing as new standards or cross-domain ontological connections, 3) effective AI-assisted

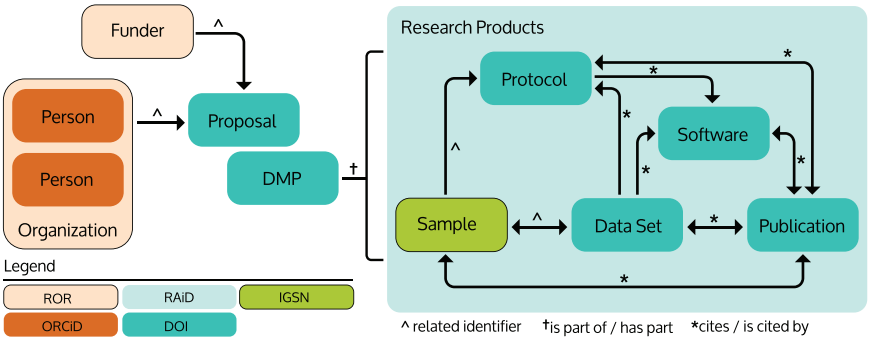


**Fig. 2 | Persistent identifiers connect FAIR data to people, places, and other research outputs.** Persistent Identifiers (PIDs) connect people, organizations, and research products, from individuals (ORCID) and organizations (ROR, Research Organization Registry) to samples (IGSN, International Geo/General Sample Number[34,53]). Related identifiers can be included in the PID metadata to establish connections in machine-readable formats. Data Management Plans (DMPs), especially those supported by the DMPTool (dmptool.org) help organize research outputs like protocols, datasets, software, and publications. These are typically given DOIs (Digital Object Identifier) and connected by relationship types, including is_part_of / has_part or cites / is_citedby. Finally, research projects can be given a Research Activity iDentifier (RAiD, https://raid.org).

## Box 1 | Examples of FAIR + COPE resources recommended by the authors

Australian BioCommons uses a nuanced engagement strategy (consultations, meetings, surveys) to understand and solve community-level data standards, analysis, and management challenges such as issues with data submission or access. Solutions – developed in partnership with infrastructures, institutes and repositories – include building or adopting technologies with ongoing community involvement and engagement. For example, Galaxy Australia's Microbiology Lab (microbiology.usegalaxy.org.au) provides tools, workflows and compute for analysis, developed with the international microbial research community[54] and the Australian Microbiome Analysis Community (www.biocommons.org.au/microbiomeanalysis).

Bacterial and Viral Bioinformatics Resource Center (BV-BRC[55]) leverages a unified data model to support a web-based AI assistant and enhanced web-based visualization and analysis tools as a service to aid in data integration for computable comparisons and workflows, which are shared back to the community for better understanding of pathogen biology and infectious diseases.

Department of Energy Systems Biology Knowledgebase (KBase[56]) has an object-based data model that enables researchers to (1) automatically convert data files into interoperable objects, (2) perform provenanced data analyses that build model predictions from complex, multiscale biological data, and then (3) immediately share with collaborators for feedback or publish them alongside a journal article.

ELIXIR[57] maintains a list of endorsed deposition databases for the submission of experimental data and supports communities in converging on shared standards, databases, and tools within their domain (e.g. Microbiome Community and Federated Human Data Community)[58,59]. For example, the 1+ Million Genomes Framework (framework.onemilliongenomes.eu) offers a shared approach for secure, ethical, and interoperable genomic data sharing across Europe relying on common privacy and legal frameworks and use cases embodied in technical infrastructure and shared data models that will be implemented in a federated network of national nodes[60].

Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) enables contributors to provide well-described and structured data through reporting formats and tools that were designed with community feedback and testing[61]. ESS-DIVE supports diverse Earth science (meta)data, including cross-domain metadata: dataset, location, and sample metadata; file-formatting guidelines: JSON, tab-deliminated, model data, etc; and domain-specific formats for biological, geochemical, and hydrological data types. Datasets containing file-level metadata and in csv formats are automatically parsed into a fusion database, which enables advanced search and discovery of data within files.

Global Biodiversity Information Facility (GBIF[62]) aggregates biodiversity data into a standardized taxonomic backbone, enabling species distribution and ecosystem models that are extended, validated, and refined through active community engagement.

International Nucleotide Sequence Database Collaboration (INSDC[28]) enforces structured sequence formats and standardized annotation pipelines, enabling integration of genomic data into countless research products, while continually expanding and updating the database through global community data submissions and curation.

MGnify[63] deploys standardized, versioned analysis workflows that enable results to be interpreted across datasets, ensures provenance linking to sample metadata and raw data in INSDC, and incorporates community annotation updates, including enhanced sample metadata extracted by machine learning[64].

UniProt Knowledgebase (UniProt[38]) provides access to important, curated data for understanding protein function and conservation. Curation of UniProtKB/Swiss-Prot is a highly engaged process, where experts use manual and ML techniques to review proposed annotations, and then update the UniProtKB database so that the knowledge is available to both humans and machines for further discovery.

National Microbiome Data Collaborative (NMDC[65]) provides sample handling and processing, standardized analysis workflows, and an interface for search and access, supported by a robust, flexible data schema that connects data across multi-omics data types and driven by community research questions.

Open Science Data Repository (OSDR[20]) contains experimental metadata linked to data files, processed data using publicly accessible pipelines to generate standardized, comparable data products, which are integrated into the visualization portal for reuse and comparison across studies.

Planet Microbe[66] leverages ontology-driven links[66–68] 'omics data with the rich, contextual, environmental data collected. Spanning >2400 ocean samples over more than 10 years of global research cruises, Planet Microbe enables the exploration of geospatially- and temporally-resolved microbial community diversity and biogeochemical patterns. Predictive models explore the distribution, variability, and trajectory of global ocean processes and microbial community dynamics.

**Table 1 | COPE extends the FAIR principles to capture the iterative, engaged process of science**

| | Recommendations | Resources | Engagement Strategies | FAIR + COPE Product |
|---|---|---|---|---|
| C | 1. Express measurements using standardized units. Document conversions.<br>2. Document methodological processes (filters, kits, primers, thresholds) in sample metadata.<br>3. Follow FAIR principles: Apply ontology-based identifiers for common labels (environment, gene, chemical). | 1. ISO, UO, UCUM<br>2. MIxS, STREAMS<br>3. ENVO, GO, ChEBI | Contribute missing terms to ontologies; share transformation scripts in public repositories; recommend standards to colleagues. | Comparable representations with explicit, re-executable mappings. |
| O | 1. Capture provenance for data objects (collection methods, processing steps, software versions).<br>2. Express experimental design parameters in machine-readable formats.<br>3. Follow FAIR principles: Link related samples, protocols, derived datasets via persistent identifiers. | 1-2. MIxS, STREAMS<br>3. RAiD, RO-Crate | Share data management plans in DMPtool; provide feedback / recommendations on data organization and application of persistent IDs. | Machine- readable provenance, experimental design, and cross-source relationships. |
| P | 1. Link annotations / models / analyses to source data, state assumptions and parameters in machine-readable formats.<br>2. Report prediction confidence and uncertainty, provide links to evidence and validation datasets, highlight discrepancies.<br>3. Re-run models as source data / methods are updated.<br>4. Follow FAIR principles: Share outputs in reusable, machine-readable formats. | 1-3. KBase, MGnify, ELIXIR<br>3. JSON-LD, RDF | Provide feedback on prediction errors; contribute validation datasets; flag outdated annotations. | Versioned predictive assertions linked to evidence, updates propagate as knowledge evolves. |
| E | 1. Participate in community standards development, data curation, model validation activities.<br>2. Track feedback, acknowledge contributions, perform validation experiments, provide annotation updates with evidence.<br>3. Follow FAIR principles: Share metadata, scripts, and workflows; deposit in public repository with a DOI. | 1. NMDC Ambassadors, OSDR AWGs, ELIXIR Communities, KBase UWGs<br>2-3. GitHub, KBase, Zenodo | Actively participate in ambassador programs and community working groups; comment on standards proposals; mentor or train peers. | Curated updates, corrections, and validation artifacts – attributed and versioned. |

FAIR + COPE enables schema linking between data from disparate sources, updates and versioning of predictive assertions, and annotation that re-execute as knowledge evolves. COPE-compliant resources should steward these updates and coordinate community engagement activities to improve predictions and correct data errors.

discovery of novel gaps in our current understanding of biology and environmental systems that can be experimentally validated, and 4) accelerated database updates that prompt existing predictive models to be rerun, or adjustment of model parameterization made, with explicit and transparent updates to confidence measurements.

As FAIR + COPE turns the FAIR principles into an iterative process (Fig. 1), we provide a generic process guide that outlines actions to operationalize COPE as a process, while also updating static FAIR datasets for AI integration, multi-study synthesis, and predictive modeling (Table 1). To check for consistency and completeness, future assessments of FAIR + COPE should model the existing FAIR assessment tools, i.e., FAIR Maturity Indicators[41] or ELIXIR's FAIR cookbook[42], but extend to COPE-specific needs based on the recommendations in Table 1. Recommendation checks should document and validate: (i) the standard(s) or ontologies are registered with OBO Foundry[43] and/or BioPortal[44], (ii) the method of normalization or unit conversion, (iii) the method of uncertainty calculation, and (iv) provenance between physical samples, data, methods, and analytical workflows (discussed in detail in[34]), and their links to relevant predictive models (e.g., Fig. 2). Similar to FAIR data maturity self-assessments, review of FAIR + COPE assessments should verify that all steps were completed and reproducible. FAIR + COPE datasets and models should be rapidly discoverable and reused by research groups and organizations, with appropriate credit and attribution[45]. Below are two examples of how the FAIR + COPE process was applied during the research process, with excerpts from the authors on the manual process and overall impact.

**Real-world demonstration of impact: Predicting bacterial phenotypic traits.** Koblitz et al.[46] provides a concrete example of how a published ML study can serve as a FAIR + COPE use case. The aim of the study was to predict bacterial phenotypic traits from genome-derived features at scale, using the BacDive knowledge base as the primary source of phenotype labels because it offers highly standardized strain-level data. Yet, despite this comparatively FAIR data starting point, making the dataset comparable and organized for modeling still required an extensive, largely manual COPE process. To organize the data and ensure diverse datasets were comparable, traits were preselected by data availability, genomes were filtered and linked to strains via consistent taxonomy identifiers and quality criteria, and ambiguous phenotype records (e.g., contradictory positive/negative labels across publications) were removed. The need for comparability became particularly visible for non-binary or mechanistically heterogeneous traits. For example, oxygen requirements – literature contradictions in BacDive meant that 7.1% of strains carried more than one oxygen state, making a multi-class approach unreliable and prompting reorganization into two separate models (AEROBE/ANAEROBE) with intermediate cases treated as negative and ambiguous cases excluded. For motility, an initial binary model failed systematically because most false predictions were gliding strains (89%), requiring iterative re-engineering of the input labels to align the target with the underlying biological mechanism. Even apparently straightforward continuous traits required explicit comparability decisions. Growth temperature as single measurements were mixed with temperature ranges, prompting additional filtering and careful definition of class boundaries.

Beyond what is described in the manuscript, and demonstrating the power of being engaged, was an additional exploratory attempt to predict pigmentation. Many strains were flagged as "pigment-producing" while the pigment name was recorded as "no pigment", presenting an internal inconsistency that rendered the labels unsuitable for ML. The issue was reported and quickly corrected in the database. This illustrates exactly how FAIR + COPE is an iterative scientific investigation, and updating FAIR data saves time: controlled vocabularies and validation rules should enforce trait/value consistency (Comparable/Organized), and structured feedback loops between data users and curators (Engaged) turned FAIR data records into FAIR + COPE, AI-ready datasets that are reproducible and easier to validate.

**Real-world demonstration of impact: Community river microbiome catalogue**. Establishing FAIR + COPE processes at the onset of a multi-study cooperative project can streamline data integration and elevate the final product. The Genome Resolved Open Watershed database (GROWdb)[7] started as a coordinated effort to align sampling designs, analytical workflows, and metadata reporting structures, while still allowing for individual researchers to pursue their research questions and hypotheses. By defining the minimum metadata, core environmental measurements, and quality control expectations, data comparability, and therefore aggregated analysis for statistical power, was possible beyond what is typically achieved by a single project effort. In addition, the GROWdb data were extensively labeled using standards and ontological terms that explicitly connect the microbes, as biological observations, to their environmental and ecosystem context. Sample-Data-Environment linkages enable rapid query, recombination, and direct use in modeling and ML workflows. Example predictions explored include: forecasting of microbial functional responses to changes in oxygen, nutrients, and hydrologic conditions; estimation of microbial contributions to biogeochemical fluxes across watersheds; and identification of conserved functional responses to shared environmental stressors. Throughout the collection, generation, and release of the GROWdb, community engagement was a key driver that refined both data quality and usability. For example, GROWdb leads aligned metadata practices with emerging community guidance for sample and metadata reporting[34]. This collaboration identified gaps, improved variable definitions, and resulted in more complete metadata, relative to what individual projects typically document. Engagement with scientists on specific projects such as lake-focused analyses that incorporated GROWdb into broader ecosystem studies (e.g., Fadum et al[47].) further ensured which data structures would support diverse reuse cases. Although this engagement required additional coordination during data generation and curation, it ultimately reduced downstream effort by minimizing post hoc data cleaning and reinterpretation, resulting in a more consistent, extensible, and broadly reusable database.

GROWdb has already demonstrated impact through reuse by multiple research groups[48,49], fostering new collaborations and enabling new studies beyond the original GROW dataset. These groups are using GROWdb to place site-specific observations and experiments into broader watershed and cross-system contexts. This illustrates how FAIR + COPE studies and databases can help microbiome research cross scales, from individual studies to regional synthesis.

These examples highlight how FAIR is necessary, but FAIR + COPE can operationalize making data directly usable for cross-study, model-driven science. FAIR + COPE allows researchers and organizations to build on their existing investments while increasing the scientific value of their data.

## Building towards a FAIR + COPE future

As a community of data generators, stewards, providers, and users across multiple institutions and infrastructures, we aim to manifest FAIR + COPE as a process. Some initial actions and solutions are outlined in Fig. 1. Critically, Engagement is not a principle – it is an embedded process within every stage of making data Comparable, Organized, and Predictive. Cross-dataset coordination, ontology alignment, and reproducible modeling all require sustained, structured collaboration between researchers, standards bodies, repositories, and tool developers. FAIR + COPE formalizes Engagement as a core operational process supported by measurable recommendations and strategies (Table 1). We acknowledge that implementing FAIR + COPE will not be easy. Data generation is already resource-intensive, and many existing COPE-ready resources (standards, ontologies, workflows) are not yet universally adopted. Yet, the effective application of standards have made transformational discoveries possible. E.g., Alphafold was possible because of standards established by the Protein Data Bank (PDB)[50]. By making Engagement explicit, we increase awareness and create training opportunities to support awareness and adoption of standards, such as NASA's Open Science 101[51], NASA's Analysis Working Groups (AWGs), and the NMDC Ambassadors[52]. And, as a community, we can support FAIR + COPE by asking funders, organizations, and publishers to: (1) support the coordination of FAIR data across platforms; (2) fund co-development of tools that make compliance easier; (3) provide free, open-source training and resources; and (4) formally recognize and reward researchers for both producing and improving FAIR + COPE datasets. Adding FAIR + COPE to our research culture will enable humans and machines to more reliably assemble and explore integrated data toward actionable insights.

Received: 28 January 2026; Accepted: 6 July 2026;
Published online: 18 July 2026

## References

1. Fleischmann, R. D. et al. Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. *Science*, (1995).
2. Hurley, A. et al. Tiny Earth: A Big Idea for STEM Education and Antibiotic Discovery. *mBio*, (2021).
3. McLoon, A. L. et al. Five draft genome assemblies from Bacillaceae isolated from a degraded wetland environment. *Microbiol. Resour. Announc.*, (2024).
4. Jordan, T. C. et al. A Broadly Implementable Research Course in Phage Discovery and Genomics for First-Year Undergraduate Students. *mBio*, (2014).
5. Branco, S. et al. Myco-Ed: Mycological curriculum for education and discovery. *PLoS Pathogens* **21**, e1013625 (2025).
6. Muth, T. R. & Caplan, A. J. Microbiomes for All. *Front. Microbiol.* **11**, 593472 (2020).
7. Borton, M. A. et al. A functional microbiome catalogue crowdsourced from North American rivers. *Nature* **637**, 103–112 (2024).
8. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotech.* **39**, 499–509 (2020).
9. Shaffer, J. P. et al. Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nat. Microbiol.* **7**, 2128–2150 (2022).
10. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 1–9 (2016).
11. Kalinin, N. A. & Skvortsov, N. A. Difficulties of FAIR Principles Implementation in Cross-Domain Research Infrastructures. *Lobachevskii Journal of Mathematics* **44**, 147–156 (2023).
12. FAIRsharing Team & Anderson L. FAIRsharing record for: Multi-omics Metadata Standards Integration Working Group landscape review collection. FAIRsharing https://doi.org/10.25504/FAIRSHARING.2FA4FB (2025).
13. Huttenhower, C., Finn, R. D. & McHardy, A. C. Challenges and opportunities in sharing microbiome data and analyses. *Nat. Microbiol.* **8**, 1960–1970 (2023).
14. Kelliher, J. M. et al. STREAMS guidelines: standards for technical reporting in environmental and host-associated microbiome studies. *Nat. Microbiol.* **10**, 3059–3068 (2025).

15. Toro, S. et al. Dynamic Retrieval Augmented Generation of ontologies using artificial intelligence (DRAGON-AI). *J. Biomed. Semantics* **15**, 19 (2024).
16. Niyonkuru, E. et al. Leveraging generative AI to assist biocuration of medical actions for rare disease. *Bioinform. Adv.* **5**, vbaf141 (2025).
17. van Reisen, M. et al. Towards the tipping point for FAIR implementation. *Data Intell* **2**, 264–275 (2020).
18. Simpson, A. et al. MISIP: a data standard for the reuse and reproducibility of any stable isotope probing-derived nucleic acid sequence and experiment. *Gigascience* **13**, (2024).
19. Meyer, F. et al. CAMI Benchmarking Portal: online evaluation and ranking of metagenomic software. *Nucleic Acids Res*. **53**, W102–W109 (2025).
20. Gebre, S. G. et al. NASA open science data repository: open science for life in space. *Nucleic Acids Res*. **53**, D1697–D1710 (2025).
21. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, (2012).
22. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, (2012).
23. Van Den Bossche, T. et al. The Metaproteomics Initiative: a coordinated approach for propelling the functional characterization of microbiomes. *Microbiome* **9**, 243 (2021).
24. Field, D. et al. The Genomic Standards Consortium. *PLoS Biol* **9**, e1001088 (2011).
25. Hedlund, B. P. et al. SeqCode: a nomenclatural code for prokaryotes described from sequence data. *Nat. Microbiol.* **7**, 1702–1708 (2022).
26. Whitman, W. B. et al. Development of the SeqCode: A proposed nomenclatural code for uncultivated prokaryotes with DNA sequences as type. *Syst. Appl. Microbiol.* **45**, 126305 (2022).
27. Parks, D. H. et al. GTDB release 10: a complete and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes. *Nucleic Acids Res*. (2025).
28. Cochrane, G., Karsch-Mizrachi, I. & Nakamura, Y. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res*. **39**, D15 (2010).
29. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res*. **38**, (2010).
30. Leinonen, R. et al. The European Nucleotide Archive. *Nucleic Acids Res.* **39**, (2011).
31. Kaminuma, E. et al. DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res*. **38**, (2010).
32. Deutsch, E. W. et al. The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res*. **51**, D1539–D1548 (2023).
33. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotech.* **34**, 828–837 (2016).
34. Damerow, J. E. et al. Opening doors to physical sample tracking and attribution in Earth and environmental sciences. *Sci. Data* **12**, 1047 (2025).
35. Carroll, S. R. et al. The CARE principles for indigenous data governance. *Data Sci. J*. **19**, (2020).
36. Schober, I. et al. BacDive in 2025: the core database for prokaryotic strain data. *Nucleic Acids Res.* **53**, D748–D756 (2025).
37. Price, M. N. et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
38. UniProt Consortium. UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res*. **53**, D609–D617 (2025).
39. Gkoutos, G. V., Schofield, P. N. & Hoehndorf, R. The Units Ontology: a tool for integrating units of measurement in science. *Database: J. of Biol. Databases Curation 2012, bas033*, (2012).
40. Chibucos, M. C. et al. An ontology for microbial phenotypes. *BMC Microbiol* **14**, 294 (2014).
41. Findability Maturity Indicators. https://fairtoolkit.pistoiaalliance.org/methods/findability-maturity-indicators/ (2026).
42. The FAIR cookbook, containing recipes to make your data more FAIR. https://github.com/FAIRplus/the-fair-cookbook (2026).
43. Jackson, R. et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database (Oxford)* **2021**, (2021).
44. Vendetti, J. et al. BioPortal: an open community resource for sharing, searching, and utilizing biomedical ontologies. *Nucleic Acids Res*. **53**, W84–W94 (2025).
45. Wood-Charlson, E. M., Crockett, Z., Erdmann, C., Arkin, A. P. & Robinson, C. B. Ten simple rules for getting and giving credit for data. *PLoS Comput. Biol.* **18**, e1010476 (2022).
46. Koblitz, J., Reimer, L. C., Pukall, R. & Overmann, J. Predicting bacterial phenotypic traits through improved machine learning using high-quality, curated datasets. *Commun. Biol.* **8**, 897 (2025).
47. Fadum, J. M., Borton, M. A., Daly, R. A., Wrighton, K. C. & Hall, E. K. Dominant nitrogen metabolisms of a warm, seasonally anoxic freshwater ecosystem revealed using genome resolved metatranscriptomics. *mSystems* **9**, e0105923 (2024).
48. Thorpe, A. C. et al. River biofilm bacteria as sentinels of national-scale freshwater ecosystems. *bioRxiv*, (2025).
49. Stach, T. L. et al. Complex compositional and metabolic response of river sediment microbiomes to multiple anthropogenic stressors. *ISME Commun*, (2025).
50. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res*. **28**, 235–242 (2000).
51. NASA TOPS Open Science 101 Curriculum Development Team et al. ARCHIVE - NASA TOPS Open Science 101 Instructor Led Training Slides. https://doi.org/10.5281/ZENODO.12752290 (2024).
52. Kelliher, J. M. et al. Cohort-based learning for microbiome research community standards. *Nature Microbiology* **8**, 751–753 (2023).
53. Davies, N. et al. Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. *Gigascience* **10**, (2021).
54. Nasr, E. et al. Microbiology Galaxy Lab: The first community-driven gateway for reproducible and FAIR analysis of microbial data. *Bioinformatics*, (2024).
55. Olson, R. D. et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res*. **51**, D678–D689 (2023).
56. Arkin, A. P. et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotech.* **36**, 566–569 (2018).
57. ELIXIR Consortium. ELIXIR Scientific Programme 2024-28. (2025) https://doi.org/10.7490/f1000research.1120280.1.
58. Garrard, C. et al. Fostering and sustaining collaborative innovation: Insights from ELIXIR Europe's life science Communities. *F1000Res* **14**, 884 (2025).
59. Finn, R. D. et al. Establishing the ELIXIR Microbiome Community. *F1000Res* **13**, 50 (2024).
60. Stark, Z. et al. A call to action to scale up research and clinical genomic data sharing. *Nat. Rev. Genet.* **26**, 141–147 (2025).
61. Varadharajan, C. et al. Launching an accessible archive of environmental data. *Eos (Washington DC)* **100**, (2019).
62. What is GBIF? https://www.gbif.org/what-is-gbif (2026).
63. Richardson, L. et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res*. **51**, D753–D759 (2023).
64. Nassar, M. et al. A machine learning framework for discovery and enrichment of metagenomics metadata from open access publications. *Gigascience* **11**, (2022).
65. Eloe-Fadrosh, E. A. et al. The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res*, (2022).
66. Ponsero, A. J. et al. Planet Microbe: a platform for marine microbiology to discover and analyze interconnected 'omics and environmental data. *Nucleic Acids Res*. **49**, D792–D802 (2021).

67. Blumberg, K., Miller, M., Ponsero, A. & Hurwitz, B. Ontology-driven analysis of marine metagenomics: what more can we learn from our data? *Gigascience* **12**, (2022).
68. Blumberg, K. L. et al. Ontology-Enriched Specifications Enabling Findable, Accessible, Interoperable, and Reusable Marine Metagenomic Datasets in Cyberinfrastructure Systems. *Front. Microbiol.* **12**, 765268 (2021).

## Acknowledgements

We would like to thank Dr. Emmanuelle Botte (Manuscribes) for thoughtful review and recommendations, Dr. Ellen Dow for help with figures, and the editor and reviewers for suggestions that improved and streamlined the manuscript.

## Author contributions

E.W.C., A.P.A. - conceptualization, writing - original, writing - review & editing, supervision, project administration; W.A., L.A., M.B., S.K.B., N.B., I.C., S.C., P.D., M.C., E.E.F., C.H., K.F., O.F., M.H., B.H., M.J., S.J., J.K., G.M., L.A.M.C., T.O.M., N.M., C.J.M., T.M.N., V.S., A.S.B., V.R.S., S.G.T., T.V.D.B. - conceptualization, writing - original, writing - review & editing.

## Funding

E.W.C., M.B., N.B., P.D., E.E.F., K.F., M.J., S.J., G.M., N.M., C.J.M., V.S., S.T., A.P.A. - U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) Contract: DE-AC02-05CH11231; C.H. - U.S. DOE BER Contract: DE-AC02-06CH11357; M.D. - U.S. DOE BER Contract: DE-AC05-00OR22725; L.A., M.H., L.A.M.C., T.O.M., V.R.S. - U.S. DOE BER Contract: DE-AC05-76RL01830; I.C. - U.S. National Institutes of Health; OF - U.S. National Institutes of Health R01 (GM155383); J.K. - Leibniz Association: Leibniz Competition K280/2019; M.H. - DOE Office of Nuclear Physics (Contract DE-AC05-76RL01830); T.V.D.B. - Research Foundation Flanders (FWO) (1286824 N); S.K.B. - RCSB Protein Data Bank Core Operations are jointly funded by the US National Science Foundation (DBI-2321666), US DOE (DE-SC0019749), National Cancer Institute, the National Institute of Allergy and Infectious Diseases, and the National Institute of General Medical Sciences of the National Institutes of Health (R01GM157729); S.V.C. - Trivedi Institute for Space and Global Medicine at the University of Pittsburgh School of Medicine; T.M.N. - Bioplatforms Australia; A.S.B. - Biological & Physical Sciences within the NASA Science Mission Directorate; W.N.A. - Science for Life Laboratory.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s42003-026-10694-y.

**Correspondence** and requests for materials should be addressed to Adam P. Arkin.

**Peer review information** *Communications Biology* thanks Daniel Berrios and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: George Inglis. A peer review file is available.



[1]Environmental Genomics and Systems Biology Division, E.O. Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [2]NBIS - National Bioinformatics Infrastructure Sweden, SciLifeLab, Uppsala University, Uppsala, Sweden. [3]Pacific Northwest National Laboratory, Richland, WA, USA. [4]Department of Food Science and Human Nutrition, Colorado State University, Fort Collins, CO, USA. [5]Rutgers, The State University of New Jersey, Piscataway, NJ, USA. [6]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [7]Office of Data Science Strategy, National Institute of Health, Bethesda, MD, USA. [8]Trivedi Institute for Space and Global Biomedicine, University of Pittsburgh, Pittsburgh, PA, USA. [9]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [10]Computing, Environment, and Life Sciences Division, Argonne National Laboratory, Lemont, IL, USA. [11]West Coast Metabolomics Center, University of California, Davis, Davis, CA, USA. [12]BIO5 Institute, The University of Arizona, Tucson, AZ, USA. [13]Estuary and Ocean Science Center, San Francisco State University, Tiburon, CA, USA. [14]Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany. [15]Australian BioCommons, The University of Melbourne, Melbourne, Victoria, Australia. [16]Amentum Services, Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA, USA. [17]Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium. [18]VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium.
✉e-mail: aparkin@lbl.gov