Data Provenance

With e-health, e-finance, 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 (3). It is often unknown how information is collected, how it is further processed, by whom, and for what purpose (8). 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" (p. 3) (9). 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, and with it, the significance of DP. 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 and transparent way.

An increasing number of businesses and organizations are beginning to utilize DP technologies in the fields of *healthcare*, *finance*, supply-chain, cloud services, scientific research, etc. (10).

Healthcare - in regard to medical treatment and patient safety, the importance of data, its origins and quality have long been recognised in clinical research [2]. 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. DP is also crucial for solving confidentiality issues with healthcare information like accidental disclosures, insider curiosity and insider subornation [].

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 growing since the emergence of e-finance. DP is of great use in investigating money laundering[], tracing donations[] or illegal funding[], however, ‘big tech’ are also venturing into financial services []. 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 [], 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.

Requirements

A data provenance approach/technology should have **[Requirement]** *that* [Description].

Actors

Data Subject - Sender/Receiver

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.

Security: Provides the necessary approaches to ensure secure exchange of private and sensitive personal data.

Requirements for *medical* Data Provenance

Actors

Patient
Physician
Institution

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 solution.

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.).

Requirements for ***financial*** Data Provenance

Actors

Consumer
Institution

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 solution.

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.).