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 "provenance", 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 provenance is to extract relatively simple explanations for the existence of some piece of data from some complex workflow of data manipulation.

In this work we define *data provenance* (DP) as an approach that can be used to record not only metadata, data origin and/or data operation, but also processes that act on data and agents that are responsible for those processes. Most importantly, this should be achieved in a secure, trustworthy and transparent way, that ensures accountability and is in accordance to international laws and regulation, with the well-being of the consumer in mind.

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 not only in investigating money laundering [Ung+06], tracing donations [Sir+19], charities [Sin+20] or illegal funding [Tei18], but also loans and financing, mortgages, trading of currencies, insurance policies and others [But20]. 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 and the processes that led to its current state, have to fulfil a number of requirements. In this section we describe the requirements derived from the available literature, as well as other, which *we think* are essential for the use cases investigated in our work. We differentiate between the following roles in our use cases:

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

General: Data Subject (Sender/Receiver)

3.1. *General* Data Provenance Requirements

Group	Requirement	Description
User	Identification	Associates each Data Subject with an unique identifier and allows identification.
	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.
	Ownership	Allows Data Subjects to get an overview, request or perform changes and deletion of the data that they own.
	Accessibility	Allows Data Subjects with access to view, store, retrieve, move or manipulate data, based on their access rights.
Data	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.
System	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.
	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.
	Trustworthiness	Ensures trust between Data Sender and Receiver with exchange of credentials, statement and certification, signatures, transparency and fulfilment of the other requirements.
Security	Confidentiality	Ensures non-disclosure of data traveling over the network to unauthorised Data Subjects.
	Integrity	Ensures that the Data Receiver may detect unauthorised changes made to the data.
	Availability	Ensuring that data and is provenance is available to Data Subjects,when and where they need it.
Other	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.

3.2. *Medical* Data Provenance Requirements

Group	Use Case Requirements
User	A patient might feel that important information should be shared, but is reluctant to do so if the information is attributed to their unique identity. Also, analysis of medical data by Institutions is an useful tool, but should not be done in a way that may link personal medical data to a specific patient. It is important that the different actors can view, store, retrieve, move, request changes/deletion or manipulate medical data based on their ownership and access rights (e.g. patients checking prescriptions, physicians issuing/altering prescriptions, institutions verifying prescriptions).
Data	Information on what transmitting principle was used, what type of medical data, for what purpose and to whom the information was sent is essential. It is important how medical data is collected; how, when, where it is stored, for incomplete data can impact decisions and put the patients' health and life at risk. Fine-grained provenance information helps achieve highly precise anomaly detection and auditing, which can improve decision making, diagnosing and patient safety.
System	e-Health is a field in which big volumes of medical data are produced, exchanged and analysed. Therefore, usage of international standards that enforces security and patient safety are essential: the quality of the patients' treatment should not depend on the quality of a specific software. It's not patients or physicians job to analyse complex data flows. The system should also provide clear interfaces and structures that display information in an understandable way (usage of icons, graphs, etc.). Trust is fundamental, for that the physician-to-patient relationship is jeopardised when patients do not trust that their personal medical data will be kept confidential, and that this information will not be utilised for purposes other than medical.
Security	There must not be any disclosure of medical data traveling over the network to unauthorised actors. Data must be accurate and changes should be detectable, otherwise patients' health and life are at risk. Also, medical data and its provenance should be available and ready for immediate use, especially in cases of emergency

3.3. *Financial* Data Provenance Requirements

Group	Use Case Requirements
User	Without ownership or access to their own information, consumers cannot be certain if their data is inaccurate, obsolete, or otherwise inappropriate. [4372] The fear of abuse alters consumer behaviour and anonymity can be misused by criminals [238168]. A balance between identification and unlinkability must be achieved. Consumers should be able to perform operations in an pseudonymous way, that ensure ownership (pseudonyms are not improperly used by others) and ensure individuals are held accountable for abuses created under any of their pseudonyms. [4372]
Data	Tracing leads to transparency among actors. It should be possible to trace messages, transactions, what information and how it has been collected, analysed or processed (e.g. if donation funds are utilized properly or not). (aid) Data must be complete, accurate and fine-grained, in order to achieve precise anomaly and fraud detection and not negatively impact decision making or put consumers, institutions and their money or financial data at risk.[fine-grained]
System	Institutions generally have an interest in maintaining good relations with consumers and share many of the same interests and concerns. [4372] To ensure trust, institutions need efficient, interlinked and, in a way, pervasive record-keeping system (fingerprint), while still providing consumers with monitoribility and control. Such systems may also have to handle a large amount of transactions.[4372] Easily scalable system can bring efficiency gains and lower entry barriers for consumers, however, there should be ways to prevent discrimination, abuse of market power, anti-competitive and monopolistic use of data. [bigtech]
Security	Where there is money related information, the actors involved are a potential subject to numerous types of crime. Non-disclosure, accuracy and availability of data, as well as state-of-the art security measures are, therefore, of great importance, in order to prevent theft, fraud, money laundering or terrorist related activity.