Research and Data Centre

Cahier 2024-21

Data lineage for the justice system

Scope, potentials, and directions

Summary

Dutch institute for Justice and Security

Cahier 2024-21

Data lineage for the justice system

Scope, potentials, and directions

Summary

Mortaza S. Bargh

Cahier
This series comprises overviews of studies carried out by or for the WODC Research and Data Centre. Inclusion in the series does not mean that the sheet's contents reflect the viewpoint of the Minister of Justice and Security.

Summary

Problem statement

Problem (context)

Data is 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 by using (advanced) algorithms to analyze 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. The way that data is collected, which is often blended with biased, partial, faulty, sensitive, and stigmatizing information about individuals, groups, and organizations; and the way that algorithms and data-driven systems are designed, implemented, interpreted, and (mis)used (are going to) impact us deeply at individual, group, and societal levels. Therefore, to capitalize on data one should be attentive of the risks of the data used for delivering personal, social, or organizational benefits. As such, gaining trust in data is a perquisite for respecting the basic human rights like privacy, liberty, autonomy, and dignity.

Also in the justice domain, particularly in the Dutch Justice System (DJS), we witness a trend of applying digital technology and data-driven systems. 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 (i.e., spreading across many administrative domains). Utilizing data in this setting requires interconnecting various information 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.

Establishing trust in data production and consumption requires, among others, mitigating the security issues of largely distributed data-reliant systems, managing the data quality issues of loosely coupled data sources, and dealing with wrongful and/or malicious use of data and algorithm outcomes. Given, on the one hand, the importance of data sharing and, on the other hand, the increased complexity involved in data sharing among (many) stakeholders and ISs, there is a need for establishing an appropriate data ecosystem. This data ecosystem establishment 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.

An efficient and effective data governance/management requires, among others, data lineage and data provenance. Data lineage refers to the process of tracking the flow of data over time, 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 (i.e., the origins) of the data and its historical changes.

Data lineage is seen as an essential instrument for enhancing data trustworthiness. It enables, for example, efficient data quality management, data change management, and data lifecycle management. A famous metaphor used to illustrate the role of data lineage for gaining trust in data is the case of gaining trust in the nutritiousness and healthiness of an apple by knowing how it is cultivated, harvested, transported, stored, distributed, and retailed. To gain this trust, one could keep track of the relevant information (lineage information) in every stage of the apple's lifecycle (i.e., the apple's journey) in the supply chain. In this metaphor, the object of interest is a physical object (commodity). Having the same type of assurance is also relevant for any digital object (e.g., a dataset, a digital picture, and a digital document). For example, citizens are increasingly subject to a growing quantity of multimedia content (i.e., data) of various types like images, videos, audio recordings, and documents. This content is often blended with misinformation (e.g., photos generated by Generative AI) or manipulated information (e.g., photos processed by the Photoshop tool), which makes it difficult for ordinary people (or even professionals) to distinguish between real and fake/manipulated content. In the case of photo sharing, data lineage can help photo makers, publishers, and consumers the ability to learn about how and by whom the photo is created and to learn about every edit made to the photo throughout its lifecycle/journey. The lineage information, which can trustfully be appended to the photo and accompany the photo through its entire journey, can easily be inspected by downstream consumers to get ensured about the origin and any edition made to the photo. This knowledge can help downstream consumers gain trust in the media content they encounter on, for example, social media and news feeds. Within the DJS, data lineage can similarly contribute to gaining trust in, for example, a standing policy by answering questions like *which datasets or documents are used as evidence for grounding the policy*.

In this contribution, we describe the results of our explorative study about data lineage (and provenance) technology, particularly about the concepts behind data lineage and the methods/tools used for data lineage. The study context relates to the data collected, shared, stored, and processed by the ISs within the DJS.

Research objective and research questions

The research objective can be specified as investigating how data lineage technology can contribute to data governance and data management within the DJS. Achieving this objective requires investigating the benefits of data lineage technology and the directions (and challenges) that one might explore (and expect) in deploying it within the DJS.

This study is a preliminary study towards the abovementioned objective. The research approach can be characterized as explorative, where we seek answers to the following research questions:

1 *What is data lineage?* For answering this question, we will also describe the context (or the data ecosystem) in which data lineage is used.
2 *Which objectives can data lineage contribute to?* For answering this research question, we will give an insight in the potential advantages of data lineage.
3 *How can data lineage tools be deployed?* For answering this question, we will elaborate on typical approaches for and challenges of data lineage deployment.
4 *What are the capabilities (and limitations) of existing data lineage tools?* For answering this question, we will sketch a framework for specifying the relevant data lineage capabilities. Further, as an example and for a limited number of existing

data lineage tools, we will give an insight in two capabilities that are of interest for this study.

Scope

These research questions will be addressed within the context of DJS, which 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. As such, the scope of the justice system is broader than courts and court procedures.

In this contribution we intend to provide an overview of some relevant aspects that can be considered for deploying data lineage in the organizational setting of the DJS. As such, we do not intend to design or prescribe a solution for data lineage deployment in this contribution. The target audience of this report is system designers and architects as well as data officers and engineers. The report aims at informing these groups about the design space within which they can design or choose a data lineage solution.

Methodology

For this study we have conducted a critical literature review, where several selected information sources are analyzed, and a reflection is done on the existing concepts, methods, and approaches. The selected information sources 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 within the DJS as well as in eliciting the needs and visions of those experts involved in data governance/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.

Main results and contributions

In this report we provide an overview of several aspects of data lineage technology and its deployment in cross organizational settings. The report can serve as a knowledge base for informing the design and deployment processes of the data lineage in the DJS setting. In the following, we summarize the answers given to the posed research questions.

What is data lineage?

Based on some existing definitions and the insights gained during the study on data lineage, we define data lineage as *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.* This definition conceptualizes not only the physical distribution of data 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 contributes to gaining trust in data and in responsible data transformation and sharing. Data lineage relies on deriving and managing metadata that is relevant for the aimed data lineage usage objective(s). Data lineage can be characterized from various aspects, which are not necessarily independent. These aspects include (a) data origin vs data flow lineage, (b) where vs how data lineage, (c) data transformation types, (d) coarse-grained vs fine-grained data lineage, (e) lazy vs eager data lineage, (f) backwards vs forward data lineage, (g) tracing vs tracking data lineage, (h) technical vs business data lineage, and (i) horizontal vs vertical data lineage. These data lineage characteristics specify the technical space in which a data lineage solution can be designed, chosen, and/or deployed.

Which objectives can data lineage contribute to?

The concept of lineage has been applied to a wide range of cases, ranging from tracking/tracing the computation flows in a single software program to tracing/tracing the data flows in distributed ISs. 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 include (a) data governance, (b) privacy protection, (c) trusting AI models, (d) data and AI explainability, interpretability and fairness, (e) data quality management, (f) data change management, (g) data ownership, (h) regulatory compliance, audit and accountability, (i) data security, (j) data modeling, and (k) data discovery. 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. Based on the required data lineage characteristics one can determine (the type of) data lineage metadata to be collected, and accordingly design data lineage (deployment) architecture to store, query, process, retrieve data lineage metadata (i.e., to decide on the architecture of data lineage metadata management).

How can data lineage tools be deployed?

Based on our literature study and expert interviews, we elucidated that data lineage within the DJS can (or is required to) contribute to data governance, data discovery, data quality management, data change management, and privacy and security mainly. Further, we draw the following conclusions for data lineage in the DJS setting.

- It is necessary to trustfully share data within and across organizational boundaries in the DJS while allowing participating organizations maintain their autonomy and have own business, conceptual and logical data models and ISs. As such, data lineage within the DJS should account for diversity at all levels, namely at technical, logical, conceptual, and business levels.
- Within the DJS, data is shared and processed for not only research and strategic purposes, but also for operational purposes. Thus, the shared data could be at various aggregation levels, i.e., at group and individual levels, which requires having data lineage at coarse-grained and fine-grained levels.
- The end-users of data lineage within the DJS can have different backgrounds with a varying set of data (science) skills and may reside at the beginning, middle or end of data pipeline. Therefore, there should be enough flexibility to serve a wide range of users (so-called, the democratization of data lineage).
- We define five abstraction levels for data lineage namely (a) physical, (b) logical, (c) conceptual, (d) business, and (e) legal and ethical levels. The legal and ethical level is particularly important in the DJS, as the DJS is responsible for overseeing and safeguarding the rule of law in the society.
- In cross organizational settings (like that of the DJS), we foresee that horizontal data lineage can be applied at not only physical level but also at business and conceptual levels. Further, we may need adopting a combined vertical and horizontal data lineage for enabling a horizontal data lineage at the physical level to deal with interoperability issues of data lineage (i.e., having an intertwined horizontal and vertical lineage configuration).

Metadata management in the DJS setting should be scalable and distributed as well as should fit the organizational structure of the DJS. Considering the cross organizational structure of the DJS, we propose 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.

For managing data lineage metadata, automizing the metadata collection process is necessary, considering the high speed and large volume at which data is collected and shared nowadays. For storing data lineage metadata, the type of data repository should be chosen according to the data lineage characteristics needed and the context in which data lineage operates. According to the type of the data repository chosen, one can opt for an appropriate query processing language. System-user interactions can be facilitated in different ways, depending on the technical skills of the users.

Being business driven is necessary as data-driven working has become a common practice nowadays. This requires involving non-technical end-users in the design process as well as the possibility of accommodating new queries by design in data lineage systems. Nevertheless, it is not feasible and efficient to collect and manage a huge amount of data lineage related metadata exhaustively in anticipation that one day some part of it would be useful for answering new business driven queries. To be cost effective, one may choose for collecting a reasonable amount of data lineage

metadata (i.e., collecting metadata coarsely) and should a new query rise, go for either a targeted search (i.e., to carry out a zoom-in search on a need-to-know basis) an/or for an approximate reply if having certain amount of uncertainty in replies is acceptable. It may be needed to deploy natural-language-based user interfaces which are capable of handling both predefined and not predefined queries. Hereby one can make data lineage become business-driven and usable for users with limited technical backgrounds (like business consultants and policymakers).

What are the capabilities (and limitations) of existing data lineage tools?

Determining the relevant capabilities of data lineage tools in each usage context is important as they can be used as criteria for evaluating the existing tools and choosing the one that fits the context the best. Based on the results of the study, we foresee several relevant capabilities for data lineage software tools in the DJS setting. The main capabilities are to have flexibility in being business driven (replying to new and unforeseen data lineage related queries), to have more granular lineage than attribute level (like being cell level), to have interoperability with other tools (from other organizations), and to offer good user experience and useability.

As part of the answer to this research question, we sketch a framework for evaluating data lineage tools. Based on this framework, one should define a set of evaluation criteria by, for example, elucidating the objectives that DJS organizations seek in deploying data lineage, elucidating the desired data lineage characteristics, and defining evaluation criteria accordingly. Subsequently, one can specify several granularity levels per each criterion. These levels can be defined qualitatively, i.e., be assigned a (meaningful) numerical value or a score. Finally, one can merge quantified criteria using numerical methods like averaging and thresholding.

Commercial software tools generally provide a wide range of data lineage functionalities together with other functionalities (like data management and data catalog). As such, they are suitable for large organizations which can afford paying the expenses of such tools and which need using a wide range of functions these tools provide. Open-source tools generally offer a limited subset of data lineage functionalities, are cost free, and are integrate-able with other (open-source) applications. As such, they are suitable for low-budget, small enterprises to use or customize these tools to their data lineage needs. Note that an extensive integration of open-source tools requires inhouse technical skills or extra budget that might not be available in small organizations. Further, the open-source tools might be useful for conducting small-scale experimentations within the DJS setting to gain some hands-on experience about data lineage technology.

Recommendations for follow-up research

Several directions are identified for future research during project execution. In this section we group them in three categories, organized from more practice-oriented research one to more applied research one.

Need for a requirement elicitation study

For choosing a (set of) data lineage tool(s) that can be experimented with (or deployed) in the DJS setting we recognize the need for eliciting the requirements with which the tool(s) should comply. To this end, we foresee the following action points:

- Identifying the typical end-users and their data lineage queries,
- Identifying whether there is a need for adjusting the end-users' queries in the future,
- Defining the desired data lineage objectives/characteristics, and
- Defining the data lineage tool evaluation method by determining the evaluation criteria and how to measure and merge them.

When the intention is to deploy data lineage in the DJS setting, there is a need for a further study of a suitable structural and functional architecture for data lineage deployment in DJS setting.

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. To this end, gaining hands on experience with data lineage tools in real world cases can be useful.

Need for democratization of data lineage

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.

Need for effective and efficient data lineage

Reducing the complexity of data lineage and the associated costs is a crucial factor in successful adoption of data lineage technology.

For data lineage related metadata collection, developing automated methods is necessity. 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 in a (cost) effective way.

An issue in data origin lineage is how to determine data origins in each setting. We suspect that there might be multiple views with different data origins (specially in cross organizational settings). If this conjecture holds, then lineaging data origin boils down to or, better said, requires lineaging data flows. It is for future research to investigate how to determine data origins (or data destinations) in each operation setting.

The WODC (Research and Data Centre), a Dutch agency in the field of Justice and Security, is an independent knowledge institute that falls under the Dutch Ministry of Justice and Security. The WODC contributes to upholding and improving the rule of law by carrying out high-quality scientific research (or commissioning others to do so on its behalf), as well as by presenting solicited and unsolicited knowledge, points for improvement and (where possible) thinking strategies.

More information:

www.wodc.nl