

# Revisiting Data Lineage and its Cross-Organizational Deployment

Mortaza S. Bargh

Research and Data Center  
Dutch Ministry of Justice and Security  
The Hague, The Netherlands  
Email: m.shoae.bargh@wodc.nl

Sunil Choenni

Research and Data Center  
Dutch Ministry of Justice and Security  
The Hague & Rotterdam, The Netherlands  
Email: r.choenni@wodc.nl

**Abstract**—Data and data sharing foster information systems that can create value for society, individuals, businesses and organizations. Data sharing and usage require establishing an appropriate data ecosystem where solid and effective data governance and management are in place to deal with associated risks like data being biased, personal, sensitive and stigmatizing, to name a few. Data lineage is a necessary means for data governance and management. In this contribution, we revisit the objectives of data lineage and investigate how it can be deployed in cross organizational settings. Specifically, we provide an overview of the objectives to which contemporary data lineage can contribute, revise the existing definition(s) of data lineage and adapt it to cross organizational settings, and propose architectural models for data lineage deployment across loosely coupled semi-autonomous organizations. Our revised definition of data lineage conceptualizes the physical distribution of data related objects as well as the semantical distribution of the related concepts that relate or apply to those data objects. The proposed architecture for data lineage related metadata management relies on the existing organizational structures, which, therefore, can scale up organically as seen in the case of federated identity management among academia.

**Keywords**—architecture, cross-organizational settings, data governance, data lineage, data management, data sharing, trust.

## I. INTRODUCTION

Data are currently being generated, collected, shared, analyzed, and distributed at a fast-growing pace. As a result of this growth, there is a rising interest (and demand) to harvest the available data and develop data-driven systems for easing our daily lives, creating additional value for businesses, providing insight into societal phenomena, and guiding policymaking processes. To capitalize on data, one should trust data, i.e., be attentive about the risks associated with data like data being blended with biased, partial, faulty, sensitive, and stigmatizing information about individuals and groups.

Given, on the one hand, the importance of data sharing and, on the other hand, the increased complexity of data sharing among (many) stakeholders and ISs, there is a need for establishing an appropriate data ecosystem. Establishing such a data ecosystem requires solid and effective data governance and data management to ensure the quality of data, secure the storage and exchange of data, optimize the tradeoff among contending values (like data utility and data privacy), and operationalize the Findable, Accessible, Interoperable and Reusable (FAIR) principles for the data. Data lineage is a necessary means for data governance and management [1][2]. Although there is no universal definition for data lineage, generally data lineage refers to the process of tracking the flow and transformations of data over time [3], i.e., during the data lifecycle/journey. It uses metadata to provide a clear description of the data origin(s), data changes in its journey, and data destination(s). As a similar case, data provenance (sometimes) refers to the sources of the data and its historical changes at the origins [4].

In this contribution, we describe the results of our explorative study about data lineage, particularly about its deployment across collaborating organizations that share data for a legitimate purpose. For example, within the Dutch Justice System (DJS), there is a growing trend of applying digital technology and data-driven systems to enhance and improve implementation of the rule of law within the Dutch society. The Information Systems (ISs) in the justice domain, which collect, store, share and processes data, are often physically distributed, have many loosely coupled subsystems, and are administrated by various organizations, spreading across many administrative domains. Utilizing data in this setting requires interconnecting many data sources and integrating their information in a trustful and responsible way. Those who share data (like judicial service providers) should entrust data consumers in using the data responsibly and those who consume data (like policymakers) should trust data sources in collecting and sharing data responsibly.

The research objective of our study can be specified as investigating how data lineage technology can contribute to data governance and data management in cross-organizational settings. Achieving this objective requires, among others, investigating the benefits of data lineage technology and the directions (and challenges) that one might explore (and expect) in deploying it across organizations. To this end, the contributions of this paper can be formulated as:

- Providing an overview of the objectives to which data lineage can contribute. This gives insight in the potentials of contemporary data lineage technology.
- Revising the existing definitions of data lineage and adapting its definition to cross organizational settings.
- Proposing two architectural models for data lineage deployment across loosely coupled semi-autonomous organizations (like that of the DJS).

For this study we have conducted a critical literature review [5], where we analyze several selected information sources and reflect on the existing concepts, models and approaches. The information sources used are not only from scholarly literature, but also from gray literature like commercial websites, whitepapers, and weblogs. The latter is because many vendors and system developers are dominantly active in the field of data lineage, who introduce many innovative concepts, features, and tools to the domain. In addition to literature study, we have conducted four semi-structured interviews with experts from different organizations within the DJS to gain insight in ongoing data lineage related (R&D) activities as well as in eliciting the needs and visions of those experts involved in data governance and management within their organizations. Further, we organized two expert focus groups with data stewards and data management experts to present our intermediary results and get early feedback. The four interviewees and the two focus

groups were chosen based on the expertise and availability of the participants rather than their representativeness. This choice is motivated by the nature of the study in being preliminary and explorative.

The organization of the contribution is as follows. We present the study background in Section II. We revisit the objectives sought within contemporary usage of data lineage in Section III. In Section IV, we present a revised definition and architecture of data lineage. In Section V, we elaborate on a federated architecture for deploying data lineage in cross organizations settings. In Section VI, we draw our conclusions and presents several avenues for future research.

## II. BACKGROUND

### A. Existing Definition(s) of Data Lineage

Although there is no universal definition for data lineage, there are many attempts to conceptualize it. A common definition of data lineage (or provenance), which stems from the academic community, defines it as: data “[l]ineage, or provenance, in its most general form describes where data came from, how it was derived, and how it was updated over time” [3]. A scan of various definitions given in (grey) literature – see [6] for some of these definitions – reveals that nowadays, especially among practitioners, data lineage includes some other aspects than those given in the conventional definition of [3]. These other aspects include data lineage being a process, expressing when, by whom, why data transformations are done, indicating where data are stored (which is relevant for stalled data), recording who has accessed data, and linking data related business concepts and technical objects along data journey paths. The latter aspect, i.e., linking data related business concepts and technical objects along data journey paths, plays an important role in cross organizational settings.

### B. Data Lineage Characteristics

Data lineage contributes to gaining trust in data and in responsible data transformation and sharing. Data lineage can be specified by various characteristics, which are not necessarily independent. These data lineage characteristics include: data origin vs data flow lineage [4], where vs how data lineage [3], data transformation types [3][7], coarse-grained vs fine-grained data lineage [3][8], lazy vs eager data lineage [3][9], backwards vs forward data lineage [10], tracing vs tracking data lineage [8][11], technical vs business data lineage [12], and horizontal vs vertical data lineage [13]. These data lineage characteristics specify the technical space in which a data lineage tool can be designed, chosen, and/or deployed [6]. For each characteristic one needs collect and manage specific metadata.

The subset of data lineage characteristics that should be included in (the design of) a data lineage tool depends on the data lineage usage objective(s). These data lineage objectives are described in Section III. Further, as mentioned in Section II.A, linking data related business concepts and technical objects along data journey paths, plays an important role in cross organizational settings. In Section IV.A, therefore, we will elaborate on technical vs business data lineage and horizontal vs vertical data lineage and their relations.

### C. Cross-Organizational Settings

Nowadays the ISs that collect, store, share and processes data are often physically distributed, have many loosely coupled subsystems, and are administrated by various

organizations (i.e., spreading across many administrative domains). For example, the DJS consists of many semi-autonomous organizations, collectively implementing the rule of law within the Dutch society. The relations between these organizations are often characterized as a linear chain (where a stage must be concluded before the next stage may begin), having sometimes loops and parallel relations. The term justice system is used to refer to the bodies in the apparatus of law, which are involved in creating data, ranging from legislative texts to judicial decisions. Thus, the scope of the DJS is broader than courts and court procedures.

## III. REVISITING THE OBJECTIVES FOR USING DATA LINEAGE

Here we review the objectives to which data lineage can contribute. We use the term “contribute” to indicate that data lineage is not the only element for realizing these objectives. Further, the mentioned objectives, which are not mutually exclusive, indicate societal relevancy of data lineage at large.

### A. To Trust in Data

Primarily, data lineage enhances *trust* in data. Data lineage does this in various ways like: knowing who has produced data and how they are changed (note that this aspect constitutes the original motivations behind data lineage adoption), showing regulatory compliance through enabling data audit, enabling data governance, facilitating data migration, discovering and mitigating privacy and security risks, enabling algorithmic explainability, interpretability, recourse, and contestability, and assuring how Machine Learning (ML) and data analytics models are trained and constructed.

### B. To Data (Analytics) Governance

Knowing data history (e.g., data origin, when and where data are transformed) contributes to data transparency. This can be helpful for various aspects of data governance [1] like enabling compliance audit, risk management improvement, accountability assurance, and compliance (i.e., ensuring data being processed in line with organizational policies, community policies, legal rules and regulations, and regulatory standards). Further, it can boost various aspects of data management such as data quality management, data migration management, data silos integration, and data gap detection and mitigation [2]. Each of these areas to which data lineage contributes, is explained in the following.

### C. To Personal Data Protection

Data lineage can contribute to various aspects of personal data protection. We review some of these aspects as specified in the EU General Data Protection Regulation (GDPR) [14] articles and recitals. Conforming with these articles and recitals requires the ability to identify which data are personal data as well as the entities from/through/to which the personal data are originated, transformed, and sent. From the viewpoint of data subjects, data lineage can contribute to realization of, among others, the GDPR articles: Art. 15 – right of access by data subjects to know about the purpose of personal data being processed, the categories of personal data being used, the (categories of) data recipients (especially those in third countries or international organizations), and the period of data being stored. Data subjects have right to get meaningful information about automated decision-making taken place based on their data, Art. 16 – right to rectification of inaccurate (and possibly incomplete) personal data, Art. 17 – right to erasure existing personal data, which is related to the right to

be forgotten, and Art. 20 – right to data portability, which enables data subjects to have the personal data transmitted directly from one data controller to another, where technically feasible. From the viewpoint of data consumers, data lineage can inform them about a dataset having privacy issues like not being pseudonymized, anonymized, fair, or consensual.

#### D. To Entrust AI Models

A precondition for entrusting models derived from data by various algorithms – like User Defined Functions (UDFs), AI/ML algorithms, and statistical algorithms – is to know which and how datasets are used for training these models. For example, users of such models should know about whether and to what extent the training data have data quality, Copy Right (CP), and bias issues. In [15][16] there are examples mentioned where ill trained AI models are developed due to not knowing about or not paying attention to the datasets used for training those models. As an example, they report about a case where researchers used datasets of adult patients with COVID-19 and of, as a control group, very-young patients without COVID-19 to train a model for detecting COVID-19. The trained model showed a strong performance as tested on those datasets but, in fact, the model “learned to identify kids, not covid” [15], “merely detecting children versus adults” [16]. Therefore, data models obtained from statistics, AI/ML and UDFs should be traceable to the data used for their training. Data lineage can contribute to this traceability.

#### E. To Explainability, Interpretability and Fairness

The increasing usage of AI systems for automated decision making or for supporting decision making (either at a personal or policy level) increases concerns over their lack of fairness, legitimacy, and accountability. Data lineage helps know how output data are derived from some input data. As such, it enables algorithmic recourse and algorithmic contestability for/by data consumers (like citizens, data journalists and civic organizations) in all stages of data lifecycle.

*Algorithmic recourse* is concerned with providing explanations and recommendations to individuals who have received unfavorable outcomes from automatic decision-making systems [17]. The explanations and recommendations should provide end-users with actionable measures whereby the outcome of an AI system can be changed to the favorable one. Making AI systems contestable by design is another way to mitigate these concerns [18]. *Algorithmic contestability* aims at making AI systems responsive to human intervention throughout the system lifecycle. Such a human intervention can ask an AI system to explain and interpret how outcomes and from which input and training data are derived. Unlike in algorithmic recourse, in algorithmic contestability the objective is not to change the algorithm outcome, but to change the outcome of the whole process from which the algorithm is part of. Answering these questions in both algorithmic recourse and algorithmic contestability can be facilitated by having data lineage in place.

#### F. To Data Quality Management

Data lineage can improve data quality by contributing to several aspects. Firstly, it can help analyzing the root causes of data quality issues observed at the data consumer side. Such a reactive detection and improvement of data quality issues can be done by back tracing the observed errors/exceptions at downstream nodes of data journey paths [2][19]. Secondly, it can assist rectifying the data quality issues from the data origin side. Such a proactive detection and correction of data quality

issues can be done by forward tracing the observed errors/exceptions at upstream nodes of data journey paths. Thirdly, it can aid exception handling by validating the accuracy and consistency of data for accurate data analytics, Business intelligence (BI) and data science usages. Fourthly, it can support identifying incorrect assumptions about data due to, for example, change of semantics along data journey paths. Finally, it can serve data cleaning via archiving data and/or deleting data whenever necessary. These may arise due to, for example, data being old or expired or data subjects requesting to delete their personal data for privacy reasons.

#### G. To Data Change Management

Data lineage allows a proactive approach for data change management [2][19], through analyzing the downstream impacts of envisioned changes made to datasets at upstream nodes. To this end, data lineage enables understanding the location of data destinations, the lifecycle of data, and the downstream IT operations. Hereby, upstream nodes get some ideas about the impacts on people and systems before propagating envisioned corrections and changes. Example usage scenarios of data lineage for data change management are to ease large data migrations (e.g., moving to clouds, implementing upgrades, and performing consolidations), and to reduce risks when implementing data process changes.

#### H. To Data Ownership

Data lineage can inform data consumers about data flows and data sources, which may include some information about the data owners. Data consumers can hereby know who is responsible for and who owns every dataset or process in the data pipeline. This knowledge can ensure data consumers about who to contact for issues or changes. Data owners, in turn, can use data lineage to know where, by whom and how their data are used in downstream. Thus, data lineage provides answers to the questions of data owners about their data [20].

#### I. To Regulatory Compliance, Audit, and Accountability

Data lineage may prevent getting fines for data processors on the data pipeline through indicating whether data transformations are done according to regulatory rules and guidelines, and whether appropriate controls and policies for containing possible threats (like security, privacy, safety, and fairness threats) are in place. Two of these regulatory regimes are EU GDPR and EU AI Act, which pertain to privacy and responsible AI usage, respectively. Data lineage, on the one hand, can assist data controllers and processors in knowing about having or not having compliance with policies, laws, and regulations for their data, data flows, and data transformations. On the other hand, it can provide a means for authorities to audit the compliance of data controllers/processors with policies, laws, and regulations.

Further, data lineage can contribute to data accountability as it can show who is/are responsible for data and its transformations along its journey path(s). To this end, data lineage should keep track of the roles and responsibilities at every stage of data journey [21]. When data lineage is used in the context of accountability it is critically important that data lineage metadata are managed securely sufficiently. Tan et al. [22] mention a minimum set of security requirements for data lineage namely confidentiality, integrity, authenticity, and reliable collection. The latter means having trustworthy and accurate data lineage related metadata collection mechanisms. We think that, for accountability, non-repudiation should also be added to the list of the security requirements mentioned.

J. To Data Security and Privacy

Data lineage can increase security and privacy posture by enabling the search of data upstream and downstream to discover anomalies, and by tracking, identifying, and correcting potential risks associated with data (flows).

When data lineage is used in the context of data security and privacy protection, like in case of accountability, it is critically important that data lineage metadata are managed securely, where confidentiality, integrity, authenticity, and reliable collection are provided adequately [22]. For example, Bertino et al. [23] discuss information leakage concerns for data provenance due to sharing lineage metadata (especially in cross organizational settings).

K. To Data Modelling

Data lineage can provide the information needed to present visual representations of differing data components and their connections. The connections between data components can be shown in a model to show the dependencies present throughout a data ecosystem [19].

L. To Data Discovery

Partly, data lineage is about discovering the whereabouts and usages of already processed/shared data. As such, it contributes to and is relevant for discovering such data. To our understanding, data lineage is not about discoverability of new data which has not been published or processed. In the sense of finding data, the scope of data discovery is wider than that of data lineage. Note that data lineage is concerned with more functionalities than just data discoverability, as listed above.

IV. REVISED DATA LINEAGE DEFINITION AND ARCHITECTURE

A. Preliminaries

To highlight the data lineage functionality within an IS, we envision a layered model as shown in Fig. 1. The data layer, which is not part of data lineage, interfaces with various (types of) ISs distributed over the whole network of organizations. The middle layer represents the data lineage related metadata, which is collected from the ISs (see the left-most arrow), processed, stored, and shared to provide data lineage information to the data lineage related interaction layer. The latter receives the data lineage related queries of various types of end-users; and provides the replies to the data lineage related queries in appropriate formats.

Although realizing the FAIR principles is an enormous challenge in practice – think of, for example, making data automatically interoperable [1] – they guide the initiatives in enterprises and organizations to (re)use and share data within and cross organizational boundaries. Making data reusable, in general, and realizing the FAIR principles, in particular, ask for, among others, a better data documentation. Technically, data documentation is possible via metadata (management). *Metadata*, which are often defined as data about data in its simplest form [24], aims "at facilitating access, management and sharing of large sets of structured and/or unstructured data" [25]. This objective aligns well with the objectives of the FAIR principles.

As mentioned in Section II, linking data related business concepts and technical objects along data journey paths, plays an important role in cross organizational settings. Therefore, we elaborate here on technical vs business data lineage and horizontal vs vertical data lineage and their relations. Both terms of technical data lineage and business data lineage are

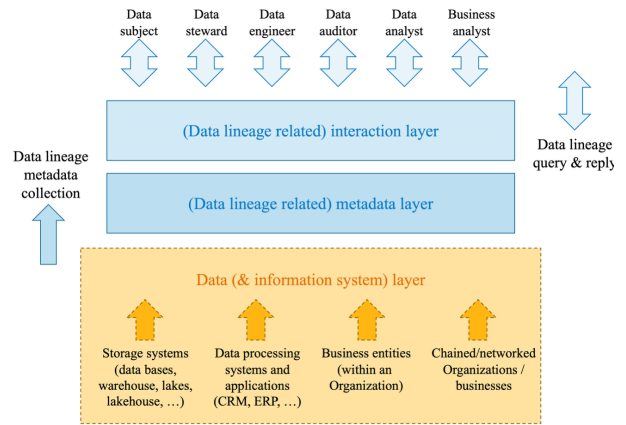


Fig. 1. A layered IS model with its data lineage component

used to refer to overseeing and monitoring data flows and transformations between ISs. The *technical data lineage* provides a physical view on stored data, data transformations, and data flows. In other words, it provides an oversight about which ISs within an organization or across organizations hold (a copy of) the data, lie on the path(s) of the data journey, and transform the data in certain ways. The *business data lineage* offers a logical view on how data flows and on how data transformations are connected to business representation of data [12] (and the references). This implies that, as we suspect, business data lineage links higher level concepts (like legal and business concepts) to lower-level data objects and transformations. Further, the link between data objects and business concepts may change across domains as data flow between domains. In principle, both technical and business lineages may spread spatially along the routes of data flows.

Linking business concepts and technical objects has become prominent in the last decade as organizations and enterprises seek to monetize data for shaping their strategies/services and/or business benefits. This linking within the data lineage field is referred to as *vertical data lineage* as it aims at establishing relationships between objects at multiple abstraction levels, ranging from business concepts to technical objects (i.e., data items). On the contrary, the traditional definitions of data lineage, e.g., that of [3], emphasize the spatial characteristics of data flows and transformations of data objects (i.e., technical data lineage) along the ISs on the data journey paths. This traditional perspective on data lineage is referred to as *horizontal data lineage*. Based on our literature review, however, we find out that business data lineage may have also a horizontal character, especially in cross organizational settings. In Table 1, we illustrate the convolution among vertical vs horizontal data lineage and technical vs business data lineage. Based on this insight, we are going to specify the architecture of data lineage within one IS (i.e., within one organization) and across ISs (i.e., within or across organizations).

TABLE 1. CONVOLUTION OF TWO DATA LINEAGE TYPES

	Horizontal lineage	Vertical lineage
Business lineage	Lineage among business / legal concepts across administrative domains (e.g., among organizations)	Lineage among business / legal concepts and data objects / transformations within an administrative domain (an organization)
Technical lineage	Lineage among data objects /transformations across administrative domains/ organizations as defined in [3]	

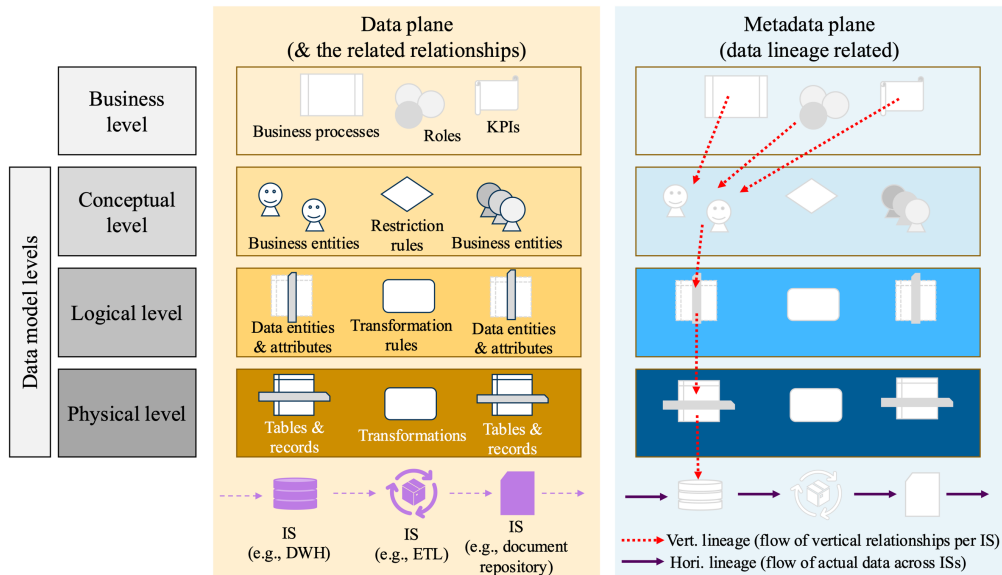


Fig. 2. The proposed layered architecture indicating horizontal vs vertical lineage together with data and metadata planes

### B. Revised Definition of Data Lineage

Based on the discussions above, we adopt the following definition of data lineage. “Data lineage is the description of data movements and transformations at various abstraction levels along data journey paths. The description encompasses those aspects that are of interest in an application context like how (i.e., by whom, when, where, which, etc.) data objects are processed (i.e., created, collected, stored, accessed, transformed, transmitted, etc.) and how they are related to high-level data concepts” [6].

The definition of data lineage touches upon the concept of “abstraction levels”, which relates to the concept of vertical data lineage (see Table 1). Vertical lineage focuses on the design of a database and relates abstract objects at different layers, objects such as a business partner (e.g., a product supplier) and physical implementation (e.g., a part of a star schema), see [13]. For the layers involved in vertical lineage we adopt five levels, shown in Fig. 2, where the top four levels are from [26]. The physical level depicts the physical artifacts on the database (like physical data models). The logical level represents data entities and data transformation rules (like logical data models). The conceptual level encompasses entities and business restriction rules (like conceptual data models). The business level includes business processes/roles. The legal and ethical level covers the legal and ethical grounds that allow using data for a certain purpose. The last level is particularly important for the organizations that oversee and safeguard the rule of law in the society like the DJS.

The definition of data lineage touches upon the concept of “along data journey paths”, which relates to the concept of horizontal data lineage. Horizontal lineage describes a physical lineage [13] through describing how data items (ranging from datasets to data tuples) flow between ISs, applications and platforms (like databases, data warehouse, data lakes, data lakehouses, data transformers) and get transformed along their paths. In short, a horizontal lineage shows how actual data flow and get transformed along its paths between ISs. These data flows and transformations can occur within an organization or across organizations. “Horizontal lineage enables understanding the dependencies

and relationships between data sources, data transformers and data consumers, and to identify potential data quality issues, risks, and gaps” [13]. According to these definitions, horizontal data lineage occurs at the physical level, as such it is called technical data lineage.

The horizontal lineage can also be applicable to higher abstraction levels (like business level), as also indicated in x Table 1. The horizontal lineage at higher abstraction levels is especially relevant in cross organizational settings, where the business logic and legal regime may change when data traverse through organizations and administrative domains.

### C. A Layered Architecture

In this section we propose a layered architecture for data lineage within one organization, which is inspired by the model in [27] and has resemblance with Zachman framework for information systems architecture (ISA) [28]. As shown in Fig. 2, the proposed architecture captures both concepts of horizontal and vertical lineage. The bottom three levels represent the typical data model levels of DAMA [29]. As mentioned before, there isn’t any established norm for defining the layers of vertical lineage. Nevertheless, it is essential to identify the higher abstractions levels too because we foresee replying some data lineage queries may require lineaging among or at these higher levels like: Which business processes do use this dataset? What are the legal grounds for sharing this dataset?

## V. DATA LINEAGE DEPLOYMENT ACROSS ORGANIZATIONS

### A. A Model for Data Lineage Metadata Management

Inspired by the strategies mentioned in [30] for storing lineage metadata of files in filesystems, we distill the following three models for data lineage storage. In *piggybacking model*, data lineage metadata accompany the data along their journey (like in-headers/footers and in-band encoding, each having a varying degree of data stickiness). In *centralized model*, data lineage metadata are stored locally in a central repository (like in a fileserver, in a local database, in a file system layer or in auxiliary files repository, each having a varying size of locality). In *distributed model*, data lineage metadata are stored locally in a central repository in a domain /organization. Further, there is an efficient built-in mechanism

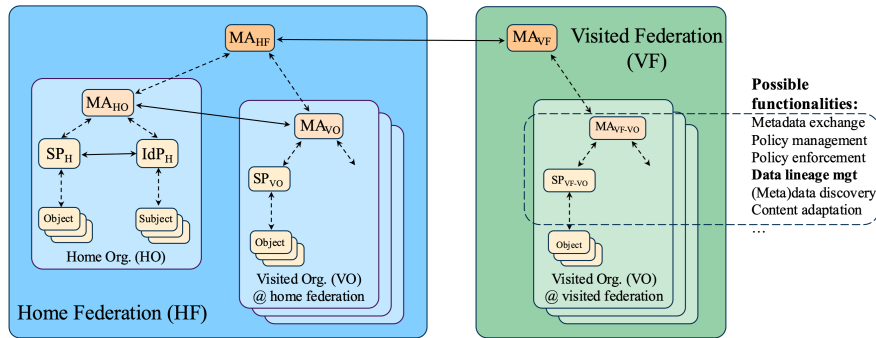


Fig. 3. A federated model for identity management that can be also used for metadata management across organizations

to establish links between repositories of different domains if a data lineage query requires it (on a need-to-know basis).

Out of these three models, we consider piggybacking and distributed models as candidate models for deploying data lineage in cross organizational setting like that of the DJS. The piggybacking model can be a valid option if one opts for data source tracking and the data are not combined with other data streams. It may not be scalable if one is interested in tracking all data transformations and the data are expected to be combined with multiple data objects. The piggybacking model alone is not helpful for the data lineage needs of the end-users who reside at the beginning of the data pipeline/journey.

The alternative option is to deploy data lineage across organizations according to the distributed model. A key issue in realizing the distributed model is to semantically describe and optimally reference and discover data lineage objects (data lineage metadata items) across domains. A relevant architecture for this purpose is that of the federated identity management used within the federation of Dutch universities, called SURFconext [31], and across the university federations of mainly European countries, called eduGAIN [32]. The architecture of these federated identity management systems is illustrated in Fig. 3, together with components Metadata Aggregator (MA), Service Provider (SP) and IdP (Identity Provider). The architecture is distributed and federated.

The federated model relies on a hierarchical structure within an organizational unit (i.e., universities in the example above) and a peer-to-peer structure across organizational units (i.e., between universities in a federation/country and across university federations/countries in the example above). The model works very well in practice, considering the success of SURFconext and eduGAIN identity federations. This success of the model can be attributed to its reliance on the existing organizational structures [33] where every organization manages its own (meta)data locally based on its internal policies and systems, partner organizations join forces and form an alliance (i.e., a federation) to collaborate on matters of common interest (in our case, data lineage metadata exchange), and alliances (i.e., federations) of organizations may also collaborate on matters of common interest (i.e., forming federation among federations or confederations). Such a reliance on the existing organizational structure works well and can scale up organically.

Inspired by these success stories, we propose considering a similar architecture – let’s call it federated data lineage metadata management architecture – for deploying data lineage in cross organizational settings such as that of the DJS. This model can be deployed in a centralized way per

organization (or per department within an organization) and in a federated way across organizations (or across departments within an organization).

The proposed architecture, shown in Fig. 3, can be used for trustful metadata exchange across collaborating organizations. The proposed architecture can be related to that of edge computing [34], in the sense that (semi)autonomous domains are connected via gateways at the boundaries of those domains. A gateway is an entry/exit point to/from a domain, which is responsible for, among others, access control to the domain, content adaption towards outside the domain, and service and data discovery within and across domains. The MAs shown in Fig. 3 are gateways that can store data lineage related metadata locally and enable exchanging (a digest of) the locally collected data lineage metadata with other peer organizations in the (con)federation. The MAs can also host other functionalities and roles like Policy Enforcement Point, Policy Decision Point, context transfer, content adaption; should the underlying components/parts of an organization or federation be unable to do so. As indicated in Fig. 3, the MAs can also host and exchange other types of metadata, like for trust establishment [35][33], data management, and data governance. These functionalities include per domain Service Discovery, Data Discovery, Data Catalog, and Data Lineage.

Note that the realization of the higher levels peer-to-peer relations in Fig. 3 (e.g., that of the confederation) can be centralized (e.g., via a trusted third party) or be made peer-to-peer (e.g., by using a blockchain). As an example of the latter, an architecture is proposed in [36] for data lineage among collaborating organizations. The proposed architecture manages data lineage metadata hierarchically in organizations and uses a blockchain for exchanging metadata among them.

### B. Intertwined Vertical and Horizontal Lineage

In cross organizational settings, we foresee an intertwined relationship between vertical and horizontal data lineage, which can be related to technical and business data lineage. As illustratively shown in Fig. 4, although horizontal data lineage is often prescribed for physical level objects, as indicated by continuous blue arrows in Fig. 4, it might also be relevant for concepts at the higher abstraction levels, as indicated by dashed red arrows in in Fig. 4. For example, in the DJS setting, some business concepts do not have one-to-one relationships across the chain of organizations. This inconsistency requires defining a horizontal lineage for these concepts at the business or conceptual level across the chain.

We expect less dynamicity and variation at the legal and ethical level within the DJS compared to those at the other levels, therefore this level is shown not predominantly in Fig.

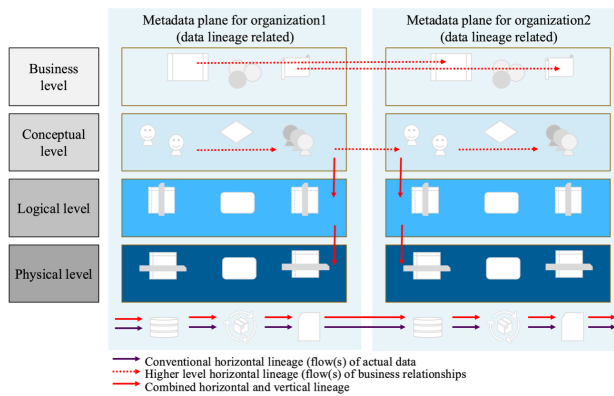


Fig. 4. Illustrating a mix of data lineage types across organizations

4. Based on this observation, we conclude that horizontal data lineage can be applicable to physical level objects (which corresponds to the technical lineage) as well as to business and conceptual level concepts in DJS settings (as also reflected in Table 1). Based on this, we also conclude that in cross organizational settings, for a horizontal data lineage at the physical level we may need adopting a combined vertical and horizontal data lineage, as conceptually indicated by continuous red arrows in Fig. 4. This is the reason to call this configuration as intertwined horizontal and vertical lineage.

## VI. CONCLUSION AND FUTURE WORK

Data lineage is a key functionality of data cataloging that contributes to gaining trust in data and in responsible data sharing. Based on some existing definitions and the body of knowledge on data lineage, we coined a definition of data lineage that conceptualizes not only the physical distribution of data related objects (like data origins, flows and transformations), but also the semantical distribution of the related concepts (like the business, legal and organizational terms that relate or apply to those data objects). Further, the definition offers a means to limit the scope of data lineage to those aspects that are of interest in each context.

Data lineage can contribute to many objectives, each of which, in turn, plays a role in enhancing trust in data, data sharing, and data-driven applications and policymaking. These objectives capture the usage areas of data lineage and, as such, they specify the societal relevancy of data lineage at large. It would be intriguing to aim at all mentioned objectives when deploying a data lineage solution. Such a versatile data lineage solution could immediately become too complex and costly, thus might become impossible to realize especially in distributed settings (e.g., among the semi-autonomous organizations of the DJS). Knowing the relevant data lineage objectives, one can determine which characteristics of data lineage are relevant in an operational setting and accordingly adopt a customized data lineage technology.

Metadata management across organizations, like in the case of the DJS, should be scalable and distributed as well as should fit the organizational structure of the DJS. Considering such cross organizational settings, we proposed considering a federated data lineage metadata management architecture for deploying data lineage. A federated architecture relies on the existing organizational structure, which, therefore, can scale up organically as seen in the case of federated identity management among European universities. The proposed architecture is cost effective as it allows collecting data lineage metadata locally (and coarsely) and, should a new query arise,

going for a targeted search (i.e., to carry out a zoom-in search on a need-to-know basis).

So far horizontal data lineage has been prescribed for physical level data objects, often coined as technical data lineage. In cross organizational settings, however, we foresee that horizontal data lineage can also be applied at higher abstraction levels to, for example, business and/or conceptual level concepts. Further, we concluded that, in cross organizational settings, we may need adopting a combined vertical and horizontal data lineage for a horizontal data lineage at the physical level. This so-called intertwined horizontal and vertical lineage configuration is necessary to deal with interoperability issues of data lineage in cross organizational settings. A combination of vertical and horizontal lineage can be needed for some real-world cases in the DJS setting.

When the intention is to deploy data lineage across organizations, there is a need for studying a suitable structural and functional architecture for data lineage deployment. Another direction for research is to investigate the requirements and ways for mixing vertical and horizontal data lineages at the boundaries of organizations. This requires mapping between data semantics at the borders of collaborating organizations and dealing with uncertainties that may be caused due to this mapping.

There is a need for further research on making data lineage become business-driven and usable for users with limited technical backgrounds (like business consultants and policymakers). To this end, investigating methods and tools for natural-language-based user interfaces is a promising direction. Large Language Model (LLMs) can be considered for mapping between natural language texts to formal database queries (e.g., SQL). This direction may require investigating ways to deal with uncertainty in the mapping between natural and formal languages.

Reducing the complexity of data lineage and the associated costs is a crucial factor in a successful adoption of data lineage technology. For data lineage related metadata collection, developing automated methods is necessary. For example, the use of LLMs can be investigated for (semi)automatically creating business level metadata (like a report in natural language that describes the technical analyses conducted on the data for business-level end-users) from technical level metadata (e.g., from data query scripts). In this way, the burden of business-level data lineage metadata creation on technical experts can be alleviated.

In conventional lineage, it is assumed that users can understand how an output is created by observing the source data and knowing that the data transformation is a sequence of simple operations like filter, join, and aggregation. However, in complex data analysis, like using AI/ML algorithms, more information about data transformation is needed. A question that may arise is which data lineage information should be provided to explain and/or influence the outcomes of very complex data transformations (like LLMs) and how this data lineage information should be managed (cost) effectively.

## ACKNOWLEDGMENT

This contribution is based on the result of our study documented in technical report [6]. We have used parts of the technical report that suit the purpose of this contribution, i.e., to present the study outcomes to the scientific community.

## REFERENCES

- [1] S. Choenni, M.S. Bargh, T. Busker and N. Netten, "Data governance in smart cities: Challenges and solution directions," *Journal of Smart Cities and Society*, 1(1), 2022, pp. 31-51, DOI: <https://doi.org/10.3233/scs-210119>
- [2] C. Stedman and D. Loshin, "What is data lineage? Techniques, best practices and tools," 2022. Retrieved on Oct. 16, 2024, from <https://www.techtarget.com/searchdatamanagement/tip/How-data-lineage-tools-boost-data-governance-policies>
- [3] R. Ikeda and J. Widom, "Data Lineage: A survey," Stanford University. 2009. Retrieved on Nov. 12, 2024, from [http://adrem.uantwerpen.be/sites/default/files/lin\\_final.pdf](http://adrem.uantwerpen.be/sites/default/files/lin_final.pdf)
- [4] T. Atlan, "Data lineage vs data provenance: Nah, they aren't same!" October 26, 2024. Retrieved on Jan. 28, 2025, from <https://atlan.com/data-lineage-vs-data-provenance/>
- [5] G. Paré, M.C. Trudel, M. Jaana and S. Kitsiou, "Synthesizing information systems knowledge: A typology of literature reviews," *Information & Management*, 52(2), 2015, pp. 183-199.
- [6] M.S. Bargh, "Data lineage for the justice system: Scope, potentials, and directions," Research and Data Center (WODC), The Hague, The Netherlands, Technical Report Cahier 2024-21, data lineage project (nr. 3482), Dec. 2024. Available: <http://hdl.handle.net/20.500.12832/3443>
- [7] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: A survey," *ACM Computing Surveys*, 37(1), 2005, pp. 1-28.
- [8] Y. Cui and J. Widom, "Lineage tracing for general data warehouse transformations," *The VLDB Journal*, 12(1), 2023, pp. 41-58.
- [9] M. Yamada, H. Kitagawa, T. Amagasa and A. Matono, "Augmented lineage: Traceability of data analysis including complex UDF processing," *The VLDB Journal*, 32(5), 2023, pp. 963-983.
- [10] Qlik, "Data lineage," 2024. Retrieved on Oct. 16, 2024, from <https://www.qlik.com/us/data-management/data-lineage>
- [11] Z. Xie, "Tracer: A machine learning based data lineage solver with visualized metadata management," Doctoral dissertation, Massachusetts Institute of Technology, 2022.
- [12] S. Karkošková and O. Novotný, "Design and application on business data lineage as a part of metadata management," In IEEE International Conf. on Computers and Automation (CompAuto), 2021, pp. 34-39.
- [13] J. Freche, M.D. Heijer and B. Wormuth, "Data lineage," *The Digital Journey of Banking and Insurance*, 3, 2021, pp. 5-19.
- [14] GDPR, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," Retrieved on Nov. 12, 2024, from <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [15] W.D. Heaven, "Hundreds of AI tools have been built to catch covid, none of them helped," *MIT Technology Review*, 2021. Retrieved on Jan. 9, 2025, from <https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/>
- [16] M. Roberts, D. Driggs, ... and C.B. Schönlieb, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, 3, 2021, pp. 199-217. DOI: <https://doi.org/10.1038/s42256-021-00307-0>
- [17] A.H. Karimi, G. Barthe, B. Schölkopf and I. Valera, "A survey of algorithmic recourse: contrastive explanations and consequential recommendations," *ACM Computing Surveys*, 55(5), 2022, pp. 1-29.
- [18] K. Alfrink, I. Keller, G. Kortuem and N. Doorn, "Contestable AI by Design: Towards a Framework," *Minds and Machines*, 33(4), 2022, 613-639. DOI: <https://doi.org/10.1007/s11023-022-09611-z>
- [19] K.D. Foote, "Data lineage use cases," June 13, 2023. Retrieved on Oct. 28, 2024, from <https://www.dataversity.net/data-lineage-use-cases/>
- [20] Informatica, "Data marketplace vs. data catalog: Benefits and key differences," 2024. Retrieved on Dec. 12, 2024, from <https://www.informatica.com/resources/articles/data-marketplace-vs-data-catalog.html>
- [21] R. Verma, P. Shrivastava and N. Merla, "Tracing the path: Data lineage and its impact on data governance," In *International Journal of Global Innovations and Solutions (IJGIS)*, 2024. Retrieved on December 12, 2024, from <https://ijgis.pubpub.org/pub/d6k8bzn0/release/1>
- [22] Y.S. Tan, R.K. Ko and G. Holmes, "Security and data accountability in distributed systems: A provenance survey," In IEEE 10<sup>th</sup> International Conference on High Performance Computing and Communications & IEEE International Conference on Embedded and Ubiquitous Computing, 2013, pp. 1571-1578. Retrieved on Dec. 12, 2024, from <https://tarjomefa.com/wp-content/uploads/2016/09/5415-English.pdf>
- [23] E. Bertino, G. Ghinita, ... and S. Xu, "A roadmap for privacy-enhanced secure data provenance," *Journal of Intelligent Information Systems*, 43, 2014, pp. 481-501.
- [24] R. Roszkiewicz, "Enterprise metadata management: How consolidation simplifies control," *Journal of Digital Asset Management*, 6(5), 2010, pp. 291-297. DOI: <https://doi.org/10.1057/dam.2010.32>
- [25] B. Kerhervé and O. Gerbé, "Models for metadata or metamodels for data?" In Proceedings of the 2<sup>nd</sup> IEEE Metadata Conference, 1997, Silver Spring, Ma, USA.
- [26] I. Steenbeek, "Choosing data management IT tools: Data lineage solutions," 2023. Retrieved on Nov. 7, 2024, from <https://datacrossroads.nl/2023/06/16/choosing-data-management-it-tools-data-lineage-solutions/>
- [27] I. Steenbeek, "Data lineage: the needs of and benefits to various stakeholders," 2022. Retrieved on Aug. 1, 2024, from <https://www.irmconnects.com/data-lineage-the-needs-of-and-benefits-to-various-stakeholders/>
- [28] J.F. Sowa and J.A. Zachman. "Extending and formalizing the framework for information systems architecture." *IBM systems journal* 31, no. 3, 1992, pp. 590-616.
- [29] Dama International, "DAMA-DMBOK: Data management body of knowledge" 2<sup>nd</sup> ed., 2017, Technics Publications, LLC.
- [30] A. Gehani, M. Kim and J. Zhang, "Steps toward managing lineage metadata in grid clusters," in 1<sup>st</sup> workshop on Theory and Practice of Provenance, 2009, pp. 1-9.
- [31] Surfconext, "Secure access everywhere with one set of credentials," 2024. Retrieved on Nov. 5, 2024, from <https://www.surf.nl/en/services/surfconext>
- [32] eduGAIN (2024). eduGAIN's technical site. Retrieved on Nov. 5, 2024, from <https://technical.edugain.org>.
- [33] M.S. Bargh, A. Omar and S. Choenni, (2024). Zero-trust security model applied to smart shipping. *Advances in Knowledge-Based Systems, Data Science, and Cybersecurity; Research*, 1(1), 5.
- [34] T. Qiu J. Chi, X. Zhou, Z. Ning, M. Atiquzzaman and D.O. Wu, "Edge computing in industrial internet of things: architecture, advances and challenges," *IEEE Communications Survey Tutorials*, 22, 2020, pp. 2462-2488.
- [35] M.S. Bargh, W. Janssen and A. Smit, "Trust and security in e-business transactions," Telematica Instituut, Enschede, The Netherlands, Technical Report: GigaTS project, 2002.
- [36] M. Matsubara, T. Miyamae, A. Ito and K. Kamakura, "Improving reliability of data distribution across categories of business and industries with chain data lineage," *Fujitsu Scientific & Technical Journal*, 56, 2020, pp. 52-59.