From Big Data to Valued Data: A Dataset Value Taxonomy for AI-Native Empirical Research

Scott Seidenberger, Anindya Maiti
University of Oklahoma
seidenberger@ou.edu, am@ou.edu

Abstract

Artificial intelligence is rapidly commoditizing many stages of empirical research, including code generation, statistical analysis, visualization, and manuscript drafting, yet its gains accrue unevenly across disciplines. As the marginal cost of these downstream tasks falls, the decisive bottleneck shifts to the data itself. The era of training and testing neural networks with ever-larger datasets is giving way to one in which the *value* of the data matters more than its volume. Building on but significantly revising earlier "Big Data" taxonomies, we introduce a *Dataset Value Taxonomy* (*DVT*) that reassesses the epistemic importance of datasets in an age where AI is commoditizing analytical labor. We argue that humans, as physically embodied agents, retain a comparative advantage in acquiring and curating observations from the world; partnered with AI, they can channel this advantage into higher-impact scholarship. To guide funders, evaluators, and researchers, we introduce a three-construct taxonomy: Temporal Investment, Scale, and Accessibility, each decomposed into measurable dimensions with ordinal class labels. We further operationalize these constructs through a calibrated scoring scheme that disciplines can adapt to field-specific conventions, enabling cross-field comparability while respecting contextual nuance. By quantifying how temporal investment, observational breadth, and access barriers jointly determine dataset salience, the framework helps allocate resources toward datasets that are most likely to advance knowledge. Ultimately, our taxonomy positions dataset creation as the new core site of human and AI cooperation and provides an actionable roadmap for recognizing, funding, and stewarding high-value data assets across the empirical sciences.

1 Introduction

Artificial intelligence is collapsing the cost of many analytic tasks that once defined empirical research across disciplines. Contemporary language and code generation models can now interpret structured and unstructured data, perform statistical analyses, draft manuscripts, and render publication-ready visualizations in minutes (Maslej et al. 2025). As the cost of this "downstream" analytic labor becomes commoditized, the decisive constraint on generating new knowledge shifts to the datasets that ground new claims. Human expertise still remains central to this endeavor. Only domain specialists can decide which phenomena are worth measuring, negotiate community norms, and curate records so that automated tools and intelligent agents can operate reliably. Partnered with AI, researchers can therefore redirect effort "upstream" toward the design, stewardship, and equitable sharing of high-value datasets.

In an effort to make this notion more concrete, we introduce the *Dataset Value Taxonomy* (*DVT*), a discipline-agnostic framework that assesses a dataset along three orthogonal constructs: Temporal investment, the irreducible time and labor embodied in longitudinal or event-triggered collection; Scale and breadth, the quantitative footprint and modal coverage that determine a dataset's representational reach; and Accessibility, the legal, technical, financial, and logistical barriers that shape who may inspect or reuse the data. Each construct is decomposed into measurable dimensions with ordinal class labels, then recomposed into a calibrated scoring scheme that any field can tune to its own evidentiary standards.

The paper makes four contributions. It synthesizes critical scholarship on "Big Data" and data governance into a compact set of evaluative currencies that transcend disciplinary silos. It operationalizes those currencies through a transparent scoring protocol that balances comparability with local calibration. It integrates an ethical lens that surfaces how temporal depth, scale, and gate-keeping decisions redistribute risk, labor, and epistemic authority. Finally, it demonstrates the framework's portability by mapping it onto exemplar corpora from astronomy, public health, climate science, finance, and digital humanities.

Section 2 motivates the taxonomy, Section 3 formalizes its scoring procedure, and Sections 4 to 6 detail each construct with illustrative calibrations and governance guidance. Section 7 discusses how financial resources modulate dataset value, while Section 8 outlines how weighted *DVT* scores can be applied across roles and suggests community-led benchmarking strategies to validate and refine the taxonomy. We close by arguing that in an AI-native empirical research economy, the most enduring contributions will come from teams that treat dataset creation as a first-order scholarly act, fusing human judgment with machine capability to build collections that are temporally deep, broadly representative, and responsibly accessible.

2 The Dataset Value Taxonomy

2.1 Motivation

The first wave of "Big Data" research framed progress as a simple function of scale. Seminal essays such as *The Unreasonable Effectiveness of Data* celebrated brute-force learning from trillion-word corpora and treated quantity as a substitute for sophisticated modeling or annotation (Halevy, Norvig, and Pereira 2009). That optimism was crystallized in the 3Vs mantra (volume, velocity, and variety) which became shorthand for data-driven discovery (Kitchin and McArdle 2016).

Critical scholarship quickly revealed that size alone does not guarantee epistemic soundness (Sambasivan et al. 2021). Boyd and Crawford cataloged unresolved questions about bias, privacy, and context loss, arguing that Big Data "changes the definition of knowledge" and therefore demands scrutiny at the level of assumptions and governance (Boyd and Crawford 2012). Ontological audits of actual datasets showed that few satisfy even the original 3Vs; Kitchin and McArdle demonstrated that velocity and exhaustibility matter more than raw volume and that datasets labeled "big" share no single trait set (Kitchin and McArdle 2016). In parallel, Leonelli's work on biology databases documented how value flows from labor-intensive curation, cross-community standards, and sustained infrastructure rather than from record counts alone (Leonelli 2014). Together, these studies repositioned data quantity as a necessary but insufficient condition for scientific insight.

The research landscape has since shifted again. Foundation-model AI now automates large portions of the empirical pipeline, from statistical modeling, figure generation, to baseline manuscript drafting (Maslej et al. 2025). This makes high-quality, well-governed data a true bottleneck in the scientific pipeline. Recent surveys of deep-learning note that model performance increasingly hinges on carefully prepared, domain-specific corpora coupled with human feedback rather than on ever-larger generic dumps (Goldblum et al. 2023). When analytical capacity is cheap and ubiquitous, the comparative advantage of a research team lies upstream in the time, breadth, and accessibility embodied in the datasets they control.

Existing scoring schemes do not meet this moment. "Big Data" traits remain descriptive, not evaluative; domain tools cannot compare a decadal ethnographic panel with a multi-petabyte sensor log (Chicco, Fabris, and Jurman 2025). A cross-disciplinary yardstick that foregrounds temporal investment, scale, and access barriers is therefore required.

2.2 Core Constructs

Dataset Value Taxonomy distills three orthogonal drivers of scholarly salience. Each captures a distinct source of epistemic leverage and is intentionally defined at a high level and Section 3 turns them into measurable dimensions and labels.

- **Temporal investment (T)**. The cumulative time and labor embodied in a corpus, which creates scarcity that technology cannot compress. Ethnographic work on large social-science repositories shows that craft-intensive curation can span years and decisively shapes downstream usability (Thomer et al. 2022).
- **Scale and breadth (S).** Scale captures the traditional definitions of volume. Breadth gauges diversity of units, modalities, and sampling frames rather than raw record count. Recent Earth-observation research finds that once models saturate, further gains hinge on representative coverage and label quality, not simply on adding more pixels or rows (Roscher et al. 2024).
- **Accessibility (A).** Legal, financial, technical, and logistical barriers govern who can inspect or reuse data. Foundational critiques of the Big-Data era emphasize that datasets remain "partial and contingent" when access is gated by format opacity, licensing, or computational cost (Thomer et al. 2022; Roscher et al. 2024); social-computing audits further demonstrate how opaque access decisions embed hidden politics into AI pipelines (Jones 2019; Scheuerman, Hanna, and Denton 2021).

These drivers are **complementary but compensatory**: excellence on one dimension can offset scarcity on another. In the next section we map T, S, and A onto concrete metrics so that diverse fields can score datasets without sacrificing cross-disciplinary comparability.

2.3 Prior Frameworks

Research on data quality and value spans three overlapping traditions, each supplying insight yet falling short of capturing how a dataset can be valued for AI-human teaming.

Descriptive ontologies. Early taxonomies framed data magnitude as a proxy for utility. These early works established the grammar for talking about big datasets. These ontologies excel at describing the characteristics of large corpora, but they remain silent on how much epistemic leverage a dataset actually delivers once modeling and curation costs are considered. These frameworks describe what big datasets look like, but they do not tell funders or reviewers how to compare a 50-year panel study with a petabyte-scale sensor log.

Normative principles and stewardship guides. In response to reproducibility crises, the FAIR guidelines codified findability and reusability, while curation scholarship documented the invisible labor behind trustworthy repositories (Wilkinson et al. 2016). Ethnographic accounts of repository work highlight the invisible craft that makes FAIR compliance possible, while critical data-studies literature reveals how governance arrangements shape who can benefit from a dataset (Leonelli 2014). Complementing these infrastructure-level guidelines, Bender and Friedman's "data statements" framework calls for publishing structured demographic and provenance metadata with every language corpus, making issues of exclusion and bias explicit and auditable (Bender and Friedman 2018). These principles tell us *how* data ought to be curated but not *how much* those efforts translate into scholarly salience relative to alternative corpora.

Domain-specific scoring rubrics. Applied communities have introduced quantitative checklists to guide dataset re-

lease and reuse. In biomedicine, the *Venus* score rates provenance and trustworthiness on a ten-point scale (Chicco, Fabris, and Jurman 2025). Data-ethics researchers have likewise proposed *Extended Data Briefs* and *Risk Labels* that pair numeric quality metrics with succinct warnings about potential harms (Rondina et al. 2025). Machine-learning researchers propose rubrics that separate syntactic from semantic validation steps (Bhardwaj et al. 2024; Gebru 2020; Zhao et al. 2024) and fairness scholars use structured surveys to benchmark demographic coverage and governance histories across widely used corpora (Fabris et al. 2022). Although these instruments advance curation within their individual niches, they stop short of furnishing a cross-domain framework for judging a dataset's overall scholarly value.

Gap. Descriptive ontologies catalog features of a dataset, governance frameworks codify how data should be handled, and field-specific rubrics can certify integrity and compliance. Yet, none of these approaches offers a cross-domain yardstick that converts such disparate signals into a single, interpretable measure of scholarly worth. *DVT* fills this gap by treating *time*, *breadth*, and *access* as orthogonal currencies of value to compare datasets across fields and to guide investment more strategically.

3 Operational Taxonomy

We translate the conceptual drivers of the *DVT* into a discipline-agnostic scoring scheme: Temporal investment (T), scale/breadth (S), and accessibility (A). Each construct is unpacked into a small set of observable *dimensions* where each dimension receives an *ordinal class label*. Those labels are purposefully generic so that subfields can recalibrate numeric break-points to field norms, regulatory regimes, and best practices without sacrificing cross-field compatibility. A corpus that took decades to assemble, spans multiple modalities, and was technically difficult to obtain *yet* is openly licensed should therefore record high scores on T, S, and A, whereas an infrequently scraped, single-modality, pay-walled dataset where data is mostly derivative of other work would not.

After dimensions are labeled, reviewers collapse them into a single score for each construct, yielding the vector:

$$\langle T,\, S,\, A \rangle$$

where higher values indicate greater scholarly salience along that axis. Many applications still benefit from a one-number headline, so we define the *Dataset Value Score*:

$$V = w_T\, T + w_S\, S + w_A\, A \quad : \quad w_T + w_S + w_A = 1$$

Equal weights ($w_T = w_S = w_A = \frac{1}{3}$) serve as a neutral default, but stakeholders are free to declare alternative weight vectors that reflect specific priorities.

This two-stage protocol, dimension labeling followed by transparent aggregation, turns the high-level constructs of the *DVT* into a reproducible practice. Subsequent subsections supply recommended dimensions and illustrative anchors for each construct, demonstrating how communities can adopt the framework while retaining the freedom to fine-tune thresholds and weightings.

4 Temporal Investment

Temporal Investment is the diachronic commitment required to bring a dataset into being. It is the sustained effort required to observe a phenomenon *through* time rather than at a single, *synchronic* moment. This timespan influences the reliability of findings, the practical and ethical costs of data collection, and the resources a study must have to sustain itself.

Longitudinal methodologists have long argued that repeated observation across extended intervals yields the only direct window into developmental, causal, and evolutionary processes (Curran and Bauer 2011). Classical panel studies follow individuals or systems over prolonged periods, often years or even decades (Caruana et al. 2015). This is in contrast to event-triggered collection, where events mobilize investigators at the instant of volcanic eruptions, market crashes, or cyber-incidents. In both cases, *time itself* becomes a material constituent of the data. Without waiting for the next election cycle, hurricane season, or birth cohort, the phenomenon under study simply does not exist in observable form. Temporal investment therefore functions as an epistemic *ante*, underwriting claims of causality, trend direction, and external validity.

4.1 Measurement Dimensions and Calibration

Table 1 formalizes these conceptual ideas by mapping Temporal Investment onto four measurable dimensions. Each dimension is paired with a set of illustrative class labels and objective indicators so that reviewers can translate narrative timelines into a reproducible coding scheme. For example, the distinction between *continuous* and *event-triggered* sampling is operationalized via the "mean inter-sample interval" or an explicit trigger rule, while the practical value of a corpus for real-time decision support is captured by the "latency-to-insight" metric. Because the table fixes only the *ordinality* of the class labels, disciplines remain free to substitute their own numeric cut-offs while maintaining the core construct.

Interactions. A dataset may be longitudinal *and* continuous (e.g., seismic waveform archives), or event-triggered yet instantaneous (e.g., a snapshot social media scrape after an earthquake). Combinatorial coding preserves nuance while enabling ordinal comparison.

The taxonomy's elasticity is illustrated in Table 2, which shows how five research areas can recalibrate to the same ladder. Astronomy treats a three-year sky survey as "longitudinal," whereas public-health epidemiology reserves that label for cohorts that span at least five years. Climate science extends the threshold to the 30-year climatology window recommended by the World Meteorological Organization, and digital-humanities scholars often view a decade of social-media records as longitudinal, but this can vary based on what specifically is being measured (Garcia, Yang, and Miceli 2025). Despite these re-scalings, the ordinal ordering of classes is preserved and can therefore be compared across fields.

These calibrations highlight two meta-insights:

- **Threshold elasticity:** what counts as longitudinal scales with the natural tempo of the phenomenon.
- **Cycle salience:** disciplines often privilege specific exogenous rhythms (e.g., El Niño-Southern Oscillation (ENSO) vs. sidereal day), which in turn dictate sampling design.

4.2 Ethics and Governance

Temporal commitment amplifies **participant vulnerability, data drift, and stewardship burden.** Longitudinal human studies must renew consent (Mascalzoni et al. 2022), mitigate attrition bias (Okpara et al. 2023), and safeguard participants' privacy as privacy norms evolve. Continuous environmental sensing raises concerns over ecological disturbance (Sethi et al. 2022) and computer science's use of AI models driving data center energy demand (Ewim et al. 2023). Event-triggered collection in crisis zones demands reflexive ethics to avoid "disaster exploitation." (Mezinska et al. 2016) The stewardship burden of continuous sensing and petabyte-scale storage raises both ecological costs and organizational liabilities (secure retention, governance, deletion). Temporal Investment should be weighed not merely as epistemic capital, but also as a vector of moral debt and a sustainability obligation. Guidelines such as staged ethics reviews and sunset clauses for study retention can help align temporal ambition with responsible practice.

Deep-time datasets can entrench a monoculture of scientific knowledge, perpetuating dominant paradigms and crowding out alternative epistemologies. To prevent this epistemic drift, longitudinal projects should undergo periodic reflexive audits that ask whose futures are being modeled and whose knowledge systems or approaches to data collection are being sidelined. Findings from these audits can trigger methodological pluralism, such as pairing quantitative panels with community-led qualitative studies, or prompt data-sovereignty negotiations that return agency to stakeholders. Viewed through this lens, Temporal Investment requires more than procedural ethics; it demands substantive equity in how epistemic rewards and societal risks unfold across the data collection time horizon.

4.3 Operational Quantification

To incorporate Temporal Investment into the composite significance score, each dimension is first mapped to a unit-interval tier. Table 3 shows the mapping for *Elapsed Duration*; analogous 0–1 scales are defined for Sampling Frequency (T_2), Cycle Dependence (T_3), and Latency (T_4).

A weighted geometric mean then yields an overall tempo-

Dimension	Class Examples	Core Indicators	Rationale
Elapsed Duration	Instantaneous, Short (hrs–days), Mid (wks–mos), Long (≥ 1 yr)	Start–freeze calendar time	Captures diachronic scope of observation
Sampling Frequency	Continuous, Periodic, Event-triggered	Mean inter-sample interval; trigger logic	Distinguishes trajectories from single snapshots
External Cycle Dependence	Natural, Social, Rare/contingent	Flag (if cycle dependent) + cycle period	Signals synchronization between data rhythm and exogenous processes
Latency to Insight	Real-time ($<$1 s), Near-real-time ($<$24 h), Batch (days–mos)	Capture-to-insight delay	Determines practicality for time-sensitive decision-making

Table 1: Measurement dimensions for Temporal Investment. Numerical cut-offs are illustrative; each discipline can tune thresholds to its own temporal scales.

Discipline	Instantaneous Example	Longitudinal Threshold	Dominant Cycle	Illustrative Dataset
Astronomy	Single-exposure FITS image	$\geq$ 3-yr	Sidereal	Gaia DR3 time-series (Vallenari et al. 2023)
Public-Health Epi.	Daily incident-case CSV	$\geq$ 5-yr	Seasonality	Framingham Study (Dawber, Meadors, and Moore Jr 1951)
Climate Science	Hourly re-analysis grid	$\geq$ 30-yr	ENSO ($\sim$4 yr)	ERA5 climate normals (Hersbach et al. 2020)
Behavioral Econ.	Millisecond limit-order book	$\geq$ 1-yr	Market regimes	LOBSTER dataset (Huang and Polak 2011)
Digital Humanities	Timestamped tweet	$\geq$ 10-yr	Election (2–4 yr)	U.S. Presidential Twitter Archive (Zimmer 2015)

Table 2: Sample recalibrations of Temporal Investment across five research areas. Thresholds shift, but the ordinal ladder (instantaneous $\rightarrow$ longitudinal) stays intact.

ral contribution where the weights α_i allow assessors to emphasize, for example, sampling cadence over latency when justified by disciplinary norms. The normalized score T is then entered in the dataset-significance function:

$$T = \Big(\prod_{i=1}^{4} T_i^{\alpha_i}\Big)^{1/\sum_{i=1}^{4}\alpha_i} \tag{1}$$

Publishing the raw T_i values alongside narrative justifications promotes transparency and supports future meta-analyses across corpora.

5 Scale and Breadth

Scale captures the *quantitative footprint* of a corpus, whereas breadth widens the lens to include *variety* (number of modalities) and *coverage* (geographic, demographic, or taxonomic reach). Together, they determine the "state space" a dataset renders observable and thus the generalizability of inferences drawn from it. A social-media sample of ten million tweets (*macro* scale) but drawn from a single city (*local* coverage) supports different claims than a smaller, multi-modal cohort spanning five regions. Table 4 formalizes these distinctions by breaking Scale/Breadth into three measurable dimensions.

Disciplinary calibration. Numeric breakpoints for "micro," "meso," and "macro" differ across research cultures; astronomy may reach the macro tier only above 50 TB, whereas ecology may cross the same threshold at 1 TB. This, of course, is dependent on the type of data being collected, such as tabular measurements vs images or audio samples. Even so, the ordinal ladder (Micro $\rightarrow$ Macro, Uni $\rightarrow$ Poly, Local $\rightarrow$ Global) remains stable, which lets reviewers compare datasets that have been rescaled. Table 5 illustrates how five disciplines reinterpret the three dimensions while still mapping them onto the common tier structure.

5.1 Ethics and Governance

We retain the same ethical dimensions as before but highlight how the mechanisms differ from temporal depth. Treating Scale/Breadth as a source of linkage risk, distributional opacity, and maintenance debt, reframes it as an ongoing governance challenge rather than a one-time engineering hurdle.

Participant vulnerability. The principal threat is the "mosaic effect" where joining even lightly de-identified records across modalities can reconstruct personal profiles with surprising accuracy (Bellovin et al. 2014). Conventional anonymization does not account for this cross-modal linkage risk. Governance should therefore move toward *fusion-aware access controls* that limit which modality combinations authorized users can query, and it should require provenance tags that record every linkage operation for later audit (Ganta 2009).

Data drift. When coverage balloons, latent heterogeneity can obscure signal and induce spurious associations. A model trained on a global corpus may appear robust yet fail when deployed in a sub-region whose patterns are underrepresented. This is less an issue of fairness than of epistemic reliability where researchers can *mistake breadth for representativeness*. Continuous drift audits (statistical tests that compare new data slices against the training distribution) should be built into dataset maintenance cycles, with thresholds that trigger re-annotation or re-sampling rather than automatic model retraining (Ackerman et al. 2021).

Stewardship burden. At macro volume and poly-modal variety, technical debt becomes a dominant risk. Every schema change, file-format migration, or metadata update can cascade through petabytes of dependent artifacts. Without rigorous versioning, researchers may cite analyses that no longer correspond to the underlying bits. Immutable, hash-addressed dataset versions and machine-readable changelogs can mitigate this burden by offering a cryptographically verifiable audit trail. Grant agencies and journals could condition funding or publication on documented lifecycle plans that detail curation workflows, governance roles, and end-of-life criteria.

5.2 Operational Quantification

Each dimension is mapped to a 0–1 tier shown in Table 6. The overall Scale/Breadth score is then computed as a weighted geometric mean:

$$S = \Big(\prod_{j=1}^{3} S_j^{\beta_j}\Big)^{1/\sum_{j=1}^{3}\beta_j} \tag{2}$$

where S_j is the score for Volume, Variety, or Coverage, and β_j allows disciplines to weight dimensions differently.

6 Accessibility

Accessibility captures the socio-technical barriers that govern who can obtain, inspect, or reuse a dataset. Whereas Temporal Investment and Scale/Breadth look inward to the effort and scope of data creation, Accessibility looks outward to the *governance perimeter* that determines circulation. Barriers arise along several axes: legal or ethical restrictions, technical infrastructure requirements, financial cost, and physical or logistical hurdles, including dependence on exquisite platforms such as orbital satellites or deep-sea submersibles. Table 7 decomposes these barriers into measurable dimensions.

Disciplinary calibration. Numeric cut-offs and dominant barriers vary by field, but we can create an ordinal ladder from open to classified. Table 8 shows representative recalibrations.

Elapsed Duration Tier	**Score** T_1
Instantaneous	0.25
Short-Term	0.50
Mid-Term	0.75
Longitudinal	1.00

Table 3: Numeric mapping for the *Elapsed Duration* dimension.

Dimension	Class Examples	Primary Indicator(s)	Rationale
Volume (Scale)	*Micro* ($< 10^3$ records or < 1 GB), *Meso* (10^3–10^7 or 1 GB–1 TB), *Macro* ($> 10^7$ or > 1 TB)	Row/observation count; byte size	Captures quantitative capacity for statistical power and model complexity
Variety (Breadth)	Uni-modal, Bi/tri-modal, Poly-modal (≥ 4 types)	Modality count; Shannon diversity index	Assesses multi-faceted representation of the target phenomenon
Coverage (Breadth)	Local, Regional / Domain-specific, Global / Multi-species	# distinct sites, regions, species, or policy domains	Signals external validity and cross-context transferability

Table 4: Measurement dimensions for the Scale/Breadth construct. Thresholds are illustrative; each field should tune them to its own data regimes. Volume, Variety, and Coverage are the three dimensions that together form the S score.

6.1 Ethics and Governance

Accessibility raises normative questions that differ from those associated with temporal depth or scale. Choices about licenses, paywalls, and specialized collection platforms are not merely administrative. **They determine who may convert raw observations into accepted knowledge, who must rely on derivative products, and who remains excluded.** Because these gatekeeping arrangements shape the distribution of epistemic authority, an ethical analysis must extend beyond individual privacy to examine systemic consequences of control, including inequities in analytic power, obstacles to independent verification, and the long-term stewardship obligations that accompany restrictive dissemination regimes.

Data and participant vulnerability. Decisions about who can access a corpus set the boundary between constructive reuse and predatory exploitation. When barriers are too low, an attacker can combine seemingly innocuous fields, like satellite imagery with shipping-log metadata to infer proprietary business operations or geopolitical movements. When barriers are too high, the same dataset may become a club good where a small circle of institutions captures the analytic surplus, leaving others to rediscover the wheel under inferior conditions. Ethics therefore demands a *fusion-aware* access model that calibrates permissions to the incremental disclosure risk created by linking with other public sources, and that pairs each access tier with audit logs so misuse can be traced retroactively.

Knowledge drift and epistemic stasis. Restricted datasets often age in the dark. Locking updates inside a restricted silo creates an uneven playing field in which the gatekeepers advance with refreshed evidence while everyone else is forced to extrapolate from a stale snapshot. This asymmetry does more than distort individual studies; it throttles collective progress by steering entire research streams down out-of-date paths. The ethical dilemma is clear, privileged access generates private advantage at the cost of public error. Mitigation requires mechanisms that surface change without compromising sensitive details, such as mandatory changelogs, cryptographically signed release

Discipline	Macro Threshold (Volume)	Poly-modal Definition	Global Coverage Proxy	Illustrative Corpus
Genomics	$>$100 TB raw FASTQ	DNA + RNA + epigenome + clinical	Multi-ethnic biobank	UK Biobank WGS release (Li et al. 2023)
Remote Sensing	$>$50 TB Cloud-Optimised GeoTIFF	Multispectral + SAR + LiDAR + AIS	Planetary landmass	NASA Harmonised Landsat/Sentinel (Claverie et al. 2018)
Social Media Analytics	$>$10 TB JSON	Text + image + network graph + metadata	$>$50 countries	Archive of Tweets 2015–2025 (Fafalios et al. 2018)
Neuroscience	$>$5 TB TIFF stacks	fMRI + EEG + behavioural logs	Multi-site consortium	Human Connectome Project (Van Essen et al. 2013)
Ecology	$>$1 TB camera-trap JPEG	RGB image + acoustic + microclimate + GPS	$\geq$3 biomes	Snapshot Serengeti extended (Swanson et al. 2015)

Table 5: Example field-specific calibrations for Scale/Breadth. Numeric cut-offs are illustrative; each community should publish its own rubric.

Dimension	Class	Score	Tier Range
Volume	Micro	0.25	$< 10^3$ records / $<$ 1 GB
	Meso	0.50	10^3–10^7 / 1 GB–1 TB
	Macro	1.00	$> 10^7$ / $>$ 1 TB
Variety	Uni-modal	0.25	Single data type
	Bi/Tri-modal	0.60	Two or three modalities
	Poly-modal	1.00	Four or more modalities
Coverage	Local	0.25	Sub-national / single site
	Regional	0.60	Multi-site or domain-specific
	Global	1.00	Cross-national / multi-species

Table 6: Numeric tiers for Scale/Breadth dimensions. Scores are normalised to the $[0, 1]$ interval; each field may adjust the numeric cut-offs while preserving ordinal order.

notes, or de-biased summary "view" layers that reveal distributional shifts while withholding raw records.

Stewardship burden. High access barriers have a double edge. On one hand, classified or NDA-protected datasets impose costly duties, including security audits, credential rotation, and declassification reviews, which can divert resources from future data collection. On the other hand, the barriers often exist because the collector assumed the financial and logistical risk of building the corpus and reasonably expects a period of privileged use. An ethically balanced governance model can honor both concerns by embedding explicit *access half-lives*: a defined window during which the data owner enjoys exclusive or premium access sufficient to recover investment, followed by an automatic step-down to a less restrictive tier unless a new harm assessment justifies an extension. This sunset approach recognizes the right to first benefit while preventing permanent lock-in, keeping accessibility strict enough to deter misuse today yet flexible enough to advance broader scientific goals tomorrow.

6.2 Operational Quantification

Accessibility begins with a legal/access ladder, scored as outlined in Table 9. Optional modifiers can adjust this base score upward by some amount for significant technical, financial, or physical barriers, capped at 1.00. The overall Accessibility contribution is:

$$A = \min(A_1 + \gamma_{\text{tech}} + \gamma_{\text{fin}} + \gamma_{\text{phys}},\ 1.00) \tag{3}$$

Table 9 assigns higher numerical values to more restrictive tiers, which can initially appear counter-intuitive. Within our framework, the Accessibility score does not reward openness in itself; instead, it serves as a proxy for the effort required to obtain primary evidence. Scarcer, tightly controlled datasets usually demand greater upfront investment, higher compliance overhead, or specialized infrastructure, attributes that make them harder to replicate and thus more "scarce" in comparative valuation. Conversely, derivatives created solely from public corpora encounter a lower access hurdle and therefore receive a lower A value, signaling that novelty or contribution must be demonstrated through other dimensions such as temporal depth or scale.

The ladder can readily be flipped when the assessment shifts from valuing inputs (how hard was the data to acquire?) to valuing outputs (how widely will the results circulate?). A funding organization or publication venue that prioritizes public benefit could rescale the tiers so that open datasets map to $A = 1.00$ and classified data to $A = 0.25$. Because the aggregation formula in Equation (3) is monotonic, either orientation preserves ordinal rankings while making the policy choice transparent.

7 Financial Resources as an Enabler

Money is not a fourth construct of *DVT*, but a fungible input that can elevate a dataset's standing along any of the three existing axes. Funding can lengthen or accelerate temporal investment by underwriting additional collection periods or by enabling higher-frequency sampling through improved equipment, such as a high-shutter-speed camera capable of capturing the biomechanics of hummingbird wings. It can expand scale and breadth through the deployment of more sensors, higher-resolution instruments, or broader sampling frames. It can also reduce access barriers by purchasing proprietary licenses, commissioning specialized collection platforms, or subsidizing the release of raw data that would otherwise remain inaccessible due to cost. However, money cannot change the pace of natural phenomena. For example, it cannot shorten the interval between El Niño events, and it cannot (and should not) override ethical or regulatory constraints on data sharing.

7.1 From Cost Accounting to Value Accounting

Traditional budget narratives focus on inputs such as staff hours, equipment costs, and travel. A *dataset-value perspective* turns that logic around. Each dollar is treated as a wager on raising one or more *DVT* dimensions. The practical test is simple: does the line item push a dimension across a tier boundary, and by how much? A modest purchase that moves latency from batch to near-real-time may be worth more than a large outlay of funds that leaves all dimensions unchanged.

Budgeting for Dataset Value. Principal investigators can operationalize value accounting by linking every expenditure to a dimension and stating its anticipated tier change. Table 10 shows illustrative mappings that reviewers can audit at proposal time and revisit at project close-out.

7.2 Incremental Uplift Ratio

To compare projects of different scale, we introduce the *Incremental Uplift Ratio*:

$$\text{IUR} = \frac{\mathbb{E}[V_{\text{post}}] - V_{\text{base}}}{\text{Budget}} \tag{4}$$

where V_{base} is the composite score that the project would achieve with zero additional funding and $\mathbb{E}[V_{\text{post}}]$ is the score expected after the requested investment.

To make $\mathbb{E}[V_{\text{post}}]$ concrete, proposals must show exactly how each budget item alters the underlying tiers and with

Dimension	Class Ladder	Primary Indicator(s)	Rationale
Legal and Ethical Access	Open public; Licensed or attribution; Controlled (IRB or DAC); Proprietary or NDA; Restricted or classified	License type; data-use agreement; security clearance	Defines baseline eligibility and compliance obligations
Technical Barrier	Downloadable; Bulky (specialized storage); High-compute (GPU cluster); Special hardware or secure enclave	Storage footprint; compute hours; required hardware	Determines practical feasibility of replication
Financial Barrier	Free; Low cost ($\sim$ \$1k USD); High cost (license or instrumentation)	Acquisition fee; subscription rate	Allocates access based on resources rather than expertise
Physical or Logistical Barrier	Online; On-site request; Remote hazardous fieldwork; Specialized platform (satellite tasking, deep-sea submersible)	Travel distance; safety or mission rating; platform availability and cost per deployment	Adds non-digital friction and high capital requirements that limit who can gather or verify data

Table 7: Measurement dimensions for Accessibility.

what likelihood. For every planned expenditure, the applicant lists (i) the specific tier upgrade it enables, (ii) the evidence supporting that upgrade, such as pilot measurements, vendor specifications, prior publications, or signed data-sharing agreements, and (iii) an explicit probability that the upgrade will succeed within the project timeline. If an extra radar satellite pass has a 0.8 chance of moving spatial coverage from regional to global, its expected gain is $0.8 \times$ tier change; if a new sensor array is guaranteed to cut latency from batch to near-real time, its gain is the full tier change. Summing these probabilistic contributions across all line items yields $\mathbb{E}[V_{\text{post}}]$.

Because the denominator is monetary, the ratio expresses how efficiently a proposal converts dollars into dataset value. Reviewers can rank proposals by IUR within thematic panels, while funding agencies can publish historical IUR benchmarks to guide applicants.

7.3 Equity and Sustainability Safeguards

Financial leverage is double-edged. A well-funded team may lock in tier superiority via prolonged exclusivity, delaying community access and skewing knowledge production. We therefore advocate three safeguards:

1. **Sunset clauses** that trigger automatic license relaxation after a predetermined interval unless a renewed harm assessment justifies extension;
2. **Scaled co-funding** incentives that match a portion of private or philanthropic contributions with public grants when the resulting data will transition to an open tier;
3. **Uplift disclosure** statements in which awardees report realized tier shifts and IUR values, creating a public ledger of cost-to-value performance across projects.

These mechanisms preserve first-mover returns while preventing permanent lock-in and encouraging transparent stewardship.

7.4 A Checklist for Value-Focused Budgets

For ease of adoption, the following pre-submission checklist can guide both applicants and reviewers:

1. Identify the base *DVT* tiers for each dimension.
2. Link every cost line to a single dimension or justify split effects.
3. State the quantitative tier change expected per line.
4. Compute pre and post-investment composite scores (V_{base} and $\mathbb{E}[V_{\text{post}}]$).
5. Report the incremental uplift ratio (IUR).
6. Describe equity safeguards (sunset date and access-tier roadmap).

7.5 Summary

Money acts as a catalyst, not an end in itself. Treating expenditures as explicit bets on *DVT* tier advancement sharpens trade-offs, aligns resource allocation with scholarly payoff, and keeps attention on the dataset qualities that underpin durable knowledge. By shifting the conversation from "How much does it cost?" to "How much value does it buy?" researchers and funders can make more transparent, accountable, and equitable decisions about the data that drive empirical discovery.

8 Stakeholder Guidance and Community Validation of Weighted *DVT* Scores

The $\langle T, S, A \rangle$ vector and composite score $V = w_T T + w_S S + w_A A$ only become useful when communities translate those numbers into concrete action. This section offers **role-specific playbooks** as well as a **benchmarking protocol** that allows empirical validation and iterative refinement of the *Dataset Value Taxonomy*.

How to Read a Weighted Score. Weights encode priorities, and publishing them *up-front* makes trade-offs explicit. A climate program that values long-horizon observation sets

Discipline	Highest Common Barrier	Typical Technical Hurdle	Cost Profile	Illustrative Dataset
Biomedicine	Controlled (HIPAA)	Secure enclave; de-identification toolkit	License fee per user	NIH dbGaP genomic panel (Mailman et al. 2007)
Astrophysics	Open public	Petabyte archive; HPC decoding	Free download	LOFAR sky survey (Shimwell et al. 2017)
Finance	Proprietary NDA	Real-time feed; low-latency API	High subscription	Consolidated limit-order book (Ntakaris et al. 2018)
Geoscience	Licensed	Bulky netCDF; remote servers	Low cost ($)	Global Digital Elevation Model (Abrams, Crippen, and Fujisada 2020)
National security	Restricted or classified	Air-gapped network	No public price	Signals intelligence corpus (Kris 2021)

Table 8: Field-specific recalibrations for Accessibility. Each community foregrounds a different primary barrier while mapping onto the same ordinal ladder.

Legal/Ethical Tier	Score A_1
Open public	0.25
Licensed or attribution	0.50
Controlled (IRB or DAC)	0.75
Proprietary or NDA	0.90
Restricted or classified	1.00

Table 9: Base scores for legal and ethical access.

$w_T = 0.4$, whereas a high-frequency finance lab might push $w_S = 0.5$. Every stakeholder should therefore (i) publish its weight vector, (ii) justify it in a sentence or two, and (iii) revisit it on a fixed cadence to reflect shifts in mission or regulation.

8.1 Funding Agencies

Funding bodies should begin by publishing their weight vector (w_T, w_S, w_A) in calls for proposals so that applicants understand the priorities guiding award decisions. Each submission can include a concise *dataset value budget* that maps every budget line to an intended tier shift in at least one dimension of the *DVT* based on prior work in the field and the research questions being investigated. Review panels should then score proposals on marginal gain per dollar, assigning the lowest ranks to projects that deliver no tier improvement. Finally, award payments ought to be tied to tier-based milestones; funds are released only after investigators document that the promised tier upgrades have been achieved.

8.2 Program Committees and Journal Editors

Embedding *DVT* scores in author checklists makes dataset value visible alongside code availability and ethics disclosures. Editors can adopt threshold rules. For instance, papers relying on corpora with $A<0.50$ and no mitigation plan are filtered early, while submissions that introduce a new dataset are expected to deliver a composite V at least 0.10 higher than the closest public alternative. Such policies streamline peer review and signal that dataset stewardship is a first-class scholarly contribution.

8.3 Researchers and Laboratories

Investigators can treat the weighted vector as a decision compass. During planning, they can simulate alternative sampling schemes, licensing arrangements, or collection timelines to maximize V under budget and ethics constraints. During collection, interim audits reveal which dimension lags (for example, if coverage sits at $S{=}0.55$ against a target of $S{=}0.75$, additional sites become the priority). Publishing the full $\langle T, S, A\rangle$ tuple with a brief rationale invites downstream users to re-weight according to their own objectives, improving transparency and reuse.

8.4 Community Benchmarking and Empirical Validation

Panels of domain experts can build confidence in the taxonomy for their use-case by rating datasets used in a broad set of research in their field. Reviewers should focus on the three factors captured by *DVT*: required time investment, breadth or scale, and difficulty of access. After ratings are complete, these significance scores can be statistically compared with venue-level metrics such as impact factor, citation counts, and altmetrics. Alignment between high-rated datasets and influential publications validates that *DVT* captures properties that matter in scholarly practice, while mismatches reveal where weighting schemes or tier thresholds need adjustment.

8.5 Maintaining Flexibility and Accountability

Stakeholders should revisit weights and tier thresholds on a regular schedule. A public change log explaining why a

Budget Line	*DVT* Dimension Affected	Expected Tier Movement
Hire local enumerators for a second survey wave	Elapsed duration	Short-term → Mid-term
Rent high-bandwidth downlink at observatory	Latency to insight	Batch → Near-real-time
Purchase PlanetScope imagery	Volume and coverage	Meso → Macro, Regional → Global
Negotiate license for industry logs	Legal / ethical	Proprietary → Licensed

Table 10: Illustrative mapping of budget allocations to anticipated tier upgrades on *DVT* dimensions.

weight vector changes (for example, stronger privacy laws increasing the salience of accessibility) maintains comparability over time. Reporting both raw dimension scores and the composite V keeps the aggregation transparent and lets others re-weight to suit their own principles.

By blending actionable guidance with a pathway for empirical benchmarking, this section positions *DVT* as both a practical governance tool and a living framework that evolves alongside evidence and community feedback.

Conclusion

Artificial intelligence continues to lower the marginal cost of analysis, coding, and visualization, making the intrinsic value of data itself increasingly decisive for scientific and societal progress. This paper introduced the *Dataset Value Taxonomy* (*DVT*), a three-construct framework that captures the temporal investment, observational scale, and accessibility constraints that determine a dataset's contribution. We provided an ordinal tier structure, a calibrated weighting scheme, and a composite score that together enable consistent comparison across disciplines. Beyond theory, we outlined actionable guidance for funders, evaluators, and researchers, showing how weighted *DVT* scores can inform grant allocation, peer review, and project planning. We also proposed a community benchmarking protocol that invites domain experts to validate the taxonomy against real-world impact metrics, ensuring that the framework stays empirically grounded and flexible when tailored to any particular field. By offering both a principled vocabulary and pragmatic tools, *DVT* positions data stewardship as a first-class scholarly activity and equips stakeholders to make transparent, accountable decisions in the era of abundant AI-powered research and analytics.

References

Abrams, M.; Crippen, R.; and Fujisada, H. 2020. ASTER global digital elevation model (GDEM) and ASTER global water body dataset (ASTWBD). *Remote Sensing*, 12(7): 1156.

Ackerman, S.; Raz, O.; Zalmanovici, M.; and Zlotnick, A. 2021. Automatically detecting data drift in machine learning classifiers. *arXiv preprint arXiv:2111.05672*.

Bellovin, S.; Hutchins, R. M.; Jebara, T.; and Zimmeck, S. 2014. When Enough is Enough: Location Tracking, Mosaic Theory, and Machine Learning.

Bender, E. M.; and Friedman, B. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*.

Bhardwaj, E.; Gujral, H.; Wu, S.; Zogheib, C.; Maharaj, T.; and Becker, C. 2024. The State of Data Curation at NeurIPS: An Assessment of Dataset Development Practices in the Datasets and Benchmarks Track. *arXiv.org*.

Boyd, D.; and Crawford, K. 2012. CRITICAL QUESTIONS FOR BIG DATA.

Caruana, E. J.; Roman, M.; Hernández-Sánchez, J.; and Solli, P. 2015. Longitudinal studies. *Journal of thoracic disease*, 7(11): E537.

Chicco, D.; Fabris, A.; and Jurman, G. 2025. The Venus score for the assessment of the quality and trustworthiness of biomedical datasets. *BioData Mining*.

Claverie, M.; Ju, J.; Masek, J. G.; Dungan, J. L.; Vermote, E. F.; Roger, J.-C.; Skakun, S. V.; and Justice, C. 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote sensing of environment*, 219: 145–161.

Curran, P. J.; and Bauer, D. J. 2011. The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual review of psychology*, 62(1): 583–619.

Dawber, T. R.; Meadors, G. F.; and Moore Jr, F. E. 1951. Epidemiological approaches to heart disease: the Framingham Study. *American Journal of Public Health and the Nations Health*, 41(3): 279–286.

Ewim, D.; Ninduwezuor-Ehiobu, N.; Orikpete, O. F.; Egbokhaebho, B. A.; Fawole, A. A.; and Onunka, C. 2023. Impact of Data Centers on Climate Change: A Review of Energy Efficient Strategies. *The Journal of Engineering and Exact Sciences*.

Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Algorithmic fairness datasets: the story so far. *Data mining and knowledge discovery*.

Fafalios, P.; Iosifidis, V.; Ntoutsi, E.; and Dietze, S. 2018. Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *European Semantic Web Conference*, 177–190. Springer.

Ganta, S. R. 2009. Fusion-Aware Privacy and Warehousing for Healthcare Databases.

Garcia, A. A.; Yang, T.; and Miceli, M. 2025. What Knowledge Do We Produce from Social Media Data and How? *Proceedings of the ACM on Human-Computer Interaction*.

Gebru, T. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. *Knowledge Discovery and Data Mining*.

Goldblum, M.; Anandkumar, A.; Baraniuk, R.; Goldstein, T.; Cho, K.; Lipton, Z. C.; Mitchell, M.; Nakkiran, P.; Welling, M.; and Wilson, A. G. 2023. Perspectives on the State and Future of Deep Learning-2023. *arXiv preprint arXiv:2312.09323*.

Halevy, A.; Norvig, P.; and Pereira, F. C. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*.

Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. 2020. The ERA5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730): 1999–2049.

Huang, R.; and Polak, T. 2011. Lobster: Limit order book reconstruction system. *Available at SSRN 1977207*.

Jones, M. R. 2019. What we talk about when we talk about (big) data. *Journal of strategic information systems*.

Kitchin, R.; and McArdle, G. 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets:. *Big Data & Society*.

Kris, D. S. 2021. The NSA's new sigint annex. *Journal of National Security Law & Policy*.

Leonelli, S. 2014. What difference does quantity make? On the epistemology of Big Data in biology. *Big data & society*.

Li, S.; Carss, K. J.; Halldorsson, B. V.; Cortes, A.; and Consortium, U. B. W.-G. S. 2023. Whole-genome sequencing of half-a-million UK Biobank participants. *medRxiv*, 2023–12.

Mailman, M. D.; Feolo, M.; Jin, Y.; Kimura, M.; Tryka, K.; Bagoutdinov, R.; Hao, L.; Kiang, A.; Paschall, J.; Phan, L.; et al. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics*, 39(10): 1181–1186.

Mascalzoni, D.; Melotti, R.; Pattaro, C.; Pramstaller, P. P.; Gögele, M.; Grandi, A. D.; Biasiotto, R.; Mascalzoni, D.; Melotti, R.; Pattaro, C.; Pramstaller, P. P.; Gögele, M.; Grandi, A. D.; and Biasiotto, R. 2022. Ten years of dynamic consent in the CHRIS study: informed consent as a dynamic process. *European Journal of Human Genetics*.

Maslej, N.; Fattorini, L.; Perrault, R.; Gil, Y.; Parli, V.; Kariuki, N.; Capstick, E.; Reuel, A.; Brynjolfsson, E.; Etchemendy, J.; Ligett, K.; Lyons, T.; Manyika, J.; Niebles, J. C.; Shoham, Y.; Wald, R.; Walsh, T.; Hamrah, A.; Santarlasci, L.; Betts Lotufo, J.; Rome, A.; Shi, A.; and Oak, S. 2025. The AI Index 2025 Annual Report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA. Accessed 9 May 2025.

Mezinska, S.; Kakuk, P.; Mijaljica, G.; Waligóra, M.; and O'Mathúna, D. P. 2016. Research in disaster settings: a systematic qualitative review of ethical guidelines. *BMC Medical Ethics*.

Ntakaris, A.; Magris, M.; Kanniainen, J.; Gabbouj, M.; and Iosifidis, A. 2018. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*, 37(8): 852–866.

Okpara, C.; Adachi, J.; Papaioannou, A.; Ioannidis, G.; and Thabane, L. 2023. Exploring participant attrition in a longitudinal follow-up of older adults: the Global Longitudinal Study of Osteoporosis in Women (GLOW) Hamilton cohort. *BMJ Open*.

Rondina, M.; Vetrò, A.; Fabris, A.; Silvello, G.; Susto, G. A.; Torchiano, M.; and Martin, J. C. D. 2025. Experience: Bridging Data Measurement and Ethical Challenges with Extended Data Briefs. *Journal of Data and Information Quality*.

Roscher, R.; Rußwurm, M.; Gevaert, C.; Kampffmeyer, M.; Santos, J. A. d.; Vakalopoulou, M.; Hänsch, R.; Hansen, S.; Nogueira, K.; Prexl, J.; and Tuia, D. 2024. Better, Not Just More: Data-centric machine learning for Earth observation. *IEEE Geoscience and Remote Sensing Magazine*.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P. K.; and Aroyo, L. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *International Conference on Human Factors in Computing Systems*.

Scheuerman, M. K.; Hanna, A.; and Denton, E. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on human-computer interaction*.

Sethi, S. S.; Kovač, M.; Wiesemüller, F.; Miriyev, A.; and Boutry, C. 2022. Biodegradable sensors are ready to transform autonomous ecological monitoring. *Nature Ecology & Evolution*.

Shimwell, T.; Röttgering, H.; Best, P. N.; Williams, W.; Dijkema, T.; De Gasperin, F.; Hardcastle, M.; Heald, G.; Hoang, D.; Horneffer, A.; et al. 2017. The LOFAR Two-metre Sky Survey-I. Survey description and preliminary data release. *Astronomy & Astrophysics*, 598: A104.

Swanson, A.; Kosmala, M.; Lintott, C.; Simpson, R.; Smith, A.; and Packer, C. 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific data*, 2(1): 1–14.

Thomer, A. K.; Akmon, D.; York, J. J.; Tyler, A. R. B.; Polasek, F.; Lafia, S.; Hemphill, L.; Yakel, E.; Thomer, A. K.; Akmon, D.; York, J. J.; Tyler, A. R. B.; Polasek, F.; Lafia, S.; Hemphill, L.; and Yakel, E. 2022. The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. *Proceedings of the ACM on human-computer interaction*.

Vallenari, A.; Brown, A. G.; Prusti, T.; De Bruijne, J. H.; Arenou, F.; Babusiaux, C.; Biermann, M.; Creevey, O. L.; Ducourant, C.; Evans, D. W.; et al. 2023. Gaia data release 3-summary of the content and survey properties. *Astronomy & Astrophysics*, 674: A1.

Van Essen, D. C.; Smith, S. M.; Barch, D. M.; Behrens, T. E.; Yacoub, E.; Ugurbil, K.; Consortium, W.-M. H.; et al. 2013. The WU-Minn human connectome project: an overview. *Neuroimage*, 80: 62–79.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.; Santos, L. O. B. d. S.; Bourne, P.; Bouwman, J.; Brookes, A.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C.; Finkers, R.; González-Beltrán, A. N.; Gray, A.; Groth, P.; Goble, C.; Grethe, J.; Heringa, J.; Hoen, P.;

Hooft, R.; Kuhn, T.; Kok, R. G.; Kok, J.; Lusher, S.; Martone, M.; Mons, A.; Packer, A.; Persson, B.; Rocca-Serra, P.; Roos, M.; Schaik, R. C. v.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G. O.; Swertz, M.; Thompson, M.; Lei, J. v. d.; Mulligen, E. V. v.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; and Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*.

Zhao, D.; Scheuerman, M.; Chitre, P.; Andrews, J. T. A.; Panagiotidou, G.; Walker, S.; Pine, K. H.; and Xiang, A. 2024. A Taxonomy of Challenges to Curating Fair Datasets. *arXiv.org*.

Zimmer, M. 2015. The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. *First Monday*.