Publication Date: 2025-09-xx
<e.g., http://dx.doi.org/10.29322/IJSRP.X.X.2018.pXXXX>

OASIS Data Provenance Standards

Ensuring Provenance and Permitted Use of Data on any Scale across Organizations

Andy Hannah
Blue Street Data
USA
andyhannah@bluestreetdata.com

Bryan Bortnick
IBM
USA
bortnick@us.ibm.com

David Kemp
NSA
USA
d.kemp@cyber.nsa.gov

Fotis Psaliidas
Microsoft
USA
fotis.psallidas@microsoft.com

Lisa Bobbit
Cisco
USA
lbobbit@cisco.com

Stefan Hagen
Individual
Switzerland
stefan@hagen.link

Abstract

Data is a core enterprise asset that underpins strategic decision-making, drives operational priorities, and supports risk management. Such heavy dependence on data creates a need for validation by understanding data's origin, quality, and permitted/developed use. This is a requirement for organizations operating at scale.

The OASIS Data Provenance Standards (DPS) were created to solve for this need. Developed through a cross-industry collaboration, they provide a consistent framework to track the origin, movement, integrity, and quality of data. The DPS use addresses the growing demand for transparency in AI, cybersecurity, supply chains, and other areas where data quality and accountability are foundational to performance and compliance - especially in regulated and high-risk environments.

Keywords

Automation, Data, Information-Model, License, Open-Source, Protection, Provenance, Verification

1. Introduction: Why Standardize Now?

Data has traditionally been siloed, managed, and governed by application (built-in controls and defined processes) and user of that application (contract or other legal basis).

In today's world where data is being sought for its value, its use in innovation, its processing alignment to regulations, and its trustworthiness via transparency and observability, data requires to be treated as an independent asset.

To be independent, data must know itself by maintaining its provenance and tracking its lineage. Standardized provenance and lineage metadata provides consistent identifiers to allow data to be independent as it traverses its processing lifecycle from creation/collection, use, storage / retention, sharing, and deletion.

The DPS use creates a common foundation to evaluate data across platforms and jurisdictions—an open, readable framework designed for interoperability across systems, teams, and industries.

The DPS use provides a framework with a common taxonomy and adoptable templates that consistently

define and document usage policies, restrictions, and lineage.

Automating the framework allows for streamlined testing of a dataset's fitness for purpose, quality, and regulatory alignment.

The DPS are tool-agnostic and thus do not require an overhaul of an existing infrastructure. The metadata descriptors can be directly integrated into existing workflows with either commercial tools or proprietary systems. This level of interoperability allows teams to focus on data evaluation at scale. With standardized metadata in place, teams can automate decisions, accelerate procurement, and reduce risk.

The DPS can be applied in a modular process depending on the organization's adoption strategy:

1. Map existing or new metadata to the framework to uncover gaps and create a clearer picture of what data the organization has, where it came from, and how it can be used.

2. Begin by identifying high-risk data products to map against these standards.

Identifying, managing, and governing your data as an independent asset today is critical for trustworthy transparency, regulatory and legal basis alignment, and use in AI. Trust in the insights and decisions coming from both traditional data and AI applications depends on understanding the origin, lineage, and rights linked with the data that drives them.

Lack of transparency has real costs, including unnecessary risks and foregone opportunities. And yet, many organizations today cannot answer basic data questions without considerable difficulty and investment of people and disparate tools.

To realize the value of data requires a reliable cross-industry baseline of data provenance (source and legal processing basis) and lineage (data processing's what, where, who, and how) as data is valued, governed, and managed across industries from customer to data processor to sub-processor.

The Data Provenance standards define a solution supporting throughout the data's life cycle. These use case scenarios showcase how the Data Provenance standards support diverse needs across the data ecosystem. The core objectives include:

- Verifiable lineage, integrated with contextual metadata, provides a transparent history of each dataset, supporting both technical assessment and executive-level oversight.

- Embedded data integrity checks and audit-readiness mechanisms to monitor for unauthorized changes and ensure compliance with regulatory obligations.

- By centralizing provenance information, one can easily evaluate data reliability, usage restrictions, and suitability for purpose—reducing the risk of deploying incomplete, manipulated, or non-authorized datasets.

- Improve operational transparency across data supply chains and AI models, providing the foundation for explainable AI and ethical data use.

2. Core Data Provenance Concepts

When organizations gain stronger command over insights into their data, the results extend far beyond governance. The adoption of standardized data provenance practices leads to sharper insight, faster time to action, and more resilient decision-making.

2.1 A Stronger Foundation for Governance and Risk Management

- Clear Lineage and Oversight: With visibility into where data originated, how it has been transformed, and where it's headed, organizations gain a defensible framework for audits and policy enforcement—governance becomes a continuous, scalable capability.

- Verified Integrity, Built-In Compliance: Knowing that data was lawfully collected and responsibly handled reduces risk across the lifecycle. This clarity directly supports incident response, consent management, and alignment with regulatory expectations.

- Procurement Built on Certainty:
  Provenance metadata equips procurement teams to assess quality, fit, and compliance upfront.
  This minimizes costly missteps and accelerates purchase decisions, all while maintaining high ethical and operational standards.

2.2 Expected Outcomes of Implementation of the Data Provenance Standards

- AI Deployment Without the Guesswork:
  Organizations can assess whether datasets are appropriate for training—representative, unbiased, and regulation-ready. This accelerates AI scalability while keeping risk in check.

- Confidence in High-Stakes Decisions:
  Traceable inputs mean fewer blind spots. Decision-makers can trust that insights are backed by transparent and reliable data, which is essential in regulated and performance-critical environments.

- Efficiency Without Compromise:
  Meeting demand for data shouldn't mean lowering standards.
  These standards allow organizations to move fast and stay compliant—reducing overhead without sacrificing integrity.

- Enabling a Trustworthy Data Ecosystem:
  AI models are only as reliable as the data that feeds them.
  Without rigorous provenance, the risk of biased outputs, legal exposure, and reputational damage climbs. These standards build shared accountability—internally and across the ecosystem.

3 Benefits of Adopting Data Provenance Metadata

- What Leadership Enables: From Intention to Implementation Enterprise adoption of the DPS doesn't require sweeping transformation from day one. It begins with strategic, high-impact actions that lay the groundwork for long-term value. Leaders play a pivotal role in enabling this shift—by clarifying priorities, empowering teams, and integrating provenance into existing governance and procurement workflows.

- Start with What Matters Most: Focus on datasets that represent the greatest risk or opportunity. Whether it's a high-value product line or a critical AI use case, select a scope where improved metadata can deliver immediate results. From there, each successive implementation becomes easier, helping to scale adoption organically across the organization.

- Build Momentum Through Practical Steps: Implement lightweight mechanisms that lower the barrier to entry. A brief onboarding survey can capture key metadata—such as whether a dataset includes personal information or its intended purpose—without requiring major infrastructure changes. These simple prompts normalize provenance practices and accelerate organizational learning.

- Visualize for Transparency, Action, and Insight: Encourage organizations to surface metadata through dashboards or overviews. When data quality, lineage, and gaps are visible, then organizations are better equipped to assess compliance, prioritize remediation, and align datasets with business objectives. This is not just transparency for transparency's sake—it's a catalyst for better, faster decisions.

- Foster Alignment Across Data, Legal, and Procurement: Cross-functional coordination is essential. Nominate key stakeholders from each function to participate in pilot programs and standards discussions. Provenance can streamline procure- ment, simplify compliance, and reduce legal exposure—benefits that resonate across departments and create a shared path to adoption.

- Elevate Vendor Expectations: When sourcing data externally, ask vendors to provide provenance metadata. Whether it's lineage, usage restrictions, or collection context, clear requests signal new expectations. Over time, this not only enhances

your own data operations—it helps shift the broader market toward more transparent and reliable data exchanges.

- Integrate Provenance Into Governance and Risk Practices: Provenance should sit alongside quality metrics and access controls as a first-class input to enterprise risk frameworks. It informs scoring models, audit documentation, and retention schedules—enabling smarter decisions about what data to keep, trust, or retire. This is not a peripheral concern; it's a core dimension of data governance.

- A Leadership Mandate for Data Confidence: Provenance adoption is no monolithic project—it's a phased, scalable evolution. By embedding it within existing practices and aligning stakeholders early, leaders create the conditions for durable, enterprise-wide impact. The standards exist. The tooling is ready. Leadership is what turns that readiness into action.

4 Provenance Information Model

The information model of the provenance metadata is described in human-readable property tables.

The Data Provenance Standards are made up of three content groups of metadata elements: Source, Provenance, and Use. In addition, other maintenance related infrastructure elements (for acknowledgements, history, notes, and versioning) support automation and verification.

The property tables first define metadata about the specification itself, then describe how a record is made of the three primary metadata elements (Source, Provenance, and Use), then describe each of the three elements in isolation.

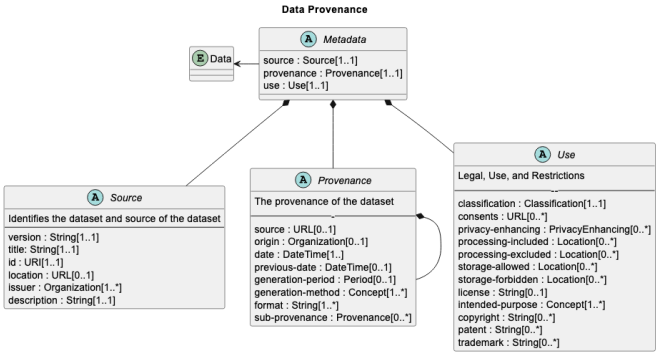

Fig. 1. DPS Information Model

The DPS can be adopted widely without requiring extensive system overhauls.

5 Conclusion and What's Next

Adopting the DPS isn't just a compliance task—it's a declaration of intent. It signals an organization's readiness to lead in a future defined by transparent, trustworthy, and ethical data practices. For leaders, it's a unique moment to drive both immediate value and long-term advantage across governance, AI, and organization operations. The DPS Technical Committee (TC) will continue to work on defining the information model, as well as providing data format specific schemata for automated exchange and verification of data provenance metadata.

In addition, the DPS TC provides descriptions of realistic use cases, describing scenarios and examples of step-by-step assessments.

6 Feedback

This work is ongoing in the DPS TC at the OASIS Open. Current members include:

- Blue Street Data
- Cisco Systems
- Data & Trust Alliance
- DataProbity
- Dell Technologies
- Dentons
- Huawei Technologies Co., Ltd.
- IBM
- Intel Corporation
- Microsoft

- National Security Agency
- Peraton
- Red Hat
- Royal Holloway University of London
- Siemens AG

Please provide feedback through the comment mailing list of the DPS TC

Notices