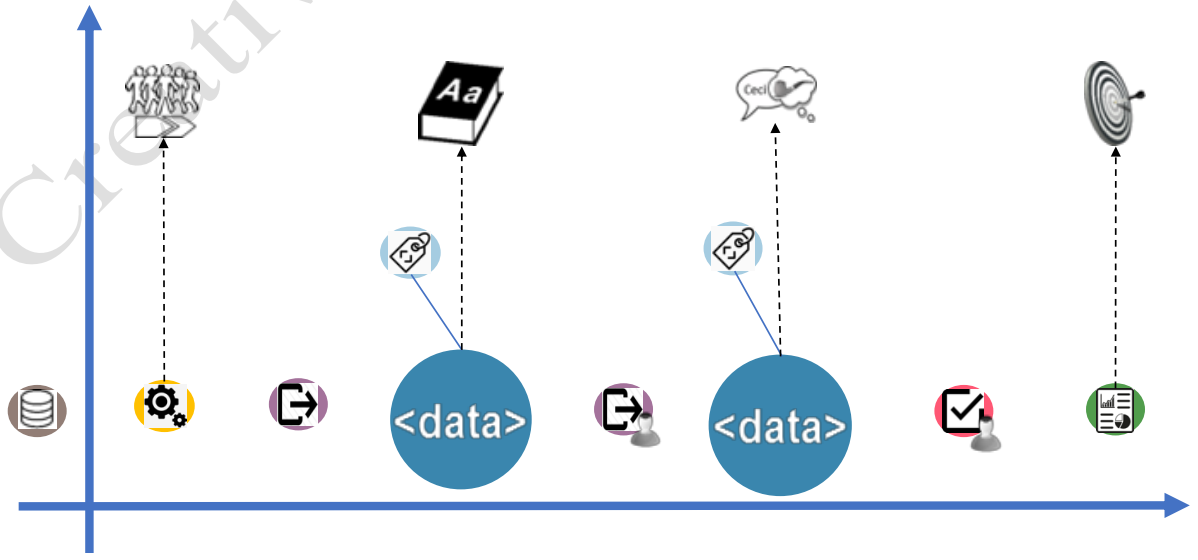
Ensuring data traceability

Topic	**Logistic aspect procedures**
Purpose of the document	Meet the analysis and optimization requirements when producing data lineage
Key words	*Data lineage, modeling, execution, data, Praxeme, method, procedure*

Reference	PxPCD-64	*Status*	Validated
Version	1.1.1	*Date*	4 May 2019
Authors, contributors	Contribution from CONIX Consulting		
Proofreader	Dominique VAUQUIER, *Translator* Joanne TOWARD		

Contents



Methodological reminders

In the context of the Praxeme method, a *procedure* is "a way of doing something, an operating mode for executing a task"[1]. It is therefore a stipulation on an individual level, in contrast to a *process,* which is a methodological response on a collective level.

The procedure sheets do not refer to possible processes in which these procedures may play a role, in order to facilitate their reuse in several contexts.

To follow the developments of the open method

- Mailing list
- LinkedIn group
- Twitter
- the wiki

To participate in the work of the Praxeme Institute

- Become a member of the Praxeme Institute

http://www.praxeme.org/communaute/

Document protection

The initiative for an open method rests on voluntary work and the pooling of investments between contributors. It aims to develop and disseminate an open, royalty-free method. Its dynamics only works if this spirit is maintained in the way the documents, which have been made available to the public, are used. This is why the documents are protected with a "creative commons"[2] license which authorizes the use or reuse of all or part of a document from the Praxeme corpus, the only condition being that the source is cited. The same conditions should also apply to any documents likely to be derived from Praxeme content. They must refer to the "creative commons" and feature the appropriate symbols:

Updates to this document

To obtain the latest version of this document, go to the Praxeme Institute website to the catalogue page: http://www.praxeme.org/telechargements/catalogue/.

Document history

Index	Date	Author	Content
0.0.0	26/11/2018	DVAU	Creation of the procedure sheet
0.0.x			Exchanges JB / DVAU
1.0.0	26/02/2019	JB, DVAU	First published
1.1.0	28/04/2019	J. TOWARD	Translation to English, additions by JB
1.1.1	4/05/2019		Fig. § 5.5
1.1.1	04/05/2019		Current version of the document

[1] Cf. Thesaurus section on the Praxeme Institute website: http://wiki.praxeme.org/index.php?n=Thesaurus.Procedure.

[2] See the philosophy and license detail at: http://creativecommons.org/.

Introductory remark

This procedure is part of a larger set: Praxeme's open method, the object of which is the controlled transformation of enterprises and complex systems. On the one hand, the procedure benefits from its links with other procedures, covering a broad spectrum of intervention on all aspects of the enterprise. On the other hand, each procedure sheet is designed to be used independently; a few reminders are included in the initial sections to this end.

The term "aspect" refers to a central notion of the method. This notion is linked to the representation framework, the Enterprise System Topology, which forms the basis of the method and organizes a multidisciplinary approach of the enterprise[3].

1. Application context of the procedure

1.1 Purpose

This procedure "Ensuring data traceability" is a practical guide for establishing a data lineage, that is to say retracing the movement of data.

The result obtained (data lineage) is a means that will be used later in response to the analysis, optimization and regulatory requirements, bringing into play data production through the description of traces (see §1.2 usage situations).

To describe data lineage, we differentiate traces at an execution level (the date at which certain data was entered and by whom) from traces at a modeling level (the processing or transformation that certain data types undergo within the context of a particular process involving (or bringing into play) a particular a business object).

The Praxeme method provides a rigorous framework that allows us to formalize data lineage (to qualify and describe it) and to include it in a broader Enterprise vision.

This procedure is a practical guide that benefits from the Praxeme framework to describe the operations needed to allow us to gather and represent data lineage. This involves gathering data about data (or traces about data), which we later refer to as metadata (the type of data processing and the name of the data source are both examples of data about data). Our aim is to reconstruct this data about data, in the best possible representation, so as to efficiently exploit the result produced. The Praxeme framework allows us to ask, in a structured way, all the questions needed for the job of collecting and reconstructing the metadata. We will see, later on, how this metadata or these traces organize themselves quite naturally in the Praxeme framework.

The effort required to establish data lineage is costly and fastidious. The purpose of this procedure is to improve the efficiency, reach and productivity of this exercise. It fixes the representation framework of data lineage as well as the related instructions to enable the industrialization of the approach.

Central concern: Carrying out data lineage – documenting the movement of data – answers the need for traceability about a data element or a dataset.

By nature, data circulates *in* and *between* information systems. When data circulates, it is subjected to changes in value, digital transformations, changes in state and in events. Gathering information about these changes and events provides us with end-to-end traceability about the data, from acquisition to use.

Controlling this traceability is becoming increasingly important due to the value that certain data has acquired, like regulatory data or indeed data linked to customer knowledge. Data has become both the revenue and the assets of the enterprise. Ensuring its traceability comes down to controlling:

[3] See, by way of introduction, the General guide, reference PxMDS-01.

- how data is produced: the data value chain (the data supply chain, combined with production traces);
- the valorization of data as assets: characterizing data as an asset, according to its origin and its quality;
- the associated risks: noncompliance with a data policy, data integrity risk at one step in the processing cycle.

The production of data lineage must allow us to respond to these stakes.

The effort required to reconstruct data lineage is often a necessary one to compensate for the failings of an information system at inception (a flaw in the traceability at the time it was built).

This effort is also necessary due to the growing complexity of processing chains (historic "piling up" of blocks coming from different systems – mergers/links between I.S., opening up of I.S., new technological layers, absence of IT city planning…).

1.2 Usage situations

In the absolute, data lineage answers an organization's need to guarantee the validity of the data it manipulates and requires to run its activity. Its purpose is to justify the accuracy of the data: data lineage contributes to the trust placed in the data.

Data lineage and the related metadata are a means of answering different issues where data is a stake (strategic, tactical, and performance-related).

Examples of issues:

- Regulations and steering (Business Intelligence): provide proof (trace) that the composition of an indicator or a key data element (for example, revenue) is the right one to comply with a regulation or to meet steering needs (examples: BCBS RDARR, BCBS 239 which require banks to prove the origin and way of building indicators linked to the Basle regulation. Or else, in a BI project, to ensure the reliability of a dashboard.)
- Analysis of observed differences on the values of a single variable (measure, aggregate, indicator), provided by different sources.
- Impact analysis: when a change happens on a processing chain, what impact will there be on data production, dashboards…?
- Understanding the source and type of data in data science exercises (data preparation loads).
- Data Quality Management (DQM): how we can improve data quality by putting improvement actions all along the processing chain (at data capture, using controls when data is integrated, etc.).
- Optimization and simplification of data production and indicators: over time, data production circuits become more complex. Using data lineage to represent them allows us to have an overall vision and enables us to optimize them (example: rationalizing management rules, optimizing the purchase of external data).
- Ensuring security rules and data access are respected along the whole circuit (in answer to a dissemination or data protection policy).

Among these usages, some come under a top-down approach: we seek to verify that the data pathway respects the mandatory rules and quality requirements, for example, of a regulation (like the GDPR or BCBS 239 for banks).

Other usages come under a bottom-up approach: from the data pathway, we seek any malfunctions or possible optimizations, for example in the context of a quality assessment.

Several factors condition how data lineage is carried out:

- the complexity of the organization (national, international, multinational,…) that data lineage is part of,
- the multiplicity of sources on a very large dataset,
- the multiplicity of ways of acquiring and storing data (in an ERP, in the cloud, in Big-Data platforms…),
- the globalization of acquisition pathways (omnichannel situation),
- the multiplicity and redundancy of information,
- the openings and exchanges with stakeholders, partners, clients…

Generally, data lineage provides an overall vision on how a single data element is processed, manipulated by multiple blocks coming from different systems and environments. This overall vision is the first contribution data lineage brings before any resolution of issues. It enables a dialogue between actors, based on a common representation.

1.3 Nominal development and deviation

The aim of the data lineage exercise is to represent the nominal situation, that is to say the normal development, planned from the outset and when the processing chain runs normally.

This representation allows us to analyze the deviation during exceptional situations. For example, during the summer, the manual transfer of a data flow became degraded.

The representation of the data lineage (nominal situation) acts as a base map to show, using deltas, the exceptional situations or issues (see chapter 4, operating mode, for an example of the idea of positioning an analysis tracing of the issues on the base map).

1.4 Positioning in the method

a. Place in the reference framework

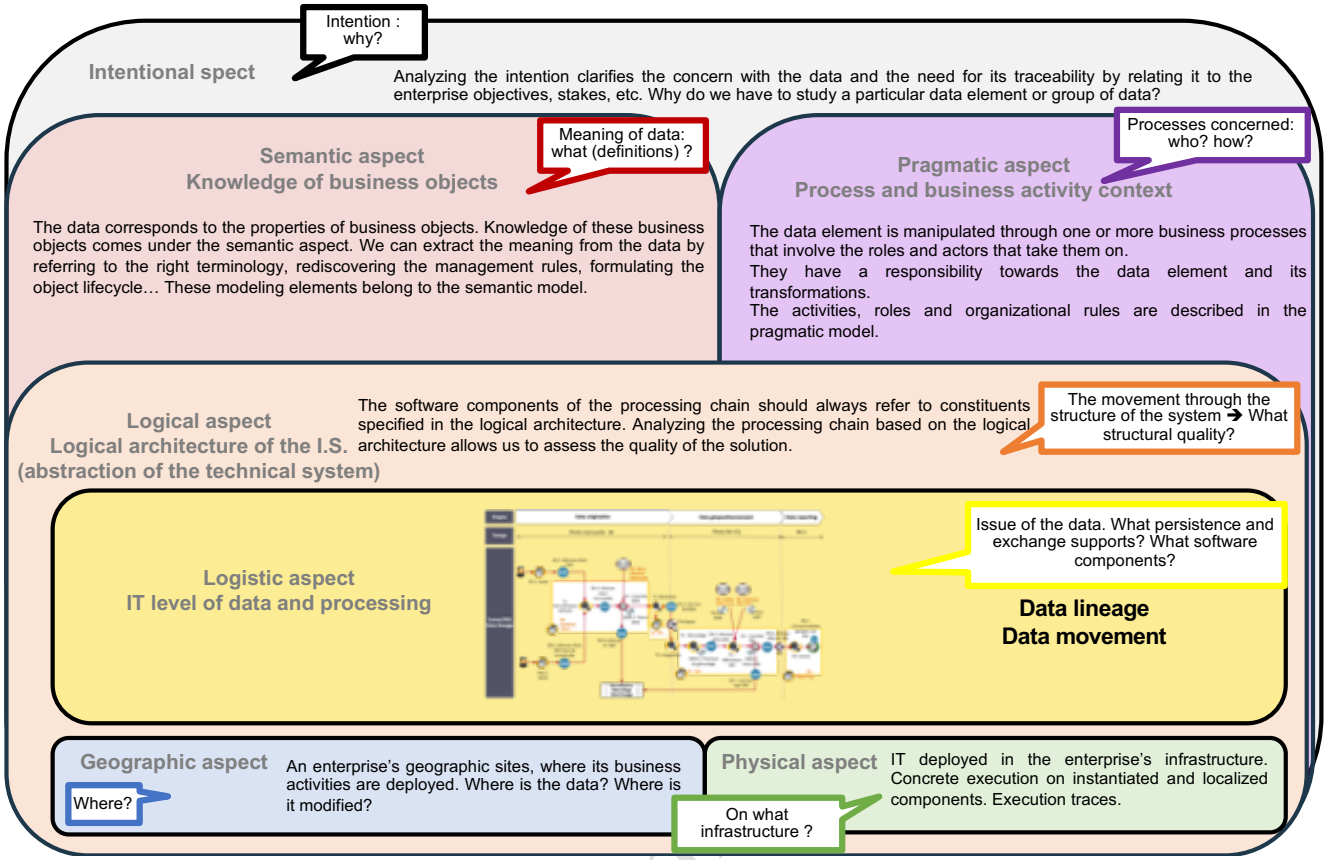
Establishing data filiation or data movement is an act that comes under the logistic aspect as defined in the Enterprise System Topology. This aspect covers the technical systems that serve the enterprise's activities, more particularly the information solutions. The request that triggers the application of this procedure is always a sign of concern about the IT system.

As a result, the bulk of the instructions proposed here concern software and the logistical model. However, additional actions enrich the procedure, enabling us to contextualize the data studied and to better deal with how it is used. In that way, the procedure has its place in a true enterprise architecture approach and has to interact with the other aspects of the enterprise system. The motivation behind this expansion lies in the desire to maximize the results of the effort undertaken. Answering the need for traceability presupposes a considerable investment given the usual state of poorly documented systems[4]. We recommend that this effort be used to lay the foundations for good documentation and to supplement the enterprise description repository. This additional effort will bring a minimal additional cost, whereas it will prevent new spending in the future.

The following figure summarizes the contributions to traceability documentation that can be obtained from the different aspects of the enterprise system. They will be the topic of the actions presented in the operating mode (chapter 4 of this data sheet).

[4] See the load indications in section 3.

Figure PCD-64_1. Contextualization of a data path, in regard to the Enterprise System aspects



b. Relations with other procedures

This procedure "Ensuring data traceability" can benefit from the deliverables produced by the following procedures:

1. **terminological procedures**, mainly "Define a term" (ref. PxPCD-14a) and "Build a thesaurus" (PxPCD-14f), because one of the first things to do is to check the correct understanding of the data expected as a result;
2. **semantic modeling procedures** (mainly PxPCD-22a), to formally express the notions involved in the traceability request;
3. **pragmatic modeling procedures**, when the procedure manages to connect the data to the activities that manipulate it (procedure PxPCD-32, concerned with the "business" perception, upstream from the IT solutions);
4. **logical aspect procedures**, including the logical architecture and IT city planning, whose products can guide those involved to find the components at stake;
5. **geographic modeling** of the enterprise, the spatial distribution of the activities and definition of the sites that may have an impact on data circulation;
6. **logistic aspect modeling**, producing the software component catalogue (if this catalogue exists and if it respects the rule book, the investigative work will be greatly facilitated);
7. **physical modeling**, whose deployment diagrams will shine a light on the issues linked to the duplication of data sources and data processing ("production" domain).

Ideally, these procedures will have been applied while the system was being built, thus enabling us to benefit from the descriptions needed to ensure the traceability. In practice, more often than not, those involved will have to produce part of this information themselves.

c. Posture

Praxeme distinguishes between the analysis and design postures that apply to all aspects of the enterprise[5].

Establishing data traceability comes under the analysis posture as it describes – *a posteriori* – the state of an existing system.

Most of the time, the procedure is applied in answer to a requirement to comply with a regulation or the need to treat quality issues or else to tackle an unexpected impact. We could say that it is a defensive gesture, calling for analysis in order to respond as quickly as possible. The action documents the existing without adding anything. It is investigative work.

However, beyond this minimal and essential posture, the work can extend the analysis and its assessments, by proposing several ideas for improvement. These can later lead to a design effort (like looking for optimizations in a processing chain).

Thus, although this procedure pertains to analysis, part of its results can spur design thinking.

2. Terminology employed

2.1 General notions

The terminology linked to the methodology is gathered together on the Praxeme wiki[6].

The table below gives the definitions of the main notions used in this data sheet and comments on them in the context of this procedure.

Figure PCD-64_2. General notions and their definition

Notion	Definition	Comment
Aspect	"Part of reality, which has been isolated for the sake of study, in accordance with its inner logic"	The results of this procedure are enriched by the multi-aspect approach[7].
Logistic aspect	"Aspect of a system made up of its logistic means"	Here, we are only talking about IT means.
Repository	"Set of elements shared by a community"	
Enterprise description repository	"Repository containing all elements gathered in the course of the work carried out to describe the Enterprise System"	See chapter 5, on tooling. Central mechanism that enables us to work more efficiently.
Business object	"Concrete or abstract object, essential to the Enterprise System's mission"	Business objects are the terms in which the fundamental business knowledge is expressed.
Process	"Set of scheduled activities"	
Procedure	"Prescribed way of doing something"	
Data	Computer-stored representation of information	
Metadata	Data about data	Details on this notion are given later on.

[5] See the white paper, ref. "SLB-02" et ibid.

[6] Wiki: http://wiki.praxeme.org/index.php?n=Thesaurus.Thesaurus.

[7] For the definition of all the aspects, see the thesaurus on the wiki. For their justification and explanation of the representation framework, see the methodological guide PxPRD-01.

2.2 Data and Metadata

From here onwards we will speak of "useful data" to designate data that is reconstructed in response to the request for traceability.

The term "metadata" has two meanings, equally involved in the lineage procedure:

1. On the one hand, the radical "meta" refers to the representation obtained through abstraction efforts. It is the data model, no longer its concrete value, but the variable positioned within a model, accompanied by its description in the form of the type, wording, comments and rules. For example, a data element is expressed as a value of an attribute in a class, this, in turn, representing a table in a database or a concept at the semantic level.
2. On the other hand, "metadata" is data that accompanies useful data and which is linked, more often than not, to a usage or an execution. Another term used to describe metadata is "envelope data". The image is explicit: useful data is the letter, slipped into the envelope, the content useful; metadata is the information written on the envelope: the addressee, date, sender, postage requirements (tariff, registered post, settlement), etc.

Both of these metadata notions are also needed to control a system's data. We will come back to them later in the procedure. They are illustrated in the following table.

Figure PCD-64_3. Examples of metadata and questions linked to data

Example	Execution metadata	Representation metadata
Information state (validated, confirmed, doubtful…)	The value potentially changes for each data element.	Variables (attributes) are added to useful data and accompany it. The model has to enrich itself. One feature of such properties is that they can be defined at a generic level, for example on the root of all business concepts (semantic aspect) or on the root of exchange structures (logical aspect).
Quality of geo-encoding		
Date of update		
Level of knowledge		
Capture campaign		
Information source. Where does the data come from? How is it used?	To designate a particular source: the person who entered the data; organization from which the data was acquired… Or a particular usage.	The answer can also be produced at the model level: in the case where the data always comes from the same source (same software component, same supplier…).
Who created the data? Who uses it? Who does it belong to? Who is responsible for its processing and maintenance?	Answer at an individual level (individual or entity), potentially different for the same data type (for example: sales rep A entered details for client X).	Answer at the model level (sales reps always enter client data; "sales rep" is a role which appears in the pragmatic model).
What is the business definition? What are the business rules? What is the degree of security? Regulatory constraints…		Answers through modeling (the questions are not about a specific data element, but about all data of the same type, the same meaning). Semantic aspect.
Where is the data stored? What are the standard denominations with the databases?		Logical aspect model, first (logical data model) and the physical aspect: the same diagram can be instantiated several times in the physical architecture.

Example	Execution metadata	Representation metadata
Why are we storing this data? What is its use and purpose? What is the business lever to use it?	Exceptionally, the answers can be individualized (specific rate for an entity, for special reasons).	Refer back to elements of intent (intentional aspect).
When was the data created, updated? What must it be deleted?	Answer possible on a data level (for example, linked to client preferences).	Answer possible at a same data-type level (for example, referring to archiving rules).
How is this data formatted? In how many databases or sources is it present?	For ad-hoc acquisitions (the format may vary).	Logistic modeling (formats and supports); physical modeling (duplication and dynamic phenomena resulting from physical redundancy).

2.3 Lexical field of traceability

In everyday use, as in industrial contexts, traceability is defined as:

> *Capacity to reconstruct a determination chain*[8].

Traceability contributes to establishing the confidence in a system be it a manufacturing or delivery process, organization, technical system…

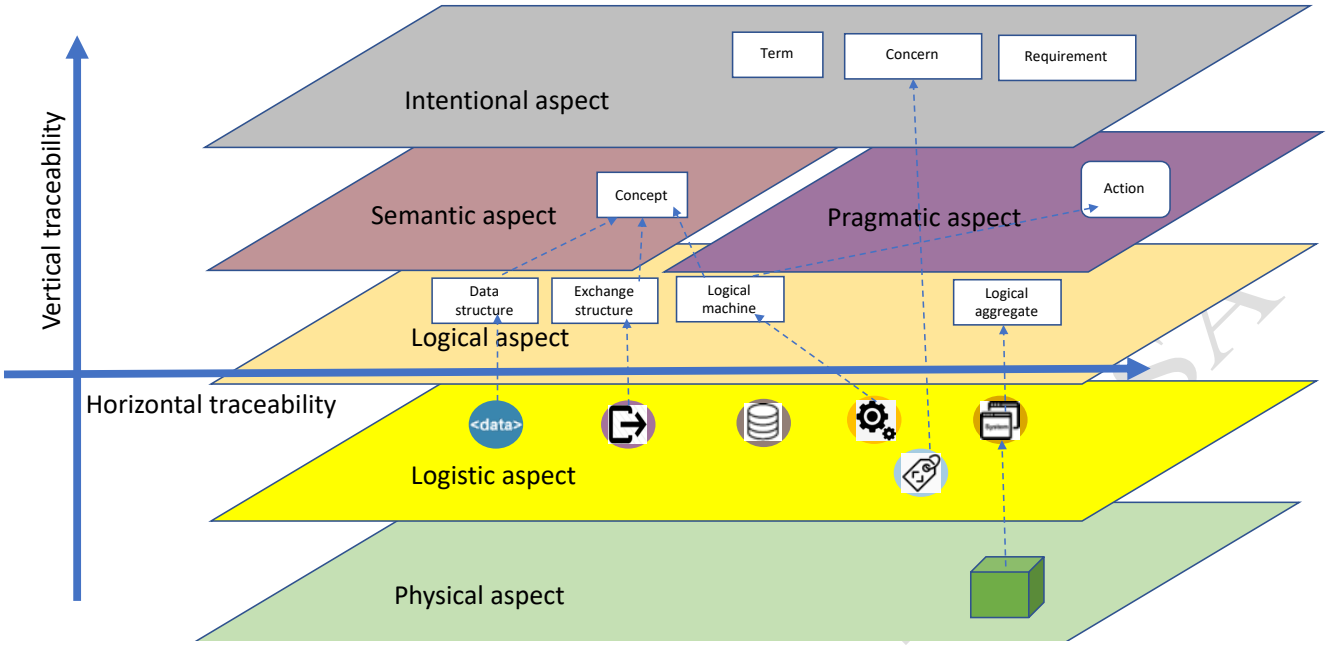
Regarding data traceability, we have to distinguish between two types of traceability:

1. *Horizontal traceability*, which is deployed on the execution plane: it is set up in the form of production chains, leading from one or more sources to a result.
2. *Vertical traceability*, developed on the construction plane: it is part of the models, linking an element from a downstream aspect to an element from an upstream aspect.

The diagram below illustrates vertical traceability with a selection of representation categories. In this sense, a traceability chain links elements belonging to different aspects. This diagram illustrates horizontal traceability only through the logistic aspect, but it can also be found in other aspects, for example in the form of a process in the pragmatic aspect.

[8] Source: http://wiki.praxeme.org/index.php?n=Thesaurus.Traceability.

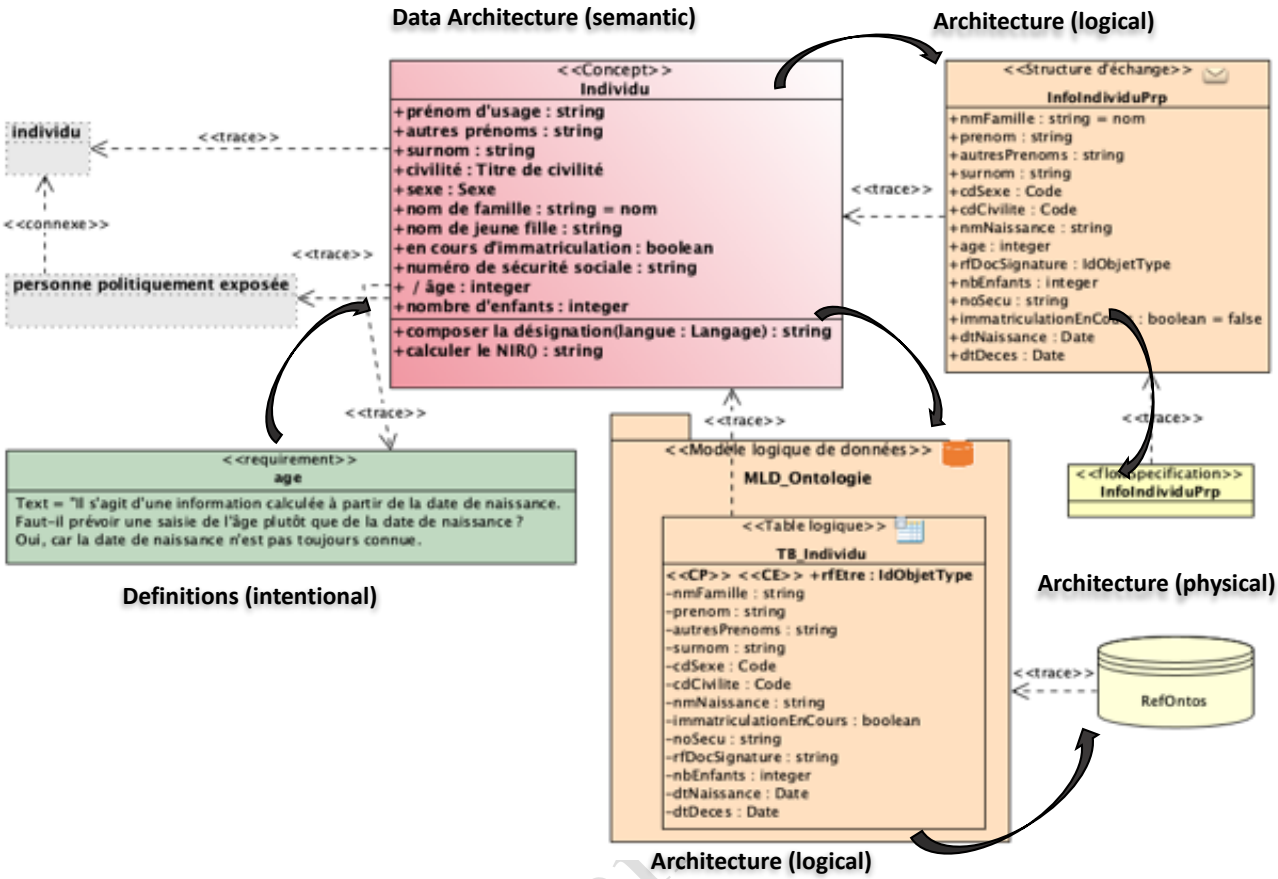
Figure PCD-64_4. The two traceability dimensions



Horizontal traceability of data is first expressed as the succession of steps needed for its production. This description comes under the logistic aspect model, and answers for one type of result – a property, a variable, an object class… If the requested traceability does not concern a type of result but a particular result – a data element, a value, an instance –, then it is appropriate to add the envelope-type metadata to the production chain, that we will call "execution traces".

Vertical traceability is made up of "construction traces", that is to say of links between modeling elements from different aspects. A complete traceability chain allows us to attach the property studied, of a logistic nature, to its logical specification, and to go back up from the latter to a business aspect element and, from there, to a requirement or an objective in the intentional aspect. These construction traces correspond to metadata in the modeling sense. They are apparent in "trace"-stereotyped dependencies according to the term used by the UML notation. Praxeme requires us to establish these traces while respecting the dependencies between the aspects of the Enterprise System Topology. This rule reduces coupling within the description repository and simplifies its exploitation.

Figure PCD-64_5. Illustration of vertical traceability: notion of an individual



Comment on the diagram

There are two types of elements of intent represented in this figure: terms taken from the dictionary; a requirement concerning the property "age". In red, the semantic class "Individual" formalizes the business concept. It provides the starting point for several projections in the logical aspect, of which two are shown: the exchange structure; the data structure (in orange, color of the logical aspect). Both of these logical elements are translated in the software level (in yellow) into an XSD schema, for example, and a database table.

When these traceability chains are perfectly documented in the enterprise description repository, the impact analysis, in the event of a change or an audit, can be dealt with in a few minutes, instead of several days.

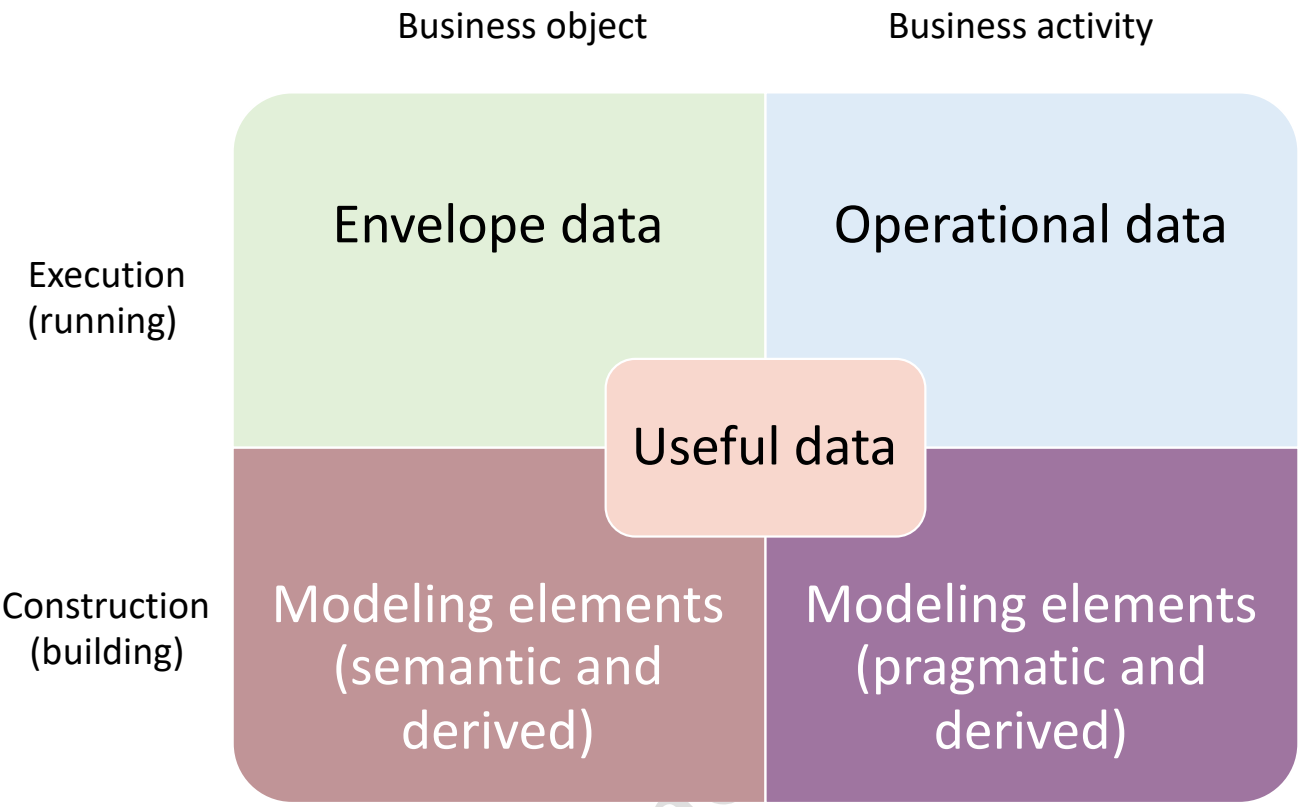
2.4 Metadata categories

Useful data – that which is the subject of the traceability request – is, almost always, essential information for the business: it concerns a "business object", that is to say a fundamental notion. Retracing its history makes us highlight the activities that manipulate it. From here, the "business activities" appear, written into the enterprise processes. Here too, both interpretations of the "metadata" term can be found:

1. The activity is represented through a model (process model, use case…), and the data production chain must refer to the activity and the process into which it is thrown (for example, an accounting process).
2. At the execution, the activity is instantiated, and it may be necessary to keep contextual information such as: the date of action, actor username, action's execution context, that it to say the "envelope" information.

Thus, by combining the object-activity couple and the envelope-model distinction, we obtain four metadata categories that accompany the data element studied, as the following figure shows.

Figure PCD-64_6. The four metadata categories that accompany useful data



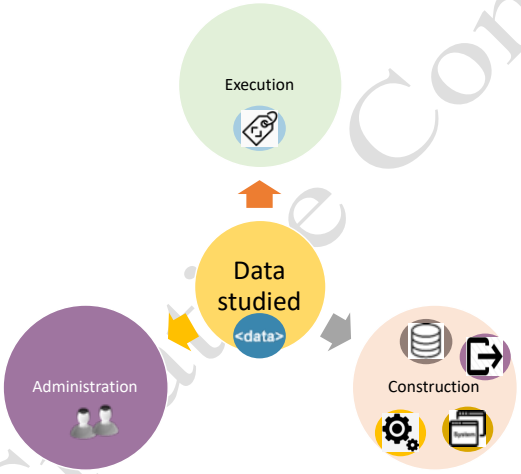
Comment on the diagram

1. By analogy with a letter, whose content would be the data element studied, the envelope data describes the execution conditions that produced or modified this data element. These execution traces allow us to analyze in detail the data element's destiny.
2. This destiny mobilizes actors whose actions may need to be memorized. We then obtain the operational metadata that tells us about the manipulations of the data element and its envelope. They are also execution or production traces. The term "production" is not limited to the meaning it has taken in IT; it also designates human, tooled or manual interventions on the data element.
3. The data element finds its definition and its meaning in a model. At the highest level is the semantic model, ensuring the understanding in business terms. The modeling element that defines the semantic data element is, more often than not, a class attribute. The class itself represents the business concept (or "business object"). From this modeling element, several derivation fields are attached that go through the logical aspect and end up in the physical aspect. Thus, modeling elements to which the data element studied is attached include: exchange and persistence (logical models) structures, their translations into software terms (logistic aspect), and finally their instances in the physical architecture. The construction traces that link these elements together are precious for the control of the system.
4. In the same way, the activities are subjected to modeling. They are described in business and organizational terms in the pragmatic model. The modeling elements are process or activity models, organization roles and rules, as well as execution contexts. These elements can also be derived and are required for data processing.
5. This analysis enables us to draw up a list of information categories that we need to collect to answer a data traceability request. By broadening the field of concerns, we will discover that other types of metadata may come into play, distributed on the enterprise system aspects (see the action "Putting data into perspective").

Metadata is all the more important as the data element can be produced in several ways and follow several paths. Traceability then consists in identifying the path taken by a data element.

Figure PCD-64_7. Notions linked to traceability

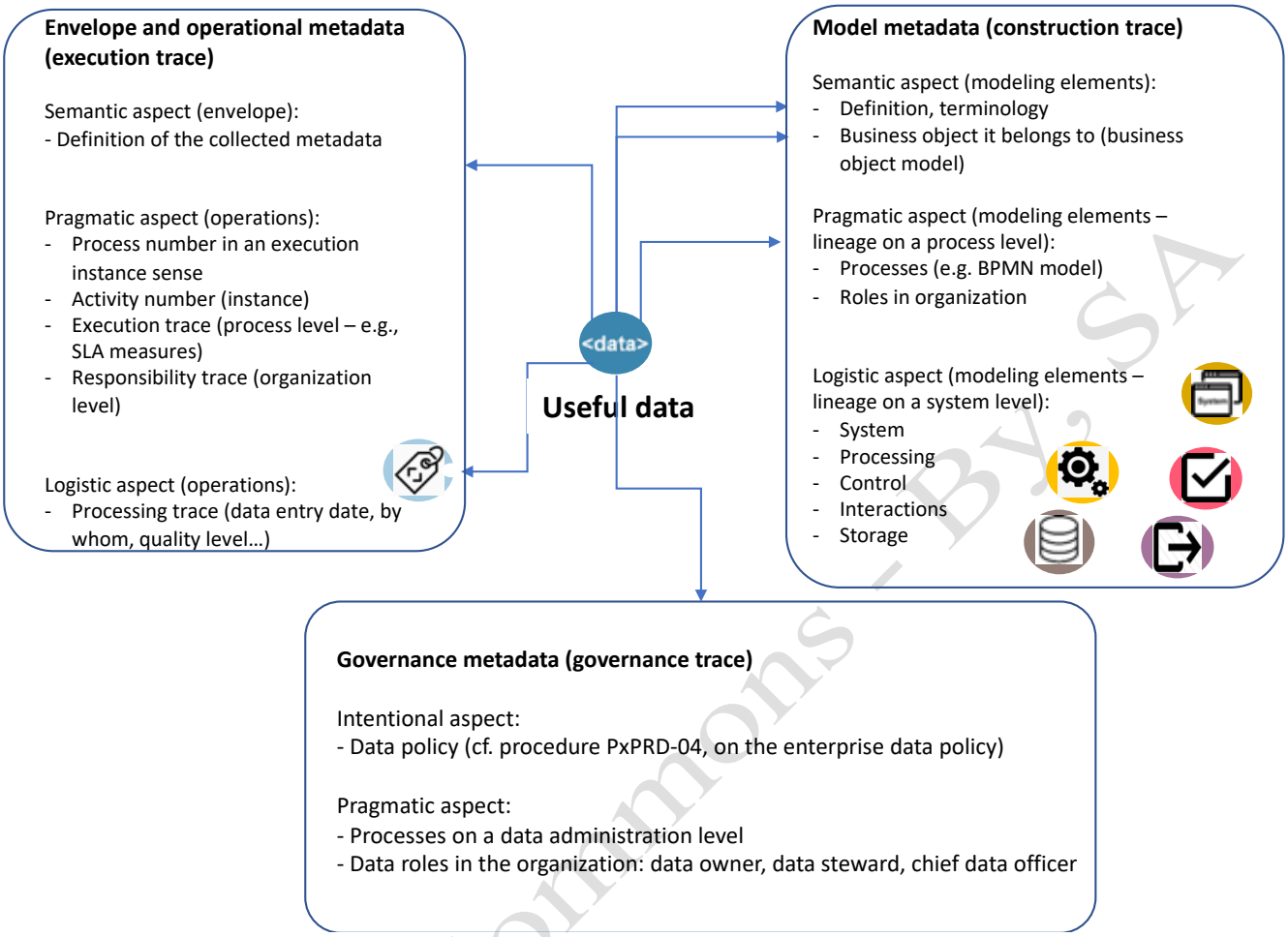
Notion	Definition	Comment
Traceability	**Capacity to reconstruct a determination chain**	The main topic of this procedure.
Trace	**Mark left by an event**	The trace is a sign; it carries information.
Execution trace	**Trace that something happened, at a given moment, in a process**	The event that the trace reveals is linked to an execution (something happened in the system, in production).
Construction trace	**Trace linking the construction choice, in one aspect, to the construction choice in an adjacent aspect** Example: derivation of the semantic model into a logical data model, then into a physical model.	The construction choice is formally expressed through a modeling element. The event is, here, a construction act of the system. The trace links two modeling elements, each one taken in a different aspect.
Metadata	**Data about data** First meaning: element of a model defining a data element or a variable. Second meaning: information on a data lifecycle.	The first meaning refers back to the model. The term "variable" is more appropriate. The data element is a value taken by a variable. The second meaning, to the envelope.
Envelope	**Dataset accompanying useful data**	Equivalent to "execution trace".



To complete our typology of information to collect, we have to take into account the organization and definition of the responsibilities around data, what is commonly called governance (covering, in the main, data administration). At construction and execution, we therefore add the administration dimension (figure opposite). In extreme cases, documentation about the data and its traceability must indeed include the description of the managers and procedures related to the data.

Figure PCD-64_8. The three metadata dimensions

Figure PCD-64_9. Link between data and metadata (traces) – segmentation of metadata according to the aspects



2.5 Data path, movement and production chain

By "data lineage" (term used in the general framework for data management processes), we understand it to mean the reconstruction of the path that a data element or dataset will take within an information system, from acquisition sources to its reconstruction for a given usage.

"Data movement" or "data production chain" are equivalent expressions:

- "Movement" makes us think more about the circulation of data, more or less unchanged, between an initial point, the source, and an end point, the supply.
- "Production chain" makes us think more about the calculation of a result from the data manipulations, by aggregation, consolidation, transformation, etc.

In every case, we have to show the steps the data element or data goes through to arrive at an expected result. These steps are detailed in terms of storage supports, streams and software components.

Establishing data lineage is, first, a job of investigation and reconstruction of the processing chains that bring into play the data. We have to map out the processing chains, covering all successive elements of data management or exploitation. In this exercise, the person involved details the data transformation rules (codification, notation, normalization, aggregation, etc. ...), from the data source (selection, filtering) to its use.

3. Required skills

This specialized procedure is part of a much larger corpus, tackling all the enterprise dimensions. Its interest also lies in bringing different skills into contact with each other and in coordinating several views on the enterprise (also a reflection of the aspects of the method mentioned previously). Beyond the purely technical view, needed to respond to concerns about traceability, the procedure calls upon disciplines that take an interest in business processes and business knowledge.

According to the profile of those involved in applying the procedure, there are two scenarios:

- either they have these different skills or this sensibility, in which case (enterprise architect role), they will be able to roll out the instructions fully;
- or they complete their action by interacting with others: business architect, business analyst, process owner, application manager, IT city planner and I.S. architect, data steward, data "consumers" – data scientist, statistician…

The data lineage exercise is predominately a collective and collaborative one, both in its design and in its use.

Praxeme coordinates these disciplines by allocating them to the aspects fixed by the representation framework. In this, the method and this procedure constitute a tool for dialogue.

Skills involved:

- legal (to take stock of the regulatory requirements),
- modeling: from the semantics to the physical (business objects, logical data modeling, physical modeling),
- architecture: from the business to the infrastructure (specialization in a minimum of three sets: business architecture, logical architecture, IT architecture),
- governance: control of the organization, roles and data governance rules (data policy[9], data administration).

4. Operating mode

4.1 Analyze the traceability need

Data lineage is rarely done in the absolute; it is always in response to a precise intention. The issues to be tackled determine the investigation and data collection work. For problems with data quality, the quality issues identified guide the search for processes, actions in a processing chain perhaps being the root causes of poor quality data. For problems of proof, we will look more specifically for the controls, risks of rupture, gaps between two points (valuation gap – one particular input amount for another output amount, coverage gap – part of the expected scope is not covered – i.e., execution traces). It is therefore important to clarify the intention overseeing the effort. This is why the procedure starts by analyzing the traceability request.

We have to answer the questions:

> *What is the real intention revealed by the traceability request? What will the result produced – the data lineage – be used for? What are the traceability requirements? Is there a regulatory dimension that requires us to provide proof or traces?*

This first action consists in an intentional analysis, in the sense of the intentional aspect of the enterprise. Working out the intention leads us to consider the following points:

- **The type of issues** and stakes for the enterprise: is it in response to a regulatory request to prove the right mode of data production, or to comply with a regulation – like the GDPR[10] – or to analyze a personal data processing chain – as with a PIA (Privacy Impact Assessment), or to deal with malfunctions which are

[9] See the form and its instructions for use PxPRD-04, on enterprise data policy (contribution from CONIX consulting firm).

[10] The General Data Protection Regulation, from the European Commission.

reflected in poor quality data with the inherent risks, or else to optimize a processing chain to be more efficient?

- **The priorities to deal with**, linked to a transformation program or an IT policy: does the planned work provide an opportunity to simplify things (for example, creating a people repository covering sensitive data)? Is the enterprise focused on certain critical data? Does the digital transformation require a better control of the informational heritage? Etc.
- **The actors and responsibilities**: From which actors does the traceability request come from? Which other actors does it involve? Which systems or subsystems are concerned? This analysis sets the scope for the work and the content of the deliverables (systems can be excluded from this).
- **The criteria** that allow us to check that we have answered the intention properly (see the chapter on "Acceptance criteria").

These expectations are entered into the enterprise description repository as elements of intent that we have to show have been satisfied. The first action consists in recording and analyzing them. The following actions in the operating mode will satisfy the criteria derived from these elements (verifying that the intentions have been satisfied).

The intentional analysis is carried out using terms that are specific to the intentional aspect. Elements of intent are formulations that we keep in the enterprise description repository (EDR). For example, the regulations are broken down into elementary formulations (paragraphs, articles…), just as we would do for the requirements. If necessary, comments can be attached to these formulations, which are the results of the analysis. In this work, it may be necessary to refer back to entries in the reference dictionary that is also stored in the EDR. The contributor may add diagrams that link the formulations to the terms (see figure PCD-64_11). If the analysis leads us to add new terms, we then apply the terminological procedures.

Figure PCD-64_10. The types of elements manipulated in the action "Analyze the traceability need"

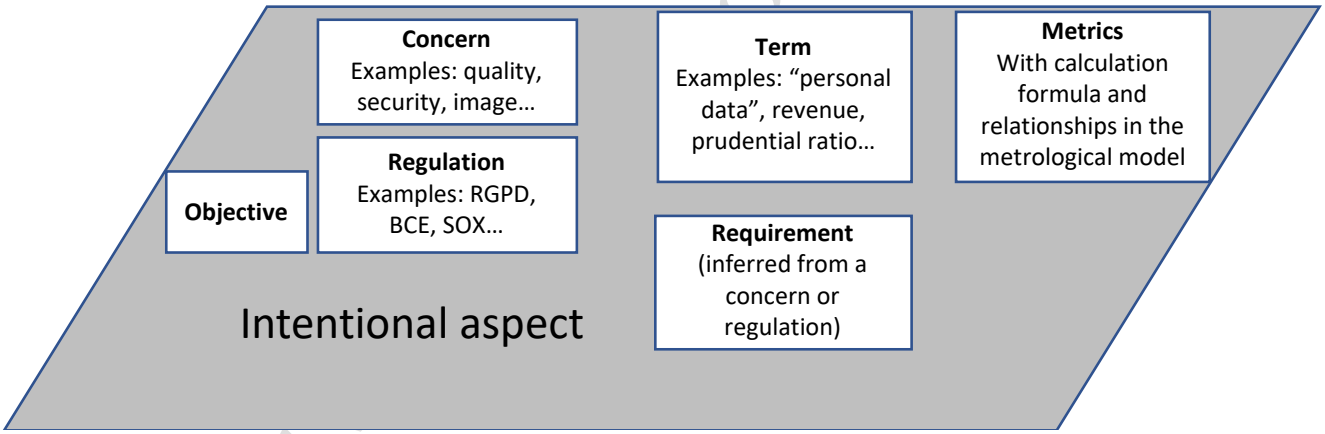
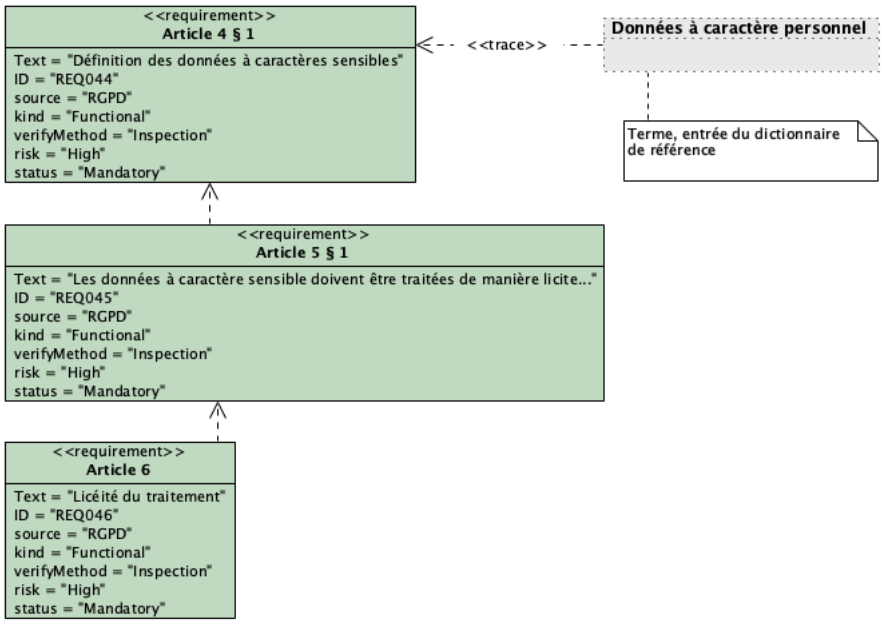


Figure PCD-64_11. Illustration: result of the action "Analyze the traceability need"



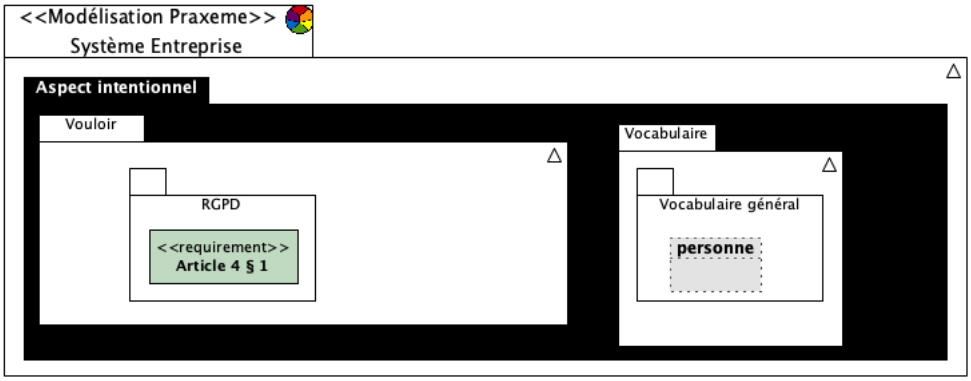
Comment on the diagram

The regulatory text is dissected into elementary formulations, represented here as requirements.

The diagram also shows a term, an entry in the reference dictionary, which provides the definition in reference to the regulation.

The following figure shows how these elements are distributed in the enterprise description repository structure.

Figure PCD-64_12. The distribution of the elements of intent in the EDR (illustration)



4.2 Locating data

a. Identifying data from the intention

The previous action allowed us to clarify the request and limit its scope. Next, we have to find the data, as it is represented in the IT system.

The traceability request can specify the data targeted (for example, revenue) or a type of data (personal data); it may also express a more general concern (data security, the pertinence of the coordinates…). From these expressions, we have to locate the corresponding data in the system.

Whereabouts in the system, under what name and in which form does the data studied exist? To what semantic content does it refer? What business objects are concerned? What is the business definition of the data analyzed? What business rules limit the data? What names are used in the databases, storage and exchange means?

---

b. Reconstruct the construction chain linked to the definition and data modeling choices

Locating data is done, in particular, in the semantic aspect, as this is where the formalization and best expression of the information has to be, from a business knowledge point of view. Upstream, terminology can be a good starting point. As illustrated previously, the traceability request absorbed into the intentional aspect is analyzed in terms and expressions, which make up entries in the reference dictionary. Each term is "projected" into the most suitable aspect, that is to say linked to a modeling element belonging to another aspect. More often than not, the traceability requirement concerns a piece of "core business" information and the term will be picked up by a semantic element. As an example, we can cite: revenue (calculated attribute, set level), personal data (properties of the concept of "individual"), energy consumption in a geographic zone… However, it may be that the object of the request refers back to an element from another aspect. One example is the performance of a process (pragmatic aspect) or an organizational unit (geographic aspect).

Figure PCD-64_13. The types of elements manipulated in the action "Locate data"

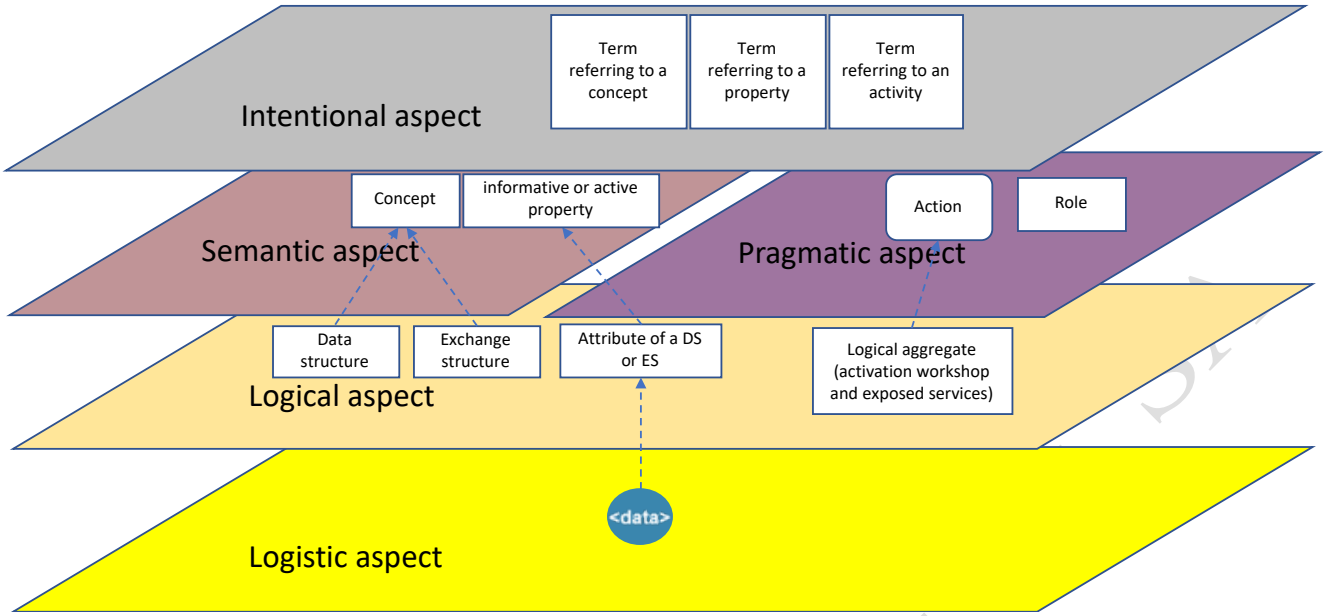
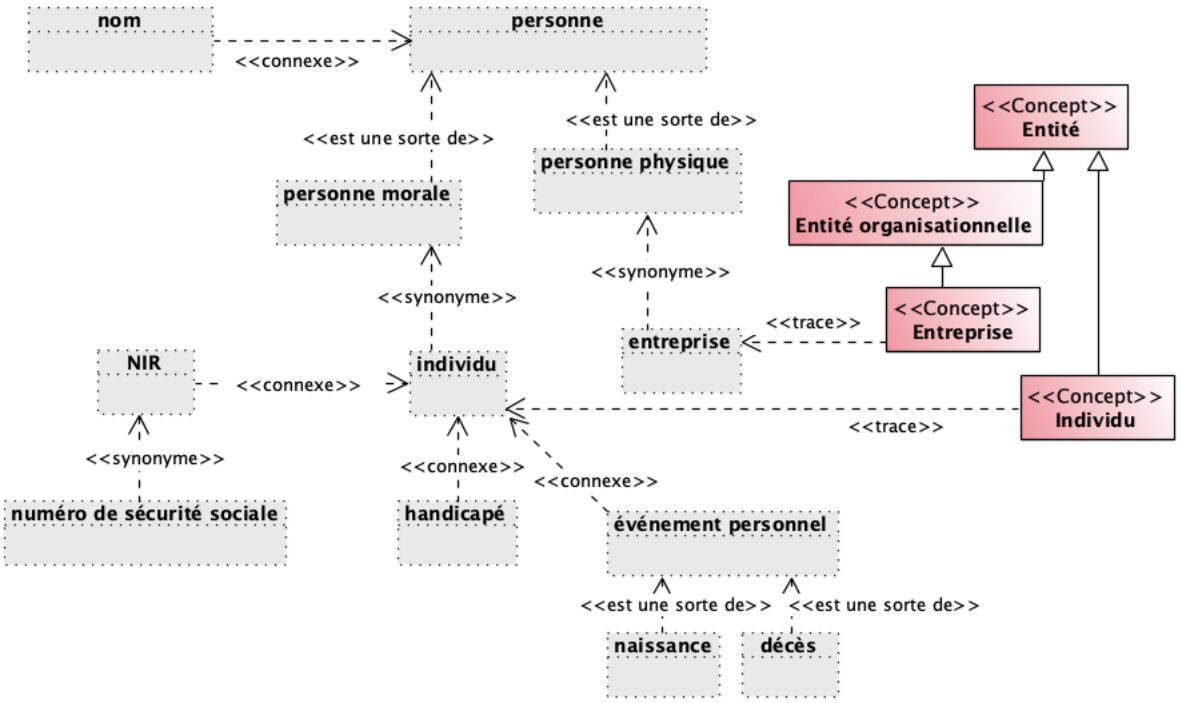


Figure PCD-64_14. Illustration of the terms projected towards semantic aspect elements



Comment on the diagram

This diagram illustrates the results that can be produced by this action (locating data). From the vocabulary found in the traceability request (intentional analysis carried out during the first action), we have to formalize the notions, here by referring back to concepts from the semantic aspect. This illustration remains at the class level. Fairly often, we have to point to the class property levels, as shown in the following figure.

Figure PCD-64_15. Illustration of a projection toward an attribute of a semantic class

The construction traces unearthed by this locating action do not stop at the semantic aspect. They follow the derivation pathways to lead to the persistence and exchange structures, as shown in the figure below.

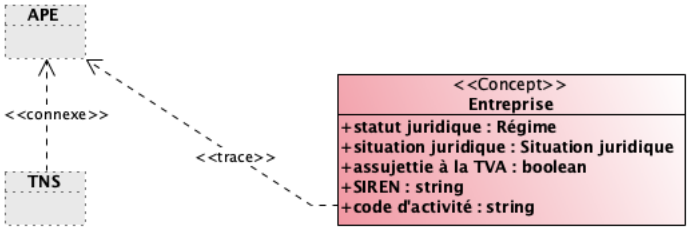


Figure PCD-64_16. The construction traceability chains (terms ➔ logical model ➔ implementation)

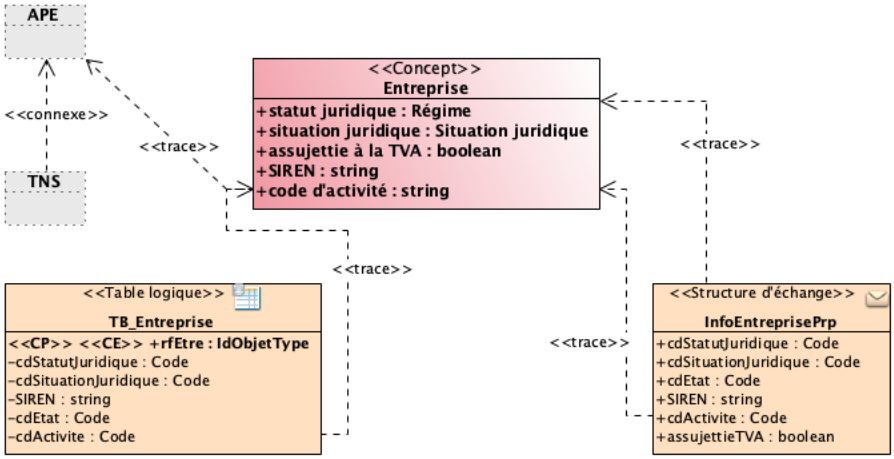


Illustration of a semantic alignment error: Data quality is no longer seen as a storage* issue but one of communication (change in view). Any data acquisition is seen as a message to other actors (human or system). So data quality is seen as an element of the quality of the communication. And this quality of communication depends on the semantic quality of the messages – meaning of the messages (readability – limited encoding so as to be read without decoding, comprehensible and formal – aligned on business standard terms and business models, fair and complete – free from error – respect of the meaning and business rules at the time of supply…).

This quality plays a key role in: exchange norms, storage and data interpretation models (like the semantic layers in data visualization tools) and, by extension, on the data movement from beginning to end.

When we describe a data element as the focus of data lineage (like an indicator), we are at the end of the chain. But, it is the semantic definition of this data element that will set the requirement in semantic alignment for the whole chain. For example, in the context of mains maintenance or a transport operator, if we seek to carry out data lineage on an indicator that wants to provide the measurement of the number of km of network monitored over a given period, respecting the definition along the whole processing chain that will result in the value measured is key. In this case, either we seek to measure the absolute number of km monitored, or we seek to measure the cumulative number of km monitored (which is not the same thing: in the first case, we will measure the section of the network which was the focus of one or more monitoring operations – even if one section of the network was the focus of several monitoring operations, we only count the km of network concerned once; in the second case, we will measure the number of km monitored – if one section of the network was monitored several times, we add up the number of km as many times as there were monitoring operations on those sections. Depending on this difference of definition, we can imagine the impact that a breakdown in the definition in data lineage of this measurement would have).

* The storage quality cannot be better than the quality coming from the "communications".

c. Locating data on the logistic aspect level

Starting from the description of the intention, we set out to describe the data that will be the focus of data lineage: a calculated indicator, a dashboard, a dataset, or elementary data (like an email address).

This description covers:

- **the identification of the data that will be the focus of data lineage**. By identification, we mean a description that allows us to formally recognize the data element (an identifier, a dashboard title, the label number of an indicator, a reference code like the ECB data codification in the BCBS 39 framework…);
- **the most formal definition possible of the data**. Note: this formal definition exercise is sometimes the first expression of a malfunction. The gap in semantic alignment (see the insert on the previous page) is a frequent cause of error that we identify through data lineage. This definition exercise is to be aligned with the semantic aspect of the method, in particular by linking the business objects and their life cycle (as presented in the previous paragraphs on this step);
- in the case of calculated or composed data, **the calculation or composition formula** with the elementary constituents that must be traced (data lineage objects);
- **the governance rules** that may have been associated with the data element that was the focus of data lineage (person responsible for the data element, policy to be applied to this data element – example regulations, confidentiality, business sensitivity…). This constitutes the gathering of the first governance metadata.

Granularity of the data that is the focus of data lineage ("useful data" of which we look for traces)

Data lineage can be carried out:

- at the elementary data level or consistent dataset level (an address, a person's identity, a transaction, a property attribute of a business objects),
- at a steering product level: indicators, dashboard, measurements. Ultimately, the exercise will move down to the elementary data level, constituents of the steering product ("useful data" is both the steering product and its components).
- at a dataset level, like the publishing of datasets for partners, a data lab, in open data,
- and sometimes at a dataflow level, when we seek to retrace the path of a dataflow from its source to its reception for processing, all along a business process (like the business event records or inventory records in accounting, customer dataflows in the enter-into-relationship process and KYC[11]).

4.3 Reconstruct the processing chain

This action consists in gathering the model's metadata, mainly the construction traces in the logistic aspect.

How are the processing chains built? How is the data valorized, transformed, enriched, filtered, merged, mapped, exchanged…?

This action concerns the logistic aspect of the enterprise system. It reconstructs the model of the processing chain and identifies the different components of the IT system that will be found along the data path.

These components are made of software bricks or manual activities:

- of data transformations (calculations),
- of data selection (controls, filters),
- of data storage (persistence, consolidation),
- of data exchange (extractions, transfers, exposed services, putting into an exchange format).

[11] Know Your Customer – banking regulations

We also identify the support systems (or applications) of these components.

Figure PCD-64_17. The types of elements manipulated in the action "Reconstruct the processing chain"

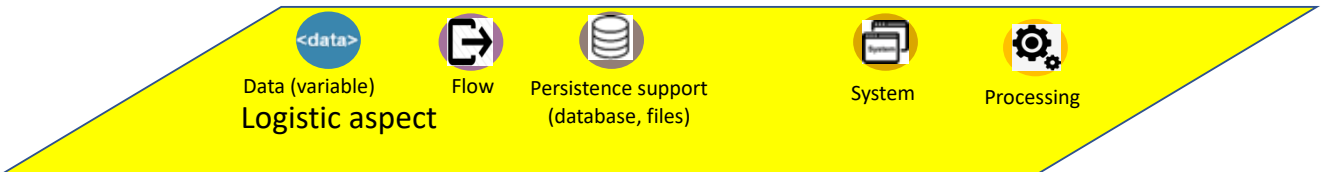
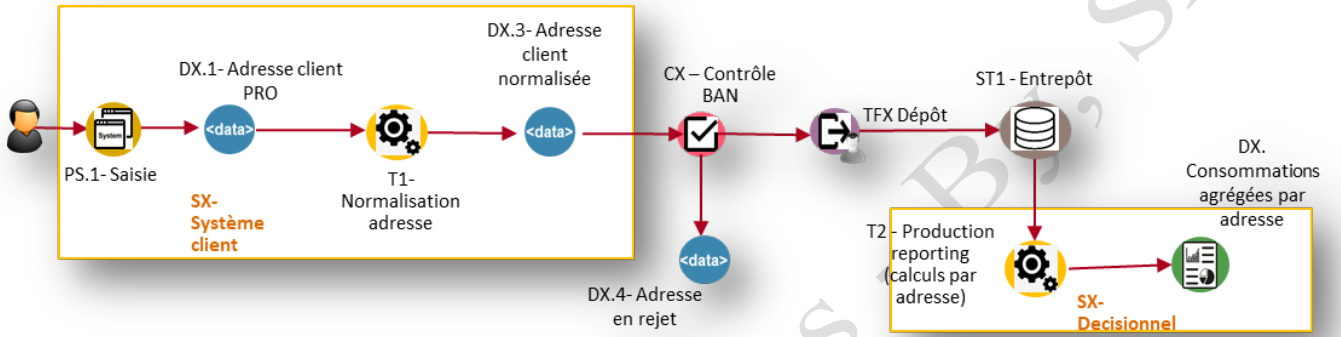


Figure PCD-64_18. An example of a processing chain



This chain describes the succession of processing stages, from entering an address to its use, to provide aggregate measurements by address.

For each component, we look for: its description, the transformation rules of the data that is the focus of data lineage ("useful data"), the event that triggers the processing linked to the component.

In the context of data lineage, by definition, we look for all processing (in the generic sense) susceptible to have an influence on the value of the data:

- application software bricks, which carry data modification processes: calculation, filter, enrichment, transcoding, applying management rules, control brought about by rejections, default values, forced value, data adjustment…
- integration software bricks: data exchange bus, service/API platform of exposed data (these bricks can incorporate transformation rules for content or format, with an impact on the value or interpretation of certain data).

In parallel, for each processing block, we set out to follow the flow of the data targeted by data lineage, by identifying the input and output data in order to follow how the data evolves and to list its transformations, along the chain.

A data lineage exercise is carried out either by starting from a result and reconstructing the successive steps that have enabled this result to be reached (how was certain data constituted?), or by starting from a data source and listing the successive processing and transformation steps of this data (how was a certain data element processed?). At the end, the objective is to collect, then to represent the paths the data will follow through the I.S.

In practice, we alternate both approaches:

1. by seeking to reconstruct where the data element comes from (we go back up the processing chain starting from the end);

2. by checking that the production of a particular data element, entering normally in the setup of the focus of data lineage (an indicator) has been properly used by the components of the processing chain (we go through the processing chain starting from the sources).

The identification of data must be linked to the semantic aspect (link with the business objects) and must respect the definitions that a data catalogue may have listed (see the section on tooling, chapter 7: data cataloguing solutions provide representation capacities to data lineage by basing themselves on the cataloguing data).

Ideally, all this information can also be found in the I.S. description repositories, support to I.S. architects and city planners.

4.4 Retracing the execution

Compared to the previous ones, this action is a change in register: it is no longer on the construction plane but on that of the execution. It gathers the envelope metadata or execution traces. Here, we are talking about instances: data values, objects in the sense of class instances, process and activity instances, individuals having carried out actions.

How are the processing chains executed and what traces can we recover? What events are produced? When? How did the execution of the processing chain go?

Depending on the issue dealt with and the context, we may need to identify the envelope metadata (execution traces).

This metadata can be found in the I.S. production activities, in the execution of activities linked to a process:

- execution period for a processing series,
- level of quality and service observed,
- coverage of the processed data, data value gap detected,
- author of an update,
- history of the successive data values (example: changes in a phone number),
- identification of the dataflow, dataset or production instance, the focus of data lineage (example: a particular dashboard at a particular date).

Figure PCD-64_19. The types of elements manipulated in the action "Retracing the execution"

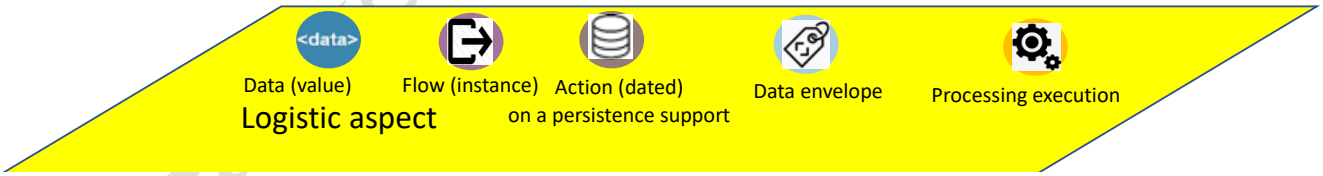
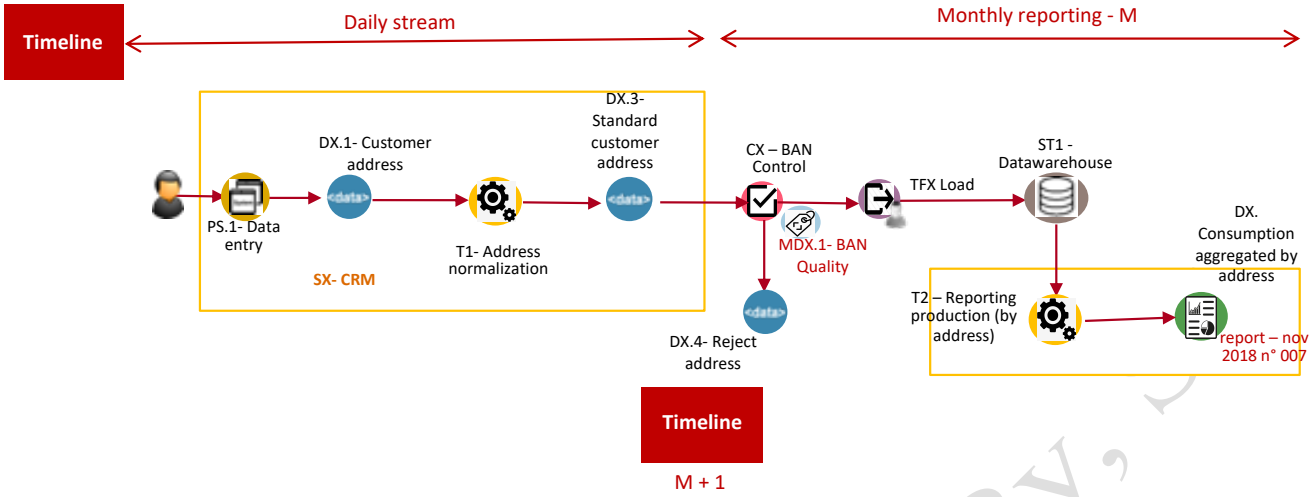


Figure PCD-64_20. An execution – dated – of the processing chain



On this diagram, the execution metadata is in red: execution time, data quality obtained – example: address quality, instance and version of a report that is the focus of data lineage…

Certain metadata can be provided directly by the processing chain, when, within this chain, it has been explicitly instructed to collect this metadata (examples: timestamp, quality level of data that is a result of processing, author of an update, value preceding the update kept… more generally the execution logs[12]).

4.5 Appreciating the conditions of data production

The question of data quality is practically intrinsic to the data lineage exercise (as we are dealing with data, its quality is a key feature).

Appreciating data quality comes under a procedure in its own right. The data lineages produced are input knowledge, essential to dealing with data quality (understanding where the failings may happen, going back to the causes of poor quality…).

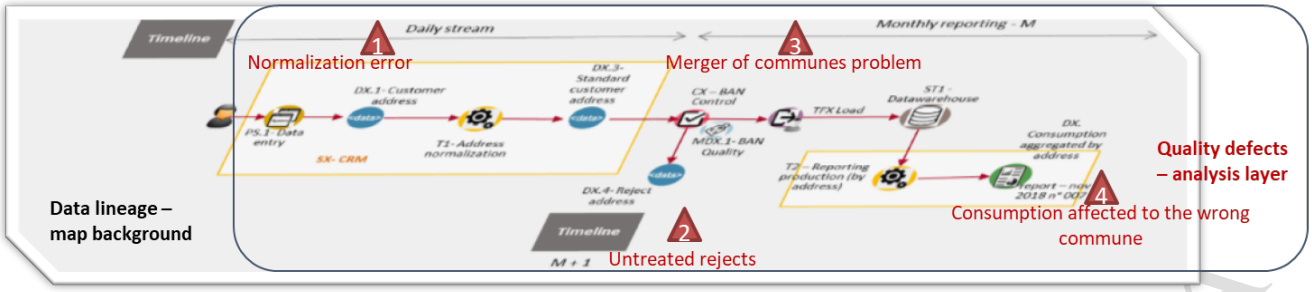
As an illustration in the context of the data lineage exercise, we can identify:

- the design failings (vertical traceability – model's metadata) such as semantic differences in the implementations (the implementation does not respect the semantic definition);
- the steps that may generate failings in the data processing and thus generate quality flaws (and with, why not, a specific processing path for rejections);
- the points in the processing chain, to measure the level of data quality (measurements that can be automatically controlled and stored in dedicated metadata, allowing the quality levels to be taken into account following on from the processing chain), or to check the absence of a value gap, occurrences between the start of processing and its output (as with accounting controls);
- in general, any situation (processing, context) that threatens data quality in terms of: exhaustiveness, accuracy, completeness, integrity, consistency, freshness.

All this information can be retranscribed on an analysis tracing that we superimpose on the representation of data lineage.

[12] A record of the steps executed.

Figure PCD-64_21. Detecting failures along the processing chain

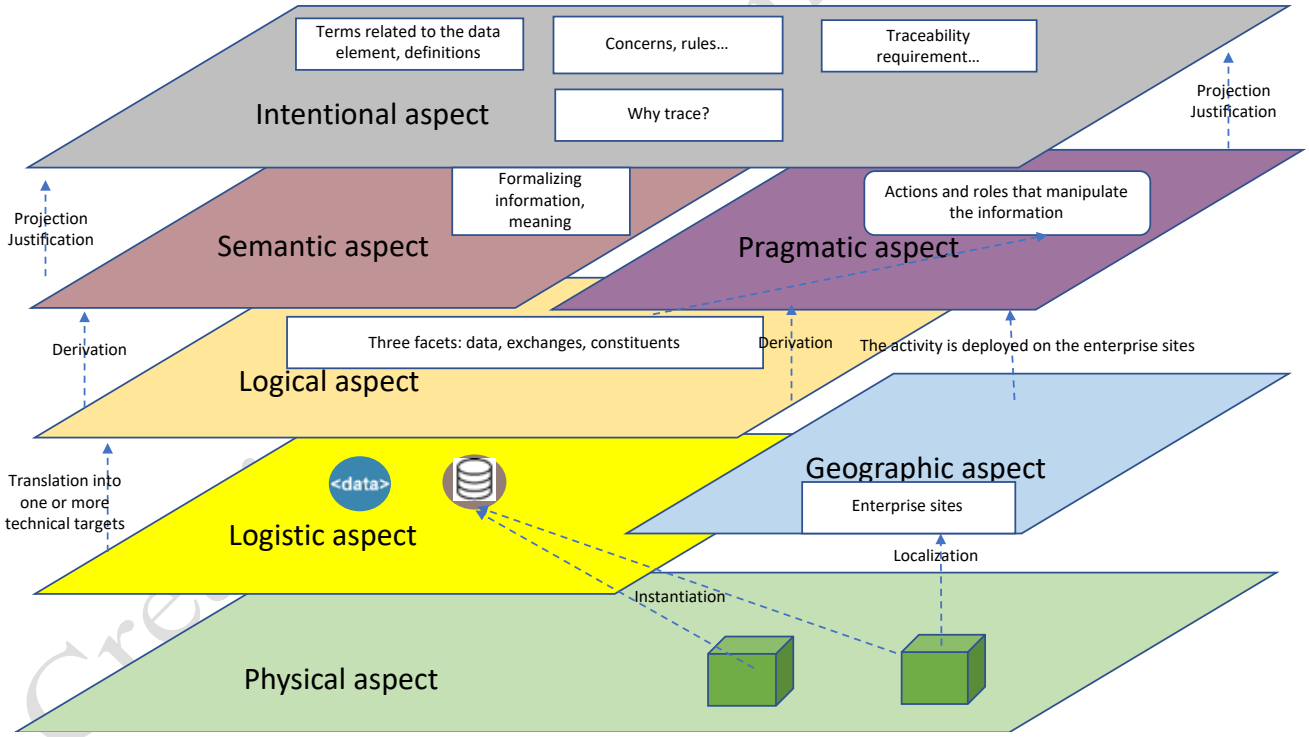


This diagram illustrates the analysis tracing idea on top of the representation map of data lineage. The tracing allows the issues to be highlighted and positioned for dealing with.

4.6 Putting data into perspective

The previous actions are enough to answer, strictly speaking, the traceability request. Putting data into perspective requires a little extra effort, with the aim of getting the most out of the procedure and preparing the future. This action consists in documenting the construction elements linked to the data studied. Ideally, it reconstructs a vertical traceability chain from beginning to end: from the intentional aspect to the physical aspect, by taking all the derivation paths from one aspect to another.

Figure PCD-64_22. Putting data into perspective: types of elements according to the aspects



Putting data into perspective consists in linking it to modeling elements belonging to upstream aspects, that is to say the "business" universes:

- Under the pragmatic aspect: potential actions on the data studied; for example: who can modify the address (role and responsibility in the processes and organization)? In which framework or context (powers, context)?
- Under the logical aspect: if the logical model reflects what exists, then it shows redundancy; from where the idea to simplify the system comes from.

- Under the logistic and physical aspects: take into account the phenomena that increase the risks to the data, notably using caches, IT solutions that come within "shadow IT", "cloud computing"…
- Under the geographic aspect: the localization of activities in the enterprise's geography is a multiplication factor as it leads us to deploy the same IT processes and solutions on several sites.
- Under the physical aspect: the deployment is obtained through instantiation of software components and localization on nodes distributed in the enterprise's geography.

a. Business context: semantic aspect

By nature and as described in the first action (locating data), the semantic aspect comes into play at the very start of the reconstruction of a data lineage. It allows us to set the type of data targeted by data lineage, then the type of data identified all along the processing chain combining to bring about the final result (a product report for example).

The elements that we consider in the semantic aspect include:

- the relation of the data with the enterprise's business objects (example: a particular data element refers to the business object Person in its role of Client),
- the object domain or object subdomain that the object in question belongs to,
- the relationships between business objects,
- the formal definitions obtained by positioning the notion (the business object) in the concept network.

In the context of data lineage, working on the semantic aspect provides the benefit of:

- contextualizing the data (to know what we are speaking about on a business sense),
- ensuring the right translation of business knowledge, through the data targeted by data lineage (vertical traceability).

b. Business context: pragmatic aspect

What are the activities, processes and roles that bring the transformations on the data? In the business processes, what attention is given to data?

Data lineage, by definition, formalizes the data processing. This processing is the reflection and support of the activities of the organization's processes.

All data lineage is therefore part of the execution of one or several processes.

The representation of this process level (in the pragmatic aspect) guarantees the proper interpretation and understanding of a data lineage.

In the context of a data lineage, the work on the pragmatic aspect allows us to:

- contextualize the data processing (to know in which activity, of which process, this processing takes place),
- ensure the pertinence of the data processing on the data targeted by data lineage compared to the activities and organization rules (vertical traceability).

Admittedly, data lineage is mainly located in the logistic aspect. It is first in the hands of the project owner and project implementation teams, and enables us to produce an application representation of the processing chain system (often by processing phase – due to the organization in silos of the I.S., more rarely from beginning to end). This part of the work intervenes as closely as possible to the reality of the IT realization and execution. However, work on the business aspects – semantic and pragmatic – attaches the data and its movement through the system to the business context (process and activity concerned). It helps us go beyond the processing silos by reconstructing a vision from beginning to end (from the data source to the data produced). We sometimes refer to this as functional data lineage.

c. Implementation context: geographic aspect and physical aspect

Geographic aspect: for some processing chains, it may be important to distinguish between the execution sites that generate different behaviors (for example, industrial sites or legal entities in different countries).

How does the localization of data processing intervene in the understanding of the processing chain, the traces?

Where is the data stored? Where does it come from – from which site? Where is it used, shared? What local / country regulatory or legal norm does it answer to? What local variations are there in management rules, data policy?

In some situations, data lineage at a processing or system level is not enough to analyze a failure. You have to go down to the physical deployment level (storage instance in such-and-such a database, located, notably by using a proxy). We complete in this way the data lineage exercise of a system or a process by studying its deployment in the physical aspect of the enterprise.

In what way can the physical deployment have an influence on how the processing chain will behave (traces)?

Where is the data physically located (relational database, NOSQL database, BI system…)?
In how many databases or sources is it present?
With what level of service?

4.7 Specify the data administration and related responsibilities

This action consists in gathering what we can call administration metadata (traces). This data can be found on two levels:

1. at the model level, the models refer to the planned roles and responsibilities, as types;
2. at the execution level, with "administration traces" mentioning the individuals involved in particular processing occurrences.

How is the administration of the target data of the data lineage managed?
Who created this data element? Who consumes it? Who does it belong to? Who is responsible for its administration? Which policy is related to it? What are the related risks?

For all data entering the production chain studied by the data lineage (resulting data, data making up the result, original/source data, transformed data), we have to identify the related administration rules:

- data owner or, more widely, RACI[13] support to the data life cycle;
- data policy (regulatory sensitivity, associated risks – example of publication – confidentiality);
- data production methodology (particularly useful for indicators, for example subject to labeling, or else in the context of statistical data production – statistical method used);
- dependency rules with reference systems (respect of nomenclatures / codifications, consistency with data repositories – example client repository)[14] ;

[13] R: responsible A: accountable C: consulted I: informed.

- reporting and data quality measurement.

This metadata is useful:

1. at the level of the actions carried out in a data administration context (should already have been identified during the processing chain reconstruction, through administration activities linked to the data lifecycle – example of data deletion),
2. at the level of the knowledge of administered data for its improved management.

5. Results produced

5.1 Representation requirements

The representation of a data lineage goes through associating textual descriptions (metadata inventories) and visual descriptions (adopting symbolic representation shapes of the metadata and the articulations between them).

Direct reading of the inventory of the metadata collected is not immediate. It is difficult to have an overall vision. An additional graphical representation is essential.

This representation aims:

- to make the processing chain visible from beginning to end (overall vision),
- to be a mapping repository to position the collected metadata and reproduce the data movement within the data processing itself, the identified rules (role of a map),
- at the moment of dealing with the issue, to position the analysis points (identified points of failure, optimization, improvement: idea of the analysis tracing mentioned in the previous examples).

Experience shows us that, in many areas of the enterprise, we can carry out data lineage. It therefore becomes interesting to define a common symbolism (a common language) on an enterprise scale.

All the metadata collected during the analysis of the processing process is recorded and will serve as a support to a data lineage representation. The way of recording metadata comes under the responsibility of the tooling of the procedure. The solutions range from using an Excel file to putting solutions of data governance into place at an enterprise level, via the use of enterprise architecture tools. This point is detailed in chapter 6.

Before deciding on the tool, an organization starting out on this work must set a level of requirement for the representations:

a) Either it adopts ad-hoc representations, especially for the lineage exercise: intuitive flowcharts with a suggestive ad-hoc symbolism or coming from an authoritative body (as with the ECB), created by using ordinary drawing or communication tools; tables or forms to present the collected metadata.
b) Or it highlights the global requirements for controlling the knowledge of the system and opts for a repository approach, imposing the use of standard notations.

The first option is appropriate as a quick response to the traceability request and to facilitate the communication between the actors involved. It reaches its limits when putting the data into perspective (action 7) and the work on the construction traces (first actions).

The second option takes advantage of the enterprise description repository, with an acceleration effect in the medium term thanks to the ability to capitalize on the knowledge of the enterprise system in all its aspects. It requires us to familiarize ourselves with notations, in particular UML, which are applied to most aspects defined by the Enterprise System Topology. Other notations complete the toolbox, in particular BPMN for process

[14] This dependency must also appear at the model metadata level (construction traces) – flows / interconnection with enterprise repositories.

modeling. The intentional aspect may, potentially, be formulated using intuitive notations, less formal, such as Archimate.

The choice of one or other of the options depends on several factors:

- Are the traceability requests limited or, on the contrary, will the work be recurrent?
- Does the enterprise benefit from a good level of maturity regarding enterprise architecture practices? Does it implement a "repository"-type approach to efficiently manage its knowledge capital?
- Do the skills mobilized cover the practices of notation and modeling techniques?
- Do the stakes linked to traceability justify the effort required to learn?
- Will the results have to be shared between several organizations (subsidiaries, partners, service providers…)? If so, it is in our interest to rely on standards, providing that we do not risk a rejection.

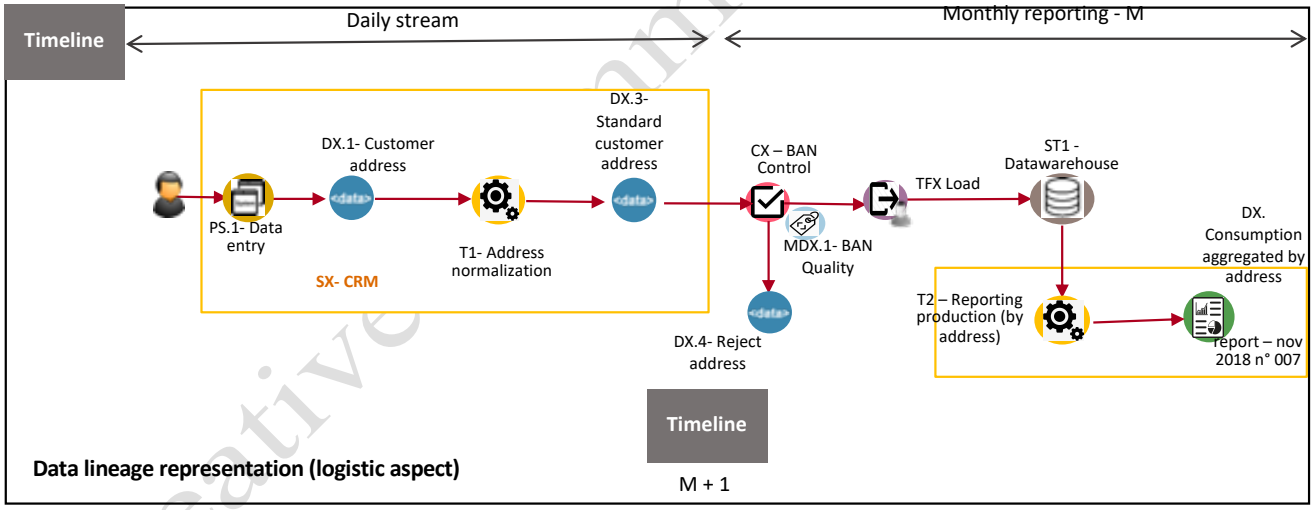
The figure below illustrates a form of representation of a data lineage. Other examples will follow later on in this chapter to illustrate the results produced.

Figure PCD-64_23. Example of a form for publishing a data lineage

Identity :
- Data lineage title : _____
- Author / Organization : _____
- Date : __/__/__
- Version : __
- Status / validation : _____
- Business Process : PX_____
- …

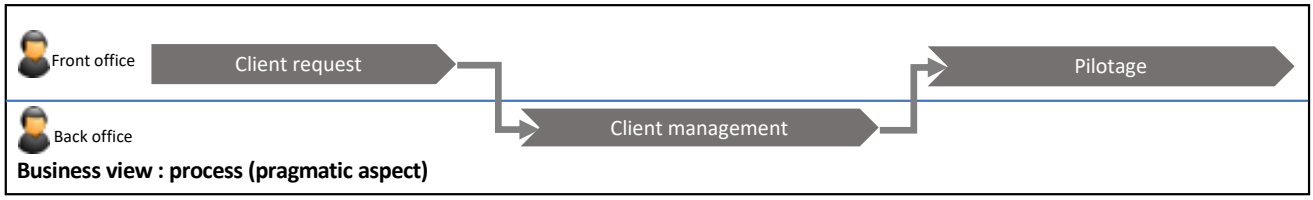
Meta-data

Nature	Code+Label	Description	Traces



Comment - description

Input step	Standardization step	Control step	Reporting step



5.2 Example of a product – case n°1: data quality

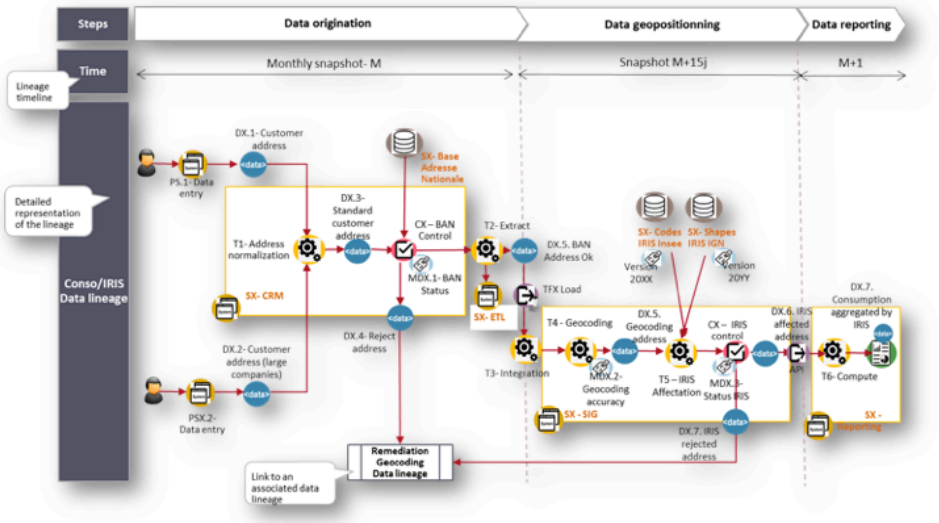
Data concerned:

- Result that is the focus of the data lineage: production of energy statistics by IRIS zones (in France, IRIS was developed by INSEE, France's Official Statistics Authority, and is an acronym for "aggregated units for statistical information" and is the fundamental unit for dissemination of infra-municipal information[15] (division Insee/Institut Géographique National).
- Data to be traced: geopositioning data (address, link to an IRIS zone) enabling the consumption data for each point to be linked to the correct IRIS zone.

Issues:

- Ensure (traceability) the exhaustiveness of the territorial coverage (all consumption points were able to be correctly geopositioned to their correct IRIS zone).
- Identify the optimization leads of the statistical production chain: improving the quality of the geopositioning of data, optimization of processing times.

Figure PCD-64_24. One form of representation of a data lineage



Key:

- Data, focus of data lineage
- Metadata generated during processing
- Processing bringing data into play
- Data transfer
- Data storage spaces
- Data control
- System (in the service/application sense – source, processing support, controls…)
- Result produced (focus of the lineage – "useful data")
- Manual interventions

Comment on the figure

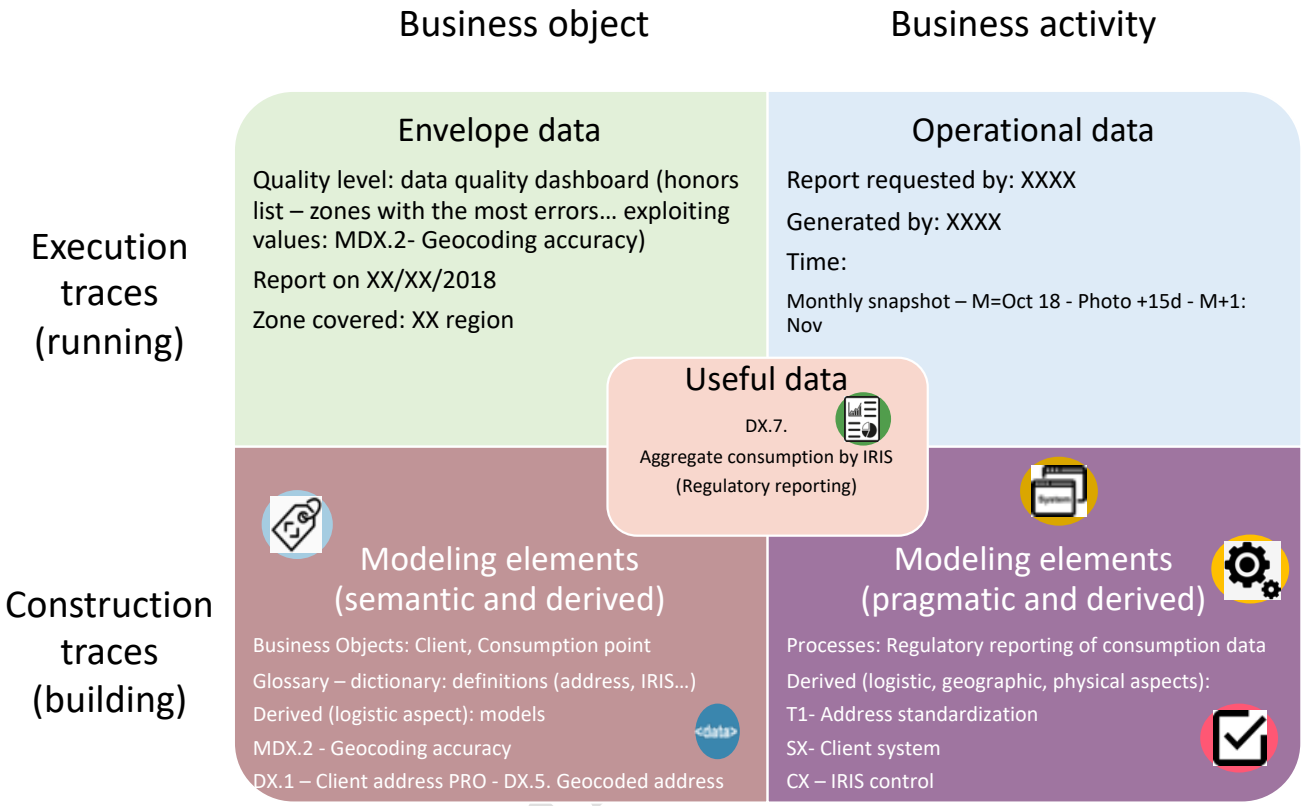
This example presents the data lineage of data concurrent with producing a regulatory report.

The intention is twofold:

[15] Source: https://www.insee.fr/en/metadonnees/definition/c1523

1. to provide proof that the expected report has been properly realized (report approved),
2. to analyze the failures that appeared during the execution of the first reports (quality errors).

Figure PCD-64_25. Typology of the metadata collected



Example of a metadata collection table:

Nature (execution, construction)	Code+Label (DX1__, T1__, SX1___…)	Description _____	Aspect	Traces (…)	Comment

5.3 Example of a product – case n°2: regulatory (GDPR, BCBS)

a. Banking domain – regulation BCBS 239

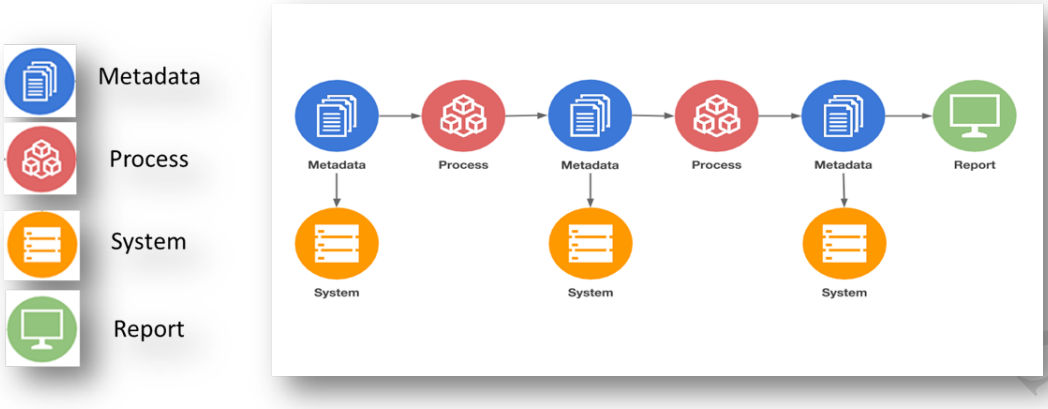
Banking establishments have been faced, since 2014, with a restrictive schedule to comply with the Basel Committee for Banking Supervision (BCBS 239), which propelled data lineage (implementing traceability) into the top 5 projects that have to be addressed and implemented to improve the banks' ability to produce and make reliable regulatory reports. These reports are subject to regular testing by the European Central Bank (ECB).

The ECB has thus produced several data lineage description frameworks concerning the production of regulatory reports (such as indicators linked to a credit risk or liquidity risk).

These description frameworks set a nomenclature of the reporting data to be produced by the banks (equivalent to a data dictionary).

They also set a minimum symbolism to aid comparisons. Banks have to focus on four types of elements: metadata, systems, processes and reports.

Figure PCD-64_26. Symbolism specific to the ECB[16]



Data lineage consists in providing the processing chain that allows the regulatory report to be produced, by identifying the systems (example: a database) and processes (activities: data processing – examples an ETL script, an algorithm) concurrent to producing the report and by highlighting the information about the data (metadata: column name in a table for example).

Instructions from the ECB for producing data lineage (extract):

"Data lineage is defined as the data lifecycle that includes the origins of the data and its movement over time. It describes what happens to the data when it passes through different processes. It helps to provide a representation support to the analysis and simplifies the tracing of errors to their source. In other words, data lineage is a process map, a useful instrument to evaluate the way in which banks implement the BCBS 239 principles on the aspects linked to aggregation and reporting of risk data."

Data lineage must enable banks to ensure their responsibility vis-à-vis their consolidation processes of risk data.

For the purposes of comparison, in this in-depth analysis, banks have to focus on four types of elements: metadata, systems, processes and reports.

Instructions provided by the ECB regarding the elements in the present procedure:

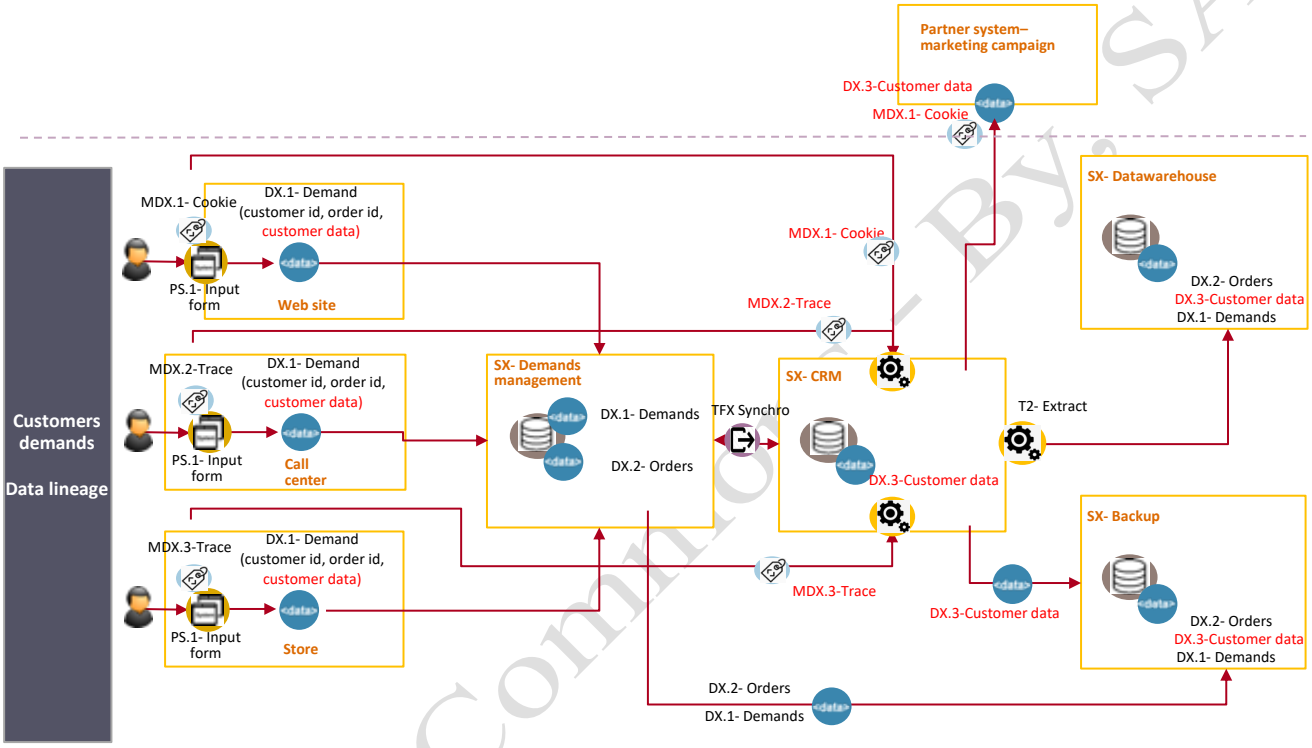
1. describe the purpose of the data lineage (example the amount of deposits and their components) -> *Procedure – "useful data" – locating data;*
2. establish the data lineage: i) how are the corresponding figures completed?; ii) by which IT systems?; and iii) if there is one or more than one source of data. For each database included in the data movement, indicate: 1) the technical solution (DBMS) and 2) the data taxonomy. -> *Procedure – "construction metadata" – reconstruct the processing chain;*
3. identify and explain the steps in the data processing process that are fully automated, partially automated or manual;
4. document the type of controls, including for partially automated and manual processes, put in place for each stage, the functions responsible for these controls and the staff incentives at all levels of the banking group (headquarters, branches and subsidiaries). The people or functions responsible should be named. In particular, explain at what level of the process the controls are carried out, what are their results, how are the errors identified, signaled and corrected (escalation), how are these figures reconciled with other sources, if the data is unambiguous in a data dictionary. -> *Procedure – construction metadata (semantic and pragmatic aspects) and governance metadata – put the data into perspective;*
5. explain what data quality indicators are in place, including the tolerance levels, and explain the process in the event that they are breached. -> *Procedure – Appreciate the data production conditions…*
6. describe if any potential modifications to data processing could improve the efficiency of the processes and the reliability of the figures -> *Procedure – usages.*

[16] There is no official symbolism repository. The ECB refers to it. It drew its inspiration from a software vendor.

Remark: via BCBS, the ECB has set its traceability requirements. These requirements were first centered on aggregate data. The current evolution, through BIRD – Banks' Integrated Reporting Dictionary and the example of Anacredit, will lead banks to no longer provide aggregate data but granular (elementary), normalized (aligned on the ECB's dictionary) data. The expected traceability will have to be at the level of this elementary data with all the consequences and new requirements to be respected in terms of controlling this traceability at its most elementary point at the banking systems level.

b. Managing personal data – the GDPR

Figure PCD-64_27. Dataflow of personal data, in the GDPR sense

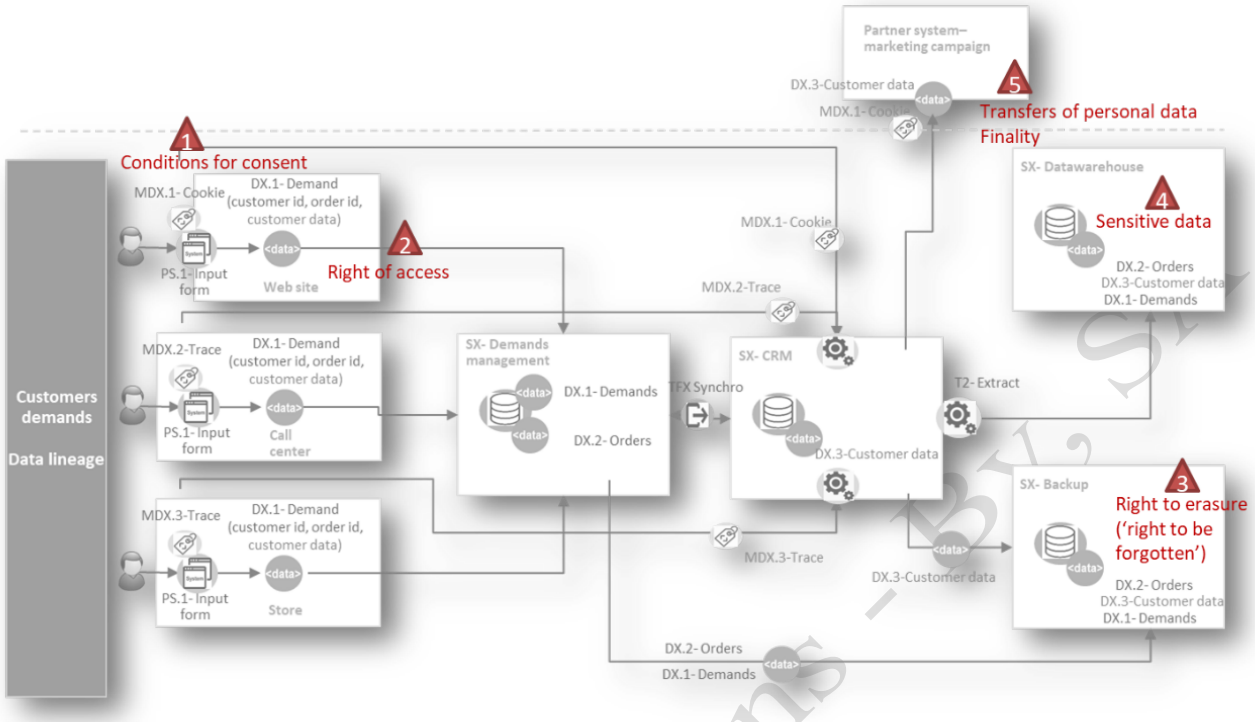


This example illustrates the movement of personal data from communication channels to back-office systems, also covering backup systems[17].

From this representation, points of vigilance vis-à-vis the GDPR are identified (see the following diagram) and should be the focus of a process.

[17] Falls within the "dataflow mapping" exercise

Figure PCD-64_28. Detection of points of vigilance on the path of sensitive data



5.4 Example of a product – case n°3: LIME Framework (Data Lineage in the Malicious Environment)

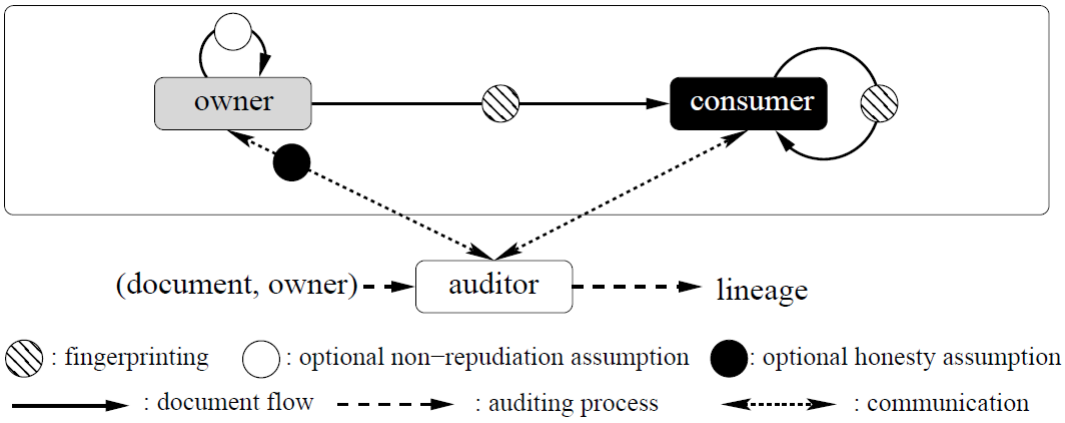
LIME is a data lineage framework for dataflow across entities that take two main roles:

- data provider and owner;
- data consumer.

The data lineage framework allows us to highlight the data security mechanisms (or lack of): identification, non-repudiation, preserving integrity, risks of data leakage (divulgation), tracing transfer flows (highlighting/trace of the provider and recipient in the data by the provider), post-leakage proof…

LIME highlights a third role of auditor, in partnership with the other two roles. The auditor is the guardian of the framework and beyond, of the expected trust.

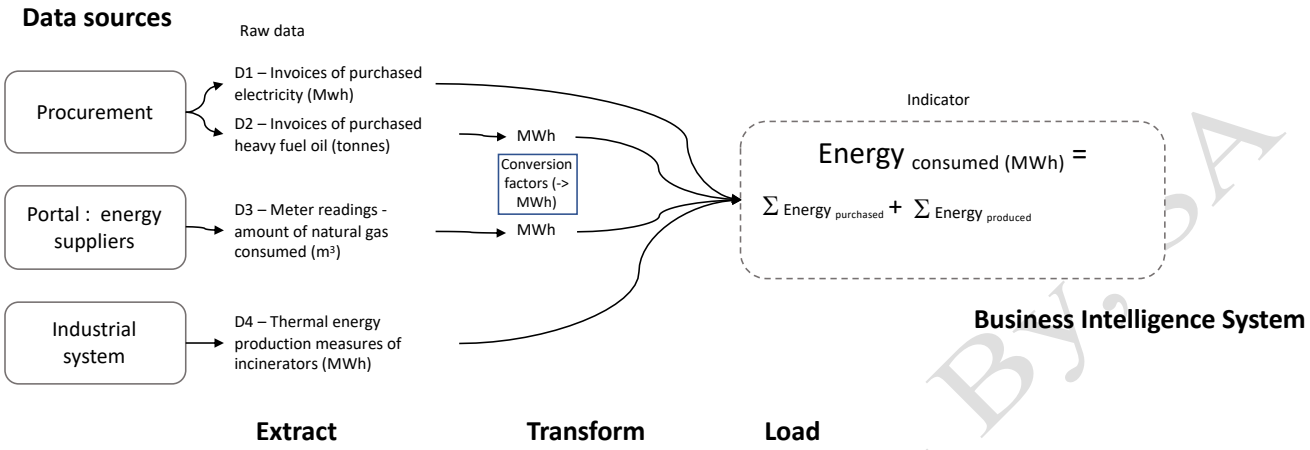
Figure PCD-64_29. Data Lineage in the Malicious Environment



Source: *LIME: Data Lineage in the Malicious Environment*, Michael Backes, Niklas Grimm, and Aniket Kate.

5.5 Example of a product – case n°4: Composition of an indicator

Figure PCD-64_30. Data lineage in a B.I. (Business Intelligence) environment – composition of an indicator

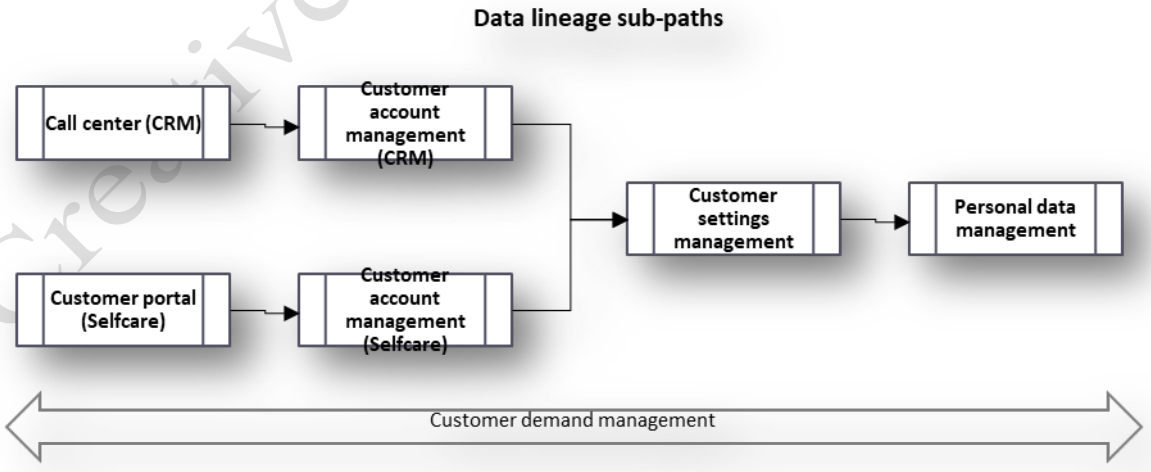


5.6 Breakdown of paths

Sometimes, the data lineage requires a considerable effort to highlight a multitude of stages whose overall view becomes difficult to see clearly. It is, then, good practice to rely on a sub-path approach (consistent set of stages around a sub-processing objective). Examples: the data lineage for data production of both the denominator and numerator of an indicator, the data lineage of the steps of data preparation before the data lineage of the calculation steps. The right data lineage division helps with its efficiency and its readability (overall macro view, reuse of pathway bricks, breakdown of the effort, etc.).

The figure below presents a macro vision of data lineage concerning the management of a customer's personal data. The breakdown in sub-paths makes the management, representation and reading of a data lineage easier.

Figure PCD-64_31. Breakdown of a path



This example, seemingly simple, is nevertheless representative. Several data acquisition channels carry different application components with their own interactions (represented by different sub-paths). And a same series of

processes is common to all channels. The exercise joins the I.S. architecture design approaches. In that capacity, I.S. architects can help by properly dividing a data lineage into sub-paths.

5.7 Consistency point

In a complex or highly distributed system, it is important to have consistency points.

Guaranteeing consistency points for data is one of the ways of dividing data lineage on more limited organizational territories (see the previous point about sub-paths).

A consistency point brings additional information. We consider that at one stage of the processing chain, the data is reliable and complies with what is expected of it (for example, with a level of quality). The data lineage analysis effort will then focus on the rest of the processing chain.

Keeping consistency points of other analyses can be a productivity factor, by avoiding the need to reanalyze one part of the chain.

5.8 Acceptance criteria

The result produced must answer the following quality requirements:

- readability of the representation (this is a key element, the data lineage produced will act as a common map between an actor facing an analysis, an issue to deal with);
- reliability of the representation (architects and application managers may need to validate the result produced);
- identification of areas of uncertainty;
- coverage of the data scope (for a dataset, ensure that we capture everything within scope: for example, make sure that a site or a type of data is not forgotten);
- keep elements of proof gathered during the work to reconstruct the data lineage (traces);
- updating the representation (the result produced was recently updated);
- responsibility to ensure the results are shared (there is an entity in charge of results and their publication in a repository that can be consulted with the expected level of reliability).

Ultimately, these criteria must allow us to have the right representation at our disposal (the right map by analogy), with the right information (the right components of the map).

The BCBS 239 regulation from the ECB, in the banking sector, illustrates these acceptance criteria.

It specifies that the data lineage must enable us to understand:

- how the expected data is given a value in the sense of being given a result;
- by what IT systems;
- if there is one or more data sources;
- what is the share between automated and manual parts…

For each database included in the data lineage, the regulation requires us to indicate:

1) the technical solution (DBMS); 2) the data taxonomy and its correspondence with different taxonomies. The ECB states that it also expects, should there be any modifications to the activities or paths identified by the data lineage exercise, that they could improve the efficiency of the processes and the reliability of the figures provided.

5.9 Managing and capitalizing on the results produced

Data lineage represents a continuous and sizeable effort. The results of this effort must be capitalized on.

As with any knowledge capital, the way in which it is formalized (as office-automation documents or in a dedicated software system – such as an enterprise repository) will have an influence on its management: maintenance, correction, sharing capacity.

This management is essential. The matter represented in a data lineage is very much a living matter.

Once the data lineage has been represented, there is a strong likelihood of finding ourselves, after a period of time, with representations that have not been updated, "ghosts" (whose existence we have almost forgotten), difficult to exploit without the presence of those who were able to produce the representation.

This is why the data lineage exercise is one of the important components of data governance.

The latter has to define the data lineage policy, as well as the necessary means for its management (organization, dedicated processes at a data governance level, tooling).

6. Tooling of the procedure

6.1 The collection of metadata

The tooling base is built on the means for collecting metadata.

Either the collection is ensured on an enterprise scale by mechanisms like the EDR (Enterprise Description Repository), with data lineage having only to exploit these mechanisms.

Or we have to proceed with an ad-hoc collecting approach that could even lead to retro-engineering parts of the system being studied. This approach for the collection of metadata goes through the investigating, gathering and describing of this metadata.

Explicitly defining the strategy for the collection of metadata is key to carrying out data lineage.

Either we consider that it is an isolated effort and the collection strategy will mobilize temporary means (task force, project resources, mobilization of a DQM (data quality management) team, data scientists in the data-preparation phase…).

Or we consider that it is a recurrent effort with stakes over time and the collection strategy has to set out to define:

- a perennial organization (dedicated activities and roles within a data management team),
- collection mechanisms planned for this task: automated gathering of metadata via connectors/probes on existing systems, via a "design by trace" planned from the outset of the design of the systems – in response to traceability requirements (like logs that we can position).

6.2 Software support solutions

The collection of execution traces (horizontal traceability, at the processing-chain level), knowledge about construction traces (vertical traceability) and the representations that we can make from them constitute a corpus rich in information.

There are several types of software solutions for managing and exploiting such a corpus:

- solutions like the enterprise repository which give a better overall vision of the different levels of representation (by aspects) of a data lineage and which can, in particular, manage the preserving of the construction traces between aspects (for example the link between choices at the pragmatic aspect level – processes and the choices at the logical aspect level then logistic one). The result of a data lineage by type completes the views/models referenced in the support tools of AE approaches (modeling processes, business objects, I.S. ...)[18];

[18] These solutions also integrate the possibility of representations using UML (the choice of a UML symbolism for a data lineage representation), allowing the interoperability and navigation between representation models (we can then easily

- data management/data governance solutions which, starting from data knowledge (business glossary, data dictionary, data catalog), extend to data lifecycles and related data lineages[19];
- specific and dedicated data lineage solutions which, beyond the dimension of data lineage representations, aim to collect the metadata required for these representations as automatically as possible [20];
- solutions like "ETL" or "data preparation tools", in the case of the application domain of business intelligence, Big Data/data lake systems[21];
- solutions built on office-automation means with all the related limits (integrity, usability, maintainability…)[22].

We recommend that you seek simplicity and the best integration possible of the tooling in the I.S. production chain, rather than adding dedicated solutions that risk communicating poorly with the rest.

Some tooling selection criteria:

- coverage of traceability aspects and axes;
- respect of representation standards (UML, BPMN);
- in contact with the I.S. to collect metadata, directly or not;
- passage to a data governance scale;
- readability of the results produced for collective analysis exercises (integration of collaborative functions).

6.3 Governance of data lineage production

The data lineage effort is always a considerable one and mobilizes important means (the accumulation of efforts, on an enterprise scale, linked to producing data lineages can be surprising).

The tooling must provide mechanisms for management, production and administration of data lineages.

Such an approach can be based on the mobilization of data designers within a centralized team. Setting up a team of data designers who circulate between data lineage activities and capitalize on them is a factor of productivity and efficiency (rather than having to reinvent everything every time data lineage is required).

Data designers are in charge of:

- carrying out data lineages;
- cross-reading data lineages in order to guarantee the right representation (respect of representation norms, homogeneity of representation) and the right readability;
- capitalizing on products through a data portfolio ("useful data") that were subject to data lineage and, on which the team of data designers can be questioned with the aim of reusing them (I have a problem with a particular data element, do you have the corresponding data lineage?);
- the relationship with the projects, project advice and follow-up on a data processing level.

This relationship enables us to initiate a virtuous circle with the data lineage representation exercise (project compliance with a data dictionary and this same reference dictionary for the representation of data lineage, information feedback on what has been implemented by the projects in terms of data processing and represented in data lineages, sharing lessons learned and expertise on certain data processing (for example, distinction between the default value and forced value, normalization rules for naming data…).

---

identify that a particular data element, the focus of the data lineage, belongs to a particular business object that is operated by one level of a particular process via particular services).

[19] These solutions often integrate a collaborative dimension (data lineage reconstruction exercise by "participative production").

[20] These solutions fall within the "metadata systems" family.

[21] These solutions have the means to represent the different processing stages that led to these systems being supplied (from source to integration in a warehouse, a Big Data platform).

[22] These solutions end, after a certain amount of time, in the risk of being confronted with a "ghost" data lineage repository (we know that it has been done, we do not know where the right version is anymore, nor if it is up-to-date; the people who managed it have left…).

7. Further reading

7.1 Correspondence with other reference frameworks

There are few practical frameworks that detail a data lineage approach (and none with an overall approach, in an enterprise architecture dimension).

The one that comes closest, being data management oriented – in the context of the DMBOK (Data Management Body Of Knowledge) initiative – Version 2[23] includes a chapter " – Chapter 12 Metadata management 12.4 Techniques 12.4.1 Data lineage and Impact Analysis".

It covers:

- Relationship between data lineage and dataflow (note: a priori both terms are interchangeable);
- Requirements related to the format and tooling of the documentation related to data lineage (note: a link to business processes and organization-role elements is made. We distinguish two high and detailed levels of representation);
- Relationship between the data lineage and data lifecycles (note: data not only has a lifecycle but also a data lineage (that is to say a path in which it moves from its point of origin to its point of use, sometimes called a "data chain");
- Concept of data lineage in different data management knowledge domains (note: one case outlined – to deal with data quality issues, with the accent on metadata, model requirements – architecture - data architecture, data modeling and design, data integration and interoperability, reference data, implementation such as DW and BI…).

The ECB in the context of BCBS 239 proposes guides (regulatory[24]) detailing its expectations in terms of data lineage reconstruction of the data that it seeks to control (Basel indicators).

The data governance solutions include data lineage modules. The software vendors of these solutions propose, for some production guides, data lineages in line with the functions of their solution.

7.2 Other similar approaches

There are some similar, even synonymous approaches.

a. Data supply chain

Context: data production is the vocation of the business process. Example: for a data lab, it might correspond to the data preparation phase before being exploited by data science algorithms. Another example is for a data marketplace that corresponds to managing the data lifecycle, from its identification to its "commercialization" on the marketplace. Data becomes a material, a finished product, distributed as in a traditional supply chain.

We have to manage this supply chain and that starts by describing it.

This description is equivalent to producing a data lineage. The difference is essentially on the fact that the stages in a supply chain are characteristic (the data supply chain is made up of three parts. First, on the supply side, the data is created, captured and collected. Then, the data is enriched, controlled and improved. Finally, on the demand side, the data is used, consumed and exploited).

b. Data biography

Certain actors speak about a data biography[25].

[23] Version 1 did not propose anything.

[24] https://www.bis.org/bcbs/index.htm

[25] https://idatassist.com/building-best-data-biography-asking/

The approach is complementary and is in line with the questions that we can ask ourselves in the context of an activity and data science objectives. It focuses on the production of datasets used by data scientists.

The approach distinguishes the upstream questions that join a data lineage type approach:

- How was the data collected? For what purpose?
- Has the collection process changed between two collection campaigns (look for collection variations)?
- What representativeness of the data collected, of the sample obtained?
- Where does the data come from, from an authority, the aggregation of different sources? Who funded the collection?
- How was the data cleaned?
- What was adopted for aberrant values (deleted, kept)?

We seek primarily to characterize the datasets obtained in terms of quality and possible bias (do not take the data used at face value).

And questions of intent: What purpose is followed: prove an analysis model, interpret behaviors to deduce a model from them?

The final objective is to keep the history (biography) of the datasets used by the data scientists (in other words the data lineage of the production of the datasets).

More widely, this approach is familiar to statisticians who associate their publication to the statistical methodology used (following the example of INSEE: https://www.insee.fr/fr/information/2838097).

It also joins the efforts to formalize the data preparation steps for data science work and how they find their way into solutions like the "data preparation tool", in which we can find the capacity to manipulate and represent forms of data lineage of the datasets used.

In a more anecdotal way, we can find approaches like "data genealogy" (What is the mother data for certain data? Which data has been generated from which data?).

7.3 Openings

Traceability is at the heart of this procedure. It should be an initial requirement for building systems.

The content of this procedure can, thus, be seen as an initial guide towards:

- the redesign of a processing chain: take advantage of the data lineage exercise to come back to an Enterprise Architecture approach, working back towards upstream aspects (logical, pragmatic and semantic) that the method proposes by exploiting the logistic level, to optimize the processing chain both locally and from beginning to end;
- the design of new processing chains in a "trace by design" logic, where the traceability requirements are in integral part of the general requirements and where explicit means of managing traces are implemented.

There are some cases for a system construction that call upon a "Metadata System Centric" logic – where the heart of the system is built on a metadata repository. Following in the footsteps of organizations whose business is data – data marketplace / data broker, statistical institutes, and who build their I.S. from a vision of the metadata (data being the product that they sell). The vision of data lineages is then an automatic one, at the heart of the business model (that also goes with the data supply chain idea seen previously). This allows us to manage approaches like "automatic change data management", that is to say the capacity to detect construction changes – changes in models, even processing changes on data.

With the idea of rethinking information systems in a data-centric logic, the obligation to think "trace by design" is essential.

7.4 Practical bibliography

Wikipedia: https://en.wikipedia.org/wiki/Data_lineage

DAMA – Data Management Association: Section on *data lineage* DMBOK V2 https://dama.org/content/body-knowledge

Mike2 - Method for an Integrated Knowledge Environment, open source methodology for Enterprise Information Management: http://mike2.openmethodology.org/wiki/Data_Lineage

Data Flow Diagrams (DFD): The Object Primer: Agile Model-Driven Development With UML 2.0 - Scott W. Ambler - 2004

Table of illustrations

Analytical contents