An IBM Office of Privacy and Responsible Technology Case Study

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

IBM

Executive summary

Building trustworthy AI depends on trustworthy data. As IBM builds AI systems for a greater breadth of use cases, we need to determine whether the increasingly large volumes of data that train and test these models align with our high standards for trust and transparency. We also need to align with those standards efficiently, quickly clearing new data sets so that our teams have ready access to an expanding and diverse catalog of quality data.

IBM's data governance program already includes a data clearance process that enables us to apply relevant controls, document lineage, and define guidelines for use and re-use. For example, IBM's Granite foundation models are some of the most transparent in the world, thanks in part to their conformity to data governance and risk criteria enabled through our data clearance review process. But to respond to an increasing volume of data clearance requests, we looked to optimize that process for greater efficiency and accuracy. To that end, we co-created and tested the Data & Trust Alliance (D&TA)'s new Data Provenance Standards, the first cross-industry standards for metadata to describe where data comes from, how it was created, and its suitability for purpose. We wanted to determine if the Data Provenance Standards could help us accelerate access to trustworthy data by enhancing transparency about the quality, origin, and rights associated with data sets.

During our testing of the Data Provenance Standards, we saw improvement in overall data clearance review time. Our initial observations also signaled that they can lead to improvement in overall data quality. Because of this, we are now working to align our Business Data Standards with the D&TA Data Provenance Standards where appropriate to further optimize enterprise data governance.

The need for greater data transparency

"AI has a massive potential for good. It will help make us more productive as people and as a society," says IBM Chief Privacy & Trust Officer Christina Montgomery. "But AI can also cause real harm if it is not built or deployed responsibly." Without ethical and quality guardrails in place, AI systems can generate biased, unrepresentative, or otherwise flawed outcomes, potentially leading to regulatory fines and reputational damage for the organizations that build or use them — not to mention the potential harms to the people who are impacted by such systems. "When organizations develop AI systems without a holistic end-to-end view, they create risk," says Montgomery.

Organizations need data transparency to assess potential risk and make informed decisions about the data they choose to use in their AI systems. "A more precise view of the makeup of a data set can enable organizations to have more confidence in the insights and decisions coming from their AI systems," explains Lee Cox, Vice President for Integrated Governance and Market Readiness within the IBM Office of Privacy and Responsible Technology. "So, it is absolutely critical that any provider of AI systems understands the provenance of the data they're using. That includes the origin of the data, the lineage of the data — in other words, how it has moved through the data pipeline and been changed over time — and usage limitations associated with the data." All these details about a data set, in the form of metadata, help users assess the overall suitability of a data set for an intended purpose. Documenting the origin, lineage, and intended uses for data sets can enable organizations to create and use AI with greater confidence and less overall risk.

> AI has massive potential for good. It will help make us more productive as people and as a society. But AI can also cause real harm if it is not built or deployed responsibly. When organizations develop AI systems without a holistic end-to-end view, they create risk.
>
> **Christina Montgomery**
> Chief Privacy & Trust Officer
> IBM Office of Privacy and Responsible Technology

IBM's opportunity: Meeting an increasing demand for trustworthy data

> It would be a game changer if organizations could agree on a consistent [data provenance] methodology and framework to use end-to-end across the data ecosystem.
>
> **Lee Cox**
> Vice President for Integrated Governance and Market Readiness
> IBM Office of Privacy and Responsible Technology

Data governance at IBM

A long-standing commitment to trust and transparency is central to IBM's work to build responsible AI systems. Enterprise data governance is a critical component of that work. "IBM's internal AI strategy is predicated on enterprise data information architecture and strong enterprise data standards and governance practices," explains Ed Lovely, IBM Vice President for Enterprise Data. A responsible approach to data governance can include documenting how data is used across the enterprise while applying relevant controls and ethical guardrails. Documenting details related to data provenance is enabling IBM to build an inventory of what goes into the AI systems we create and use.

The challenge of incomplete, inconsistent data set metadata

Tracking and verifying data provenance is an important yet often time- and resource-consuming aspect of data governance. Public provenance information can be incorrect or missing, requiring manual follow-up to collect needed details. Even when full provenance details are provided, manual verification is sometimes required due to inconsistent metadata terms and definitions. The lack of data provenance consistency from one data set to another is a pain point for IBM and other organizations that build and use AI. "It would be a game changer if organizations could agree on a consistent methodology and framework to use end-to-end across the data ecosystem," says Cox.

Like others, IBM is experiencing ever-increasing internal demand for data as we develop and deploy new AI capabilities and use cases and expand our AI solutions across industries. Optimizing data clearance processes to meet that demand with greater efficiency — and without sacrificing standards for responsible data acquisition — would help make more quality data available to teams more quickly.

Developing and testing the Data & Trust Alliance Data Provenance Standards

Why IBM helped co-create the Data Provenance Standards

Universal, cross-industry data transparency standards that foster trust for data sets do not currently exist. To address this gap, the Data & Trust Alliance (D&TA) enlisted IBM and 18 other enterprises to co-create the Data Provenance Standards, the first cross-industry standards for data set metadata. As a not-for-profit consortium, D&TA is focused on developing practices for the responsible use of data and AI across all industries.

“These practical standards, co-created by senior practitioners across industry, are designed to help evaluate whether AI workflows align with ever-changing regulations while also helping generate increased business value,” says Rob Thomas, Senior Vice President, IBM Software and Chief Commercial Officer and chair of the D&TA Data Provenance initiative. “While the standards may not address every application of AI, we believe they fill an important, longstanding need.”

The goal of the Data Provenance Standards is to help organizations determine the suitability, representativeness, and trustworthiness of data sets through a common metadata taxonomy. The metadata associated with the Data Provenance Standards provides context so that organizations can assess trustworthiness and make more informed decisions about third party data they aim to use.

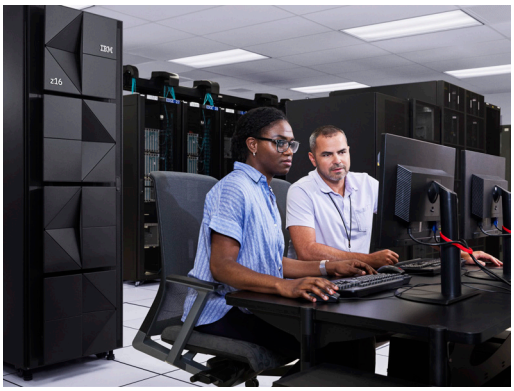
D&TA believes that for AI to fulfill its promise of creating new value for business and society, the data used to train it must be evaluated for transparency. “This belief is quickly becoming a reality,” explains Cox. “AI Acts and other regulatory activities around the world are driving policies that govern the use of AI systems with required data origin disclosures.” A common language for driving data transparency across companies and industries, and between data producers and consumers, is a critical first step toward facilitating trust and meeting current and anticipated AI regulations.

How IBM tested the Data Provenance Standards

As a D&TA member organization and early collaborator on the development of the Data Provenance Standards, IBM led the comprehensive testing and review of the Data Provenance Standards. Our testing centered on two key performance indicators (KPIs):

1. Quality of data: Are the Data Provenance Standards contributing to improvements in data quality?

2. Review processing time: Are the Data Provenance Standards contributing to a reduction in data clearance submission and review processing times?



First, we evaluated the comprehensiveness of the Data Provenance Standards. To do this, we compared the Data Provenance Standards to our own data intake requirements for data sets used to develop foundation models and assessed how adequately their metadata taxonomy enabled us to validate data suitability for four broadly applicable intended uses:

- Pre-Training
- Fine Tuning & Alignment
- Evaluation
- Synthetic data generation

Next, we evaluated the straightforwardness and comprehensibility of the Data Provenance Standards. To do this, we asked IBM data set developers and researchers of various levels of experience to apply the Data Provenance Standards to several common types of data sets, including:

- Data sets that have no third-party data (for example, data developed and owned by IBM)
- Data sets that include third-party proprietary data (for example, includes commercially licensed third-party data)
- Data sets that include HAP (hate speech, abusive language, and profanity) material or other explicit material

Lastly, experts from IBM's AI Ethics teams examined the completeness and accuracy of the metadata submissions in accordance with the Data Provenance Standards, reviewing the submissions with the developers and researchers to better understand any pain points or confusion. We observed that when there was difficulty in applying the Data Provenance Standards, it was generally not related to lack of knowledge or expertise, but rather how the Data Provenance Standards and their related guidance were presented. This enabled us to identify terms, definitions, and guidance that might be unclear or ambiguous and provide specific feedback and recommendations back to D&TA.

Throughout our testing, we translated our findings into actionable feedback, sometimes sharing our own taxonomies and data intake requirements to help inform revisions to the Data Provenance Standards and their accompanying guidance. For example, IBM recommended that the *Privacy and protection* standard should require the name of the specific tool(s) used to enhance data set privacy instead of requiring an indication of whether data is anonymized. We made this recommendation because providing an accurate answer requires an understanding of various legal definitions of, and regulatory requirements for, data anonymization. Another recommendation we shared was to make *Intended data use* a mandatory field to make suitability for purpose and license compliance easier to evaluate.

Driving improvements to quality and efficiency with the Data Provenance Standards

During our testing of the Data Provenance Standards, we saw improvement in overall data clearance review time. Our initial observations also signal that the Data Provenance Standards can lead to improvement in overall data quality. While concurrent technology and process enhancements also influenced these improvements, the Data Provenance Standards were a meaningful contributing factor. Tracking the lineage of a data set can be a prolonged process that requires diligence from all involved. We found that the Data Provenance Standards simplify that process because they enable trust by driving transparency to reduce the overall effort and resources required to help assess data lineage.

Because of the value we saw through testing, IBM is now working to more closely align its Business Data Standards with the Data Provenance Standards where appropriate. “Standardizing and expanding the taxonomy we use to describe and document data set metadata will continue to help facilitate more efficient data clearance review and improved content quality, enabling us to respond even more rapidly to increasing demand for data transparency,” says Cox.

Although it is too early in our testing to quantify other types of value, we anticipate that aligning with the Data Provenance Standards could help create operational efficiencies across the enterprise. “Greater transparency across the data ecosystem is a win for all,” summarizes Montgomery. For example, when more data sets have robust metadata attached, we anticipate that:

During our testing of the Data Provenance Standards, we saw improvement in overall data clearance review time.

- Developers could more easily compare data sets to determine which one best meets the requirements of their use case.
- Governance and compliance officers could more readily assess data sets against current or anticipated regulatory requirements because of the clearer and more complete auditing trails enabled by an expanded metadata taxonomy.
- Cybersecurity teams could more comprehensively assess and mitigate potential risk when they have a clearer view of the data protection measures

Conclusion

IBM believes that all technology, including AI, must be transparent and explainable. In practice, this means that organizations should bring clarity to who trains their AI systems, what goes into their algorithms' recommendations, and what data was used in training. This objective can be furthered when organizations have visibility into the provenance of the data sets used to train and test their AI systems — which can be achieved more efficiently through a standard, common, comprehensive metadata taxonomy.

By filling a critical gap, D&TA's Data Provenance Standards foster greater trust across the data ecosystem, helping organizations make informed choices about data that will ultimately contribute to the development of more trustworthy AI systems. "We really believe that IBM's role is not just developing practices for us to use, but also to help find solutions so that more organizations across the globe can be responsible stewards of technology," says Montgomery. IBM is proud to contribute to the development of the Data Provenance Standards and welcomes the transparency they will foster across the ecosystem.

By filling a critical gap, D&TA's Data Provenance Standards foster greater trust across the data ecosystem, helping organizations make informed choices about data that will ultimately contribute to the development of more trustworthy AI systems.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America

June 2024

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

All client examples cited or described are presented as illustrations of the manner in which some clients have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions. Contact IBM to see what we can do for you.

It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Statement of Good Security Practices: IT system security involves protecting systems and information through prevention, detection and response to improper access from within and outside your enterprise. Improper access can result in information being altered, destroyed, misappropriated or misused or can result in damage to or misuse of your systems, including for use in attacks on others. No IT system or product should be considered completely secure and no single product, service or security measure can be completely effective in preventing improper use or access. IBM systems, products and services are designed to be part of a lawful, comprehensive security approach, which will necessarily involve additional operational procedures, and may require other systems, products or services to be most effective.

IBM DOES NOT WARRANT THAT ANY SYSTEMS, PRODUCTS OR SERVICES ARE IMMUNE FROM, OR WILL MAKE YOUR ENTERPRISE IMMUNE FROM, THE MALICIOUS OR ILLEGAL CONDUCT OF ANY PARTY.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.