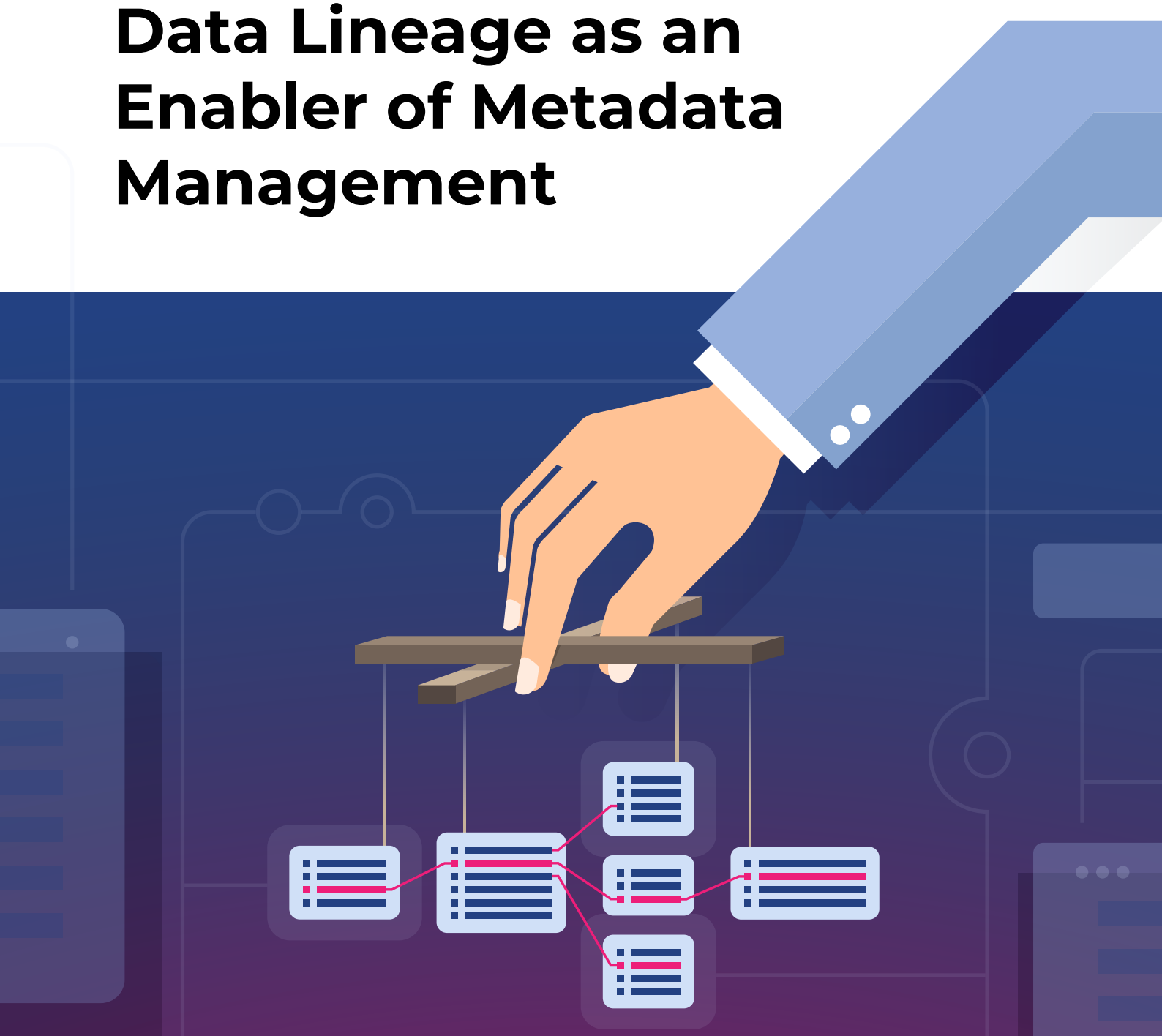
MANTA

EXPERT ANALYSIS

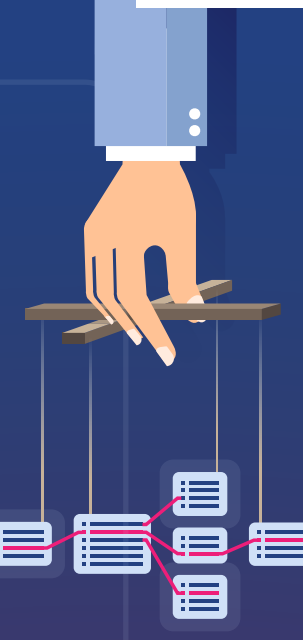
Data Lineage as an Enabler of Metadata Management



Written by Irina Steenbeek, Ph.D.

Data Management Practitioner & Coach | Data Lineage

Data Lineage as an Enabler of Metadata Management



In my book "Data Lineage from a Business Perspective," I investigated the relationships between the concepts of metadata and data lineage and came to the following conclusion: data lineage is a comprehensive metadata construct that consists of various complex metadata objects. Later, my introduction to the functionality of modern data lineage solutions gave me new insight: data lineage enables metadata management.

In this expert analysis, we will discuss:

- The concepts of metadata and data lineage, and the relationship between them
- The concept of metadata management
- Metadata stakeholders and their needs
- The manner in which data lineage enables metadata management

Concepts of Metadata and Data Lineage

Metadata

The terms "data" and "metadata" have a complicated relationship. Data is the physical or electronic representation of facts or signals "in a manner suitable for communication, interpretation, or processing by human beings or by automatic means." A data model is an example of metadata. Metadata puts data in a particular context. For example, if you had a dataset with customer financial information, you couldn't use the dataset without a detailed description. The description of a dataset is metadata that defines data and explains the context in which data can be used. The challenging aspect of defining metadata is that the same data can be recognized as either data or metadata, depending on the context. For example, data models are metadata for business users. For data modelers, on the other hand, data models can be considered data that will in turn require other metadata to describe data models. Different sources contain different approaches to classifying metadata. In this analysis I use the metadata classification proposed by The DAMA Guide to the Data Management Body of Knowledge, second edition (DAMA-DMBOK2):

Business metadata

"Business metadata focuses largely on the content and condition of the data and includes details related to data governance." This is a challenging definition as the term "data governance" has different meanings in different contexts.

Technical metadata

"Technical Metadata provides information about technical details of data, the systems that store data, and the processes that move it within and between systems."

Operational metadata

"Operational Metadata describes details of the processing and accessing of data."

Data lineage

The classical definition of data lineage is relatively simple: "A description of the pathway from the data source to their current location and the alterations made to the data along that pathway". What this definition does not say is "HOW" to describe this pathway. In fact, you can describe the pathway of data movements by means of metadata. I have drawn a number of conclusions regarding the definition and model of data lineage through my daily work with data lineage.

Several other concepts have definitions similar to that of data lineage. These concepts are data value chain, data chain, data flow, integration architecture, and information value chain.

Data lineage can be documented at four levels of abstraction:

- Business level
- Data model at:
  - Conceptual level
  - Logical level
  - Physical level
- The key components of a data lineage construct/metamodel are:
  - Business processes and roles
  - IT assets like systems, applications, databases, networks
  - Data models at the conceptual, logical, and physical levels
  - Business rules and their technical implementation in the form of an ETL (Extract, Transform, Load) processes

Each of these components is a complex metadata object by itself. A simplified concept map in Figure 1 demonstrates this complex metamodel of data lineage:

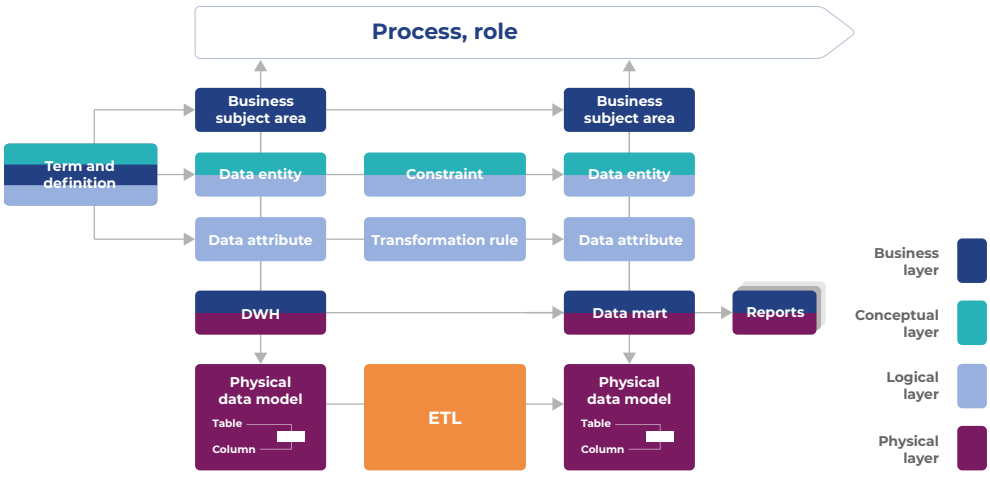


Figure 1: Concept map of the data lineage metamodel.

Data lineage classification

In another context, data lineage could be classified differently. Classification assists in identifying a company's data lineage needs and the scope of a data lineage-related initiative. Limiting this initiative to a reasonable scope ensures its successful implementation. Four factors play roles in this classification, as demonstrated in Figure 2:

- Level of documentation: business, conceptual, logical, physical
- Subject of documentation: metadata and data value lineage
  It is worth a reminder that we discuss only metadata lineage in this article. Metadata lineage describes the process of data transformation by a means of metadata. Data lineage documented at each of four above-mentioned levels remains metadata lineage.
- Direction of documentation: horizontal and vertical
  Horizontal data lineage describes the pathway of data transformation along data chains and can be documented at each of the four layers. Vertical data lineage links data lineage components between various layers.
- Method of documentation: descriptive and automated
  Descriptive data lineage is a method of recording metadata data lineage manually in a repository. Automated data lineage records metadata lineage by implementing automated processes for scanning and ingesting metadata into a repository.

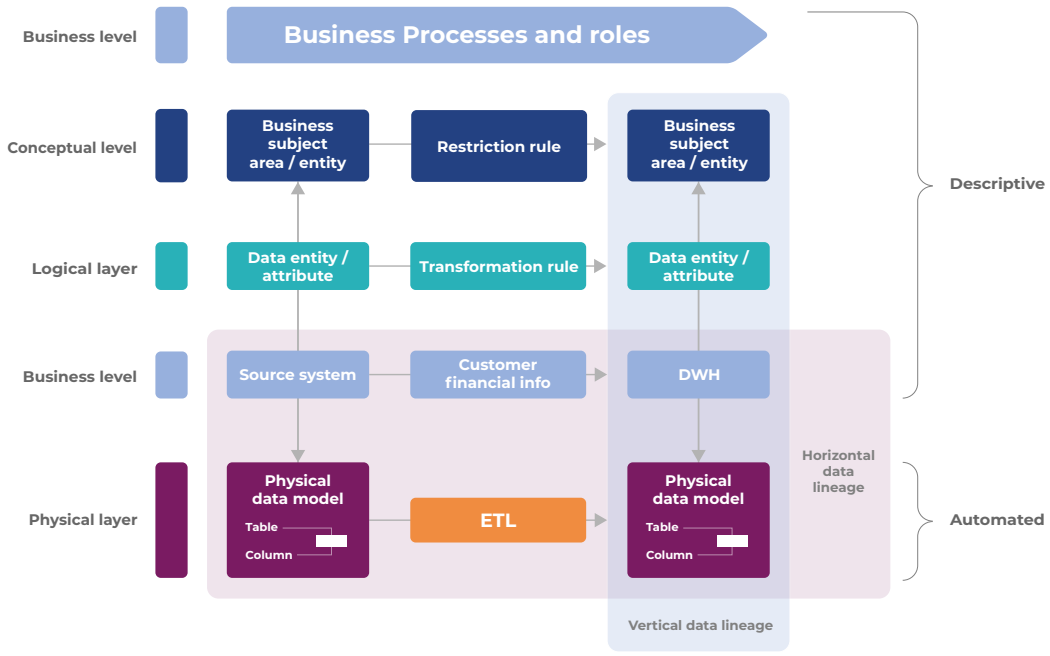


Figure 2: Various types of metadata lineage.

Certain data lineage types have interdependencies among each other.

The relationship between the concepts of metadata and data lineage

To identify the relationship between metadata and data lineage, we first need to compare objects of various types of metadata with the components of data lineage discussed above (See Figure 1). For this comparison, I used a list of metadata objects provided by DAMA-DMBOK2. Table 1 illustrates the results of the analysis: specifically, the relationships between data lineage layers (column 1), data lineage components (column 2), and various examples of metadata retrieved from DAMA-DMBOK2.

The analysis of these relationships leads to multiple important conclusions:

1. Data lineage is a comprehensive metadata construct that consists of other various complex metadata objects. One remarkable fact: data lineage itself is recognized as one metadata object that simultaneously pertains to two differerent metadata types (business and technical).

2. Data lineage unites and maps objects of various types of metadata.

DATA LINEAGE LAYER	DATA LINEAGE COMPONENT	METADATA TYPE		
		BUSINESS	TECHNICAL	OPERATIONAL
Data lineage as one metadata object		Data provenance and lineage	Data lineage including upstream and downstream change	
Vertical linkage between models and physical assets				
Business layer	Business process			
	Business function and role			Technical roles and agreements
	IT asset		Names and descriptions	Data sharing rules and agreements
Conceptual layer	Data set	Definitions and descriptions		
	Conceptual data model			
	Data entities			
	Business rules			Data archiving and retention rules
Logical level	Logical data model			
	Data entities and data attributes			
	Transformation rules			
Physical level	Physical data model			
	Table, column		Physical names, keys, properties	
	Calculations, derivations, mapping			
	ETLjob			Execution and error logs
	Data quality rules	Rules and measurement results		
Operational information about data lineage				History of extracts, reports and query access patterns, execution time, backup, volumetric and usage patterns

Table 1: Relationships between data lineage and metadata types.

The Concept of Metadata Management

For this analysis, I did research on publications about metadata management. Strangely enough, I found only two recently published books on this topic. This is primarily due to the complexity and variety of the concept of metadata. When people speak about metadata management, they usually mean metadata at the physical level. Even ISO/IEC 11179 Metadata Registry Standard provides recommendations for the metadata repository for the documentation of metadata at the physical level. However, we've just now seen that metadata can also refer to a much wider variety of layers and corresponding objects. This diversity causes challenges in defining the scope of and tools for metadata management. DAMA-DMBOK2 remains one of the primary sources of the definition of metadata management: metadata management is "Planning, implementation, and control activities to enable access to high quality, integrated metadata". Based on the DAMA's definition and the results of our analysis in the previous paragraph, we can state the following: metadata management should plan, implement, and integrate business, technical, and operational metadata represented by diverse metadata objects at various abstraction levels.

There are two key goals of metadata management that derive from this definition:

Plan, implement, and maintain metadata

To realize this goal, a company should have several operational metadata-related capabilities like metadata governance, modeling, architecture, quality, lifecycle management, and security. Different repositories store and manage different types of metadata. For example, a business process modeling solution is an example of a repository for business processes, while a data modeling solution is a repository for data models at different abstraction levels, and so on.

Map and integrate metadata of various types

The integration of various types of metadata is the most challenging task. Data lineage is one of the most effective solutions (maybe the only solution) that can be deployed to realize this goal and enable metadata management to successfully integrate. All of the above information brings us to the realization of the relationship between metadata management and data lineage: data lineage is a metadata construct that enables the integration of various types of metadata stored in different repositories.

Metadata Stakeholders and their Needs

It is worth mentioning that different business stakeholders have diverse needs and requirements for different types of metadata and data lineage; these different needs lead to varying metadata/data lineage initiatives. In this analysis, I discuss only a company's internal stakeholders. I use the classification of data/metadata management stakeholders demonstrated in Figure 3:

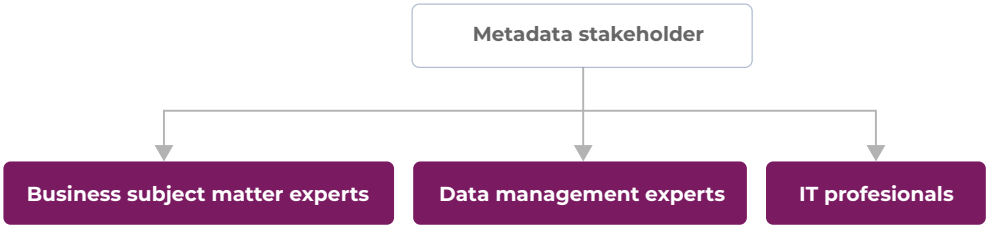
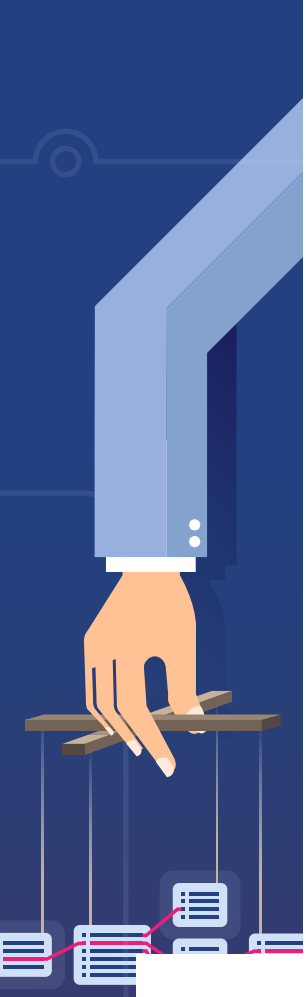


Figure 3: Classification of internal metadata stakeholders.



Business subject matter experts are business executives and professionals. They represent the interests of the company's management, commercial, risk, finance, compliance, marketing, and sales functions. They are mainly interested in business metadata and data lineage at the business and conceptual levels. However, some stakeholders, like business analysts, may be interested in more technical metadata, like vertical data lineage. Business stakeholder groups will be interested in data management initiatives focused on compliance with various regulations, optimization of business models, and architecture landscapes.

Data management professionals have knowledge, skill, and experience in one or more data management domain, excluding those related to IT. They may have interests in all three types of metadata and data lineage at all four levels of abstraction for performing data management- related initiatives (e.g., data quality and master and reference data).

IT-professionals have knowledge, skill, and experience in one or more IT domain. These professionals will be interested mainly in the technical and operational metadata focusing on DevOps and migration initiatives.

Table 2 summarize all said above:

DATA LINEAGE LAYER	DATA LINEAGE LAYER OBJECTS	METADATA TYPES & STAKEHOLDERS BUSINESS TECHNICAL OPERATIONAL			DATA MANAGEMENT INITIATIVE
		BUSINESS	TECHNICAL	OPERATIONAL	
Business	Business processes, functions, roles; IT assets	1. Business subject matter expert 2. Data management professionals	1. Data management professionals 2. IT professionals	1. Data management professionals 2. IT professionals	1. Compliance with regulations 2. Optimization of architecture landscapes 3. Business model optimization
Conceptual	Data sets/flows; conceptual data model; Data entities and business rules	1. Business subject matter expert 2. Data management professionals			1. Compliance with regulations 2. Optimization of architecture landscapes 3. Business model optimization
Logical	Logical data models; data entities & attributes; transformation rules	1. Data management professionals 2. IT professionals			1. Compliance with regulations 2. Data management initiatives
Physical	Physical data models; tables, columns; ETLjobs; Calculations, derivations, mapping, validation rules	1. Data management professionals 2. IT professionals	1. Data management professionals 2. IT professionals	1. Data management professionals 2. IT professionals	1. Compliance with regulations 2. Data management initiatives 3. DevOps and migration initiatives
Operational data about data lineage				1. IT professionals	1. Data management initiatives 2. DevOps and migration initiatives

Table 2: Relationships between data lineage, metadata, and stakeholders' concerns and initiatives

Conclusion: The Manner in which Data Lineage Enables Metadata Management

We will demonstrate the enabling role of data lineage by analyzing the metadata sources or repositories and linking them to lineage layers and components. Table 3 compiles the results of this analysis:

DATA LINEAGE LAYER	DATA LINEAGE LAYER OBJECTS	METADATA TYPES			METADATA REPOSITORIES/ SOURCES
		BUSINESS	TECHNICAL	OPERATIONAL	
Data lineage as one metadata object	Relations between various data lineage objects				1. Graph database solution 2. Integrated metadata solution 3. Relationship repository
Business	Business processes, functions, roles; IT assets				1. Business process modeling solution 2. Enterprise architecture solution 3. Data governance solution 4. IT asset catalog
Conceptual	Data sets/flows; conceptual data model; Data entities and business rules				1. Enterprise architecture solution 2. Data modeling solution 3. Business glossary 4. Data (set) catalog
Logical	Logical data models; data entities & attributes; transformation rules				1. Data modeling solution 2. (Meta)data dictionary/repository 3. Business rules repository
Physical	Physical data models; tables, columns; ETL jobs; Calculations, derivations, mapping, validation rules				1. Data modeling solution 2. (Meta)data dictionary/repository 3. Application, DWH, data lake databases 4. Data integration and mapping tool 5. IT asset/Database catalogs 6. Data quality tool 7. Event messaging tools 8. Service registers 9. Business rules repository 10. Configuration management tools
Operational data about data lineage					

Table 3: Analysis of relationships between data lineage, metadata types and sources.

As an example, let's use the conceptual layer of data lineage to explain the content of Table 3 and the conclusions that can be derived from it. Data lineage at the conceptual level requires documentation of the following objects: business processes, functions, roles, and IT assets. Business, technical, and operational metadata describes these objects. For example, business processes and IT asset owners are examples of business metadata. The names and descriptions of IT assets are technical metadata. The execution time of a business process and data sharing rules and agreements between different IT assets both represent operational metadata. The above-mentioned business, technical, and operational metadata can be found in the following solutions: business process modeling, enterprise architecture, data governance, and IT asset catalog.

In Table 3, column "Metadata repositories/sources," I referenced 15 different types of metadata repository, but this list could be expanded. The biggest challenge is the integration of these various repositories. No one integration solution can handle such a wide range of metadata or host multiple metadata repositories. However, the capabilities of automated data lineage systems can assist with scanning, integrating, and storing relations between various metadata repositories.

Figure 4 schematically illustrates the idea of tasks being distributed between metadata management at four abstraction levels and data lineage capabilities (marked grey). The relationships between different metadata repositories are shown only for demonstration purposes and can be much more complex in "real life."

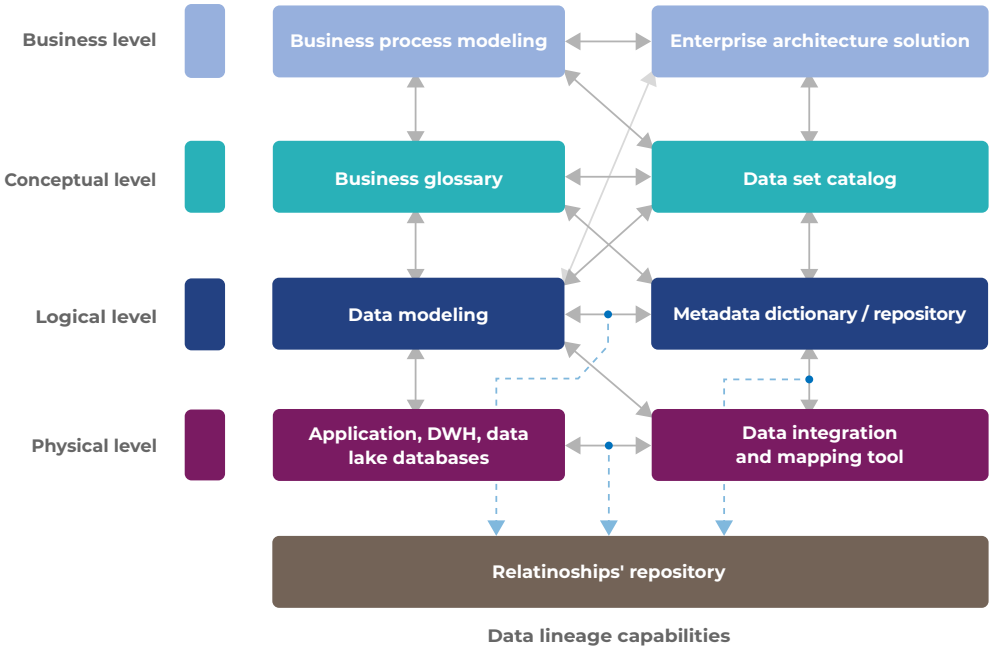


Figure 4: Distribution of tasks between metadata management and data lineage capabilities.

In this way, the distribution of tasks between metadata management and data lineage demonstrates the enabling role of data lineage:

- Metadata management plans, designs, implements, and maintains metadata of various types in corresponding metadata repositories.
- Automated data lineage capabilities plan, design, implement, and maintain integration and visualization of relationships between various metadata repositories.

MANTA is committed to bringing you expert guidance on the power of data lineage. Bookmark our resource library for complimentary whitepapers, market guides, case studies, and more.