Genefold Data Governance: A Semantic Governance Platform for AI Safety, Drift Monitoring, and Federated Data Asset Management

Lorenzo Moriondo, Tommaso Moriondo

Abstract

AI governance increasingly depends on an organization's ability to manage data assets across internal departments and external partner organizations without losing control over identity, access, provenance, or accountability. This extended abstract presents Genefold Data Governance, a platform that combines Solid-based federated authentication and authorization, ArrowSpace-powered semantic retrieval, AI safety and drift monitoring, and standards-based dataset publishing. The core claim is that AI governance should operate directly on governed data assets and their representation spaces, while remaining enforceable across companies and among departments inside the same company. In this design, Solid provides the trust and access-control layer, ArrowSpace provides the semantic and topology-aware evidence layer, and the surrounding governance stack supports auditability, controlled reuse and publication. This has been collected into a platform to provide governance of datasets to potential interested parties like research institutions, research-intensive companies, public institutions or any organization that has data assets managements as relevant operation.

Keywords

AI governance, Solid, federated identity, semantic retrieval, data governance, dataset publishing

1. Motivation

Three practical concerns motivate the platform design. First, enterprises need stronger *data governance* for AI-ready assets that are reused across training, evaluation, retrieval, and publication workflows [5]. Second, *AI safety* depends on the ability to connect policy, intended use, and evidence about the data and representation spaces used by models [26, 7]. Third, *data drift monitoring* is essential because embedding corpora, derived datasets, and retrieval collections evolve over time, potentially moving away from the conditions under which they were originally approved or assessed [7, 5]. These three concerns suggest that governance should be integrated with retrieval and monitoring rather than implemented as separate enterprise tools [5].

Recent literature strengthens this motivation by showing that AI safety is increasingly a *data quality* problem as much as a model or policy problem. A 2025 data-centric safety study argues that low-quality, unsafe, or unethical datasets can directly increase the risk of harmful downstream capabilities and therefore calls for responsible dataset design across the full AI lifecycle [1]. The 2026 International AI Safety Report likewise emphasizes that safety cannot be assessed only at the level of model outputs, because capability, robustness, and misuse risk are shaped by the quality, composition, and traceability of the data used to build and adapt systems [2]. More targeted empirical work also shows that imperfect data can significantly affect model behavior in realistic deployments: recent results on tabular foundation models report that common real-world data defects remain pervasive and materially alter performance and internal representations [3], while other 2025 safety frameworks explicitly incorporate dataset quality measures and internal test data as core components of safety assurance [4]. Taken together, these studies suggest that poor labeling, duplication, bias, missing provenance, and uncontrolled distribution shift are not peripheral data engineering issues but primary determinants of whether an AI system remains safe, reliable, and governable in practice [1, 3, 4]. For this reason, a governance platform for

NXDG 2026: NeXt-generation Data Governance Workshop, co-located with SEMANTiCS 2026, September 2026, Ghent, Belgium
lorenzo@genefold.ai (L. Moriondo); tommaso@genefold.ai (T. Moriondo)
https://www.genefold.ai/ (L. Moriondo); https://www.genefold.ai/ (T. Moriondo)
0000-0002-8804-2963 (L. Moriondo); 0000-000x-xxxx-xxxxx (T. Moriondo)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

AI-ready assets must support continuous inspection of data quality and representation quality, rather than assuming that safety can be guaranteed solely through model-side evaluation after deployment.

2. Solid as federated governance layer

A central contribution of the platform is the use of **Solid** as a federated authentication and authorisation layer for governed data assets. Solid relies on WebIDs to identify agents and on Solid-OIDC and Web Access Control (WAC) to authenticate and authorise access to resources across domains [8, 9, 10]. This makes it possible for users and groups hosted by one organisation to receive access to data assets hosted by another organisation without requiring local account duplication on the target server [8]. In enterprise terms, this means departments inside a company, and companies inside a consortium or supply chain, can authenticate as distinct agents and receive explicitly scoped access to datasets, governed subsets, or published data products.

This federated model is particularly relevant to AI governance because governed assets are rarely confined to a single team. A training dataset may be produced by one department, curated by another, audited by a governance or legal unit, and then shared with an external partner for evaluation [5]. Conventional sharing through exports, local platform accounts, or replicated copies makes revocation, auditing, and policy enforcement more difficult. By contrast, Solid supports a resource-centric governance model in which assets remain governed where they are hosted while authenticated external agents can be granted role-specific access under explicit policies [8, 9]. In Genefold, Solid therefore functions not only as a standards-based protocol substrate, but also as the trust layer for cross-company and cross-department data asset management [5].

Origins: from the Read-Write Web to Social Linked Data

The intellectual roots of Solid reach back to MIT CSAIL research funded by DARPA in 2000, which produced the early RDF tooling and the *Tabulator* — a JavaScript client for exploring linked data on the open web [11]. By 2009 the focus had shifted from read-only exploration to a *Read-Write Linked Data* model: the principle that any resource a user can read and has permission to write should also be writable, directly challenging the data-silo architecture of emerging social platforms. Tim Berners-Lee articulated this vision in a 2009 TED talk and in a series of W3C Design Issues notes on "Socially Aware Cloud Storage" [11, 12].

The name *Solid*—standing for *Social Linked Data*—was coined by Sandro Hawke during a collaboration between MIT and the Qatar Computing Research Institute (QCRI), which also produced the first formal access-control model grounded in RDF resources [11]. The platform was publicly demonstrated at the World Wide Web Conference in 2016 [13], and in 2015 Mastercard Labs funded two further years of specification and open-source development at MIT [11]. Berners-Lee co-founded **Inrupt** in 2017 to provide commercial infrastructure for Solid deployments; a W3C Community Group was formally chartered in 2018 to govern the protocol specifications through an open, royalty-free standardisation process [11, 14].

Standardisation trajectory

The protocol matured through a sequence of W3C Community Group releases: version 0.9 in December 2021, version 0.10 in December 2022, and a series of companion specifications covering Solid-OIDC, WebID, Access Control Policies (ACP), and the Solid Notifications Protocol [8, 9]. In 2024 the Open Data Institute (ODI) became the steward of the Solid project, broadening governance beyond the founding company and anchoring the specification in a public-interest institution with an explicit mandate for responsible data innovation [15]. Real-world deployments have validated the model at scale: Athumi in Flanders uses Solid to manage the data of 6.5 million citizens across government departments; NHS pilots in Greater Manchester have explored patient-controlled health-record sharing; and US construction-industry deployments leverage Solid for supply-chain document management [15].

3. ArrowSpace as semantic evidence layer

On top of this federated governance substrate, we introduce **ArrowSpace** [21] as a fully-fledged semantic retrieval and structural inspection layer [20, 19]. The implementation materials describe ArrowSpace as supporting lambda-aware similarity, topology-aware search, and spectral organization of embedding corpora rather than only brute-force semantic similarity [19]. Evaluation materials further describe topology-aware open retrieval in which the method can surface useful additional documents beyond a cosine nearest-neighbor baseline, emphasizing novelty, coverage, and diversity in retrieved results [20]. This behavior is useful for governance because auditors and stewards often need to inspect representative cases, boundary cases, drifted clusters, or structurally informative alternatives rather than only the nearest matches.

ArrowSpace therefore acts as a semantic evidence layer for governed assets. It enables governance workflows that are exploratory and diagnostic, allowing reviewers to ask not only "what is similar?" but also "what is structurally related, unusual, drifting, or underrepresented?" [20, 7]. This is important in AI governance settings where semantic search alone may be insufficient for safety review, risk investigation, or evidence collection. This is possible by the novel approach that leverage a novel graph Laplacian-based algorithm to produce a semantic web automatically out of the vector space; this is increasingly relevant because of the semantically dense spaces produced by modern embedding models.

This dual-space perspective (geometric space as defined by vectors and semantic space as defined by the feature-space graph Laplacian) is important because ArrowSpace does not treat retrieval as a problem confined to a single embedding geometry but as a controlled interaction between the *data space* in which assets, records, and governance units are organized and the *vector space* in which semantic similarity is computed [21, 20, 19]. In practice, this means that governance operations can be applied at two levels at once: first, over the underlying data assets and their partitions, provenance scopes, organizational ownership, and publication constraints; second, over the embedding manifold itself, where one may inspect local neighborhoods, spectral smoothness, cluster boundaries, and deviations from expected structure [21, 19]. The ArrowSpace formulation introduces spectral indexing precisely to bridge these two levels, adding bounded structural scores to conventional semantic representations so that retrieval can be steered not only by content affinity but also by how an item sits within the broader topology of the corpus [21, 20]. This enables a form of vector-space control that is highly relevant for AI governance: stewards can constrain or prioritize searches according to manifold coherence, diversity, structural novelty, or drift-sensitive regions, instead of relying only on nearest-neighbor proximity. At the same time, it enables data-space control because these structural signals can be attached back to governed datasets, subsets, and publication units, making it possible to decide which assets are stable enough for reuse, which regions require further review, and which slices of the corpus should be exposed, restricted, or monitored more closely [21, 20]. In this sense, the dual-space semantic approach turns ArrowSpace into more than a search index: it becomes a governance instrument for coordinating semantic access, structural diagnostics, and policy-aware control over AI-relevant data assets [21, 19].

4. Platform architecture

The cited strategic plan frames Genefold's product as a three-plane architecture: an identity and authorization plane, a data plane, and a governance and monetization plane [5]. In this design, the identity and authorization plane is implemented through Solid-based federation; the data plane is centered on chunked N-dimensional arrays, initially using Zarr v3 for embedding and ML-relevant data; and the governance plane is exposed through a web console for dataset registration, contract-aware access, auditability, and publishing workflows [5]. This separation is useful because it keeps identity, storage, and governance concerns interoperable but distinct. More specifically, the architecture is intended to support federated management of data assets across companies and across departments within the same company, so that ownership, stewardship, and access decisions remain attached to the originating organizational context rather than being flattened into a single central platform tenant

Table 1
Genefold platform layers and governance roles

Layer	Function	Relevance to AI governance
Federated identity and authorization	Solid uses WebIDs, Solid-OIDC, and WAC to identify agents and govern access across domains [8, 9, 10].	Enables cross-company and cross-department authentication, revocable sharing, and auditable access to data assets without local account duplication [8, 5].
Semantic retrieval and inspection	ArrowSpace provides semantic, spectral, and topology-aware retrieval over governed corpora [20, 19].	Supports evidence discovery, structural review, anomaly investigation, and governance-oriented exploration [20, 7].
Safety and monitoring	Structural artifacts expose drift and out-of-distribution changes in governed representation spaces [7, 5].	Connects AI safety oversight with ongoing governance decisions and reviews [7].
Publishing and reuse	Standards-based dataset descriptions support discovery and publication under governance constraints [5].	Preserves provenance, discoverability, and accountability as assets move across organizational boundaries [5].

[5, 8, 9]. In that sense, the identity plane does not merely authenticate users; it establishes a trust fabric in which departments, subsidiaries, suppliers, and external collaborators can appear as distinct governed agents with separate rights and responsibilities. The data plane correspondingly preserves the locality and structure of the governed assets, allowing arrays, embeddings, and derived artifacts to remain where they are managed while still being made discoverable and accessible through common protocols [5]. The governance plane then provides the institutional layer needed to turn this federation into an operational model for enterprise data assets, including traceable contracts, auditable sharing, controlled publication, and selective reuse across organizational boundaries. Taken together, the three-plane design positions Genefold not as a centralized repository, but as a federated platform for data sovereignty, interoperable access, and policy-aware management of AI-relevant assets.

The resulting platform can be summarized through four interoperating layers shown in Table 1.

Storage layer: Zarr v3 arrays served via `arro-server`

The physical storage foundation of the platform is a **Zarr v3** array hierarchy [25], served over HTTP through `arro-server` [**?**], an open-source FastAPI service developed by Genefold. Zarr organises n-dimensional arrays into chunked, compressed stores that map naturally onto standard object-storage or file-system backends, with each chunk independently addressable via a key derived from its coordinates. This layout enables efficient partial reads—a client fetches only the chunks that intersect a requested window or multi-axis slice—without any server-side cursor state, making the store intrinsically suited to governed, metered streaming. `arro-server` exposes a clean REST interface over one or more labelled Zarr roots: `GET /api/datasets` enumerates available arrays and groups; `/metadata` returns shape, dtype, chunk geometry, and user attributes; `/data` provides paginated row-window access over the leading axis; and `/slice` accepts numpy-style multi-axis slice expressions (`start:stop:step` per axis, comma-separated) for precise sub-array retrieval. A hard per-request element budget (configurable via `arro-server_MAX_WINDOW`, default 10 000 elements per leading-axis row) enforces metering at the HTTP layer without requiring changes to downstream consumers.

ArrowSpace metadata is integrated through a two-tier adapter: if the `pyarrowspace` package is importable it is used directly; otherwise the server reads a sidecar directory `<dataset>/_arrowspace/` containing `manifold.json`, `stats.json`, and `index.json`. This sidecar mechanism allows a dataset to advertise its ArrowSpace manifold, spectral statistics, and semantic index to governance tooling even in deployment environments where the full Genefold stack is not installed. The `/manifold` and

`/stats` endpoints expose these artefacts over HTTP, while `/search` provides a search seam whose backend upgrades automatically from naive sidecar substring matching to full vector and manifold search once `pyarrowspace` is wired up. New storage backends—S3, GCS, Apache Parquet, Apache Iceberg—can be added by implementing the `StorageBackend` protocol and registering in the multi-backend multiplexer, keeping the governance and metadata layers independent of the physical storage medium. The server is packaged as a `Containerfile` compatible with both Podman and Docker and ships a `compose.yaml` for local and orchestrated deployments, consistent with the platform's Podman-first container strategy.

5. AI safety, drift, and publishing

The platform's contribution to **AI safety** lies in keeping semantic evidence, governance metadata, and structural diagnostics close enough that a dataset can be both discovered and evaluated through the same governance surface [7]. Genefold publicly positions its platform as producing reusable structural artifacts for search, diagnostics, drift detection, and data valuation over embedding corpora [7]. This turns drift monitoring into a governance concern rather than an isolated MLOps afterthought. A dataset or retrieval corpus that was previously acceptable may evolve in ways that affect its suitability for downstream AI use, so governance systems need to observe representation spaces over time, not only control initial access [5, 7]. Recent literature reinforces this position by showing that AI safety failures are often rooted in data quality problems rather than only in model architecture or deployment policy. A 2025 data-centric safety framework argues that low-quality, unsafe, or unethical datasets directly increase the risk of harmful downstream capabilities and calls for responsible dataset design across the entire AI lifecycle [1]. The 2026 International AI Safety Report similarly emphasizes that safety evaluation must consider the quality, composition, provenance, and traceability of the data used to build and adapt general-purpose AI systems [2]. Recent work on safety datasets further shows that poor-quality benchmarks can create a misleading sense of safety: one 2026 study finds that widely used adversarial safety datasets over-rely on unrealistic triggering cues, and that models previously judged reasonably safe become unsafe once these cues are removed [22]. Broader reviews of AI safety benchmarking also document substantial technical and epistemic shortcomings in current benchmark practice, suggesting that weak data and weak evaluation data can jointly undermine reliable safety claims [23]. Together, these findings support the view that AI safety governance must include continuous inspection of dataset quality, representation quality, and evaluation quality, rather than treating them as peripheral concerns.

The platform also treats **dataset publishing** as part of governance rather than as a terminal export step. The roadmap describes STAC-compatible catalog metadata and standards-based discovery for Genefold-managed datasets [5]. In this model, publication means exposing a governed resource to broader networks of departments, institutions, and collaborators while continuing to control authentication, inspection, evaluation, and reuse through the federated Solid layer [5, 8]. This is especially relevant for research consortia and regulated sectors where discoverability and accountability must coexist with sovereignty and revocability.

Recent analyses of foundation-model risk also stress that inadequate or weakly governed training data can propagate factual errors, harmful content, and disproportionate harm to under-represented populations, while limited external oversight makes those weaknesses harder to detect before deployment [24]. From this perspective, governed publication is not simply a dissemination feature; it is part of the safety case for a dataset, because publication metadata, access constraints, provenance records, and downstream reuse conditions all contribute to whether a data asset can be responsibly reused in AI pipelines [2, 1]. By linking publication with federated access control and structural diagnostics, the platform aims to ensure that discoverability does not come at the cost of safety, and that reusable datasets remain inspectable, accountable, and revocable after release.

6. Visualising the flow

A simple UX demo application can show the platform as an end-to-end governance workflow over AI-relevant data assets. A dataset can be registered by one organisational unit, exposed through governed metadata, discovered through semantic search, accessed by authorised agents from other departments or partner companies via Solid-based federation, inspected for structural or drift-related issues, and then prepared for standards-based publication [5, 7]. This sequence directly addresses the workshop interest in smarter and enforceable data governance for AI-powered data ecosystems. This flow allows the monitoring of risks related to data quality and other AI safety-related supervision in an integrated fashion with the functions needed by data operators in large organisations.

The demonstration is a static site served at https://www.genefold.ai/data-governance.genefold.github.io/, whose source is publicly available at https://github.com/Genefold/data-governance.genefold.github.io. Each page corresponds to a distinct actor role or platform function, and together they form the complete governance lifecycle. The five panels are described below in the order a practitioner would traverse them.

Catalog (/)

The landing page presents a searchable, filterable catalog of data products across the federated mesh. Each card shows the dataset title, owning organisation, domain, modality, geographic jurisdiction, applicable license, and AI training eligibility. Faceted filters expose the full governance vocabulary: domain (medical imaging, earth observation, computational biology, and others), geography, license (CC-BY-4.0, ODbL-1.0, research-only, partner-only), and training eligibility (approved, partner-approved, approved-with-DPIA). A free-text search queries title, owner, and purpose together, enabling semantic discovery without exposing raw data payloads. Datasets are streamed—never bulk-transferred—using HTTP Range requests against Cloud-Optimised GeoTIFF (COG) or BigTIFF storage; catalog cards therefore carry provenance and contract metadata rather than data copies.

Data owner workflow (`/owner/`)

The owner panel implements a four-step publication lifecycle: *Register*, *Access mode*, *Publish*, and *Live*. In the registration step the data steward declares the dataset title, the owning Solid WebID or DID, the organisation, a natural-language purpose statement, jurisdiction (EU/GDPR, UK NHS-DSPT, US HIPAA, or public), sensitivity class, domain, and a typed schema. The second step exposes a multi-select access-mode contract: preview, research-only, train-only, inference-only, internal-only, partner-only, and paid external. These modes are first-class contract terms that directly parameterise the capability tokens issued downstream.

On submission, a publish checklist is evaluated automatically against the technical guarantees of the platform: BigTIFF/COG layout validation, tile size and overview levels ($1\times$–$32\times$), HTTP `Accept-Ranges` availability, ArrowSpace graph index construction (default $k = 24$), DPIA reference attachment, Solid-OIDC DPoP nonce binding, capability key minting, and STAC catalog entry generation. A live log panel reports each step with PASS/WARN/INFO status, making governance state immediately auditable by the publisher. The underlying mechanics are made explicit in a "What just happened?" summary: the catalog entry is minted as a Solid resource on the owner's pod; permissions, lineage, and contracts are linked Solid resources that remain governable and revocable; every approved consumer receives a DPoP-bound capability JWT scoped to area-of-interest, resolution, purpose, and time window.

Data consumer workflow (`/consumer/`)

The consumer panel exposes three sequential steps: *Search*, *Preview*, and *Request access*. The search step mirrors the catalog filters and allows a consumer to narrow datasets by domain, modality, and training eligibility. The preview step issues a minimal HTTP Range fetch—64 KB at a selected COG overview

level ($4\times$ or $16\times$ downscale) and area-of-interest tile—so consumers can validate spatial coverage and data quality before committing to a full access request; this preview is itself governed, operating under the Discovery pricing tier and logging the byte-range read to the audit ledger.

The access request form captures the stated purpose, contract duration, legal basis (e.g. GDPR Art. 6(1)(e) public task, GDPR Art. 9(2)(j) research, or contractual), intended model usage (train weights, fine-tune, inference-only, evaluation-only), and the requesting WebID or DID. On submission the platform mints a signed capability token encoding all parameters as machine-readable scopes. A streaming session log then replays the byte-range sequence: COG header fetch, overview read at the requested downscale level, full-resolution tile fetch, and a metering counter update, illustrating how every access event is ledgered without centralised data movement.

Dataset detail panel (`/dataset/`)

Navigating from any catalog card opens the dataset detail page, which aggregates the full governance dossier for a single data product: a tile-level COG preview, provenance metadata, sensitivity and jurisdiction badges, allowed-use contract terms, and a lineage timeline. This panel is the artefact that a data operator or external auditor would inspect to verify that an AI training run was permitted, properly metered, and reproducibly documented. The STAC catalog entry generated at publish time appears here as a structured metadata block, enabling standards-based interoperability with external catalogues and downstream pipeline tooling.

Compliance and monetisation console (`/console/`)

The operator console provides a real-time governance dashboard across five tabs. The *Access products & pricing* tab surfaces the full pricing taxonomy, making the financial governance of AI training data transparent and auditable.

The *Audit & policy* tab is the most directly AI-safety-relevant panel: it presents a live, append-only audit ledger of every byte-range read, policy decision, and actor event, alongside six toggleable policy rules—require DPoP for all reads, block unknown DIDs, require DPIA for sensitivity $\geq$ high, auto-revoke on cap breach, watermark previews, and region-pin EU datasets. These controls operationalise the consent and access-revocation requirements of GDPR and emerging AI Act provisions, directly connecting to the data-drift monitoring and AI safety supervision motivations discussed in Section **??**.

The *Downstream mesh* tab visualises the federated consumer graph as a node-link diagram: the Genefold hub connects to labelled consumer nodes (academic institutions, healthcare providers, commercial partners) via edges that represent active governed capabilities, not data transfers. This topology is the runtime expression of the Solid-based federation model. Finally, the *Usage & frequency* tab presents bytes streamed per dataset over the previous 30 days and a time-series of mesh-wide reads per day, giving data operators the signals needed to detect anomalous consumption patterns—an operational proxy for data-drift and misuse detection.

7. Conclusion

Genefold Data Governance proposes a platform approach to AI governance in which **Solid** serves as the federated authentication and authorization layer across companies and departments, **ArrowSpace** serves as the semantic and topology-aware evidence layer, and the surrounding governance stack supports safety monitoring and standards-based dataset publication [8, 9, 20, 19, 5]. The key proposition is that next-generation AI governance requires both semantic understanding of AI-relevant data assets and federated control over who may access, inspect, and reuse them across organizational boundaries [5].

Declaration on Generative AI

During the preparation of this work, Generative AI tools were used to support drafting and language refinement. The authors reviewed and edited the content as needed and take full responsibility for the publication's content. This declaration is included in line with the CEUR-WS Policy on AI-Assisting Tools [26].

References

[1] A. Raghuvanshi, A. Qureshi, M. Riedlinger, and collaborators. Data-Centric Safety and Ethical Measures for Data and AI Foundation Models. arXiv preprint arXiv:2506.10217, 2025. https://arxiv.org/abs/2506.10217. Accessed 2026-06-22.

[2] Yoshua Bengio, Stephen Clare, Clara Prunkl, et al. International AI Safety Report 2026. arXiv preprint arXiv:2602.21012, 2026. https://arxiv.org/abs/2602.21012. Accessed 2026-06-22.

[3] Authors to be confirmed from the proceedings PDF. Out in the Wild: Investigating the Impact of Imperfect Data on a Tabular Foundation Model. In *Proceedings of the VLDB 2025 Workshops, QDB 2025*, 2025. https://www.vldb.org/2025/Workshops/VLDB-Workshops-2025/QDB/QDB25_1.pdf. Accessed 2026-06-22.

[4] Authors as listed on arXiv. Domain-Agnostic Scalable AI Safety Ensuring Framework. arXiv preprint arXiv:2504.20924, 2025. https://arxiv.org/abs/2504.20924. Accessed 2026-06-22.

[5] Genefold. Genefold: Governed Array Intelligence Platform — Strategic Plan. Internal planning document provided by the authors, 2026. document

[6] Genefold. Genefold: Governed Array Intelligence Platform — Presentation. Presentation document provided by the authors, 2026. presentation

[7] Genefold. Genefold — Demo of AI Governance Platform. https://www.genefold.ai/data-governance.genefold.github.io/. Accessed 2026-06-22.

[8] Solid Project. Solid Specification. https://github.com/solid/solid-spec/blob/master/README.md. Accessed 2026-06-22.

[9] Solid Community Group. Solid-OIDC. https://solidproject.org/TR/oidc. Accessed 2026-06-22.

[10] Solid Community Group. Solid WebID Profile. https://solid.github.io/webid-profile/. Accessed 2026-06-22.

[11] T. Berners-Lee, "A Short History of the Solid Protocol," W3C Design Issues, 2023. Available: https://www.w3.org/DesignIssues/history/AShortHistoryoftheSolidProtocol.html Accessed: 2026-06-22.

[12] T. Berners-Lee, "The year open data went worldwide," TED, 2009. Available: https://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide Accessed: 2026-06-22.

[13] T. Berners-Lee, N. O'Sullivan, and collaborators, "Solid demonstration at the International World Wide Web Conference," WWW Conference Demo, 2016. Referenced in: T. Berners-Lee, "A Short History of the Solid Protocol." Available: https://www.w3.org/DesignIssues/history/AShortHistoryoftheSolidProtocol.html Accessed: 2026-06-22.

[14] W3C, "Solid Community Group Charter," World Wide Web Consortium. Available: https://www.w3.org/community/solid/charter/ Accessed: 2026-06-22.

[15] Open Data Institute, "Solid's next chapter – returning the web to its people-first roots," ODI Blog, 2025. Available: https://theodi.org/news-and-events/blog/solids-next-chapter-returning-the-web-to-its-people-first-roots/ Accessed: 2026-06-22.

[16] Open Data Institute, "Solid Symposium 2026," Event page, 2026. Available: https://theodi.org/news-and-events/events/solid-symposium-2026/ Accessed: 2026-06-22.

[17] Open Data Institute, "Solid26 Preview: Building personal apps and AI agents," Event page, 2026. Available: https://theodi.org/news-and-events/events/solid26-preview-building-personal-apps-and-ai-agents/ Accessed: 2026-06-22.

[18] L. Moriondo "pyarrowspace: Python bindings for arrowspace-rs." GitHub Repository, 2026. code

[19] L. Moriondo "arrowspace-rs: A database for vectors supported by a graph representation and a key-value store." GitHub Repository, 2026. code

[20] Tuned. "pyarrowspace: Python bindings for arrowspace-rs." GitHub Repository, 2026. https://github.com

[21] Moriondo, L. "ArrowSpace: introducing Spectral Indexing for vector search." *Journal of Open Source Software*, 10(113), 9002, 2025. https://doi.org

[22] Shahriar Golchin and Marc Wetter. Intent Laundering: AI Safety Datasets Are Not What They Seem. arXiv preprint arXiv:2602.16729, 2026. https://arxiv.org/abs/2602.16729. Accessed 2026-06-22.

[23] Authors as listed on arXiv. How Should AI Safety Benchmarks Benchmark Safety? arXiv preprint arXiv:2601.23112, 2026. https://arxiv.org/abs/2601.23112. Accessed 2026-06-22.

[24] Anne L. Washington. Fragile Foundations: Hidden Risks of Generative AI. Bertelsmann Stiftung, 2025. https://www.bertelsmann-stiftung.de/en/publications/publication/did/fragile-foundations-hidden-risks-of-generative-ai. Accessed 2026-06-22.

[25] Zarr Developers, "Zarr core specification," Zarr Specifications Documentation, Version 3. Available: https://zarr-specs.readthedocs.io/en/latest/v3/core/index.html Accessed: 2026-06-22.

[26] Angelo Salatino, Fabrizio Fornari, and Jelena Zdravkovic. CEUR-WS Policy on AI-Assisting Tools. CEUR-WS, 2024. https://ceur-ws.org/GenAI/Policy.html. Accessed 2026-06-22.