The FAIRSCAPE AI-readiness Framework for Biomedical Research

Authors: Sadnan Al Manir[1], Maxwell Adam Levinson[1], Justin Niestroy[1], Christopher Churas[2], Nathan C. Sheffield[1], Brynne Sullivan[1], Karen Fairchild[1], Monica Munoz-Torres[3] , Sarah J. Ratcliffe[1], Jillian A. Parker[2], Trey Ideker[2], & Timothy Clark[1]*

[1] University of Virginia School of Medicine; [2] University of California San Diego School of Medicine, [3] University of Colorado Anschutz

Corresponding author: twc8q@virginia.edu

Abstract

Objective: Biomedical datasets intended for use in AI applications require packaging with rich pre-model metadata to support model development that is explainable, ethical, epistemically grounded and FAIR (Findable, Accessible, Interoperable, Reusable).

Methods: We developed FAIRSCAPE, a digital commons environment, using agile methods, in close alignment with the team developing the AI-readiness criteria and with the Bridge2AI data production teams. Work was initially based on an existing provenance-aware framework for clinical machine learning. We incrementally added RO-Crate data+metadata packaging and exchange methods, client-side packaging support, provenance visualization, and support metadata mapped to the AI-readiness criteria, with automated AI-readiness evaluation. LinkML semantic enrichment and Croissant ML-ecosystem translations were also incorporated.

Results: The FAIRSCAPE framework generates, packages, evaluates, and manages critical pre-model AI-readiness and explainability information with descriptive metadata and deep provenance graphs for biomedical datasets. It provides ethical, schema, statistical, and semantic characterization of dataset releases, licensing and availability information, and an automated AI-readiness evaluation across all 28 AI-readiness criteria. We applied this framework to successive, large-scale releases of multimodal datasets, progressively increasing dataset AI-readiness to full compliance.

Conclusion: FAIRSCAPE enables AI-readiness in biomedical datasets using standard metadata components and has been used to establish this pattern across a major, multimodal NIH data generation program. It eliminates early-stage opacity apparent in many biomedical AI applications and provides a basis for establishing end-to-end AI explainability.

Keywords

Artificial Intelligence, Data Provenance, FAIR Principles, AI-readiness, Metadata, Pre-Model AI Explainability

1. Introduction

Artificial intelligence (AI) readiness is a critical acceptance factor for biomedical datasets intended for use in ethical, FAIR (Findable, Accessible, Interoperable, Reusable) [1] AI applications, whether in the clinic or the laboratory. At the conclusion of an AI analysis, we must be able to explain and interpret the results [2–5]. This requires full transparency of the data preparation and analysis pipeline, including complete pre-model explainability of the input data and its various transformations, from patient or laboratory instrument to model training and execution [6]. Without substantiated explainability to establish epistemic validation for the foundational data, all further model building and analysis are subject to potentially catastrophic epistemic failure modes. Leonelli, who argues strongly against treating data as “ground truth”, stresses the role of understanding data provenance and characterization in determining its meaning, and hence its epistemic value in supporting subsequent interpretations [7].

FAIRSCAPE is a containerized client-server framework in Python, JavaScript, and React, with a human-in-the-loop AI assist mode. It was developed to provide transparency and validation with every biomedical dataset destined for Artificial Intelligence (AI) and Machine Learning (ML) analysis. It operationalizes AI-readiness criteria developed by the NIH Bridge to Artificial Intelligence (Bridge2AI) Standards Working Group [8] and integrates MLCommons Croissant [9], selected Croissant Responsible AI (RAI) [10], and LinkML schemas [11] for structured metadata representation. The framework continues to evolve in coordination with the Bridge2AI Standards Working Group, representing investigators from forty leading institutions, and through active collaboration with Global Alliance for Genomics and Health (GA4GH) [12], the NIH Generalist Repository Ecosystem Initiative (GREI) repositories consortium [13], the RO-Crate initiative at the University of Manchester [14], and the LinkML development team at Lawrence Berkeley National Laboratory to address emerging AI-readiness requirements across the biomedical AI community.

FAIRSCAPE creates deep metadata, including provenance graphs and dataset schemas, providing FAIR-compliant persistent identifiers (PIDs) for biomedical data and software. It creates and provides rich, human-readable Datasheets in HTML, extending the basic conceptions of Gebru et al. 2021 [15] to cover additional metadata required for biomedical AI-readiness. It also captures Model Cards [16] for cases where AI models are integrated with a data preparation pipeline, as in the Bridge2AI Functional Genomics Grand Challenge (Cell Maps for Artificial Intelligence - CM4AI) [17]. The framework supports a simple, direct upload process to any instance of the Harvard Dataverse, a repository within the NIH Generalist Repository Ecosystem Initiative (GREI). Interfaces to other GREI repositories are planned for the near-term.

Here we present the FAIRSCAPE client-server framework in detail. FAIRSCAPE establishes and validates pre-model AI explainability and AI-readiness of biomedical datasets by capturing rich information about the datasets and the software components involved in extracting and computing them, and managing this information in standardized exchange packages. The system has been validated by managing several large multimodal data releases in a major NIH program (Bridge2AI) in full conformance to that program's AI-readiness criteria.

2. Related Work

2.1 FAIRness

The FAIR Principles established "Level 0" (Findable, Accessible, Interoperable, Reusable) requirements for reusable biomedical metadata and data. They are widely referenced and mandated but insufficiently specified for AI applications. FAIR Reusability subproperty R1.1 requires metadata describing a "plurality of accurate and relevant attributes" but does not specify which attributes are required. R1.2 states that metadata must be "associated with detailed provenance," but does not specify which provenance model to employ and how detailed it must be. R1.3 states metadata must "meet domain-relevant community standards," but does not specify what those standards are and places the burden of development on each domain, which is not appropriate for an overarching general standard.

Our approach operationalizes ethical, FAIR, and reusable metadata for the biomedical AI domain, guided by a well-defined set of criteria.

2.2 Digital Commons Environments and Generalist Repositories

Multiple digital commons environments have been developed over the past several decades. Specialist repositories, such as Sequence Read Archive (SRA), Database of Genotypes and Phenotypes (dbGaP), Gene Expression Omnibus (GEO), NHLBI BioData Catalyst, NHGRI

Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL), Genomic Data Commons (GDC), and others developed by NIH [18–21], and a multitude of resources developed at EMBL-EBI [22] are essential and widely used. They cannot, however, be used for datasets containing Protected Health Information (PHI).

One notable provenance-enabled data commons based on executable research objects was the NSF-funded Whole Tale project, developed by Chard et al. at the University of Chicago [23]. Whole Tale was designed to capture all data and software used in a research study for reproducibility. Unfortunately, this system does not appear to be actively maintained.

Robust repositories in the NIH GREI, including Dataverse, Zenodo, Figshare, OSF, Vivli, Dryad, and Mendeley [24–31], are widely used, including for multimodal data. The GREI consortium has published recommendations on adopting common metadata standards [32] based on the DataCite metadata schema [33]. We incorporate these key DataCite metadata elements into our framework and push FAIRSCAPE-packaged CM4AI datasets to the University of Virginia Dataverse instance for the long-term sustainability required by NIH. Unfortunately, at present, there is no mechanism to view our detailed metadata in Dataverse, other than viewing release-level information. Datasets and their schemas and provenance graphs must be exported from Dataverse for access to that information.

2.3 Provenance Metadata

Two notable and widely used specifications for provenance are the W3C PROV [34] and PAV (Provenance, Authoring, and Versioning) [35] ontologies. Developed by workflow experts to align outputs across heterogeneous pipelines, the W3C PROV ecosystem is grounded in twelve formally defined ontology standards of the World Wide Web Consortium [36–38].

PROV defines provenance as formal interactions among core Entities, Activities, and Agents, while allowing flexibility for domain practitioners to specify these components more deeply. PAV is intentionally lightweight and omits activities, resulting in a single-level provenance graph that specifies only “pav:derivedFrom” predicates on digital objects.

Seminal work leading to the EVI profile of PROV was the Micropublications ontology [39], which formalized claims and their support in biomedical articles as forms of bimodal defeasible argumentation. EVI reduced that scope to focus on datasets and their preceding computations, software, and input datasets, treating them as assertions justified by their provenance.

2.4 CEDAR Biomedical Metadata Templates

The Center for Expanded Data Annotation and Retrieval (CEDAR) metadata ecosystem [40,41] supports the creation of domain-specific metadata templates (reporting guidelines) for biomedical samples, protocols, and experimental entities, and is integrated with practically all relevant domain ontologies via the National Center for Biomedical Ontology’s BioPortal [42]. CEDAR has experienced widespread adoption for formalizing ontology-based characterization of biomedical experiments using domain-specific templates. CEDAR templates have been deployed to HuBMAP (Human BioMolecular Atlas Program) [43,44] to build over 30 metadata templates for use in that NIH program. Most notably, the NIH HEAL initiative [45] mandates its use to characterize experiments in over 1,000 projects. Overall, CEDAR has highly impressive adoption, with more than 5,500 registered users and 4,700 metadata templates created [46].

CEDAR’s strengths are complementary to FAIRSCAPE. While CEDAR does not connect its templates to deep provenance graphs in the data preparation pipeline and was not designed for AI-readiness preparation and evaluation against formal readiness criteria, it may offer an attractive complementary solution for templating rigorous reporting guidelines for ontology annotation of experimental materials and methods in the data preparation pipeline. Current

practice in our program is to add ontology references of these entities ad hoc. This suggests a valuable role for CEDAR templates in a future integration project.

2.5 Other Domain-Specific Metadata Efforts

Other domain-specific metadata efforts in the life sciences include: the dmdScheme R package [47], ISA-Tab [48], and the Ecological Metadata Language (EML) [49]. The dmdScheme package was archived in 2023 and has not had the significant impact of CEDAR. ISA-TAB is used for metadata submission in several biomedical journals, but does not address AI-readiness, nor does it use a contemporary packaging standard such as RO-Crate. EML is a dataset-level documentation standard in the domain of ecosystem research that provides XML descriptions of datasets: who collected them, where, and by what methods. It forms the backbone of several important ecological repositories.

None of these three additional methods provides or evaluates AI-readiness, packages datasets in a portable exchange format, or provides resolvable deep provenance graphs.

2.6 Croissant and Croissant Responsible AI Metadata

The Croissant metadata vocabulary developed by the industry-sponsored MLCommons consortium provides essential metadata limited to computational AI-readiness. Croissant capabilities for computation-related metadata enable machine learning models to acquire **information** needed to train and be implemented on appropriately annotated datasets. However, Croissant views pre-model explainability merely as simple data engineering, as if the data provided were ground truths. This is demonstrably not the case, particularly in biomedical research due to its ethical requirements and often complex data extraction, transformation, analysis, and inference data preparation pipelines.

Although the Croissant RAI's vocabulary introduces elements such as recommended and prohibited use cases to address this gap, it is a general-purpose specification that often lacks adaptation to the nuances of biomedical use cases. We therefore use it selectively. While some RAI properties can be repurposed for biomedical AI, a currently non-existent domain profile is required for several. The development of biomedical-specific extensions represents an important direction for future work. Most critically, RAI fails to deal comprehensively with data characterization and, critically, *treats data as presumptively given ground truths without detailed history*. As noted above, this can lead to substantial epistemic errors in making inferences, classifications, and predictions from data, particularly of the pervasive "Clever Hans" type [50–53], with an impact on the scientific integrity of findings and ethical applicability in the clinic.

2.7 The Language Engineering Approach

A related but substantially different approach to documenting pre-model dataset characteristics is DescribeML [54], a domain-specific machine-readable language with associated tooling, provided as a plugin to Visual Studio Code. DescribeML structures textual metadata descriptions across dimensions of structure, provenance, and ethical/social risks to enable automated analysis. However, its provenance representation is relatively flat and non-resolvable, typically confined to statements such as "source: CIFAR10 from Kaggle" or "collectedBy X from Y". While suitable for simply derived datasets, such as those used for online benchmarking challenges (e.g., Kaggle), this representation does not support biomedical applications or deep biomedical provenance and does not permit resolution to specific datasets, software, models, or responsible agents. The framework also includes metadata elements such as annotator demographics, which raise broader questions about consistency and scope in demographic reporting across roles in the research lifecycle. These elements were subsequently incorporated into Croissant RAI, which drew conceptually from DescribeML.

2.8 Machine Learning Specific Metadata Profiles and Checklists

As pointed out and analyzed in detail by Edmunds et al. 2026 [55], there is a growing, confusingly large body of work specifying standardized domain-specific ML metadata standards and checklists. These models, including the Data, Optimization, Model, and Evaluation (DOME) recommendations for ML in biology, focus primarily on AI/ML model design, training, and deployment rather than on data preparation pipelines or AI-readiness. While this body of work is essential for responsible model development, it does not address the foundational epistemic validation and data provenance that underpin those models – issues that FAIRSCAPE is designed to comprehensively support.

2.9 RO-Crate Packaging Tools

While tools like Describo [56] and its successor Crate-o [57] have successfully lowered the barrier to entry for building RO-Crates through general-purpose, user-friendly interfaces, they primarily rely on manual data entry and lack domain-specific validation.

FAIRSCAPE is not primarily about building generic RO-Crates. It is centrally concerned with producing ethical, FAIR, epistemically valid pre-model AI data. RO-Crates, in this context, are an affordance, not a *raison d'être*. FAIRSCAPE builds upon the RO-Crate standard but introduces an automated, enterprise-level architecture specifically tailored to generate and enforce the rigorous epistemic and ethical requirements of biomedical AI-readiness.

3. Methods

3.1 Development Approach

FAIRSCAPE was developed iteratively in multiple versions using agile methods over the past eight years, working closely with clinician-scientists, biostatisticians, laboratory investigators, ethics experts, and ontologists. We began by adapting a previously proven provenance-aware analytics framework for machine learning in the University of Virginia Neonatal Intensive Care Unit (NICU) [58,59] to support the distributed forty-institution research program Bridge2AI. The initial focus was on supporting CM4AI's four separate contributing data production laboratories across three University of California campuses. This required client-side data and metadata packaging, for which we adopted the RO-Crate 1.2 specification while it was still in draft form, and initiated ongoing discussions with the RO-Crate development team. In Year 3 of the program, we began alignment with the AI-readiness criteria during their development.

3.1.1 Iterative Standards Aligned Development

The AI-readiness criteria, which served as the basis for our metadata, were developed iteratively by experts in genomics, proteomics, clinical informatics, ontologies, machine learning, and metadata standards, across 17 institutions. We evolved our software in close alignment with the criteria and metadata mappings as they were developed. We also worked in close collaboration with data production groups in multiple institutions to validate the practicality of proposed metadata instantiations of the criteria. Software was iteratively validated across successive complex data releases in major NIH-funded projects.

Croissant and selected Croissant Responsible AI (RAI) mappings were added to facilitate interoperability with ML ecosystems. LinkML mappings were added to enable later semantic enhancement.

We found Croissant RAI's "one-size-fits-all" metadata to be too general in many instances for adoption in biomedical research, requiring a domain profile, which we intend to develop. A leading example of over-generality would be the property rai:personalSensitiveInformation, which is intentionally broad to cover diverse sectors like computer vision, NLP, and social

science, but lacks the granular precision required for legally mandated de-identification and security protection of subjects in clinical research.

Evaluation of metadata across the final seven dimensions and twenty-eight criteria of AI-readiness [8] was initially implemented as a self-evaluation. An automated evaluation based on the RO-Crate package metadata was later implemented, improving the scoring rigor and addressing the problems inherent in self-assessment [60].

3.2 RO-Crate Metadata Architecture

RO-Crates compliant with the version 1.2 specification [61] serve as the packaging format for FAIRSCAPE archival and distribution, supporting multiple relevant standards. This version was adopted while still in draft form to support multi-modal datasets in file trees with dataset references to external repositories. This is important to preserve deposition of specialist data modalities in domain-specific repositories where these exist, while retaining the integrative AI-readiness metadata and packaging.

RO-Crates are lightweight standardized packages ("crates") consisting of a file tree of 1 to n levels with defined slots for metadata in specific serializations and vocabularies. They constitute cross-domain Boundary Objects, allowing sufficient plasticity to meet the local needs and constraints of the various parties implementing, exchanging, and using them while preserving robustness across the biomedical practice and communications ecosystem.

Metadata in RO-Crate's required *ro-crate-metadata.json* file is serialized in JSON-LD [62] using schema.org [63]. FAIRSCAPE adds provenance characterization in W3C PROV [34] and the EVI domain profile of PROV [64,65] to these principal vocabularies. EVI domain profiles subclass the principal PROV classes and predicates for biomedical research. For example, *prov:Entity* is profiled by *evi:Dataset*, *Software, Model, Reagent, Sample*, and *Instrument* subclasses; *prov:Activity* as evi:*Computation*, *Experiment*, and *Annotation;* and *prov:Agent* as *evi:Person, SoftwareAgent,* and *Organization.* All *Entity* classes are resolvable to archived digital components and may be annotated with domain ontology terms.

FAIRSCAPE's RO-Crates are composed of a top-level Release Package containing standard *ro-crate-metadata.json* and *.html* files, and a human-readable Datasheet that substantially extends the Gebru Datasheets for Datasets concept [15]. Along with the Datasheets, an AI-readiness evaluation histogram evaluates the seven dimensions of AI-readiness [8] across the full depth of the RO-Crate, and a LinkML schema translation to enable later deep semantic characterization via ontology mapping. The human-readable Datasheet is a textual and graphical rendering of the *ro-crate-metadata.json* file.

Sub-crates (sub-directories) in the file tree are organized by data modality, e.g., imaging, clinical records, genomics, proteomics, etc. Within each sub-crate are 1 to n dataset folders, each containing 1 to n files, all with the same format and construction. Dataset structure is characterized using JSON Schema [66] with Frictionless Data [67] validation.

Each component includes additional associated descriptive metadata using the schema.org vocabulary, plus PROV/EVI and optional ontology terms. Users may extend the metadata they associate with RO-Crates. FAIRSCAPE-CLI uses the Frictionless Data framework to generate JSON Schema definitions for tabular and HDF5 files, which are associated with their referenced datasets. Validation utilizing Frictionless ensures that datasets conform to their provided schemas. Every RO-Crate component receives a locally unique key. Data may be packaged directly or simply referenced using a Uniform Resource Identifier (URI) [68]. Because RO-Crate packaging occurs client-side and does not require server mediation for metadata construction, FAIRSCAPE avoids tight coupling to a specific server instance, enabling portability, archival independence, and long-term sustainability of AI-ready datasets. Once an RO-Crate is

packaged, it may be uploaded directly to the server, where the local keys become resolvable ARK persistent IDs [69,70].

FAIRSCAPE manages the instantiation, transfer, archival, persistent identification, permissioning, and search of the RO-Crates, which are pre-validated on the client side using a Pydantic schema [71] and Frictionless Data [67,72]. Alternatively, if desired, the standard RO-Crate validator may be used with a preprocessor.

Figure 1 illustrates an example RO-Crate from a functional genomics dataset.

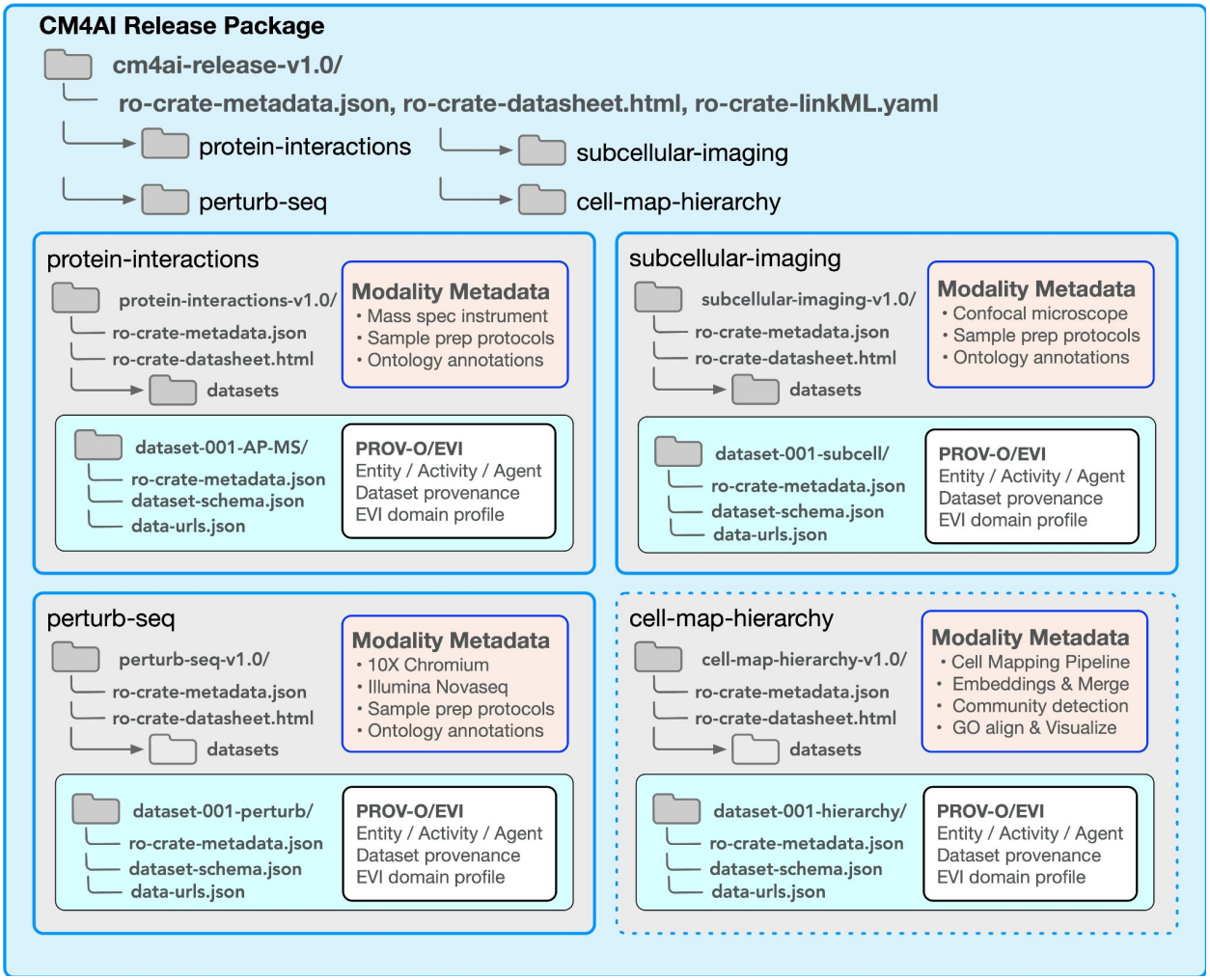

Figure 1: Cell Maps for Artificial Intelligence (CM4AI) RO-Crate schematic. *All CM4AI release packages follow the RO-Crate 1.2 specification with some optional component additions. The top-level (root) crate is the "Release Crate" holding general information for the entire release. It includes a standard ro-crate-metadata.json file serialized in JSON-LD with schema.org vocabulary plus annotations from biomedical ontologies; an HTML rendered Datasheet; a LinkML translation of the Datasheet metadata; and links to sub-crates (shown in grey) for each modality in the release. CM4AI releases currently consist of 3 modalities (protein interactions, subcellular imaging, and perturbSeq), with the 4th modality (cell map hierarchy) planned for late 2026 shown in dotted outline. Each modality sub-crate contains 1-n dataset directories, shown in turquoise. These may contain 1-n files, a JSON Schema for the files, and the provenance graphs for the files.*

4. Results

4.1 Features

The FAIRSCAPE framework generates, packages, and integrates descriptive metadata to robustly operationalize the AI-readiness criteria [8]. These metadata fully cover criteria across FAIRness, Provenance, Characterization, Pre-model Explainability, Ethics, Sustainability, and Computability. All metadata are represented in extended, human- and machine-readable Gebru-style Datasheets sealed with cryptographic signatures and packaged in lightweight standardized RO-Crate version 1.2 packages containing both the metadata and either (a) a link to associated datasets housed in an external sustainable repository, or (b) for smaller datasets, where allowed, the datasets themselves. RO-Crates are organized hierarchically as described above in Section 3.2, to support single-modality or multiple-modality data at shallow or deep levels of organization. The datasets are prepared client-side using either of three client packages, a command-line interface, or a simpler GUI application. Release-level metadata describing the entire package may be generated manually or via a human-in-the-loop AI assist.

The FAIRSCAPE framework comprises a set of Python and JavaScript packages, including alternative client versions (a command-line (CLI) Python 3 version, a GUI electron app, and a React-based JavaScript version) and a pip-installable, cloud-ready server that prepares datasets for AI analytics. AI-ready data packages present human- and machine-readable metadata characterizing the dataset(s) and link to, or include, those datasets. The electron GUI app is the simplest, intended mainly for less complex use with straightforward REDCap exports and analyses. The React-based JavaScript GUI provides all the necessary support for complex multimodal datasets, including human-in-the-loop AI assistance. The CLI client is designed for embedding in scripted pipeline code or in direct python calls by software engineers.

Data may be uploaded in simplified RO-Crate packages with minimal metadata, in compliance with DataCite and the GREI consortium recommendations. An automated evaluator notes deficiencies for various levels of requirement from “Basic” (DataCite schema only) to full AI-readiness, including Croissant metadata and LinkML translation. Datasets are automatically recognized by Frictionless Data libraries, and basic schemas are assigned with the option to expand them with user-defined data element names and descriptions. Alternatively, schemas may be imported from REDCap Data Dictionaries or specified in an attached JSON Schema description. FAIRSCAPE users specify deep provenance graphs on the client side using objects and predicates from W3C PROV and PROV’s EVI domain profile. The framework performs feature validation on uploaded data, software, and computations for all datasets.

FAIRSCAPE translates ethical, regulatory, schema, statistical, and semantic characterization of dataset releases, licensing, and availability information into formal human- and machine-readable metadata. It provides an automated AI-readiness evaluation, displayed as a histogram of the twenty-eight AI-readiness criteria, organized into seven axes.

4.2 Software Implementation

The FAIRSCAPE framework architecture is shown in **Figure 2**.

4.2.1 RO-Crate Packaging

RO-Crate packaging utilizes a combination of existing standards, as described in Section 3.2. The packaging hierarchy is flexible, except that the top-level or root RO-Crate must contain general “release” information describing the release and crate hierarchy, as well as a human- and machine-readable datasheet and a LinkML translation. We recommend including first-level sub-crates within the release crate, organized by data modality, and within each modality, by “dataset”, where a dataset is a folder of files with the same schema.

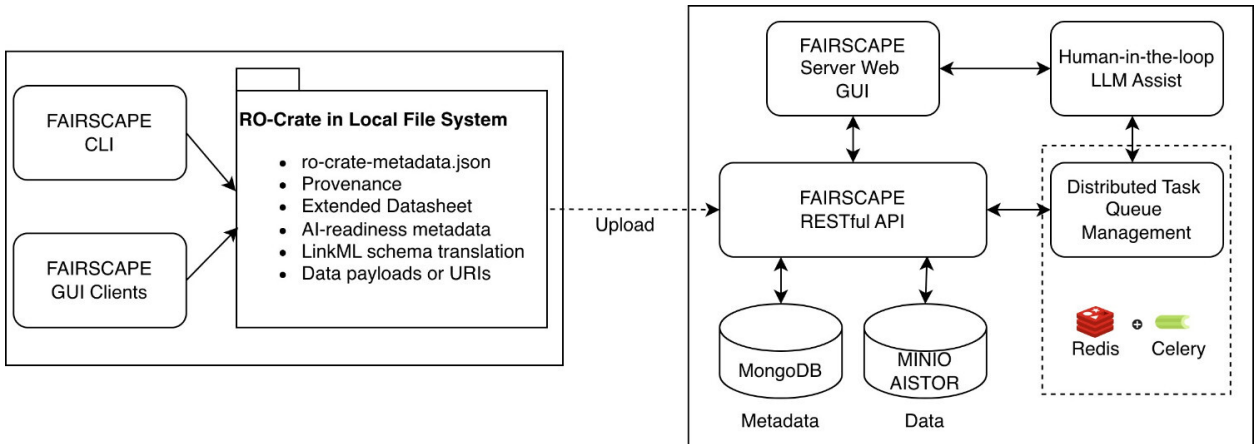

Figure 2. *FAIRSCAPE framework architecture. The client-side applications package (meta)data in the canonical exchange layer as RO-Crates and submit to the server as compressed packages. The server REST API manages user sign-on with valid credentials, accepts the compressed packages, and passes management operations to the Redis message broker, which pushes metadata to the MongoDB store and datasets to the MinIO* AIStor. *Access to the metadata and data is controlled by user, group, project, and organization-level permissions stored in MongoDB.*

Packages may be prepared for a single computation or for a long series as required. It is not necessary to package each step for trivial or unimportant operations, as long as the complete software script performing the series is referenced and resolvable for inspection. In this way, FAIRSCAPE and EVI operate differently from a traditional workflow engine. RO-Crate-packaged computations may be connected in series, with Crate A as input to Crate B's computations - the framework will then determine Crate A's output dataset and connect it as input to the computation in Crate B to create a merged package.

4.2.2 Client Software

Users may select the most appropriate client package for their use case: the Electron app is the simplest, the React GUI app is more complex and full-featured, and the command-line (CLI) version is more complex but supports in-pipeline package construction via calls from the command line or direct invocation as Python functions. Together, these tools create, manage, and upload RO-Crate packages to the server with detailed provenance graphs based on W3C PROV and the EVI Evidence Graph Ontology's domain profile of PROV. FAIRSCAPE supports extending the metadata package, which is serialized in JSON-LD.

4.2.2.1 FAIRSCAPE Command Line Interface Client

The FAIRSCAPE-CLI client is pip-installable and may be called either from the command line or directly as Python functions. It builds an RO-Crate and incrementally adds its components in a set of JSON-LD graphs using the following pattern:

- Computation <uses> Software
- Computation <uses> Model
- Dataset(1) <usedBy> Computation <generates> Dataset(2)

The underlying data model follows a defined provenance pattern in which a *Computation* entity *uses* specific *Software* (and optionally, *Model*) components, while a *Dataset* serves as input to the *Computation*, which subsequently *generates* an output *Dataset*. The tool allows users to define, infer, validate, and register schemas for common tabular data formats. Features for importing external datasets, such as NCBI BioProjects and Portable Encapsulated Projects

(PEPs) [73] into RO-Crate formats, are also implemented. Additionally, it generates detailed evidence graphs that represent the provenance relationships among components, along with HTML datasheets for datasets in the release-level exchange package (top-level RO-Crate). Finally, the client provides mechanisms for RO-Crate release management by linking related sub-crates to support the packaging of multi-modal datasets and by facilitating publication to external repositories, including FAIRSCAPE and any instance of Harvard's Dataverse.

4.2.2.2 Electron Client

The FAIRSCAPE GUI client, built with Electron and JavaScript, guides the user through RO-Crate initialization and component upload. At each step, clients display a form to collect the required metadata, and the resulting JSON-LD is displayed on the side of the application. After completing all the required forms, users can review their created RO-Crate and its contents, package it into a zip file, and upload it to a FAIRSCAPE instance.

4.2.2.3 JavaScript / React Client

An improved GUI client in React and JavaScript is also available, supporting both client human-in-the-loop AI-assisted and fully manual metadata packaging. The AI-assist feature instantiates significant release-level metadata in approximate form by providing documentation, other dataset metadata, publications describing data and methods, and a designed prompt, to a selected LLM. All metadata produced by the LLM are fully editable and require human approval with digital signoff before final acceptance.

4.2.3 FAIRSCAPE Server

The FAIRSCAPE server is a cloud-ready Python application for managing metadata and data uploads in compressed format, with user-, group-, and project-based permissioning. It supports and manages coordinated ingress, access to, and publication of datasets via GUI and/or REST API. Enterprise-level FAIRSCAPE runs in a Kubernetes instance across multiple pods in our research IT environment, but it is flexible in its deployment targets and can be run on much smaller systems, including single machine deployments.

The FAIRSCAPE server receives, catalogs, indexes, and stores uploaded RO-Crate zip packages, extracts and registers their components and associated metadata, and stores that information. It uses the FastAPI framework and provides REST API access.

Storage is coordinated between a Mongo NoSQL database for metadata and a MinIO AISTOR object store for datasets. A worker task intercepts and schedules all requests to these data stores using a Redis cache as an in-memory message broker, allowing non-blocking submission of very large datasets.

All packaged datasets and external datasets or software referenced in another repository are managed in the MinIO AISTOR S3-compliant database.

The FAIRSCAPE Server provides a Web GUI for viewing metadata and downloading RO-Crates uploaded to FAIRSCAPE. After logging in, a dashboard displays all of the researcher's uploaded RO-Crates, with links to download them or visit their landing page. Each landing page presents multiple views of an RO-Crate:

- A table summary of its required JSON-LD metadata and the files it contains.
- Serializations of its metadata in JSON-LD, RDF, and Turtle.
- An interactive visualization of dataset evidence graphs, built upon the React-Flow library, for exploring dataset provenance.

For cloud installations, MinIO access may be replaced by another S3 API. Metadata is managed in the Mongo NoSQL database. The server uses Redis as an in-memory cache and message

broker to pass information and commands from the API to the internal Worker process for execution. Multi-user and group permissioning is handled as metadata. Objects stored in FAIRSCAPE may be pushed directly to any instance of the Dataverse academic repository system.

5. Discussion

As noted above, FAIRSCAPE was developed in close collaboration with laboratory researchers, clinician scientists, computer scientists, and standards experts across multiple programs using an agile method. Its design is highly responsive to the unique needs of multimodal biomedical datasets. The metadata elements of the framework are expected to evolve as additional needs are encountered.

One area of active development is adaptation to Croissant RAI. Many elements of this general-purpose standard, while widely promoted by the MLCommons consortium, are too general for biomedical research and thus require a domain-specific profile and further evaluation for proper use. For instance, current Croissant RAI properties suffer from implementation ambiguity, and *rai:annotatorDemographics* embeds ethical risks, such as promoting demographic essentialism (demography as determinant of scientific merit) and reinforcing power imbalances by over-emphasizing contributor identities. To address these shortcomings, a forthcoming biomedical domain profile will introduce formal subclassing, property selection, and alignment with domain-specific controlled vocabularies to ensure technical and ethical rigor.

An additional limitation and area for development concerns the AI-assisted evaluation of metadata. A number of release-level metadata elements are textual and therefore require qualitative human assessment to determine how well they align with the data they describe. Because such human assessments may introduce variability, we address this challenge in two stages. First, metadata components at the dataset release level undergo formal human review and digital signing by an authorized, PI-delegated curator (or the PIs themselves in smaller projects). Second, all criteria-mapped metadata are automatically evaluated using presence/absence and completeness metrics. Future work will incorporate AI-based assistants to suggest topic-specific improvements to descriptive metadata.

More broadly, AI explainability (XAI) is among the most rapidly expanding areas in the AI literature, yet it remains largely focused on post hoc explanations of trained models, with increasing but still limited attention to in-model approaches. This emphasis leaves a substantial epistemic gap, or “black box,” in the pre-model data preparation phase, which, in biomedical contexts, can involve complex and heterogeneous processes including variation in reagents, samples, instruments, lab software, and analytic pipelines. The growing integration of AI components directly into these upstream pipelines further amplifies the need for transparent, provenance-aware documentation prior to model training and deployment.

FAIRSCAPE closes this gap by providing substantial, verifiable pre-model explainability for biomedical AI projects, thereby laying the foundation for end-to-end explainability. In this sense, it extends Shapin’s concept of “virtual witnessing” [74,75], the practice established during the scientific revolution of publishing detailed methodological accounts to enable independent scrutiny, in contrast to the pre-scientific scholastic model of validation by simply citing authorities [76,77]. The rise of large-scale, automated data pipelines has complicated this norm by rendering many preparatory steps effectively invisible. FAIRSCAPE reconstitutes virtual witnessing in machine-readable form through detailed, resolvable documentation of data preparation processes. FAIRSCAPE’s RO-Crate packaging provides a robust Boundary Object [78–83] that mediates between domain practice groups in biomedical research, computer science, ethics, and ontology development. The deep, resolvable provenance graphs it provides allow data preparation pipelines to be replicated outside their initial context.

An additional development we are currently exploring is the provision of audience- or persona-focused textual explanations derived from FAIRSCAPE's provenance graphs and related information in the exchange package.

Finally, GREI repositories are just beginning to consider AI-readiness issues. Integration of FAIRSCAPE tools, packaging, and viewers to support GREI AI-readiness for non-PHI information would be an extremely promising area for future development.

6. Conclusion

The FAIRSCAPE AI-readiness framework provides a rigorous, scalable, common platform for creating, managing, and distributing ethical, FAIR, AI-ready biomedical datasets. It was initially developed to provide deep provenance on computations in clinical predictive analytics applications in the University of Virginia Neonatal ICU. Subsequently, it was re-architected and significantly extended to support complete AI-readiness metadata packaging for diverse biomedical datasets, and validated across multimodal laboratory and clinical datasets in a major NIH-funded program, Bridge2AI.

FAIRSCAPE's overall significance to biomedical research is in closing the epistemic gap in the pre-model data preparation phase — typically seen by purely CS-aligned AI researchers as limited to rote data engineering for model consumption. Taking biomedical data "as provided" as ground truth is not appropriate for biomedical research. The Bridge2AI program provided our team with an exemplary laboratory to develop methods to close this gap. FAIRSCAPE now provides foundational pre-model explainability metadata for this important purpose.

FAIRSCAPE remains under active development with the goal of broad applicability and adoption. The framework supports the production of FAIR, ethical, and epistemically robust AI-ready data. It supports all 28 practice-based AI-readiness criteria [8] across diverse biomedical contexts, with the potential to significantly strengthen the reliability and reusability of pre-model data for AI, and to increase the translational impact of AI applications across many research and clinical settings.

CRediT Author Statement

SAM: Software, Supervision, Validation, Writing - Original Draft, Writing - Review and Editing; MAL: Software, Validation, Writing - Review and Editing; JN: Software, Validation, Writing - Review and Editing; CC: Software, Validation, Writing - Review and Editing; NCS: Software, Validation, Writing - Review and Editing; BS: Writing - Review and Editing; KF: Writing - Review and Editing; MMT: Funding acquisition, Methodology, Software, Validation, Writing - Review and Editing; SJR: Methodology, Writing - Review and Editing; JAP: Project Administration, Funding Acquisition, Methodology, Writing - Review and Editing; TI: Funding Acquisition, Methodology, Project Administration, Writing - Review and Editing; TC: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing - Original Draft, Writing - Review and Editing.

Declaration of Competing Interest

NCS is a consultant for InVitro Cell Research, LLC. Other authors declare no competing interests.

Acknowledgements

This work was funded by the U.S. National Institutes of Health Bridge2AI program (OT2OD032742, OT2OD032701, 5U54HG012513-04), by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (5R01HD072071-10), and by the University of Virginia Frederick Thomas Fund. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

We thank the RO-Crate engineering team at the University of Manchester for very helpful discussions on RO-Crate architecture; the Bridge2AI Standards Working Group for collaboration on the metadata architecture; and Chris Mungall and staff at the Lawrence Berkeley National Laboratories Biosystems Data Science Department for developing the LinkML model and collaborating on a version of human-in-the-loop AI assist.

We are grateful to Katy Krahn for organizing the many effective meetings and discussions of the Center for Advanced Medical Analytics at UVA, which led to the initiation of this platform.

And we especially thank Carol Goble, Maryann Martone and Randall Moorman for many helpful discussions and interactions during platform design and development, and Mark Musen for discussions about potential CEDAR integration.

Ethical Approval

The study conducted in University of Virginia Medical Center's Neonatal Intensive Care Unit (NICU) used chart-reviewed Electronic Health Record (EHR) data. It was conducted under Institutional Review Board (IRB) Exemption 4 (45 CFR 46.104) because it involved secondary use of previously collected clinical data. All data in that study was anonymized.

Declaration of Generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author(s) used Gemini 1.5 Pro in order to brainstorm potential image layouts, as well as to review the manuscript for compliance with journal editorial policies and to suggest potential improvements. The authors used Claude Opus 4.6 to point out grammar and capitalization issues in the pre-final copy. After using these tools/services, the author(s) created the original images, reviewed and edited the content as needed, and take full responsibility for the content of the published article.

Data and Software Availability

Source code and Pydantic models for the FAIRSCAPE framework in this study are open-source, provided under Apache 2.0 license, and freely available on Github via Zenodo at the DOIs indicated below.

The LinkML code and documentation is freely available under Apache 2.0 license at the DOI indicated below.

Patient privacy constraints and Institutional Review Board (IRB) restrictions prevent the raw Electronic Health Record (EHR) data used for the NICU prediction task from being made publicly available.

A de-identified public dataset and an evidence graph for the NICU analyses are available on the University of Virginia Dataverse instance at the indicated DOIs below.

Software and Tooling

- **FAIRSCAPE AI-readiness Framework:**
  - Levinson, M.A., et al. 2026 - FAIRSCAPE CLI (v1.1.21). Zenodo. https://doi.org/10.5281/zenodo.18714417 [84]
  - Levinson, M.A., et al. 2026 - FAIRSCAPE Server (v1.0.32). Zenodo. https://doi.org/10.5281/zenodo.18244417 [85]
- **RO-Crate Validation Classes:**
  - Niestroy, J., et al. 2026 - FAIRSCAPE Pydantic Models (v1.0.1). Zenodo. https://doi.org/10.5281/zenodo.18234523 [72]

- Leo, S., et al. 2025 - ro-crate-py (0.14.2). Zenodo. https://doi.org/10.5281/zenodo.17342107 [86]

- **LinkML-based Semantic Modeling and Translation:**
  - S.A.T. Moxon, H. Solbrig, N.L. Harris, P. Kalita, M.A. Miller, S. Patil, K. Schaper, C. Bizon, J.H. Caufield, S. Cirujano Cuesta, C. Cox, F. Dekervel, D.M. Dooley, W.D. Duncan, T. Fliss, S. Gehrke, A.S.L. Graefe, H. Hegde, A. Ireland, J.O.B. Jacobsen, M. Krishnamurthy, C. Kroll, D. Linke, R. Ly, N. Matentzoglu, J.A. Overton, J.L. Saunders, D.R. Unni, G. Vaidya, W.-M.A.M. Vierdag, T. Putman, LinkML Community Contributors, O. Ruebel, C.G. Chute, M.H. Brush, M.A. Haendel, C.J. Mungall, LinkML: A Linked Open Data Modeling Language, (2026). https://doi.org/10.5281/ZENODO.5703670. [87]

Datasets

NICU Highly Comparative Time Series Analysis. This data and its provenance (evidence) graph were packaged by an early version of FAIRSCAPE, before RO-Crate Packaging was added to bind provenance and other metadata to the datasets. The datasets and provenance were exported from FAIRSCAPE to the University of Virginia Dataverse instance for long-term archival.

- Niestroy, J, et al. 2021. Evidence For: Discovery of Signatures of Fatal Neonatal Illness In Vital Signs Using Highly Comparative Time-series Analysis. University of Virginia, 2021-03-31, 2021. https://doi.org/10.18130/V3/HHTAYI [88]
- Niestroy, J., et al. 2021. *Replication Data For: Discovery of Signatures of Fatal Neonatal Illness In Vital Signs Using Highly Comparative Time-series Analysis*. University of Virginia, 2021-03-26, 2021. https://doi.org/10.18130/V3/VJXODP [89]

- **Bridge2AI Functional Genomics (Cell Maps for Artificial Intelligence - CM4AI) data releases.** The RO-Crate packages were exported from FAIRSCAPE to the University of Virginia Dataverse instance for long-term archival. Inspection of the packages shows the evolution of both data and metadata completeness during the program.
  - Clark T, Schaffer L, Obernier K, et al. Cell Maps for Artificial Intelligence - Data Release V1. University of Virginia Dataverse. https://doi.org/10.18130/V3/DXWOS5 [90]
  - Clark T, Parker J, Al Manir S. Cell Maps for Artificial Intelligence - March 2025 Data Release (Beta) V1. University of Virginia Dataverse. https://doi.org/10.18130/V3/B35XWX [91]
  - Clark T, Parker J, Al Manir S. Cell Maps for Artificial Intelligence - June 2025 Data Release (Beta) V3. University of Virginia Dataverse. https://doi.org/10.18130/V3/F3TD5R [92]
  - Clark T, Parker J, Al Manir S, et al. Cell Maps for Artificial Intelligence - October 2025 Data Release (Beta) V3. University of Virginia Dataverse. https://doi.org/10.18130/V3/K7TGEM [93]

REFERENCES

[1] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P.

Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. https://doi.org/10.1038/sdata.2016.18.

[2] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Information Fusion 99 (2023) 101805. https://doi.org/10.1016/j.inffus.2023.101805.

[3] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J.H. Moore, M. Zitnik, J.H. Holmes, A manifesto on explainability for artificial intelligence in medicine, Artificial Intelligence in Medicine 133 (2022) 102423. https://doi.org/10.1016/j.artmed.2022.102423.

[4] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

[5] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A Review of Machine Learning Interpretability Methods, Entropy 23 (2020) 18. https://doi.org/10.3390/e23010018.

[6] D. Minh, H.X. Wang, Y.F. Li, T.N. Nguyen, Explainable artificial intelligence: a comprehensive review, Artif Intell Rev 55 (2022) 3503–3568. https://doi.org/10.1007/s10462-021-10088-y.

[7] S. Leonelli, Data Governance is Key to Interpretation: Reconceptualizing Data in Data Science, Harvard Data Science Review (2019). https://doi.org/10.1162/99608f92.17405bb6.

[8] T. Clark, H. Caufield, J.A. Parker, S. Al Manir, E. Amorim, J. Eddy, N. Gim, B. Gow, W. Goar, M. Haendel, J.N. Hansen, N. Harris, H. Hermjakob, M. Joachimiak, G. Jordan, I.-H. Lee, S.K. McWeeney, C. Nebeker, M. Nikolov, J. Shaffer, N. Sheffield, G. Sheynkman, J. Stevenson, J.Y. Chen, C. Mungall, A. Wagner, S.W. Kong, S.S. Ghosh, B. Patel, A. Williams, M.C. Munoz-Torres, AI-readiness for Biomedical Data: Bridge2AI Recommendations, bioRxiv (2024) 2024.10.23.619844. https://doi.org/10.1101/2024.10.23.619844.

[9] M. Akhtar, O. Benjelloun, C. Conforti, L. Foschini, J. Giner-Miguelez, P. Gijsbers, S. Goswami, N. Jain, M. Karamousadakis, M. Kuchnik, S. Krishna, S. Lesage, Q. Lhoest, P. Marcenac, M. Maskey, P. Mattson, L. Oala, H. Oderinwale, P. Ruyssen, T. Santos, R. Shinde, E. Simperl, A. Suresh, G. Thomas, S. Tykhonov, J. Vanschoren, S. Varma, J. van der Velde, S. Vogler, C.-J. Wu, L. Zhang, Croissant: A Metadata Format for ML-Ready Datasets, in: Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning, 2024: pp. 1–6. https://doi.org/10.1145/3650203.3663326.

[10] N. Jain, M. Akhtar, J. Giner-Miguelez, R. Shinde, J. Vanschoren, S. Vogler, S. Goswami, Y. Rao, T. Santos, L. Oala, M. Karamousadakis, M. Maskey, P. Marcenac, C. Conforti, M. Kuchnik, L. Aroyo, O. Benjelloun, E. Simperl, A Standardized Machine-

readable Dataset Documentation Format for Responsible AI, (2024). https://doi.org/10.48550/arXiv.2407.16883.

[11] S.A.T. Moxon, H. Solbrig, N.L. Harris, P. Kalita, M.A. Miller, S. Patil, K. Schaper, C. Bizon, J.H. Caufield, S.C. Cuesta, C. Cox, F. Dekervel, D.M. Dooley, W.D. Duncan, T. Fliss, S. Gehrke, A.S.L. Graefe, H. Hegde, A.J. Ireland, J.O.B. Jacobsen, M. Krishnamurthy, C. Kroll, D. Linke, R. Ly, N. Matentzoglu, J.A. Overton, J.L. Saunders, D.R. Unni, G. Vaidya, W.-M.A.M. Vierdag, O. Ruebel, C.G. Chute, M.H. Brush, M.A. Haendel, C.J. Mungall, LinkML: An Open Data Modeling Framework, GigaScience (2025) giaf152. https://doi.org/10.1093/gigascience/giaf152.

[12] H.L. Rehm, A.J.H. Page, L. Smith, J.B. Adams, G. Alterovitz, L.J. Babb, M.P. Barkley, M. Baudis, M.J.S. Beauvais, T. Beck, J.S. Beckmann, S. Beltran, D. Bernick, A. Bernier, J.K. Bonfield, T.F. Boughtwood, G. Bourque, S.R. Bowers, A.J. Brookes, M. Brudno, M.H. Brush, D. Bujold, T. Burdett, O.J. Buske, M.N. Cabili, D.L. Cameron, R.J. Carroll, E. Casas-Silva, D. Chakravarty, B.P. Chaudhari, S.H. Chen, J.M. Cherry, J. Chung, M. Cline, H.L. Clissold, R.M. Cook-Deegan, M. Courtot, F. Cunningham, M. Cupak, R.M. Davies, D. Denisko, M.J. Doerr, L.I. Dolman, E.S. Dove, L.J. Dursi, S.O.M. Dyke, J.A. Eddy, K. Eilbeck, K.P. Ellrott, S. Fairley, K.A. Fakhro, H.V. Firth, M.S. Fitzsimons, M. Fiume, P. Flicek, I.M. Fore, M.A. Freeberg, R.R. Freimuth, L.A. Fromont, J. Fuerth, C.L. Gaff, W. Gan, E.M. Ghanaim, D. Glazer, R.C. Green, M. Griffith, O.L. Griffith, R.L. Grossman, T. Groza, J.M. Guidry Auvil, R. Guigó, D. Gupta, M.A. Haendel, A. Hamosh, D.P. Hansen, R.K. Hart, D.M. Hartley, D. Haussler, R.M. Hendricks-Sturrup, C.W.L. Ho, A.E. Hobb, M.M. Hoffman, O.M. Hofmann, P. Holub, J.S. Hsu, J.-P. Hubaux, S.E. Hunt, A. Husami, J.O. Jacobsen, S.S. Jamuar, E.L. Janes, F. Jeanson, A. Jené, A.L. Johns, Y. Joly, S.J.M. Jones, A. Kanitz, K. Kato, T.M. Keane, K. Kekesi-Lafrance, J. Kelleher, G. Kerry, S.-S. Khor, B.M. Knoppers, M.A. Konopko, K. Kosaki, M. Kuba, J. Lawson, R. Leinonen, S. Li, M.F. Lin, M. Linden, X. Liu, I.U. Liyanage, J. Lopez, A.M. Lucassen, M. Lukowski, A.L. Mann, J. Marshall, M. Mattioni, A. Metke-Jimenez, A. Middleton, R.J. Milne, F. Molnár-Gábor, N. Mulder, M.C. Munoz-Torres, R. Nag, H. Nakagawa, J. Nasir, A. Navarro, T.H. Nelson, A. Niewielska, A. Nisselle, J. Niu, T.H. Nyrönen, B.D. O'Connor, S. Oesterle, S. Ogishima, V. Ota Wang, L.A.D. Paglione, E. Palumbo, H.E. Parkinson, A.A. Philippakis, A.D. Pizarro, A. Prlic, J. Rambla, A. Rendon, R.A. Rider, P.N. Robinson, K.W. Rodarmer, L.L. Rodriguez, A.F. Rubin, M. Rueda, G.A. Rushton, R.S. Ryan, G.I. Saunders, H. Schuilenburg, T. Schwede, S. Scollen, A. Senf, N.C. Sheffield, N. Skantharajah, A.V. Smith, H.J. Sofia, D. Spalding, A.B. Spurdle, Z. Stark, L.D. Stein, M. Suematsu, P. Tan, J.A. Tedds, A.A. Thomson, A. Thorogood, T.L. Tickle, K. Tokunaga, J. Törnroos, D. Torrents, S. Upchurch, A. Valencia, R.V. Guimera, J. Vamathevan, S. Varma, D.F. Vears, C. Viner, C. Voisin, A.H. Wagner, S.E. Wallace, B.P. Walsh, M.S. Williams, E.C. Winkler, B.J. Wold, G.M. Wood, J.P. Woolley, C. Yamasaki, A.D. Yates, C.K. Yung, L.J. Zass, K. Zaytseva, J. Zhang, P. Goodhand, K. North, E. Birney, GA4GH: International policies and standards for data sharing across genomic research and healthcare, Cell Genomics 1 (2021) 100029. https://doi.org/10.1016/j.xgen.2021.100029.

[13] S. Barbosa, L. Curtin, H. Cousijn, Generalist Repository Ecosystem Initiative Introductory Brochure, (2023). https://doi.org/10.5281/ZENODO.8350509.

[14] RO-Crate Community, Research Object Crate (RO-Crate), (2023). https://www.researchobject.org/ro-crate/.

[15] T. Gebru, J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H.D. Iii, K. Crawford, Datasheets for datasets, Commun. ACM 64 (2021) 86–92.

https://doi.org/10.1145/3458723.

[16] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji, T. Gebru, Model Cards for Model Reporting, Proceedings of the Conference on Fairness, Accountability, and Transparency (2019) 220–229. https://doi.org/10.1145/3287560.3287596.

[17] T. Clark, J. Mohan, L. Schaffer, K. Obernier, S. Al Manir, C.P. Churas, A. Dailamy, Y. Doctor, A. Forget, J.N. Hansen, M. Hu, J. Lenkiewicz, M.A. Levinson, C. Marquez, S. Nourreddine, J. Niestroy, D. Pratt, G. Qian, S. Thaker, J.-C. Bélisle-Pipon, C. Brandt, J. Chen, Y. Ding, S. Fodeh, N. Krogan, E. Lundberg, P. Mali, P. Payne-Foster, S. Ratcliffe, V. Ravitsky, A. Sali, W. Schulz, T. Ideker, Cell Maps for Artificial Intelligence: AI-Ready Maps of Human Cell Architecture from Disease-Relevant Cell Lines, (2024). https://doi.org/10.1101/2024.05.21.589311.

[18] E.W. Sayers, J. Beck, E.E. Bolton, J.R. Brister, J. Chan, R. Connor, M. Feldgarden, A.M. Fine, K. Funk, J. Hoffman, S. Kannan, C. Kelly, W. Klimke, S. Kim, S. Lathrop, A. Marchler-Bauer, T.D. Murphy, C. O'Sullivan, E. Schmieder, Y. Skripchenko, A. Stine, F. Thibaud-Nissen, J. Wang, J. Ye, E. Zellers, V.A. Schneider, K.D. Pruitt, Database resources of the National Center for Biotechnology Information in 2025, Nucleic Acids Research 53 (2025) D20–D29. https://doi.org/10.1093/nar/gkae979.

[19] S. Ahalt, P. Avillach, R. Boyles, K. Bradford, S. Cox, B. Davis-Dusenbery, R.L. Grossman, A. Krishnamurthy, A. Manning, B. Paten, A. Philippakis, I. Borecki, S.H. Chen, J. Kaltman, S. Ladwa, C. Schwartz, A. Thomson, S. Davis, A. Leaf, J. Lyons, E. Sheets, J.C. Bis, M. Conomos, A. Culotti, T. Desain, J. Digiovanna, M. Domazet, S. Gogarten, A. Gutierrez-Sacristan, T. Harris, B. Heavner, D. Jain, B. O'Connor, K. Osborn, D. Pillion, J. Pleiness, K. Rice, G. Rupp, A. Serret-Larmande, A. Smith, J.P. Stedman, A. Stilp, T. Barsanti, J. Cheadle, C. Erdmann, B. Farlow, A. Gartland-Gray, J. Hayes, H. Hiles, P. Kerr, C. Lenhardt, T. Madden, J.O. Mieczkowska, A. Miller, P. Patton, M. Rathbun, S. Suber, J. Asare, Building a collaborative cloud platform to accelerate heart, lung, blood, and sleep research, Journal of the American Medical Informatics Association 30 (2023) 1293–1300. https://doi.org/10.1093/jamia/ocad048.

[20] M.C. Schatz, A.A. Philippakis, E. Afgan, E. Banks, V.J. Carey, R.J. Carroll, A. Culotti, K. Ellrott, J. Goecks, R.L. Grossman, I.M. Hall, K.D. Hansen, J. Lawson, J.T. Leek, A.O. Luria, S. Mosher, M. Morgan, A. Nekrutenko, B.D. O'Connor, K. Osborn, B. Paten, C. Patterson, F.J. Tan, C.O. Taylor, J. Vessio, L. Waldron, T. Wang, K. Wuichet, A. Baumann, A. Rula, A. Kovalsy, C. Bernard, D. Caetano-Anollés, G.A. Van Der Auwera, J. Canas, K. Yuksel, K. Herman, M.M. Taylor, M. Simeon, M. Baumann, Q. Wang, R. Title, R. Munshi, S. Chaluvadi, V. Reeves, W. Disman, S. Thomas, A. Hajian, E. Kiernan, N. Gupta, T. Vosburg, L. Geistlinger, M. Ramos, S. Oh, D. Rogers, F. McDade, M. Hastie, N. Turaga, A. Ostrovsky, A. Mahmoud, D. Baker, D. Clements, K.E.L. Cox, K. Suderman, N. Kucher, S. Golitsynskiy, S. Zarate, S.J. Wheelan, K. Kammers, A. Stevens, C. Hutter, C. Wellington, E.M. Ghanaim, K.L. Wiley, S.K. Sen, V. Di Francesco, D. S Yuen, B. Walsh, L. Sargent, V. Jalili, J. Chilton, L. Shepherd, B.J. Stubbs, A. O'Farrell, B.A. Vizzier, C. Overbeck, C. Reid, D.C. Steinberg, E.A. Sheets, J. Lucas, L. Blauvelt, L. Cabansay, N. Warren, B. Hannafious, T. Harris, R. Reddy, E. Torstenson, M.K. Banasiewicz, H.J. Abel, J. Walker, Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space, Cell Genomics 2 (2022) 100085. https://doi.org/10.1016/j.xgen.2021.100085.

[21] A.P. Heath, V. Ferretti, S. Agrawal, M. An, J.C. Angelakos, R. Arya, R. Bajari, B. Baqar, J.H.B. Barnowski, J. Burt, A. Catton, B.F. Chan, F. Chu, K. Cullion, T. Davidsen, P.-M. Do, C. Dompierre, M.L. Ferguson, M.S. Fitzsimons, M. Ford, M. Fukuma, S. Gaheen, G.L. Ganji, T.I. Garcia, S.S. George, D.S. Gerhard, F. Gerthoffert, F. Gomez, K. Han, K.M. Hernandez, B. Issac, R. Jackson, M.A. Jensen, S. Joshi, A. Kadam, A. Khurana, K.M.J. Kim, V.E. Kraft, S. Li, T.M. Lichtenberg, J. Lodato, L. Lolla, P. Martinov, J.A. Mazzone, D.P. Miller, I. Miller, J.S. Miller, K. Miyauchi, M.W. Murphy, T. Nullet, R.O. Ogwara, F.M. Ortuño, J. Pedrosa, P.L. Pham, M.Y. Popov, J.J. Porter, R. Powell, K. Rademacher, C.P. Reid, S. Rich, B. Rogel, H. Sahni, J.H. Savage, K.A. Schmitt, T.J. Simmons, J. Sislow, J. Spring, L. Stein, S. Sullivan, Y. Tang, M. Thiagarajan, H.D. Troyer, C. Wang, Z. Wang, B.L. West, A. Wilmer, S. Wilson, K. Wu, W.P. Wysocki, L. Xiang, J.T. Yamada, L. Yang, C. Yu, C.K. Yung, J.C. Zenklusen, J. Zhang, Z. Zhang, Y. Zhao, A. Zubair, L.M. Staudt, R.L. Grossman, The NCI Genomic Data Commons, Nat Genet 53 (2021) 257–262. https://doi.org/10.1038/s41588-021-00791-5.

[22] M. Thakur, N. Bosc, C. Brooksbank, C. Ernst, M.A. Freeberg, K.T. Gurwitz, H. Hermjakob, D.G. Hulcoop, M.J. Martin, E.M. McDonagh, A. Mithani, N.M. O'Boyle, D. Ochoa, T. Payne, Y. Perez-Riverol, U. Sarkans, A. Sokolov, N. Staudt, J.D. Stephenson, E. Tzampatzopoulou, J.A. Vizcaíno, B. Zdrazil, J. McEntyre, EMBL's European Bioinformatics Institute (EMBL-EBI) in 2025, Nucleic Acids Research 54 (2026) D10–D19. https://doi.org/10.1093/nar/gkaf1078.

[23] Chard Kyle, Gaffney Niall, Hategan Mihael, Kowalik Kacper, Ludäscher Bertram, McPhillips Timothy, Nabrzyski Jarek, Stodden Victoria, Taylor Ian, Thelen Thomas, Turk Matthew J., Willis Craig, Toward Enabling Reproducibility for Data-Intensive Research Using the Whole Tale Platform, in: Advances in Parallel Computing, IOS Press, 2020. https://doi.org/10.3233/APC200107.

[24] L. Curtin, L. Feri, J. Gautier, S. Gonzales, G. Gueguen, D. Scherer, R. Scherle, K. Stathis, A. Van Gulick, J. Wood, GREI Metadata and Search Subcommittee Recommendations_V01_2023-06-29, (2023). https://doi.org/10.5281/ZENODO.8101957.

[25] G. King, An Introduction to the Dataverse Network as an Infrastructure for Data Sharing, Sociological Methods & Research 36 (2007) 173–199. https://doi.org/10.1177/0049124107306660.

[26] V. Gurav, S.R. Nagarkar, ZENODO: A PLATFORM FOR OPEN ACCESS AND SUSTAINABLE DIGITAL RESEARCH REPOSITORY, (2025). https://doi.org/10.2139/ssrn.5246452.

[27] J. Singh, FigShare, Journal of Pharmacology and Pharmacotherapeutics 2 (2011) 138–139. https://doi.org/10.4103/0976-500X.81919.

[28] E.D. Foster, A. Deardorff, Open Science Framework (OSF), Jmla 105 (2017). https://doi.org/10.5195/jmla.2017.88.

[29] B.E. Bierer, R. Li, M. Barnes, I. Sim, A Global, Neutral Platform for Sharing Trial Data, N Engl J Med 374 (2016) 2411–2413. https://doi.org/10.1056/NEJMp1605348.

[30] J. Greenberg, H.C. White, S. Carrier, R. Scherle, A Metadata Best Practice for a Scientific Data Repository, Journal of Library Metadata 9 (2009) 194–212. https://doi.org/10.1080/19386380903405090.

[31] J. Singh, Mendeley: A free research management tool for desktop and web, 2010. http://www.jpharmacol.com/article.asp?issn=0976-500X;year=2010;volume=1;issue=1;spage=62;epage=63;aulast=Singh.

[32] M. Hahnel, J. Chodacki, S. Neumann, L.H. Nielsen, J. Gautier, A. Hamelers, G. Gueguen, M. Call, D. Scherer, GREI Metadata Recommendations from DataCite schema version 4.6, (2025). https://doi.org/10.5281/ZENODO.16953588.

[33] DataCite Metadata Working Group, DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.6, (2024). https://doi.org/10.14454/MZV1-5B55.

[34] Y. Gil, S. Miles, K. Belhajjame, H. Deus, D. Garijo, G. Klyne, P. Missier, S. Soiland-Reyes, S. Zednik, PROV Model Primer: W3C Working Group Note 30 April 2013, (2013). https://www.w3.org/TR/prov-primer/.

[35] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. Gray, C. Goble, T. Clark, PAV ontology: Provenance, Authoring and Versioning, J Biomed Semantics (2013). Nov 22;4(1):37. doi: 10.1186/2041-1480-4-37.

[36] L. Moreau, P. Groth, J. Cheney, T. Lebo, S. Miles, The rationale of PROV, Journal of Web Semantics 35 (2015) 235–257. https://doi.org/10.1016/j.websem.2015.04.001.

[37] L. Moreau, P. Missier, K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, Curt Tilmes, PROV-DM: The PROV Data Model: W3C Recommendation 30 April 2013, World Wide Web Consortium, 2013. http://www.w3.org/TR/prov-dm/.

[38] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, PROV-O: The PROV Ontology W3C Recommendation 30 April 2013, (2013). http://www.w3.org/TR/prov-o/.

[39] T. Clark, P.N. Ciccarese, C.A. Goble, Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications, J Biomed Semantics 5 (2014) 28. https://doi.org/10.1186/2041-1480-5-28.

[40] R.S. Gonçalves, M.J. O'Connor, M. Martínez-Romero, A.L. Egyedi, D. Willrett, J. Graybeal, M.A. Musen, The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments, in: C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, J. Heflin (Eds.), The Semantic Web – ISWC 2017, Springer International Publishing, Cham, 2017: pp. 103–110. https://doi.org/10.1007/978-3-319-68204-4_10.

[41] M.A. Musen, C.A. Bean, K.-H. Cheung, M. Dumontier, K.A. Durante, O. Gevaert, A. Gonzalez-Beltran, P. Khatri, S.H. Kleinstein, M.J. O'Connor, Y. Pouliot, P. Rocca-Serra, S.-A. Sansone, J.A. Wiser, and the CEDAR team, The center for expanded data annotation and retrieval, Journal of the American Medical Informatics Association 22 (2015) 1148–1152. https://doi.org/10.1093/jamia/ocv048.

[42] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, M.A. Musen, BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications,

Nucleic Acids Research 39 (2011) W541–W545. http://nar.oxfordjournals.org/content/39/suppl_2/W541.abstractN2.

[43] M.J. O'Connor, J. Hardi, M. Martínez-Romero, S. Somasundaram, B. Honick, S.A. Fisher, A. Pillai, M.A. Musen, Ensuring Adherence to Standards in Experiment-Related Metadata Entered Via Spreadsheets, Sci Data 12 (2025) 265. https://doi.org/10.1038/s41597-025-04589-6.

[44] HuBMAP Consortium, Writing Group, M.P. Snyder, S. Lin, A. Posgai, M. Atkinson, A. Regev, J. Rood, O. Rozenblatt-Rosen, L. Gaffney, A. Hupalowska, R. Satija, N. Gehlenborg, J. Shendure, J. Laskin, P. Harbury, N.A. Nystrom, J.C. Silverstein, Z. Bar-Joseph, K. Zhang, K. Börner, Y. Lin, R. Conroy, D. Procaccini, A.L. Roy, A. Pillai, M. Brown, Z.S. Galis, Caltech-UW TMC, L. Cai, J. Shendure, C. Trapnell, S. Lin, D. Jackson, Stanford-WashU TMC, M.P. Snyder, G. Nolan, W.J. Greenleaf, Y. Lin, S. Plevritis, S. Ahadi, S.A. Nevins, H. Lee, C.M. Schuerch, S. Black, V.G. Venkataraaman, E. Esplin, A. Horning, A. Bahmani, UCSD TMC, K. Zhang, X. Sun, S. Jain, J. Hagood, G. Pryhuber, P. Kharchenko, University of Florida TMC, M. Atkinson, B. Bodenmiller, T. Brusko, M. Clare-Salzler, H. Nick, K. Otto, A. Posgai, C. Wasserfall, M. Jorgensen, M. Brusko, S. Maffioletti, Vanderbilt University TMC, R.M. Caprioli, J.M. Spraggins, D. Gutierrez, N.H. Patterson, E.K. Neumann, R. Harris, M. deCaestecker, A.B. Fogo, R. Van De Plas, K. Lau, California Institute of Technology TTD, L. Cai, G.-C. Yuan, Q. Zhu, R. Dries, Harvard TTD, P. Yin, S.K. Saka, J.Y. Kishi, Y. Wang, I. Goldaracena, Purdue TTD, J. Laskin, D. Ye, K.E. Burnum-Johnson, P.D. Piehowski, C. Ansong, Y. Zhu, Stanford TTD, P. Harbury, T. Desai, J. Mulye, P. Chou, M. Nagendran, HuBMAP Integration, Visualization, and Engagement (HIVE) Collaboratory: Carnegie Mellon, Tools Component, Z. Bar-Joseph, S.A. Teichmann, B. Paten, R.F. Murphy, J. Ma, V.Yu. Kiselev, C. Kingsford, A. Ricarte, M. Keays, S.A. Akoju, M. Ruffalo, Harvard Medical School, Tools Component, N. Gehlenborg, P. Kharchenko, M. Vella, C. McCallum, Indiana University Bloomington, Mapping Component, K. Börner, L.E. Cross, S.H. Friedman, R. Heiland, B. Herr, P. Macklin, E.M. Quardokus, L. Record, J.P. Sluka, G.M. Weber, Pittsburgh Supercomputing Center and University of Pittsburgh, Infrastructure and Engagement Component, N.A. Nystrom, J.C. Silverstein, P.D. Blood, A.J. Ropelewski, W.E. Shirey, R.M. Scibek, University of South Dakota, Collaboration Core, P. Mabee, W.C. Lenhardt, K. Robasky, S. Michailidis, New York Genome Center, Mapping Component, R. Satija, J. Marioni, A. Regev, A. Butler, T. Stuart, E. Fisher, S. Ghazanfar, J. Rood, L. Gaffney, G. Eraslan, T. Biancalani, E.D. Vaishnav, NIH HuBMAP Working Group, R. Conroy, D. Procaccini, A. Roy, A. Pillai, M. Brown, Z. Galis, P. Srinivas, A. Pawlyk, S. Sechi, E. Wilder, J. Anderson, The human body at cellular resolution: the NIH Human Biomolecular Atlas Program, Nature 574 (2019) 187–192. https://doi.org/10.1038/s41586-019-1629-x.

[45] R.G. Baker, W.J. Koroshetz, N.D. Volkow, The Helping to End Addiction Long-term (HEAL) Initiative of the National Institutes of Health, JAMA 326 (2021) 1005. https://doi.org/10.1001/jama.2021.13300.

[46] M.A. Musen, M.J. O'Connor, J. Hardi, M. Martínez-Romero, Knowledge Engineering for Open Science: Building and Deploying Knowledge Bases for Metadata Standards, AI Magazine 47 (2026) e70048. https://doi.org/10.1002/aaai.70048.

[47] R.M. Krug, O.L. Petchey, Metadata Made Easy: Develop and Use Domain-Specific Metadata Schemes by following the dmdScheme approach, Ecology and Evolution 11 (2021) 9174–9181. https://doi.org/10.1002/ece3.7764.

[48] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. Tong, S.-A. Sansone, ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level, Bioinformatics 26 (2010) 2354–2356. https://doi.org/10.1093/bioinformatics/btq415.

[49] E.H. Fegraus, S. Andelman, M.B. Jones, M. Schildhauer, Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation, Bulletin of the Ecological Society of America 86 (2005) 158–168. https://doi.org/10.1890/0012-9623(2005)86%5B158:MTVOED%5D2.0.CO;2.

[50] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking Clever Hans predictors and assessing what machines really learn, Nat Commun 10 (2019) 1096. https://doi.org/10.1038/s41467-019-08987-4.

[51] J. Kauffmann, J. Dippel, L. Ruff, W. Samek, K.-R. Müller, G. Montavon, Explainable AI reveals Clever Hans effects in unsupervised learning models, Nat Mach Intell 7 (2025) 412–422. https://doi.org/10.1038/s42256-025-01000-2.

[52] W. Ye, L. Jiang, E. Xie, G. Zheng, Y. Ma, X. Cao, D. Guo, D. Qi, Z. He, Y. Tian, M. Coffee, Z. Zeng, S. Li, Ting-hao, Huang, Z. Wang, J.M. Rehg, H. Kautz, A. Zhang, The Clever Hans Mirage: A Comprehensive Survey on Spurious Correlations in Machine Learning, (2025). https://doi.org/10.48550/arXiv.2402.12715.

[53] A.K. Pathak, M. Gupta, G. Jain, Unmasking the Clever Hans effect in AI models: shortcut learning, spurious correlations, and the path toward robust intelligence, Front. Artif. Intell. 8 (2026) 1692454. https://doi.org/10.3389/frai.2025.1692454.

[54] J. Giner-Miguelez, A. Gómez, J. Cabot, DescribeML: A dataset description tool for machine learning, Science of Computer Programming 231 (2024) 103030. https://doi.org/10.1016/j.scico.2023.103030.

[55] S.C. Edmunds, N. Nogoy, Q. Lan, H. Zhang, Y. Fan, H. Zhou, C. Armit, Integrating Machine Learning Standards in Disseminating Machine Learning Research, Data Science Journal 25 (2026) 1. https://doi.org/10.5334/dsj-2026-001.

[56] Marco La Rosa, Contributors, Describo, (2023). https://describo.github.io/.

[57] Language Data Commons of Australia (LDaCA), Crate-O: A browser-based editor for Research Object Crates (RO-Crate), (2024). https://language-research-technology.github.io/crate-o/.

[58] J.C. Niestroy, J.R. Moorman, M.A. Levinson, S.A. Manir, T.W. Clark, K.D. Fairchild, D.E. Lake, Discovery of signatures of fatal neonatal illness in vital signs using highly comparative time-series analysis, Npj Digit. Med. 5 (2022) 6. https://doi.org/10.1038/s41746-021-00551-z.

[59] M.A. Levinson, J. Niestroy, S. Al Manir, K. Fairchild, D.E. Lake, J.R. Moorman, T. Clark, FAIRSCAPE: a Framework for FAIR and Reproducible Biomedical Analytics, Neuroinform 20 (2022) 187–202. https://doi.org/10.1007/s12021-021-09529-4.

[60] D. Dunning, C. Heath, J.M. Suls, Flawed Self-Assessment: Implications for

Health, Education, and the Workplace, Psychol Sci Public Interest 5 (2004) 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x.

[61] P. Sefton, E. Ó Carragáin, S. Soiland-Reyes, O. Corcho, D. Garijo, R. Palma, F. Coppens, C. Goble, J.M. Fernández, K. Chard, J.M. Gomez-Perez, M.R. Crusoe, I. Eguinoa, N. Juty, K. Holmes, J.A. Clark, S. Capella-Gutierrez, A.J.G. Gray, S. Owen, A.R. Williams, G. Tartari, F. Bacall, T. Thelen, H. Ménager, L. Rodríguez-Navas, P. Walk, brandon whitehead, M. Wilkinson, P. Groth, E. Bremer, L.J. Castro, K. Sebby, A. Kanitz, A. Trisovic, G. Kennedy, M. Graves, J. Koehorst, S. Leo, M. Portier, P. Brack, M. Ojsteršek, B. Droesbeke, C. Niu, K. Tanabe, T. Miksa, M. La Rosa, C. Decruw, A. Czerniak, J. Jay, S. Serra, R. Siebes, S. de Witt, S. El Damaty, D. Lowe, X. Li, S. Gundersen, M. Radifar, R. Wittner, O. Woolland, P. De Geest, D. Fils, F. Wetzels, R. Sirvent, A. Miller, J. Emerson, D. Fucci, B.P. Kinoshita, M. Bąk, J. Hollunder, M. Weise, V. Bisht, T.N. Hiraki, B. Ulrichts, M. Falk, E. Chadwick, D. Bauer, J. Love, E. Adamidi, J. Moore, L. Schöbitz, A. Meier, J. Fuentes, E. Bainglass, B.E. Pataki, RO-Crate Metadata Specification 1.2.0, (2025). https://doi.org/10.5281/ZENODO.13751027.

[62] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, P.-A. Champin, N. Lindström, JSON-LD 1.1: A JSON-based Serialization for Linked Data, W3C Recommendation 16 July 2020 (2020). https://www.w3.org/TR/json-ld/.

[63] R.V. Guha, D. Brickley, S. Macbeth, Schema.org: evolution of structured data on the web, Communications of the ACM 59 (2016) 44–51. https://doi.org/10.1145/2844544.

[64] S. Al Manir, J. Niestroy, M.A. Levinson, T. Clark, Evidence Graphs: Supporting Transparent and FAIR Computation, with Defeasible Reasoning on Data, Methods, and Results, in: B. Glavic, V. Braganholo, D. Koop (Eds.), Provenance and Annotation of Data and Processes, Springer International Publishing, Cham, 2021: pp. 39–50. https://doi.org/10.1007/978-3-030-80960-7_3.

[65] S. Al Manir, J. Niestroy, M. Levinson, T. Clark, EVI: The Evidence Graph Ontology, OWL 2 Vocabulary, (2021). https://doi.org/10.5281/zenodo.4630931.

[66] Wright, A., Andrews, A., Hutton, B., Dennis, G., JSON Schema: A Media Type for Describing JSON Documents, (2022). https://json-schema.org/draft/2020-12/json-schema-core.

[67] D. Fowler, J. Barratt, P. Walsh, Frictionless Data: Making Research Data Quality Visible, IJDC 12 (2018) 274–285. https://doi.org/10.2218/ijdc.v12i2.577.

[68] T. Berners-Lee, R. Fielding, L. Masinter, IETF RFC 3986: Uniform Resource Identifier (URI): Generic Syntax, 2005. http://tools.ietf.org/html/rfc3986.

[69] J. Kunze, R. Rodgers, The ARK Identifier Scheme, (2008). https://escholarship.org/uc/item/9p9863nc.

[70] N. Juty, S.M. Wimalaratne, S. Soiland-Reyes, J. Kunze, C.A. Goble, T. Clark, Unique, Persistent, Resolvable: Identifiers as the foundation of FAIR, Data Intelligence 2 (2020) 30–39. https://doi.org/http://doi.org/10.5281/zenodo.3267434.

[71] Pydantic, Pydantic: Documentation for version: v2.12.5, (n.d.). https://docs.pydantic.dev/latest/.

[72] Justin Niestroy, Max Levinson, Sadnan Al Manir, T. Clark, Fairscape Pydantic Models, (2026). https://doi.org/10.5281/ZENODO.18234523.

[73] N.C. Sheffield, M. Stolarczyk, V.P. Reuter, A.F. Rendeiro, Linking big biomedical datasets to modular analysis with Portable Encapsulated Projects, Gigascience 10 (2021) giab077. https://doi.org/10.1093/gigascience/giab077 (accessed August 8, 2021).

[74] R. Cunningham, Virtual Witnessing and the Role of the Reader in a New Natural Philosophy, Philosophy and Rhetoric 34 (2001) 207–224. https://doi.org/10.1353/par.2001.0013.

[75] S. Shapin, Pump and Circumstance: Robert Boyle's Literary Technology, Social Studies of Science 14 (1984) 481–520. http://sss.sagepub.com/content/14/4/481.abstractN2.

[76] S. Shapin, S. Schaffer, Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life, Princeton University Press, 2011. https://doi.org/10.1515/9781400838493.

[77] L. Principe, In retrospect: The Sceptical Chymist, Nature 469 (2011) 30–31. https://doi.org/10.1038/469030a.

[78] J.T. Hummel, H. Berends, P. Tuertscher, From Boundary Objects to Boundary Infrastructure: A Process Study of Collaboration between Big Science and Big Business, J Management Studies 62 (2025) 1644–1679. https://doi.org/10.1111/joms.13118.

[79] A.M. Pedersen, C. Bossen, Cultivating Data Practices Across Boundaries: How Organizations Become Data-driven, Comput Supported Coop Work 33 (2024) 1177–1221. https://doi.org/10.1007/s10606-024-09489-8.

[80] G.C. Bowker, S. Timmermans, A.E. Clarke, E. Balka, eds., Boundary Objects and Beyond: Working with Leigh Star, The MIT Press, 2016. https://doi.org/10.7551/mitpress/10113.001.0001.

[81] S.F. Akkerman, A. Bakker, Boundary Crossing and Boundary Objects, Review of Educational Research 81 (2011) 132–169. https://doi.org/10.3102/0034654311404435.

[82] E.G. Arias, G. Fischer, Boundary Objects: Their Role in Articulating the Task at Hand and Making Information Relevant to It, in: 2000.

[83] S.L. Star, J.R. Griesemer, Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39, Social Studies of Science 19 (1989) 387–420. http://ezp-prod1.hul.harvard.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=11446079&site=ehost-live&scope=site.

[84] M.A. Levinson, S. Al Manir, T. Clark, FAIRSCAPE CLI, (2026). https://doi.org/10.5281/ZENODO.18714417.

[85] M.A. Levinson, S. Al Manir, J. Niestroy, T. Clark, FAIRSCAPE Server, (2026). https://doi.org/10.5281/ZENODO.18244417.

[86] Simone Leo, Stian Soiland-Reyes, Ignacio Eguinoa, Bert Droesbeke, Daniel Bauer, Eli Chadwick, Bruno P. Kinoshita, Toshiyuki NH, Laurent Thomas, Laura

Rodríguez-Navas, albangaignard, Paul De Geest, Sebastiaan Huber, Matthias Hörtenhuber, Luca Pireddu, Raül Sirvent, ro-crate-py, (2025). https://doi.org/10.5281/ZENODO.17342107.

[87] S.A.T. Moxon, H. Solbrig, N.L. Harris, P. Kalita, M.A. Miller, S. Patil, K. Schaper, C. Bizon, J.H. Caufield, S. Cirujano Cuesta, C. Cox, F. Dekervel, D.M. Dooley, W.D. Duncan, T. Fliss, S. Gehrke, A.S.L. Graefe, H. Hegde, A. Ireland, J.O.B. Jacobsen, M. Krishnamurthy, C. Kroll, D. Linke, R. Ly, N. Matentzoglu, J.A. Overton, J.L. Saunders, D.R. Unni, G. Vaidya, W.-M.A.M. Vierdag, T. Putman, LinkML Community Contributors, O. Ruebel, C.G. Chute, M.H. Brush, M.A. Haendel, C.J. Mungall, LinkML: A Linked Open Data Modeling Language, (2026). https://doi.org/10.5281/ZENODO.5703670.

[88] J. Niestroy, M.A. Levinson, S. Al Manir, Timothy. Clark, Evidence Graph for: Discovery of signatures of fatal neonatal illness in vital signs using highly comparative time-series analysis, (2021). https://doi.org/10.18130/V3/HHTAYI.

[89] J. Niestroy, M.A. Levinson, S. Al Manir, T. Clark, J.R. Moorman, K.D. Fairchild, D.E. Lake, Replication Data for: Discovery of signatures of fatal neonatal illness in vital signs using highly comparative time-series analysis, V2, (2021). https://doi.org/10.18130/V3/VJXODP.

[90] T. Clark, L. Schaffer, K. Obernier, S. Al Manir, C. Churas, A. Dailamy, Y. Doctor, A. Forget, J. Hansen, M. Hu, J. Lenkiewicz, M. Levinson, C. Marquez, J. Mohan, S. Nourreddine, J. Niestroy, D. Pratt, G. Qian, S. Thaker, J.-C. Belisle-Pipon, C. Brandt, J. Chen, Y. Ding, S. Fodeh, N. Krogan, E. Lundberg, P. Mali, P. Payne-Foster, S. Ratcliffe, V. Ravitsky, A. Sali, W. Schulz, T. Ideker, Cell Maps for Artificial Intelligence - May 2024 Data Release, (2024) https://doi.org/10.18130/V3/DXWOS5.

[91] T. Clark, J. Parker, S. Al Manir, Cell Maps for Artificial Intelligence - March 2025 Data Release (Beta), (2025) 31852, 1024071, 2777891100, 3419964605, 3058024443, 2989. https://doi.org/10.18130/V3/B35XWX.

[92] T. Clark, J. Parker, S. Al Manir, Cell Maps for Artificial Intelligence - June 2025 Data Release (Beta), (2025) https://doi.org/10.18130/V3/F3TD5R.

[93] T. Clark, J. Parker, S. Al Manir, U. Axelsson, F. Ballllosero Navarro, B. Chinn, C. Churas, A. Dailamy, Y. Doctor, J. Fall, A. Forget, J. Gao, J. Hansen, M. Hu, A. Johannesson, H. Khaliq, Y. Lee, J. Lenkiewicz, M. Levinson, C. Marquez, C. Metallo, M. Muralidharan, S. Nourreddine, J. Niestroy, K. Obernier, E. Pan, B. Polacco, D. Pratt, G. Qian, L. Schaffer, A. Sigaeva, S. Thaker, Y. Zhang, J. Bélisle-Pipon, C. Brandt, J. Chen, Y. Ding, S. Fodeh, N. Krogan, E. Lundberg, P. Mali, P. Payne-Foster, S. Ratcliffe, V. Ravitsky, A. Sali, W. Schulz, T. Ideker, Cell Maps for Artificial Intelligence - October 2025 Data Release (Beta), (2025) https://doi.org/10.18130/V3/K7TGEM.