Towards Traceability in Data Ecosystems using a Bill of Materials Model

Iain Barclay, Alun Preece, Ian Taylor
Crime and Security Research Institute,
Cardiff University,
Cardiff, UK
Email: BarclayIS@cardiff.ac.uk

Dinesh Verma
IBM TJ Watson Research Center,
1110 Kitchawan Road,
Yorktown Heights,
NY 10598, USA

Abstract

Researchers and scientists use aggregations of data from a diverse combination of sources, including partners, open data providers and commercial data suppliers. As the complexity of such data ecosystems increases, and in turn leads to the generation of new reusable assets, it becomes ever more difficult to track data usage, and to maintain a clear view on where data in a system has originated and makes onward contributions. Reliable traceability on data usage is needed for accountability, both in demonstrating the right to use data, and having assurance that the data is as it is claimed to be. Society is demanding more accountability in data-driven and artificial intelligence systems deployed and used commercially and in the public sector. This paper introduces the conceptual design of a model for data traceability based on a Bill of Materials scheme, widely used for supply chain traceability in manufacturing industries, and presents details of the architecture and implementation of a gateway built upon the model. Use of the gateway is illustrated through a case study, which demonstrates how data and artifacts used in an experiment would be defined and instantiated to achieve the desired traceability goals, and how blockchain technology can facilitate accurate recordings of transactions between contributors.

References

[1] M. I. S. Oliveira, G. d. F. B. Lima, and B. F. Lóscio, "Investigations into data ecosystems: a systematic mapping study," *Knowledge and Information Systems*, pp. 1–42, 2019.

[2] N. Diakopoulos, "Accountability in algorithmic decision making," *Communications of the ACM*, vol. 59, no. 2, pp. 56–62, 2016.

[3] L. Byron, "Graphql: A data query language." [Online]. Available: https://code.facebook.com/posts/1691455094417024/graphql-a-data-query-language

[4] D. M. Lambert, M. C. Cooper, and J. D. Pagh, "Supply chain management: implementation issues and research opportunities," *The international journal of logistics management*, vol. 9, no. 2, pp. 1–20, 1998.

[5] L. U. Opara, "Traceability in agriculture and food supply chain: a review of basic concepts, technological implications, and future prospects," *Journal of Food Agriculture and Environment*, vol. 1, pp. 101–106, 2003.

[6] T. Kelepouris, K. Pramatari, and G. Doukidis, "Rfid-enabled traceability in the food supply chain," *Industrial Management & data systems*, vol. 107, no. 2, pp. 183–200, 2007.

[7] J. N. Petroff and A. V. Hill, "A framework for the design of lot-tracing systems for the 1990s," *Production and Inventory Management Journal*, vol. 32, no. 2, p. 55, 1991.

[8] M. H. Jansen-Vullers, C. A. van Dorp, and A. J. Beulens, "Managing traceability information in manufacture," *International journal of information management*, vol. 23, no. 5, pp. 395–413, 2003.

[9] C. Van Dorp, "A traceability application based on gozinto graphs," in *Proceedings of EFITA 2003 Conference*, 2003, pp. 280–285.

[10] P. Missier, K. Belhajjame, and J. Cheney, "The w3c prov family of specifications for modelling provenance metadata," in *Proceedings of the 16th International Conference on Extending Database Technology*. ACM, 2013, pp. 773–776.

[11] S. B. Davidson and J. Freire, "Provenance and scientific workflows: challenges and opportunities," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 1345–1350.

[12] J. Singh, J. Cobbe, and C. Norval, "Decision provenance: Harnessing data flow for accountable systems," *IEEE Access*, vol. 7, pp. 6562–6574, 2019.

[13] M. Hind, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, A. Olteanu, and K. R. Varshney, "Increasing trust in ai services through supplier's declarations of conformity," *arXiv preprint arXiv:1808.07261*, 2018. [Online]. Available: https://arxiv.org/pdf/1808.07261.pdf

[14] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumeé III, and K. Crawford, "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018. [Online]. Available: https://arxiv.org/abs/1803.09010

[15] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 220–229.

[16] S. Schelter, J.-H. Böse, J. Kirschnick, T. Klein, and S. Seufert, "Automatically tracking metadata and provenance of machine learning experiments," in *Machine Learning Systems Workshop at NIPS*, 2017.

[17] "Node-red: Flow-based programming for the internet of things." [Online]. Available: https://nodered.org/

[18] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. F. Da Silva, M. Livny *et al.*, "Pegasus, a workflow management system for science automation," *Future Generation Computer Systems*, vol. 46, pp. 17–35, 2015.

[19] S. Nakamoto *et al.*, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[20] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, pp. 1–32, 2014.

[21] D. Tapscott and A. Tapscott, "How blockchain will change organizations," *MIT Sloan Management Review*, vol. 58, no. 2, p. 10, 2017.

[22] L. Luu, D.-H. Chu, H. Olickel, P. Saxena, and A. Hobor, "Making smart contracts smarter," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 254–269.

[23] I. Barclay, A. Preece, I. Taylor, and D. Verma, "A conceptual architecture for contractual data sharing in a decentralised environment," *arXiv preprint arXiv:1904.03045*, 2019.

[24] J. Benet, "Ipfs-content addressed, versioned, p2p file system," *arXiv preprint arXiv:1407.3561*, 2014.

[25] E. Jonas, J. Schleier-Smith, V. Sreekanti, C.-C. Tsai, A. Khandelwal, Q. Pu, V. Shankar, J. Carreira, K. Krauth, N. Yadwadkar *et al.*, "Cloud programming simplified: A berkeley view on serverless computing," *arXiv preprint arXiv:1902.03383*, 2019.