Karlsruhe Institute of Technology

Title of the Thesis

Master's Thesis/Bachelor's Thesis of

My Name

at the Department of Informatics, Institute of Information Security and Dependability (KASTEL)
Decentralized Systems and Network Services Research Group

Reviewer:	Prof. A
Second reviewer:	Prof. B
Advisor:	M.Sc. C
Second advisor:	M.Sc. D

xx. Month 20XX – xx. Month 20XX

2. Data Provenance

With e-health [Eys01], e-finance [AMS02], cloud services, 'Internet of Things', social media, etc. spreading and growing by the day, data exchanged, analysed or produced by intelligent devices become more and more difficult to trace [17]. It is often unknown how information is collected, how it is further processed, by whom, and for what purpose [Zub15]. This kind of information is often referred to as data provenance (DP), where "The provenance of a data item includes information about the processes and sources that lead to its creation and current representation" [GD07, p. 3]. The purpose of DP is to extract relatively simple explanations for the existence of some piece of data from some complex workflow of data manipulation.

With digitalisation, the concern with potential exposure of private and sensitive personal information is rising [TQV21], and with it, the significance of DP [BT19]. Also, information is not only personal and private, but also proprietary. Consumers should know if their data had been manipulated and how, in a network, that provides interoperability and connects actors in a secure, trustworthy, transparent and 'user friendly' way [Sun+14].

An increasing amount of research is being done to utilize DP technologies [BT19] in the fields of *healthcare* [Mar+20; LAC19; Le 18; HK21; Rah+20; Sun+14], *finance* [Sin+20; Liu+21; SAD19; Sir+19], supply-chain [Man+18], cloud services [Xia+17], scientific research [SPG05], etc.

Healthcare: in regard to medical treatment and patient safety, the importance of data, its origins and quality have long been recognised in clinical research [Cur+17] [Muh14]. Creating trust relationships among the various actors is vital - e.g., evidence-based medicine and healthcare-related decisions using third-party data are essential to patient safety [Mar+20]. DP is also crucial for solving confidentiality issues with healthcare information like accidental disclosures, insider curiosity and insider subornation [Rin97].

Finance: in online banking, digital money and digital financial services, the importance of information about transactions, money flow, money origin, credit scores and financial decisions is becoming bigger and bigger since the emergence of e-finance [AHS02]. DP is of great use in investigating money laundering [Ung+06], tracing donations [Sir+19], charities [Sin+20] or illegal funding [Tei18]. However, 'big tech' are also venturing into financial services [Boi+21]. While being accused for abuse of market power and anti-competitive behaviour, they are also famous for not giving extensive information on how personal data is analysed, processed or interacted with by third parties and international or government organisations [, RV19], which has a negative impact on the consumers ability to trace their personal data.

On the other hand, in European data protection law, everybody has the right to know where the organisation accountable got his data from, what the data was used for, where it was transferred to and how long it is stored, regardless of location [, GDPR].

3. Requirements

Data Provenance approaches/technologies, suitable for tracing the origin and source of personal data, have to fulfil a number of requirements. In this section we describe the requirements derived from the available literature, which we think are essential for the use cases investigated in our work. We differentiate between the following roles in our use cases:

General: Data Subject (Sender/Receiver)

Medical: Patient, Physician, Institution
Financial: Consumer, Institution

3.1. *General* Data Provenance Requirements

Requirement	Description
Identification	Associates each Data Subject with a unique identifier.
Policies	Enforce laws (GDPR, etc.) and regulations such as purpose limitation, data minimisation, Data Subject access rights.
Logging	Provides mechanisms to log and timestamp the transfer of the data between Data Subjects.
Accessibility	Allows Data Subjects with access to view, store, retrieve, move or manipulate data, based on their access rights.
Availability	Ensuring that data and is provenance is available to Data Subjects, when and where they need it.
Ownership	Allows Data Subjects to get an overview, request or perform changes and deletion of the data that they own.
Integrity	Ensures that the Data Receiver may detect unauthorised changes made to the data.
Confidentiality	Ensures non-disclosure of data traveling over the network to unauthorised Data Subjects.
Anonymity/ Unlinkability	Give the possibility to send, receive or access data in an anonymous or pseudonymous way. However, provenance is an example for a possible conflict between transparency and unlinkability.
Traceability/ Transparency	Give information on what transmitting principle was used, what type of data, for what purpose and to whom the information was sent. How data is collected; how, when, where it is stored.
Completeness	Collecting complete provenance information can fully take the advance to track data and actions for identity management, error detection, etc. Incomplete provenance information may lead to detection missing and suppression of abnormal behaviors.
Granularity	Not only the process derivation of a data file should be traced, but also the components of files such as paragraphs, shapes and images should be traced with regard to their origins. In short, fine-grained provenance information helps achieve highly precise anomaly detection and auditing.
Scalability	With the increase of the data volume and the number of operations, it should be possible to store complete provenance information without risks of information loss.
Trustworthiness	Ensures trust between Data Sender and Receiver with exchange of credentials, statement and certification, signatures, transparency and fulfilment of the other requirements.
Interoperability	By definition - the capability to communicate, execute programs, or transfer data between various systems in a manner that requires Data Subjects to have little or no knowledge of the unique characteristics of those systems.
Usability	Provides clear interfaces and structures that display security aspects, required data, digital traces, policies, possible threats in an understandable way (usage of icons, graphs, etc.). Also managing security (and privacy) is not the primary task of the user.

3.2. *Medical* Data Provenance Requirements

Requirement	Description
Identification	An unique identifier is important in differentiating between medical data of patients, physicians and institutions.
Policies	Medical data should be subject to laws and regulations to ensure patient safety, medical data confidentiality and trust in Physicians and medical Institutions.
Logging	Logging and timestamping transfer of medical data is essential in clinical research and patient treatment.
Accessibility	It is important that the different actors can view, store, retrieve, move or manipulate medical data based on their access rights. (e.g. patients checking prescriptions, physicians issuing/altering prescriptions, institutions verifying prescriptions).
Availability	Medical data and its provenance should be accessible and ready for immediate use, especially in cases of emergency.
Ownership	Allows patients to get an overview, request or perform changes and deletion of personal medical data that they own (e.g. patient loses trust in or changes an Institution).
Integrity	Data must be accurate and changes should be detectable, otherwise patients' health and life are at risk.
Confidentiality	Ensures non-disclosure of medical data traveling over the network to unauthorised actors.
Anonymity/ Unlinkability	A patient might feel that important information should be shared, but is reluctant to share if the information is attributed to them. Also, analysis of medical data by Institutions is useful but this should not be done in a way that may link personal medical data to a specific patient.
Traceability/ Transparency	Give information on what transmitting principle was used, what type of medical data, for what purpose and to whom the information was sent. How medical data is collected; how, when, where it is stored.
Completeness	Incomplete data can impact decisions and put the patients' health and life at risk.
Granularity	fine-grained provenance information helps achieve highly precise anomaly detection and auditing, which can improve decision making, diagnosing and patient safety.
Scalability	e-Health is a field in which big volumes of medical data are produced, exchanged or analysed.
Trustworthiness	Physician-to-patient relationship is jeopardised when patients do not trust that their personal medical data will be kept confidential, and that these data will not be utilised for purposes other than medical.
Interoperability	Usage of international standards enforces security and patient safety: the quality of the patients' treatment is not depending on the quality of a specific software.
Usability	Provides clear interfaces and structures that display information in an understandable way (usage of icons, graphs, etc.).

3.3. *Financial* Data Provenance Requirements

Requirement	Description
Identification	An unique identifier is important in differentiating between Consumers and Institutions.
Policies	Consumer and Institutions, as well as financial documents, monetary transactions should be subject to laws and regulations.
Logging	Logging and timestamping transfer of money and exchange of personal financial data between actors is essential to ensure trust and transparency.
Accessibility	It is important that the different actors can view, store, retrieve, move or manipulate financial data and money based on their access rights. (e.g. Consumer checking their financial data, Institutions managing financial data of many customers).
Availability	Financial data and its provenance should be accessible and ready for immediate use, especially in cases of fraud.
Ownership	Allows Consumers to get an overview, request or perform changes and deletion of personal financial data that they own (e.g. Consumer loses trust in or changes an Institution).
Integrity	Data must be accurate and changes should be detectable, otherwise Consumers' and Institutions' financial data and money are at risk.
Confidentiality	Ensures non-disclosure of financial data traveling over the network to unauthorised actors.
Anonymity/ Unlinkability	Consumers should be able to anonymously/pseudonymously transfer/donate money, maintaining the traceability of the operation, while ensuring unlinkability to personal information.
Traceability/ Transparency	Give information on what transmitting principle was used, what type of financial data or money, for what purpose and to whom it was sent. Also, how financial data is collected; how, when, where it is stored.
Completeness	Incomplete data can impact decisions and put Consumers and Institutions at risk.
Granularity	fine-grained provenance information helps achieve highly precise anomaly and fraud detection and auditing.
Scalability	e-Finance is a field in which big volumes of financial data are produced, exchanged or analysed.
Trustworthiness	Institution-to-Consumer relationship is jeopardised when Consumer do not trust that their personal financial data will be kept confidential, and that this data will not be utilised for purposes other than desired.
Interoperability	Usage of international standards enforces security and safety of both actors: the quality of the financial operations and their provenance is not depending on the quality of a specific software.
Usability	Provides clear interfaces and structures that display information in an understandable way (usage of icons, graphs, etc.).

Bibliography

[] *Chapter 3 (Art. 12-23) Archives*. en-US. URL: https://gdpr.eu/tag/chapter-3/ (visited on 2021-06-05).

[] *Getting my personal data out of Facebook*. en. URL: https://ruben.verborgh.org/facebook/ (visited on 2021-06-05).

[17] "IoT Data Provenance Implementation Challenges". en. In: *Procedia Computer Science* 109 (Jan. 2017). Publisher: Elsevier, pp. 1134–1139. ISSN: 1877-0509. DOI: 10.1016/j.procs.2017.05.436. URL: https://www.sciencedirect.com/science/article/pii/S1877050917311183 (visited on 2021-06-05).

[AHS02] Helen Allen, John Hawkins, and Setsuya Sato. "Electronic trading and its implications for financial systems". In: *Technology and Finance*. Routledge, 2002, pp. 213–247.

[AMS02] Franklin Allen, James McAndrews, and Philip Strahan. "E-Finance: An Introduction". In: *Journal of Financial Services Research* 22.1 (Aug. 2002), pp. 5–27. ISSN: 1573-0735. DOI: 10.1023/A:1016007126394. URL: https://doi.org/10.1023/A:1016007126394.

[BKR09] Steffen Becker, Heiko Koziolek, and Ralf Reussner. "The Palladio Component Model for Model-driven Performance Prediction". In: *Journal of Systems and Software* 82 (2009), pp. 3–22. DOI: 10.1016/j.jss.2008.03.066. URL: http://dx.doi.org/10.1016/j.jss.2008.03.066.

[Boi+21] Frederic Boissay et al. "Big techs in finance: on the new nexus between data privacy and competition". In: *The Palgrave Handbook of Technological Finance*. Springer, 2021, pp. 855–875.

[BT19] Peter Buneman and Wang-Chiew Tan. "Data Provenance: What next?" In: *ACM SIGMOD Record* 47.3 (Feb. 2019), pp. 5–16. ISSN: 0163-5808. DOI: 10.1145/3316416.3316418. URL: https://doi.org/10.1145/3316416.3316418 (visited on 2021-06-06).

[Cur+17] Vasa Curcin et al. "Templates as a method for implementing data provenance in decision support systems". In: *Journal of biomedical informatics* 65 (2017), pp. 1–21.

[Eys01] G. Eysenbach. "What is e-health?" In: *J Med Internet Res* 3.2 (June 2001), e20. ISSN: 1438-8871. DOI: 10.2196/jmir.3.2.e20. URL: http://www.ncbi.nlm.nih.gov/pubmed/11720962.

[GD07] B. Glavic and K. R. Dittrich. “Data provenance: A Cctegorization of existing approaches”. eng. In: *BTW ’07: Datenbanksysteme in Buisness, Technologie und Web* 103 (Mar. 2007). Ed. by A. Kemper et al. Conference Name: 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" ISBN: 9783885791973 Meeting Name: 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" Number: 103 Place: Bonn Publisher: Gesellschaft für Informatik (GI), pp. 227–241. DOI: 10.5167/uzh-24450. URL: http://www.btw2007.de/paper/p227.pdf (visited on 2021-06-05).

[HK21] Taylor Hardin and David Kotz. “Amanuensis: Information provenance for health-data systems”. In: *Information Processing & Management* 58.2 (2021), p. 102460.

[LAC19] Gary Leeming, John Ainsworth, and David A Clifton. “Blockchain in health care: hype, trust, and digital health”. In: *The Lancet* 393.10190 (2019), pp. 2476–2477.

[Le 18] Tran Le Nguyen. “Blockchain in Healthcare: A New Technology Benefit for Both Patients and Doctors”. In: *2018 Portland International Conference on Management of Engineering and Technology (PICMET)*. 2018, pp. 1–6. DOI: 10.23919/PICMET.2018.8481969.

[Liu+21] Wei Liu et al. “A donation tracing blockchain model using improved DPoS consensus algorithm”. In: *Peer-to-Peer Networking and Applications* (2021), pp. 1–12.

[Man+18] Suruchi Mann et al. “Blockchain technology for supply chain traceability, transparency and data provenance”. In: *Proceedings of the 2018 International Conference on Blockchain Technology and Application*. 2018, pp. 22–26.

[Mar+20] Andrea Margheri et al. “Decentralised provenance for healthcare data”. en. In: *International Journal of Medical Informatics* 141 (Sept. 2020), p. 104197. ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2020.104197. URL: https://www.sciencedirect.com/science/article/pii/S1386505619312031 (visited on 2021-06-05).

[Muh14] Jill C Muhrer. “The importance of the history and physical in diagnosis”. In: *The Nurse Practitioner* 39.4 (2014), pp. 30–35.

[Rah+20] Mohamed Abdur Rahman et al. “Secure and provenance enhanced Internet of health things framework: A blockchain managed federated learning approach”. In: *Ieee Access* 8 (2020), pp. 205071–205087.

[Rin97] Thomas C. Rindfleisch. “Privacy, Information Technology, and Health Care”. In: *Commun. ACM* 40.8 (Aug. 1997), pp. 92–100. ISSN: 0001-0782. DOI: 10.1145/257874.257896. URL: https://doi.org/10.1145/257874.257896.

[SAD19] Hadi Saleh, Sergey Avdoshin, and Azamat Dzhonov. “Platform for tracking donations of charitable foundations based on blockchain technology”. In: *2019 Actual Problems of Systems and Software Engineering (APSSE)*. IEEE. 2019, pp. 182–187.

[Sin+20] Aashutosh Singh et al. “Aid, Charity and donation tracking system using blockchain”. In: *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*. IEEE. 2020, pp. 457–462.

[Sir+19] N Sai Sirisha et al. “Proposed solution for trackable donations using blockchain”. In: *2019 International Conference on Nascent Technologies in Engineering (ICNTE)*. IEEE. 2019, pp. 1–5.

[SPG05] Yogesh L Simmhan, Beth Plale, and Dennis Gannon. “A survey of data provenance in e-science”. In: *ACM Sigmod Record* 34.3 (2005), pp. 31–36.

[Sun+14] Ali Sunyaev et al. “Availability and quality of mobile health app privacy policies”. In: *Journal of the American Medical Informatics Association* 22.e1 (Aug. 2014), e28–e33. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2013-002605. eprint: https://academic.oup.com/jamia/article-pdf/22/e1/e28/34145987/amiajnl-2013-002605.pdf. URL: https://doi.org/10.1136/amiajnl-2013-002605.

[Tei18] Fabian Maximilian Johannes Teichmann. “Financing terrorism through cryptocurrencies–a danger for Europe?” In: *Journal of Money Laundering Control* (2018).

[TQV21] Ofir Turel, Hamed Qahri-Saremi, and Isaac Vaghefi. “Special Issue: Dark Sides of Digitalization”. In: *International Journal of Electronic Commerce* 25.2 (2021), pp. 127–135. DOI: 10.1080/10864415.2021.1887694. eprint: https://doi.org/10.1080/10864415.2021.1887694. URL: https://doi.org/10.1080/10864415.2021.1887694.

[Ung+06] Brigitte Unger et al. “The amounts and the effects of money laundering”. In: *Report for the Ministry of Finance* 16.2020.08 (2006), p. 22.

[Xia+17] QI Xia et al. “MeDShare: Trust-less medical data sharing among cloud service providers via blockchain”. In: *IEEE Access* 5 (2017), pp. 14757–14767.

[Zub15] Shoshana Zuboff. “Big other: surveillance capitalism and the prospects of an information civilization”. In: *Journal of Information Technology* 30.1 (Mar. 2015), pp. 75–89. ISSN: 1466-4437. DOI: 10.1057/jit.2015.5. URL: https://doi.org/10.1057/jit.2015.5.