Data & Trust Alliance

Thank you for your interest in the Data Provenance Standards.

We are currently in the process of testing our proposed standards with Member companies.

As part of this process, please take our short survey to provide input on our metadata and help us to build out our values library.

TAKE THE SURVEY

Access the survey on mobile or share the link:

https://bit.ly/DataProvSurvey

© 2023, The Data And Trust Alliance. All rights reserved.

Standards, applied at the dataset level

SET IDENTIFIER

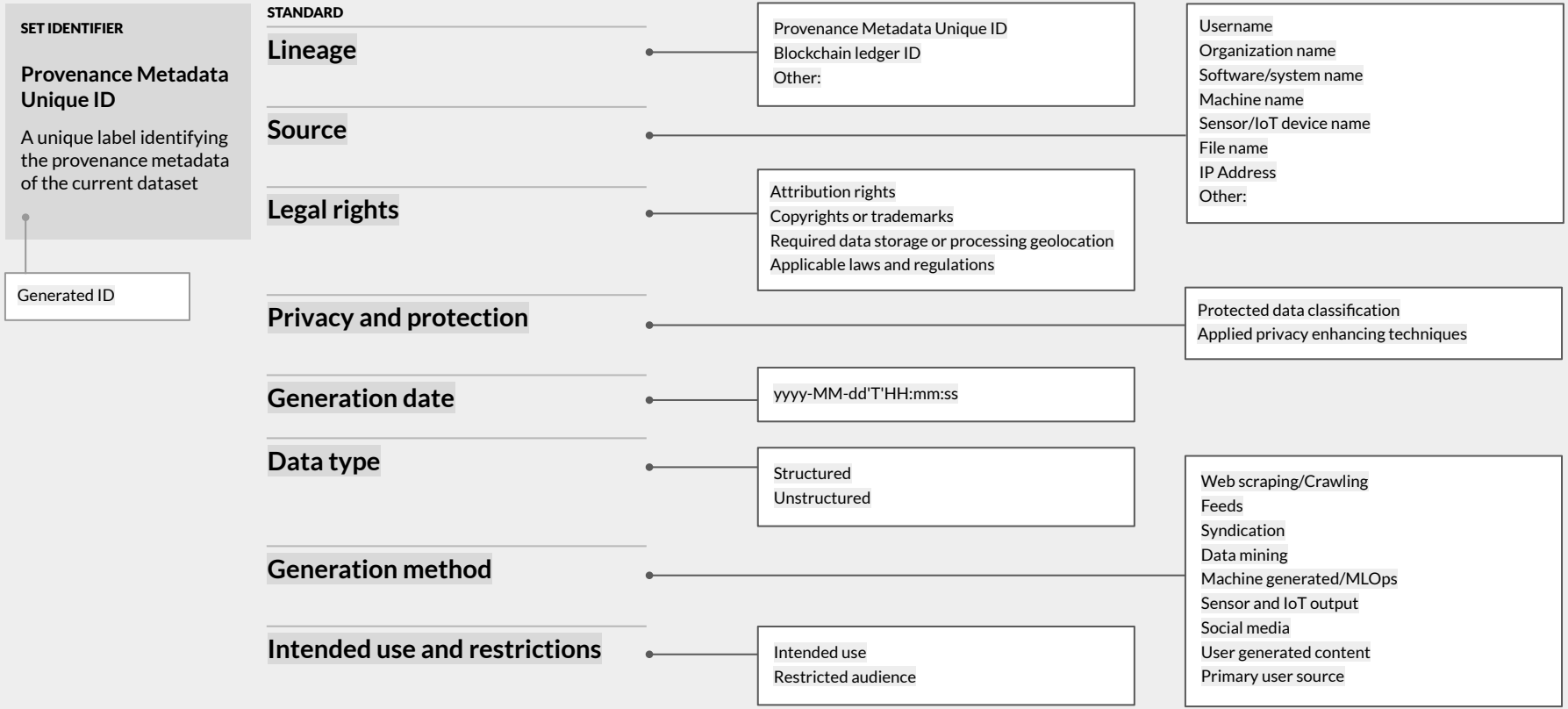
Provenance Metadata Unique ID

A unique label identifying the provenance metadata of the current dataset

STANDARD	DESCRIPTION
Lineage	Identifiers or pointers of metadata representing the data which comprise the current dataset
Source	Identifies the origin (person, organization, system, device, etc.) of the current dataset
Legal rights	Identifies the legal or regulatory framework applicable to the current dataset, along with the required data attributions, associated copyright or trademark, and localization and processing requirements
Privacy and protection	Identifies any types of sensitive data associated with the current dataset and any privacy enhancing techniques applied
Generation date	Timestamp marking the creation of the current dataset
Data type	Identifies the data type contained in the current set, and provides insights into how the data is organized, its potential use cases, and the challenges associated with handling and using it
Generation method	Identifies how the data was produced (data mining, machine-generated, IoT sensors, etc.)
Intended use and restrictions	Identifies the intended use of the data and which downstream audiences should not be allowed access to the current dataset

© 2023, The Data And Trust Alliance. All rights reserved.

Each standard has metadata



© 2023, The Data And Trust Alliance. All rights reserved.

Use Cases and Scenarios

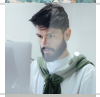
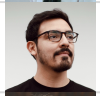
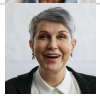
The following pages share a fictitious use case and seven scenarios.

Each showcases **what decisions the standards can inform, and what they can't.**

© 2023, The Data And Trust Alliance. All rights reserved.

Use Case (fictitious)

- **Humboldt Inc.** is researching the correlation between Hepatitis C in homeless women and Hepatitis C in homeless children in Humboldt County, California.
- The company uses predictive AI modeling to find undiagnosed patients and shares data with local hospitals, pharmaceutical companies, academic institutions, and government agencies to raise awareness and solve the health crisis.
- Humboldt Inc. is owned by Humboldt German GmbH and is using proprietary ML algorithms and tools to perform its AI modeling.
- The company sources data regularly for its predictive analysis calibration and is currently evaluating a dataset offered by Humboldt College.

		Role	Use of data
1		Data scientist	Develop predictive models
2		Data engineer	Intra-company data transfer for modeling
3		Product manager	Identify the best AI data supplier
4		Analytics and insights director	Address complex LLM data needs
5		Procurement specialist	Secure healthcare reference datasets
6		Data governance manager	Improve the quality and value of core data assets
7		Assistant general counsel	Address regulatory protections and business strategic needs

© 2023, The Data And Trust Alliance. All rights reserved.

Scenario 1: Maya, a data scientist, wants to build a predictive model

KEY CONCERNS

1. Do we have the appropriate rights in place to feed this data into the Hepatitis C predictive model and share the outputs with local hospitals, such as the local federal government clinic?
2. Who is the data supplier?
3. Is the data supplier also the data owner or did the data supplier source the data from other agents, such as local clinics and hospitals which may not have used the same type of data collection standards?
4. How fresh is the data?
5. How was the data generated?
6. What is the intended use of this data?

How the Standards increase transparency

STANDARD	METADATA & VALUES		HELPS WITH
Provenance ID	6f0a1eea-11bd-4a2b-8038-21b1829c72e0		4
Lineage	604ED310-5A03-4808-8076-97A9DB889FD1 CBC237E0-0C1F-47C4-834B-3882B98F9F53 E7630C09-6B98-4AEA-85F2-44EFA88C4671		2 3 4
Source	Humboldt College		1 2
Legal rights	Attribution rights	None	1
	Copyrights or trademarks	None	1
	Required data storage or processing geolocation	None	1
	Applicable laws and regulations	HIPAA	1
Privacy & protection	Protected data classification	Protected Health Information (PHI)	1
	Applied privacy enhancing techniques	Data transformation (homomorphic encryption)	1
Generation date	2023-02-03 20:17:46		4
Data type	Structured	sql csv	5
	Unstructured	txt	5
Generation method	Data mining - Anomaly detection	CS decision intelligence engine	5
	Feeds - Interval timed database info	InfluxDB 234512 Unx 1600000000	5
	User generated - Survey/Textual	Humboldt Survey_human01_series2 Humboldt Survey_human02_series5	5
Intended use & restrictions	Intended use	Machine learning	6
	Restricted audience	Federal government	1 6

Outcome

STANDARDS INFORM DECISION

Maya chose not to use this data because it has components of unstructured data, contains PHI and she expected it to be inconsistent, requiring extra cleaning.

Her decision came from the visibility into the numerous ways the data was generated, along with finding previous datasets used for this one had their own multiplicity of data sources.

Furthermore, Maya realized that she couldn't share this data with the Federal government, which is a business impediment as Humboldt shares results with local hospitals including the local federal government clinic.

© 2023, The Data And Trust Alliance. All rights reserved.

Scenario 2: Clark, a data engineer, wants to make an intra-company data transfer for modeling

KEY CONCERNS

1. Can we legally use the data from our parent company?
2. Can we use the data in the US?
3. Can we share the outputs of the model with our stakeholders, which include local hospitals, such as the local Veterans Affairs Clinic?

How the Standards increase transparency

STANDARD	METADATA & VALUES		HELPS WITH
Provenance ID	{E64A6804-1E21-48F3-BB43-21CD82490B14}		
Lineage	None		
Source	DB_Humboldt_Master_Model123		
Legal rights	Attribution rights	None	1
	Copyrights or trademarks	Humboldt German GmbH	2
	Required data storage or processing geolocation	EU	1 2
	Applicable laws and regulations	GDPR Germany: BDSG	1
Privacy & protection	Protected data classification	Protected Health Information (PHI)	1 2
	Applied privacy enhancing techniques	Data generation (synthetic)	1 2
Generation date	2023-07-23 08:17:06		
Data type	Structured	Oracle DB	
Generation method	Machine generated/ MLOps - Synthetic	Adaptive Server Enterprise	
Intended use & restrictions	Intended use	Expert systems Machine learning Natural language processing	
	Restricted audience	None	3

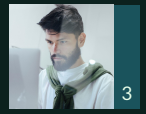
Outcome

STANDARDS INFORM DECISION

Clark chose to utilize this data due to its single source, synthetic generation, allowing it to be freely transferred to the US without special safeguards.

While Humboldt's parent company holds copyrights in the database, the model's outcomes can be shared with stakeholders.

© 2023, The Data And Trust Alliance. All rights reserved.



Scenario 3: Peter, a product manager, wants to identify the best AI data supplier

KEY CONCERNS

1. How many sources of data contributed to the compilation of the existing dataset?
2. How fresh is the dataset?
3. What is the generation method of the data?
4. What legal rights do we have to use the dataset?
5. In what way can I use the data being offered?
6. Is there sensitive data in the set?
7. Are there downstream data sharing restrictions?

How the Standards increase transparency

STANDARD	METADATA & VALUES		HELPS WITH
Provenance ID	16488554-0EE5-460C-BFDF-5606174B24A9		
Lineage	44f1557992574ce78ab8fdd60630a97e 36c735701aa8432d8a029d24edbca044 4c42b409420046f89ba759e3512ceafc {ae5027957c494e46b4465cb59d75a513} 170141adf02e49e5b2174d2107baba4f F4BB8E04-8086-4F24-9B89-9483D7FB2A43 1caf703d68c740158b8247d3062f0eab		1
Source	AGBO		
Legal rights	Attribution rights	None	4
	Copyrights or trademarks	WebMD	4
	Required data storage or processing geolocation	None	4
	Applicable laws and regulations	None	4
Privacy & protection	Protected data classification	None	4 6
	Applied privacy enhancing techniques	None	6
Generation date	2020-07-23 08:17:06		2
Data type	Structured	MySq	
	Unstructured	txt	
Generation method	Machine generated/MLOps - Generative Social Media - Reviews and ratings		3
Intended use & restrictions	Intended use	Machine learning (ML)	5 7
	Restricted audience	None	5 7

Outcome

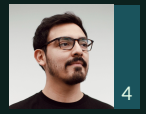
STANDARDS INFORM DECISION

Peter concludes the data supplier is a data broker and not the original data producer, as evident from the lineage sources.

The data was created using generative AI and web scraping, both inadequate for understanding Hepatitis C patterns locally. The stringent usage restrictions raised more concerns.

The outdated data source and the copyright owner eroded Peter's trust in the data. **Consequently, he's exploring other suppliers.**

© 2023, The Data And Trust Alliance. All rights reserved.



Scenario 4: Sal, an analytics and insights director, is addressing complex LLM data needs

KEY CONCERNS

1. How was this data generated?
2. When was the dataset generated?
3. How many parameter points does each dataset offer?
4. How many sources of data have been merged into this single set?
5. Has the data been tailored to our vertical or use cases?
6. Is this dataset a fit for my LLM needs?

How the Standards increase transparency

STANDARD	METADATA & VALUES		HELPS WITH
Provenance ID	F883E328-076C-424B-9F67-4B4EE30F5999		
Lineage	Hash1109a556751d8582dac62ef0d9fbaeaa4656		1 2 4
Source	70.106.244.139		
Legal rights	Attribution rights	Ryan Smith (OmniSol Extract Script)	
	Copyrights or trademarks	OmniSol Klipfolio	
	Applicable laws and regulations	Health Insurance Portability and Accountability Act California Consumer Privacy Act Children's Online Privacy Protection Act	
Privacy & protection	Protected data classification	Personally Identifiable Information (PII) Protected Health Information (PHI)	
	Applied privacy enhancing techniques	Data anonymization	
Generation date	2023-08-01 02:01:02		2
Data type	Structured	MySql	1
	Unstructured	txt	1
Generation method	Feeds - API source	OmniSol ERM	1
	Feeds - File feed info	Klipfolio ERM	
	Web scraping - Textual	HealthWeb chat groups	
Intended use & restrictions	Intended use	Machine learning Natural language processing Expert systems Vision Speech Planning Robotics	6
	Restricted audience	None	

Outcome

STANDARDS DO NOT INFORM DECISION

While Sal was able to understand how and when the data set was generated, and how many sources of data have been merged into the single data set, he was unable to answer:

3. How many parameter points does the data set offer?
5. Has the data been tailored to our vertical or use cases?

The standards are not currently designed to answer these questions.

© 2023, The Data And Trust Alliance. All rights reserved.

Scenario 5: Fei, a procurement specialist, is securing healthcare reference datasets for the billing department

KEY CONCERNS

1. Is the data single sourced or generated from multiple sources?
2. When was the data generated?
3. How was the data generated?
4. Does the dataset contain any sensitive or restricted data?
5. Are there any special laws or regulations that apply to this data?

How the Standards increase transparency

STANDARD	METADATA & VALUES		HELPS WITH
Provenance ID	CE804E5D-A8AD-41C8-B118-42B855B33168		
Lineage	none		1
Source	Centillion		
Legal rights	Attribution rights	None	
	Copyrights or trademarks	Centillion	
	Required data storage or processing geolocation	None	
	Applicable laws and regulations	None	5
Privacy & protection	Protected data classification	None	4
	Applied privacy enhancing techniques	None	
Generation date	2023-06-06 06:08:05		2
Data type	Structured	AmazonRDS	
Generation method	Feeds - API source	https://api.centillion.com	3
Intended use & restrictions	Intended use	Data processing	
	Restricted audience	None	

Outcome

STANDARDS INFORM DECISION

Fei bought the healthcare reference dataset for the billing department. The data is typical data processing information, not subject to any laws or regulations.

Centillion, the data supplier, is the original source with no additional lineage. The data, obtained via an API, is structured, fitting common billing data methods.

© 2023, The Data And Trust Alliance. All rights reserved.

Scenario 6: Daksh, a data governance manager, is improving the quality and value of core data assets

KEY CONCERNS

1. What is the unique ID of this dataset?
2. Which elements within this dataset are duplicative of data that already exists in our data lake?
3. If I encounter a patient record in this dataset and the record is identical to one in our data lake but the metadata differs, should I replace the patient record with this new data?
4. The source of the dataset is August 6, 2023, but when were the individual data elements collected?

How the Standards increase transparency

STANDARD	METADATA & VALUES		HELPS WITH
Provenance ID	e389f87a-41fa-4f06-ab31-8a62e2d5d541		1
Lineage	{42d73792-6882-427a-973b-4af9b0729321} {32877DED-800E-491E-96DF-543F0221296B} B50FF8CE-4E17-4802-A1E0-19D06CB68AE7		
Source	Attain		
Legal rights	Attribution rights	None	
	Copyrights or trademarks	None	
	Required data storage or processing geolocation	None	
	Applicable laws and regulations	None	
Privacy & protection	Protected data classification	None	
	Applied privacy enhancing techniques	None	
Generation date	2023-08-06 01:01:01		
Data type	Structured	xml DynamoDB json	
Generation method	Feeds - RSS source Feeds - API source Feeds - Real time database info Feeds - Interval timed database info Feeds - File feed info	https://rss.centillion.com https://api.centillion.com	
Intended use & restrictions	Intended use	Machine learning Natural language processing Expert systems Vision Speech	
	Restricted audience	None	

Outcome

STANDARDS DO NOT INFORM DECISION

Daksh couldn't use the standards to enhance core data quality and value.

These standards aim to show data origins and supplier trustworthiness based on usage rights and methods. They don't serve as operational quality controls.

© 2023, The Data And Trust Alliance. All rights reserved.

7

Scenario 7: Valentina, assistant general counsel, is addressing regulatory protections and business strategic needs

KEY CONCERNS

1. When we acquire new data is it from known sources?
2. Do we acquire the minimum amount of data necessary for our business purposes?
3. Are we acquiring data that allows us to own the intellectual property rights associated with products built with that data?
4. Are we associating ourselves with third parties that pose a brand reputation risk based on consumer sentiment?

How the Standards increase transparency

STANDARD	METADATA & VALUES		HELPS WITH
Provenance ID	{EDB89C41-3BAC-4C52-927C-26DC151F7D42}		
Lineage	6af8d988-a8f7-4587-b784-280f2da47c2e 2b34f476-ba07-41be-877d-0fb64447a736 Cf03837f-203f-444c-ae05-946bdc83508f {42d9aa7b-24b3-40cf-b91f-acea57b5abf9}		1 4
Source	Acxiom		1 4
Legal rights	Attribution rights	None	3
	Copyrights or trademarks	PatientPop	3
	Required data storage or processing geolocation	EU	
	Applicable laws and regulations	Argentina: Personal Data Protection Act (PDPA) European Union: General Data Protection Regulation (GDPR)	3
Privacy & protection	Protected data classification	Personal Financial Information	2
	Applied privacy enhancing techniques	None	
Generation date	2023-04-06 12:13:11		
Data type	Structured	DB2 Microsoft Access html	
	Unstructured		2
Generation method	Feeds - API source Feeds - Interval timed database info Syndication - Social media		
Intended use & restrictions	Intended use	Machine learning Natural language processing	
	Restricted audience	None	

Outcome

STANDARDS INFORM DECISION

Valentina evaluated the risks and benefits of using data from this supplier. She collaborated with the data governance manager to identify primary and secondary data suppliers, assessing potential reputational and regulatory risks.

She observed that the company retains unnecessary consumer financial data, which raises exposure in case of a breach.

Though the data's provenance metadata didn't provide a complete risk assessment, **Valentina quickly grasped the layout to pinpoint areas needing more audits.**

© 2023, The Data And Trust Alliance. All rights reserved.

About the Data & Trust Alliance

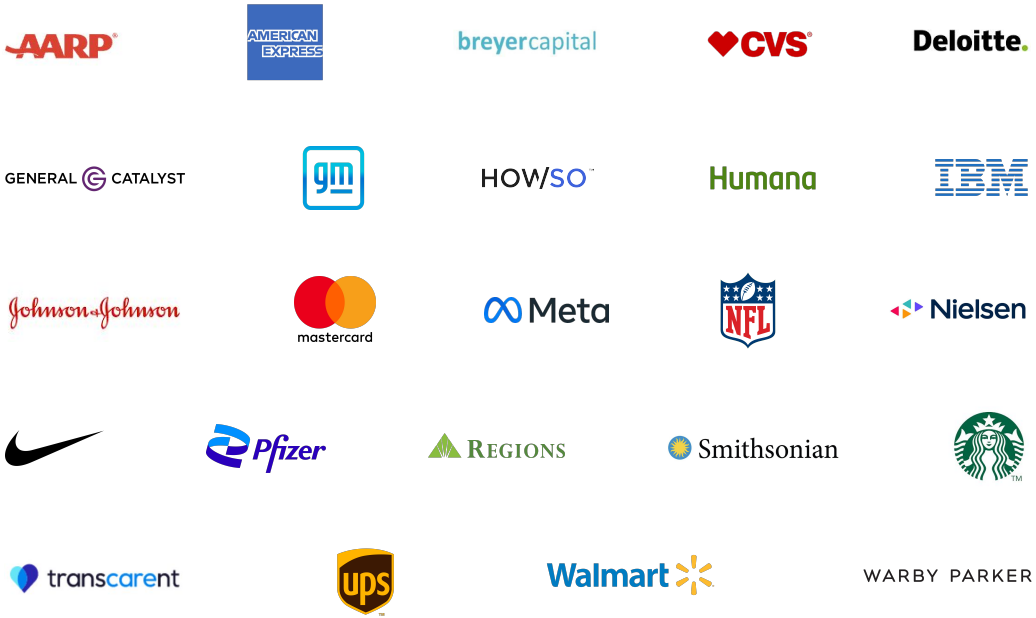
In 2020, a group of CEOs of major companies established the Data & Trust Alliance based on a shared conviction: *the future of business will be powered by the responsible use of data and AI.*

Since then, the Alliance has leveraged the collective expertise and influence of its member organizations to learn, develop and adopt responsible data and AI practices.

4.6M
employed by Alliance member organizations

$3.5T+
market capitalization of Alliance companies

$1.9T+
revenue of Alliance companies in 2022



© 2023, The Data And Trust Alliance. All rights reserved.

Learn more at www.DataAndTrustAlliance.org

November 2023