DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications

Metadata in Trustworthy AI: From Data Quality to ML Modeling

Jian Qin
Syracuse University, USA
jqin@syr.edu

Bei Yu
Syracuse University, USA
byu@syr.edu

Abstract

Metadata play a significant role in making AI models trustworthy by providing information on input, output, models, pipelines, and other artifacts to meet the requirements for trustworthy AI. This concept paper focuses on what role metadata play in an AI lifecycle and how metadata research can ride out this AI wave with innovative creations. Specifically, we explore metadata's role and potential related to data quality and ML models. The multidimensionality of metadata for data in AI is driving metadata to be micro-specific, embedded in data and models, highly computational, and fast-moving or agile. While there are no universally agreeable metadata schemas for documenting the artifacts in ML model development, there are some common areas or types of metadata for ML models. Data quality and ML models are tightly connected and can impact one another in significant ways. Trustworthy AI must rely on quality data and responsible, ethical, reproducible, verifiable ML models, and the assurance of these data and ML model properties relies on metadata. The complex, fast paced, and highly computational nature of metadata for AI artifacts (datasets, models, pipelines, algorithms, lineages, etc.) is making conventional metadata development processes and methods outdated, but meanwhile has prompted some innovative metadata creations.

Keywords: metadata for AI; trustworthy AI; data quality; ML model metadata

1. Introduction

The rapid development in artificial intelligence (AI) tools and applications brings excitements across academia and industry sectors. While experts in computing and data sciences are embracing the tremendous opportunities brought about by AI, policy makers and researchers are concerned with the negative impact or risks that could possibly result in harms to people, organizations, and ecosystems (Tabassi, 2023). The European Commission (EC) convened a group of high-level experts in AI to address the concerns and developed a framework for achieving trustworthy AI. This framework includes seven key requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability (European Commission, 2020b). In the assessment list for trustworthy AI that followed the EC framework, many assessment questions can only be answered by metadata. For example, detection and response mechanisms are one of the assessment questions for human oversight requirement. In order to facilitate the detection of errors or biases in data and/or machine learning (ML) models, there must be metadata to allow human or machine to trace and analyze the lineages of data processing and transformation as well as machine learning pipelines and models (European Commission, 2020a).

Many factors can contribute to the trustworthiness of AI. Validity and reliability are essential conditions of trustworthiness and on which other trustworthiness characteristics are built: safety, security and resilience, accountability and transparency, explainability and interpretability, privacy-enhanced, and fairness with harmful bias managed (Tabassi, 2023). In the lifecycle and key dimensions of an AI system (Fig. 1), metadata play a significant role in making AI models trustworthy, whether it is in describing the entities—actors such as people and computer agents, artifacts such as datasets and AI models, or documenting methods for building and using models.

https://doi.org/10.23106/dcmi.953354037

Metadata enhances traceability, one of the key requirements for trustworthy AI, because it is related to methods used for designing, developing, testing, and validating the algorithmic system as well as the outcomes of the algorithms or the subsequent decisions on these outcomes (Mora-Cantallops et al., 2021). Metadata also helps verify and enhance data quality, one of the challenges many organizations face but having difficulty in addressing. Research has found that training data related issues are constraining AI project success – 96% of respondents encountered data quality and labeling challenges (Dimensional Research, 2019). Data is a major obstacle for 90% of firms to scale AI across their enterprises (Forrester Consulting, 2020).

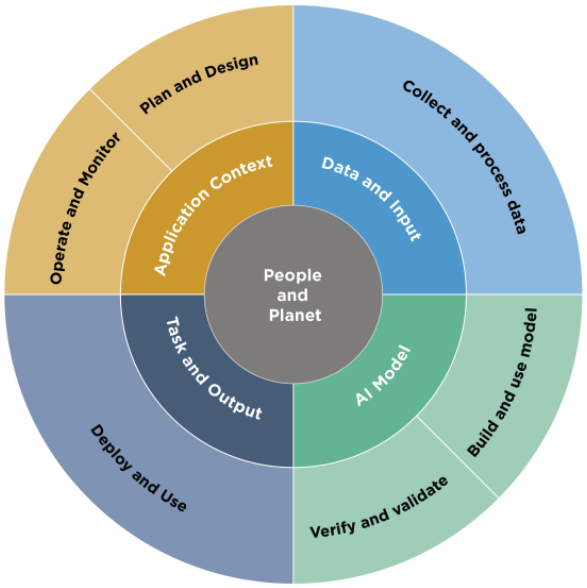

FIG 1. Lifecycle and Key Dimensions of an AI System. The two inner circles show AI systems' key dimensions and the outer circle shows AI lifecycle stages (Tabassi, 2023)

This paper is a concept paper that focuses on what role metadata plays in an AI lifecycle and how metadata research can ride out this AI wave with innovative creations. As such, we will explore metadata's role and potential related to data quality and ML models.

2. Policy and Research in Trustworthy AI

While AI has tremendous potential for benefiting human, society, and environment, it is also a double-edged sword that can generate harms if not regulated and controlled by policy and legislation. As mentioned earlier in this paper, the European Commission issued guidelines for trustworthy AI, in which lawful, ethical, and robust AI is emphasized (European Commission, 2020a & 2020b). In the U.S., the Biden administration has taken steps to promote responsible AI innovations, including the Blueprint for an AI Bill of Rights issued by the White House Office of Science and Technology Policy (2022) and the AI risk management framework (Tabassi, 2023) developed at the National Institute of Standards and Technology. The Blueprint is "a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intelligence" (Tabassi, 2023, p. 4). These five principles require AI systems to be safe and effective, protect humans from algorithmic discrimination, protect data privacy, keep users notified that an automated system is being used and let them understand how and why it contributes to outcomes that impact them, and allow human users alternatives, consideration, and fallback (White House Office of Science and Technology Policy, 2022).

These policies and guidelines issued by governments at national and international levels are the fruit of research and advocates from experts/scholars in the computing community. In the computing community, trustworthy means a set of properties:

- reliability—does the system do the right thing?
- safety—does the system do no harm?
- security—how vulnerable is the system to attach?
- privacy—does the system protect a person's identity and data?
- availability—is the system up when I need to access it?

https://doi.org/10.23106/dcmi.953354037

- usability—can a human use it easily? (Wing, 2021)

Wing continues to indicate that AI systems raised these properties beyond trustworthy computing. The trustworthy AI would desire properties such as accuracy, robustness, fairness, accountability, transparency, interpretability/explainability, and ethical, among others (Wing, 2021). Claims made by AI developers should be verifiable through institutional, software, and hardware mechanisms (Brundage et al., 2020) so that whether AI systems satisfy all of the trustworthy AI requirements can be tested and validated (Kaur et al., 2023).

There are quite consistent views and opinions on what trustworthy AI means and requires in the policy arena and research communities. However, how to enforce the guidelines and enact the requirements from the start of AI development lifecycle is still an area to be explored. For the metadata community, this is an opportunity to innovate and build new metadata models, theories, tools, and practices. Researchers in metadata have already started investigating metadata's role in supporting computing research, for example, Leipzig and colleagues examined the metadata standards in the biomedical domain by metadata level and content and noted the embeddedness of metadata in the data files and connection between metadata and reproducibility (Leipzig et al., 2021).

This paper will focus on two important areas for trustworthy AI, data quality and ML models, and discuss the role of metadata and approaches in developing metadata that can satisfy the requirements of trustworthy AI.

3. Data Quality and Metadata

The quality of training data has great impact on the AI models trained on them. However, many details about training data were lost in scholarly communication. For example, Sentiment140 is a famous data set for training sentiment classifiers (Go et al., 2009). This data set was developed by a group of students at Stanford, who downloaded 1.6M tweets and annotated their sentiments using emoticons as proxy labels. Although the authors explicitly explained that the labels are "noisy", the dataset was commonly used as a benchmark for evaluating sentiment classification models and has been cited more than 4,000 times with rare discussions on the source of the labels and potential data quality issues. This is just a mild example of the lack of metadata for describing training data quality. The problem is more acute for high-stake tasks such as detecting fake news and moderating toxic comments.

The development of AI has shown two kinds of approaches: the model-centric and data-centric. The former takes data as a given and emphasizes on improving or optimizing the model based on the data, while the latter focuses on scalable methods to systematically improve the data pipeline (Liang et al., 2022). A data-centric approach to AI involves a series of steps in designing, sculpting, and strategizing data for model development and testing. Data design involves identifying and documenting the sources of data, which is critical for mitigating bias and ensuring. From data to algorithm in an AI development lifecycle, there could be biases in data related to measurement, variables, representation or coverage, aggregation, sampling, longitudinal data fallacy, and linking (Mehrabi et al., 2022). Data bias in representation and coverage can be mitigated by synthetic data, i.e., data generated by using AI techniques. Examples include using human faces produced by deep generative AI models to protect individual's privacy and synthetic medical records to avoid disclosing information that could be used to identify patients (Nikolenko, 2019). An accurate and full documentation of, or metadata for data design in relation to data context, measurement, variable, coverage, and sampling will be critical for trustworthy AI, because AI models' performance as well as reproducibility and traceability are highly dependent on data quality.

The lack of metadata for datasets used in AI model development and testing has been a concern among computing researchers and strategies have been developed to address the concern. Examples include the templates for reporting datasets for standardizing the report on data analysis methods and results (Holland et al., 2018), Data Version Control for datasets and ML models (https://dvc.org/), repositories for curating datasets with human-created metadata, data annotation

https://doi.org/10.23106/dcmi.953354037

tools such as Label Studio (https://github.com/heartexlabs/label-studio), data quality assurance frameworks and tools such as TensorFlow Data Validation (https://github.com/modAL-python/modAL), and data-centric benchmarks such as (https://dynabench.org/). In the natural language processing (NLP) community, datasets rarely provide detailed data statement about the speakers whose data is captured such as age, gender, race/ethnicity, native language, socioeconomic status, and presence of disordered speech (Bender & Friedman, 2018). Bender and Friedman (2018) argue that the data statement (a.k.a. metadata) for NLP datasets should capture the rationale for curating the language datasets, language variety, speaker demographic, annotator demographic, speech situation, text characteristics, recording quality and others. IBM proposed "factsheets" to help increase trust in AI services. A factsheet describes the purpose, performance, safety, security, and provenance information for AI service consumers (Arnold et al., 2019). Borrowing the data provenance concept from the database community, the "Datasheets for Data Sets" framework asked dataset creators and consumers to examine the motivation, composition, collection process, pre-processing/cleaning/labeling, uses, maintenance, distribution (Gebru et al., 2021). The benefits that metadata for datasets in AI model development brings are twofold: on the one hand, metadata helps AI model developers to maintain a balance between data-centric and model centric approaches to enhance the trustworthiness of AI models generated. On the other hand, the availability of metadata for data sources facilitates the discovery, selection, and use/reuse of data and better understand the data characteristics and quality, which can help save time and improve AI model development effectiveness and efficiency.

Although much of the effort in data quality control for AI is interpretable in metadata terms, the terminology used, and strategies deployed have gone beyond traditional metadata workflows and practices where the development of and compliance to metadata standards play a central role. Datasets used in ML model development, testing, and validating are in constant sculpting (training) through selection, cleaning, and annotation until they are suitable to be evaluated on the model's generalizability and trustworthiness. The multidimensionality of metadata for datasets in AI is driving metadata to be micro-specific, embedded in data and models, highly computational, and fast-moving or agile. Metadata graph, for instance, is a new approach to capture and present metadata for very large datasets. In the MetaShift project, Liang and Zou (2022) created subsets within a large image collection with annotation graph. They employed the network science method to identify nodes (e.g., cat) and various contexts the nodes are associated with, e.g., cat with sink, cat with faucet, based on the similarity/distance between two subsets (edge weight). The resulting meta-graph can be used to identify a collection of nodes (subsets, e.g., cat) for training (e.g., cats with outdoor) inside the large collection to enhance model's accuracy and trustworthiness (Liang & Zou, 2022). Google Cloud also employs graph method to represent metadata produced and consumed from ML workflows. This metadata graph connects the artifacts, executions, events, and contexts related to ML models in one presentation for tracking, analyzing, managing, and using metadata (Google Cloud, 2023). These examples illustrate the changes in metadata creation and use in the AI era and the need for innovating metadata methods and practices.

4. ML Models and Metadata

So far the guidelines and policies issued by government and learned societies have been broadly referring to ML models primarily. ML models as the outcomes from the orchestration of data, algorithms, parameters, and methods are the main target for scrutinization against the trustworthy AI requirements. Metadata for ML models therefore are designed to answer questions about the artifacts, methods, and processes. Who built the model by using which artifacts or datasets? How do I compare experiments, and which is the winning model? When things break, which model can I roll back to? How can I identify the model's lineage? What are the hyperparameter settings? Which pipeline is being used? And the list can go on. Since metadata generation requires significant amount of time and effort, there is also a need to determine which metadata items are the most critical for trustworthy AI.

https://doi.org/10.23106/dcmi.953354037

To document properties of AI models for transparency and ethical use, Mitchell et al. (2019) proposed "model cards" for reporting performance characteristics of released models. A model card includes model details (such as version, developer, citation, and contact person), intended use (such as primary use and out-of-scope use scenarios), performance-relevant factors (such as use groups, instrumentation, and environment), training and evaluation data, ethical concerns, etc. Model-centric metadata, combined with data-centric metadata, can then facilitate the inference of trustworthiness. For example, whether a disease prediction model is generalizable from one patient group to another. It remains an open question whether model cards can be one-size-fit-all, or domain-oriented customization is worth considering. For example, what constitutes similar and different metadata items for an AI model to predict greenhouse emission of factories and an AI model to predict patients' cancer risks?

While there are no universally agreeable metadata schemas for documenting the artifacts in ML model development, there are some agreements on what areas in an ML modeling lifecycle should be focused. In a review of management systems for ML lifecycle artifacts, Schlegel and Sattler classified the systems they reviewed into five categories: lifecycle management, pipeline management, experiment management, model management, and dataset & feature management. Although the review do not provide what metadata are included in the management systems, many of the systems in their review have a metadata layer for capturing artifacts from ML lifecycle (Schlegel & Sattler, 2023). Another way to look at the ML model metadata is based on the tracking needs: audit metadata, experiment metadata, metadata for datasets, lineage metadata, metadata on decision-making metrics, and metadata for versioning management (Manish Kumar, 2022).

Currently the effort for capturing metadata in ML modeling lifecycle is largely led by frontrunners in the industry such as Google, Amazon, and IBM. For example, the Vortex ML metadata is developed by Google Cloud to track the lineage of ML artifacts and downstream usage of these artifacts as well as to analyze production ML systems and experiments (Google Cloud, 2023). The metadata schema developed at Amazon has modules for dataset metadata, training run metadata, model metadata that includes hyperparameter and transform metadata, prediction metadata, and evaluation metadata. Within each of these modules, there are specific metadata items, for example, model metadata includes id, name, version, fromFramework, fromFrameworkVersion, trainedOnDatasetId, learningAlgorithm, hyperparameters (which is an array data type), location, transforms (array), and annotations (Schelter et al., 2017).

5. Conclusion

In this short paper we discussed the requirements of trustworthy AI and how metadata might contribute to meeting the requirements. The discussion focused on data quality and ML models as these two are tightly connected and can impact one another in significant ways. Such interdependency between data quality and ML models has been reiterated in research and industry publications. Trustworthy AI must rely on quality data and responsible, ethical, reproducible, verifiable ML models, and the assurance of these data and ML model properties relies on metadata. The complex, fast paced, and highly computational nature of metadata for AI artifacts (datasets, models, pipelines, algorithms, lineages, etc.) is making conventional metadata development processes and methods outdated, but meanwhile has prompted some innovative metadata creations as mentioned earlier in the paper. The metadata for trustworthy AI is moving in multiple directions with different flavors depending on the application domain and industry sector. Yet for the metadata community, there is much to be learned and explored. Should there be some sort of metadata standard(s) for AI artifacts in support of trustworthy AI? If so, what level of generality and specificity would be appropriate so that the AI developer community is willing to adopt it without dragging down the productivity? If any standard is to be developed for AI artifacts, what role will metadata community play in this process? Generative AI has made amazing progress in the last few years. Will metadata make similar breakthroughs while riding the AI wave? These questions offer great inspirations and opportunities for the metadata community to innovate in the coming years.

https://doi.org/10.23106/dcmi.953354037

References

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., … Anderljung, M. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. https://doi.org/10.48550/ARXIV.2004.07213

Dimensional Research. (2019). *Artificial Intelligence and Machine Learning Projects Are Obstructed by Data Issues: Global Survey of Data Scientists, AI Experts and Stakeholders*. Alegion. https://content.alegion.com/dimensionalresearchs- survey

European Commission. (2020a). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-ass.* https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

European Commission. (2020b). *Ethics and Guidelines for Trustworthy AI.* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Forrester Consulting. (2020). *Overcome obstacles to get to AI at scale: Invest in and scale AI to become an industry leader*. Forrester Consulting. https://www.ibm.com/downloads/cas/VBMPEQLN

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision* (CS224N project report, Stanford). https://www-cs-faculty.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Google Cloud. (2023). *Intorduction to Votext AI. Votex ML Metadata*. https://cloud.google.com/vertex-ai/docs/ml-metadata

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards*. https://doi.org/10.48550/ARXIV.1805.03677

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2023). Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, *55*(2), 1–38. https://doi.org/10.1145/3491209

Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. *Patterns*, *2*(9), 100322. https://doi.org/10.1016/j.patter.2021.100322

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Author Correction: Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, *4*(10), 904–904. https://doi.org/10.1038/s42256-022-00548-7

Liang, W., & Zou, J. (2022). *MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts*. https://doi.org/10.48550/ARXIV.2202.06523

Manish Kumar. (2022, March 24). Everything about Metadata and its meaning for Machine Learning practitioners. *InfuseAI*. https://blog.infuseai.io/everything-about-metadata-and-its-meaning-for-machine-learning-practitioners-cd353ca160d1

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M.-A. (2021). Traceability for Trustworthy AI: A Review of Models and Tools. *Big Data and Cognitive Computing*, *5*(2), 20. https://doi.org/10.3390/bdcc5020020

Nikolenko, S. I. (2019). *Synthetic Data for Deep Learning*. https://doi.org/10.48550/ARXIV.1909.11512

Schelter, S., Joos-Hendrik Böse, Johannes Kirschnick, Thoralf Klein, & Stephan Seufert. (2017, December 8). Automatically Tracking Metadata and Provenance of Machine Learning Experiments. *Workshop on Machine Learning Systems*. Neural Information PRocessing Systems 2017. http://learningsys.org/nips17/assets/papers/paper_13.pdf

Schlegel, M., & Sattler, K.-U. (2023). Management of Machine Learning Lifecycle Artifacts: A Survey. *ACM SIGMOD Record*, *51*(4), 18–35. https://doi.org/10.1145/3582302.3582306

Tabassi, E. (2023). *AI Risk Management Framework: AI RMF (1.0)* (NIST AI 100-1; p. error: NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

https://doi.org/10.23106/dcmi.953354037

White House Office of Science and Technology Policy. (2022). *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. White House Office of Science and Technology Policy. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

Wing, J. M. (2021). Trustworthy AI. *Communications of the ACM*, *64*(10), 64–71. https://doi.org/10.1145/3448248

https://doi.org/10.23106/dcmi.953354037