

Data Provenance Standards Executive Overview

The first cross-industry metadata standards to bring transparency to the origin and use of datasets for AI and traditional data applications.

01

Why Data Provenance

“AI is all about the data.
In fact, data may be the
only sustainable source
of competitive advantage.”

— Rob Thomas, IBM, SVP Software and Chief Commercial Officer and
Chair of the D&TA Data Provenance Standards initiative

Trust in data requires understanding its provenance: where it comes from, how it was created, and whether it may legally be used.

“Transparency and accuracy around the origin of food, water, raw materials and capital are fundamental prerequisites for society, essential to establishing trust and defining quality. We've always felt the same standard must apply to data.”

Neil Blumenthal,
Warby Parker, CEO

Transparency into data provenance is critical for business value

Today, many organizations are not making progress on AI adoption because they cannot answer basic data provenance questions—on rights, restrictions, and sourcing.

61% of CEOs cite lack of clarity on data lineage and provenance as a top barrier to adoption of generative AI. [Source]

Lack of provenance already results in high costs: overpaying for low quality data and investing heavily in data cleanup.

~40% of data scientists time spent working on data preparation and cleansing tasks. [Source]

Visibility into provenance can not only increase efficiency, accuracy, and quality assurance but also lead to more informed decision making, speeding up innovation.

AI regulation now demands higher data transparency

Emerging regulation includes provisions on transparency, provenance, and the need to thoroughly understand the input data to AI models.





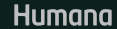







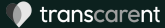


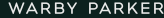
| | |
|----|---|
| US | Executive Order on AI AI Foundation Model Transparency Act |
| EU | AI Act |
| UK | Bletchley declaration |
| G7 | Hiroshima AI Process |
| SG | AI Governance Framework |

All organizations will need clearer data transparency to comply.

Today, there is no cross-industry standard for data transparency and trustworthiness.

We developed the first cross-industry standards for data provenance to meet the business and regulatory needs for better, trustworthy data.

CREATED BY EXPERT TEAMS FROM INDUSTRY-LEADING COMPANIES

| | | | | | | | |
|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

“Companies like ours feel a deep responsibility to ensure new value creation, as well as trust and transparency of data with all of our customers and stakeholders. Data provenance is critical to those efforts.”



Ken Finnerty, UPS,
President, IT & Data
Analytics

“We are committed to the adoption of Responsible AI, and an important component of that is trust in the data and approach used to train and deploy AI models. It starts with standards, and this is an important step to ensure transparency and responsible innovation.”



Greg Ulrich,
Mastercard, Chief AI
and Data Officer

02

How the the standards were created

Built by practitioners across the Alliance, starting from business-critical use cases

Assembled a working group of CDOs, CIOs, and leads of data strategy, enterprise data governance, data acquisition, compliance and legal from across 15 industries.

We started with 25 use cases that were critical to each company's business strategy and showcased challenges around provenance.

| | | | | | |
|------------------|-------------------------|--|--------------|--------------------------|--|
| AARP | Jaye Campbell | SVP, Legal - Corporate, Media, IP & Privacy | MASTERCARD | Travis Carpenter | SVP, Data Quality and Sources |
| AMERICAN EXPRESS | Laurel Shifrin | Vice President, Enterprise Data Governance | NIELSEN | Christine Pierce | Chief Data Officer, Audience Measurement |
| HOWSO | Chris Hazard | Co-Founder & CTO | PFIZER | Genta Spahiu | Director, Enterprise Data Governance Lead |
| | Michael Meehan | General Counsel and Chief Legal Officer | TRANSPARENT | Thi Montalvo | SVP Reporting and Analytics |
| HUMANA | Genevy Dimitrion | VP, Data Strategy & Governance | | Thomas Birchfield | Technical Program Manager |
| IBM | Lee Cox | VP, Integrated Governance & Market Readiness, Office of Privacy and Responsible Technology | UPS | Mallory Freeman | PhD; VP, Enterprise Data and Analytics |
| | Bryan Bortnick | Counsel, IBM Data Governance | | Zeenat Syed | Director of Data & AI Governance |
| | Bryan Kyle | Sr. Technical Staff Member, Platform Architect, IBM Enterprise Data | WALMART | Gregory Schaffer | Chief Counsel, Cyber Security & VP, Digital Trust Compliance |
| | Orla Flannery | Privacy Program Manager, Chief Privacy Office | WARBY PARKER | Peter Cross | Head of Data |
| KENVUE | Bernardo Tavares | Chief Technology & Data Officer | | | |
| | Ajay Dhau | SVP, Global Data, Applied AI and Digital Business Transformation | | | |

Iterated with business value and practicality in mind, and designed for adoption.

The standards were iterated through 150+ deep dives across the Alliance.

They were designed for ease of implementation and to realize multiple sources of business value:

Increased efficiency in data usage (buy once, reuse many times)

Ability to estimate level of cleanup effort for dataset

Clear flags on PI, PII, PCI or PHI in a data set, makes it easier to use/reuse

Faster data procurement turnaround

Cleaner data entering the enterprise and less reliance on SLAs

Faster audits as a result of understanding where data comes from

Faster time to use or decline to use data, due to understanding origin

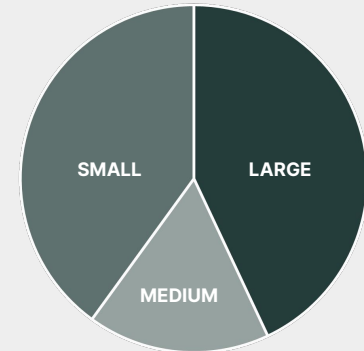
Not paying for stale data

Refined with feedback from 55+ organizations—small and large—outside the Alliance

Expertise represented in external feedback

| | |
|----------------------|--------------------|
| Aviation | Education |
| Ad Technology | Energy |
| AI Provenance | Financial Services |
| AI Governance | Healthcare |
| Corporate Governance | Human Rights |
| Cybersecurity | Insurance |
| Data Analytics | Media |
| Data Policy | Real Estate |
| Data Privacy | Telecommunications |

Org. size



Data suppliers, tool providers and the data management community believe adoption across the ecosystem is necessary

“Dun & Bradstreet is pleased to have partnered with the Alliance to test the Data Provenance Standards. We believe the Data Provenance Standards will help organizations establish trust in solutions and experiences that leverage data and AI technologies through increased transparency, interoperability and compliance insights to support accountability—all of which are essential building blocks in this rapidly evolving space to help everyone achieve better outcomes.”



Gary Kotovets
Dun & Bradstreet
Chief Data & Analytics
Officer

“The Data Provenance Standards are the missing piece for trustworthy AI. It shifts the focus beyond ‘what’ the data says to ‘why’ and ‘how’ it came to be. This unlocks a new wave of innovation built on compliance, trust, safety and privacy.”



Tim Wagner
Vendia
Co-founder and CEO

“Oasis is very pleased to be the future host of D&TA Data Provenance Standards, vital for tackling today’s crucial data and AI challenges. Bringing these standards to OASIS will help drive their global advancement and adoption, and we look forward to enhancing the interoperability, transparency, and effectiveness of these standards through collaborative efforts across diverse sectors.”



Francis Beland
OASIS Open
Executive Director

“Placing trust in data often begins with knowing the source. The EDM Council is grateful to the Data & Trust Alliance for seeking input from our global data management community on standards for addressing this critical issue. We are pleased to announce our intention to adopt D&TA’s standards for tracing the origin of datasets into the next versions of our flagship data management capability frameworks (DCAM and CDMC). Industry collaboration for advancing data management and analytic capabilities is core to our mission, and we look forward to further collaboration with D&TA.”



Jim Halcomb
EDM Council
Global Head of
Product Management

03

What are the Data Provenance Standards

Version 1.0.0 contains 22 metadata fields, grouped into 3 standards.

The metadata is intended to travel with a dataset as it is shared and transformed.

| STANDARD | METADATA |
|---------------------------|--|
| SOURCE | Standards version used |
| | Dataset title/name |
| | Unique metadata identifier |
| | Metadata location (unique URL of the current dataset) |
| | Dataset issuer |
| | Description of the dataset |
| PROVENANCE | Source metadata for dataset |
| | Source (if different from Issuer) |
| | Data origin geography |
| | Dataset issue date |
| | Date of previously issued version of the dataset (if applicable) |
| | Range of dates for data generation |
| | Method |
| | Data format |
| USE | Confidentiality classification |
| | Consent documentation location |
| | Privacy enhancing technologies (PETs) or tools applied? |
| | Data processing geography inclusion/exclusion |
| | Data storage geography inclusion/exclusion |
| | License to use |
| | Intended data use |
| Proprietary data presence | |

Download our Use Case Scenarios to understand how the standards inform decision making across different business scenarios

[View Use Case Scenarios](#)

Scenario 1 Healthcare insurance data procurement

Evaluating a new dataset that contains comprehensive patient and insurance payment information, for use in predictive analytics

Scenario 2 Media consumption pattern dataset for consumer behavior insights

Curating a high-quality dataset that tracks media consumption habits across diverse platforms for content personalization

Scenario 3 Financial services customer product enablement

Evaluating a new dataset for refining AI algorithms used in customer credit card offerings

Scenario 4 Enhancing global logistics efficiency through AI-driven tariff harmonization

Managing data to refine AI systems for accurately predicting tariff costs across countries and categories

Visit our Data Provenance Standards Metadata Generator to learn more about the standards and metadata fields

Visit the generator

This generator is designed to help people create and download standardized metadata files in JSON, CSV, or CML format to meet the Data Provenance Standards and facilitate data sharing

Source

Dataset title/name

Required See description

Enter name here

Unique metadata identifier

Required See description

Enter unique identifier (such as UUID)

Metadata location

See description

Enter unique URL of the current dataset

Data issuer

Required See description

+ Add data issuer

Description of the dataset

Required See description


Enter description

04 Adopting the Standards

Case study: How IBM is using the Data Provenance Standards

In early 2024, IBM tested the standards as part of their clearance process for datasets used to train foundational AI models. They saw increases in both efficiency (time for clearance) and overall data quality, and this is laying the path for enterprise-wide adoption.

[Read more in their case study →](#)

A graphic with a black background and white text. At the top left, it says 'Case study' and 'IBM Office of Privacy and Responsible Technology'. The main text reads 'Optimizing data governance with the Data & Trust Alliance Data Provenance Standards'. At the bottom left is the IBM logo. To the right of the graphic is a quote in black text on a light gray background, and below that is the name and title of Lee Cox.

Case study **IBM Office of Privacy and Responsible Technology**

Optimizing data governance with the Data & Trust Alliance Data Provenance Standards

IBM

“Standardizing and expanding the taxonomy we use to describe and document data set metadata will continue to help facilitate faster data clearance review and improved content quality, enabling us to respond even more efficiently to increasing demand for trustworthy data.”

Lee Cox, VP Integrated Governance & Market Readiness Office of Privacy and Responsible Technology, IBM

Adoption Kit: A free set of tools to support Standards adoption

Access adoption tools

Data Provenance Standards Metadata Generator

Create standardized metadata files in a variety of formats

Use Case Scenarios

Showcases how the standards inform decision making across different scenarios

Data Provenance Standards MVP

MS Excel file for capturing metadata in a familiar tool

Technical Resource Center

Technical standards specs, code snippets, and other implementation assets, hosted on GitHub

Community of practice

Are these standards applicable to all industries?

Yes, they are designed to apply to all industries. While we believe they are especially beneficial for sectors like healthcare, finance and technology, any industry that handles data and uses it for AI can benefit from implementing these standards.

What size companies should implement these standards?

All companies, regardless of size, are encouraged to adopt these standards. Implementing them may not only strengthen trust in your data integrity within the broader data ecosystem, but also can signal that your business is a leader in data transparency and reliability efforts. Adoption can serve as a competitive advantage, showcasing your commitment to best practices in data management and increasing the value of your data and AI in the marketplace.

How do these standards improve data security and compliance?

By providing a clear history of data provenance and its appropriate use, these standards help in auditing and monitoring data use, thus supporting data security and regulatory compliance efforts.

What are the benefits of implementing Data Provenance Standards?

Adopting these standards enhances data transparency, which produces a range of business value. At minimum, the transparency can lead to efficiency and cost savings by decreasing time to data acquisition, cleanup and pre-processing. It also improves quality assurance and security, which are foundational to innovating and building new value with data. Finally, this data transparency can help organizations seeking to comply with existing and emerging data protection laws and provisions in AI regulation. In sum, these standards are designed to increase trust in data use and sharing between organizations and with consumers.

How do these standards align with GDPR, CCPA, and other data protection laws?

These standards complement data protection laws by enhancing the ability to test compliance through clear data lineage and permitted use, including data storage and processing requirements and documentation. They do not guarantee compliance with any particular data protection law.

What are the technical prerequisites for implementing the standards?

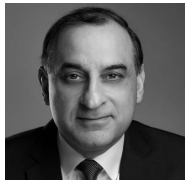
For data suppliers, the primary technical requirement is a website where you can publish metadata, which can be captured either through our web interface or a standalone spreadsheet. As a data consumer, there is no need for any technical infrastructure to read the metadata. However, if you intend to pass data downstream to other consumers, you will need the same infrastructure as a data producer, including capabilities for data logging and tracking to ensure compliance and maintain the integrity of the data provenance.

Are there specific technologies or platforms required?

No specific technologies are required, but for future-proofing, you might consider making systems capable of integrating with APIs and services that support metadata management and audit trails.

Why D&TA member company experts believe adoption is necessary

“As part of our commitment to healthy people and planet, our commitment to well-established data provenance standards is critical to our work providing the correct information to our customers and consumers for everything from sourcing to ingredients. The Data & Trust Alliance has helped us enable cross-industry importance of data provenance and has spearheaded common standards and policies for broad adoption. We are excited to be a member company and look to continue to contribute and leverage these new high standards across companies.”



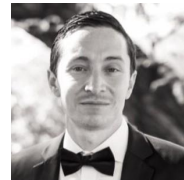
Ajay Dhau
Kenvue
Senior Vice President,
Global Data, Applied
AI and Digital Business
Transformation

“The newly announced Data Provenance Standards represent a substantial step forward for companies committed to sharing data with greater traceability and trust. This trust, which extends to AI insights and decisions, is bolstered when companies better understand data lineage and associated rights, allowing them to make informed, ethical decisions to grow their business and help consumers.”



Bernardo Tavares
Kenvue
Chief Technology &
Data Officer

“Content creators justifiably are entitled to be acknowledged for their contributions, especially as businesses require quality data to develop AI applications and run business tasks effectively. The Data Provenance Standards provide industry value for content creators by ensuring that creators’ rights and terms of use are known and respected. Moreover, for businesses, these attributes, along with the values, are critical to making informed choices about sourced data, including suitability for various purposes.”



Bryan Bortnick
IBM
Legal Counsel,
Intellectual Property
Lawyer

“Data is central to everything we do. Understanding where data came from, how it was acquired, and what it contains is essential to trusting what’s built on top.”



Bryan Kyle
IBM
Senior Technical
Staff Member, Data
Engineering, Chief
Data Office

Why D&TA member company experts believe adoption is necessary

“For data to have value, it must be trusted. For data to be trusted, it must have provenance and lineage back to trusted sources. This has never been more true than it is today, as new capabilities such as generative AI flood the business landscape. The Data Provenance Standards are an important foundational tool to help ensure that organizations can continue to make meaningful data-driven decisions.”



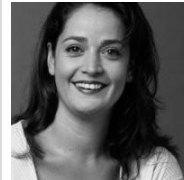
Chris Hazard
Howso
Co-founder and CTO

“As technology and AI are rapidly transforming industries, organizations need a blueprint for evaluating the underlying data that fuels these algorithms. Through the collaboration of experts across multiple industries and disciplines, the D&TA Data Provenance Standards meet this need. The standards promote trust and transparency by surfacing critical metadata elements in a consistent way, helping practitioners make informed decisions about the suitability of data sources and applications.”



Christine Pierce Nielsen
Chief Data Officer,
Audience
Measurement

“I am excited to see version 1.0.0 of the Data & Trust Alliance’s Data Provenance Standards, which mark a significant milestone in ensuring data transparency and accountability. At Humana, we are committed to upholding the highest standards of data integrity, and these standards will enhance the trust and reliability of the data we produce and consume across the enterprise to allow us to deliver value to the individuals we serve.”



Genevy Dimitrion
Humana
VP, Data Strategy &
Governance, Humana

“Participating in the development of the Data Provenance Standards provided AARP an excellent platform to encourage companies across diverse industries to consider the impact that advances in data and AI technologies have on people over 50.”



Jaye Campbell
AARP
SVP, Legal - Corporate,
Media, IP & Privacy

Why D&TA member company experts believe adoption is necessary

“The lack of data provenance consistency from one dataset to another is a pain point for organizations that build and use AI. This will be further accentuated as regulatory frameworks around the world require data origin disclosures. It is a game-changer to have organizations agree on a consistent methodology to use end-to-end across the data ecosystem.”



Lee Cox
IBM
VP Integrated
Governance & Market
Readiness, IBM Office of
Privacy and Responsible
Technology

“The new Data Provenance Standards are key to making data more reliable, not just for us at UPS, but for our customers and their supply chains. We’ve strengthened our own standards while collaborating with forward-thinking leaders across industries, and companies and consumers around the world will benefit from this work.”



Dr. Mallory Freeman
UPS
VP, Enterprise Data
and Analytics

“Data provenance standards are important for the entire data ecosystem. Beyond simplifying ingestion and use of data, use of the D&TA Data Provenance Standards, particularly by upstream data providers, will allow analysis of appropriateness, consent, and quality of aggregated datasets in a way that we have not previously had.”



Mike Meehan
Howso
General Counsel
& Chief Legal Officer

“The Data Provenance Standards will enhance transparency about the quality, origin, and intended uses and restrictions of datasets, which will help enterprises more rapidly access trustworthy data.”



Orla Flannery
IBM
Privacy Program
Manager, Chief Privacy
Office

Why D&TA member company experts believe adoption is necessary

“These Data Provenance Standards are so important for Transcient to be able to establish trust. We see them as another layer of quality assurance processes that are now required, at least within Transcient, when we’re looking at the data to ensure not only the accuracy of data, but the usability of it.”



Thi Montalvo
Transcient
SVP Reporting and
Analytics, Transcient

“Safe adoption of future AI tools will require trust and transparency in the data powering them. Cross-industry collaboration toward a universal set of data provenance standards is a key component of leveraging data effectively and responsibly.”



Thomas Birchfield
Transcient
Technical Program
Manager

“Trust in the data is based on our knowing that the data was sourced appropriately, is of good quality and has the consents necessary to be used. These Data Provenance Standards are an important step forward to ensure metadata about the sourcing, quality, and permissions are provided in a consistent manner, eliminating manual efforts which can introduce business risk.”



Travis Carpenter
Mastercard
SVP, Data Quality and
Sources



Learn more on our website at
dataandtrustalliance.org