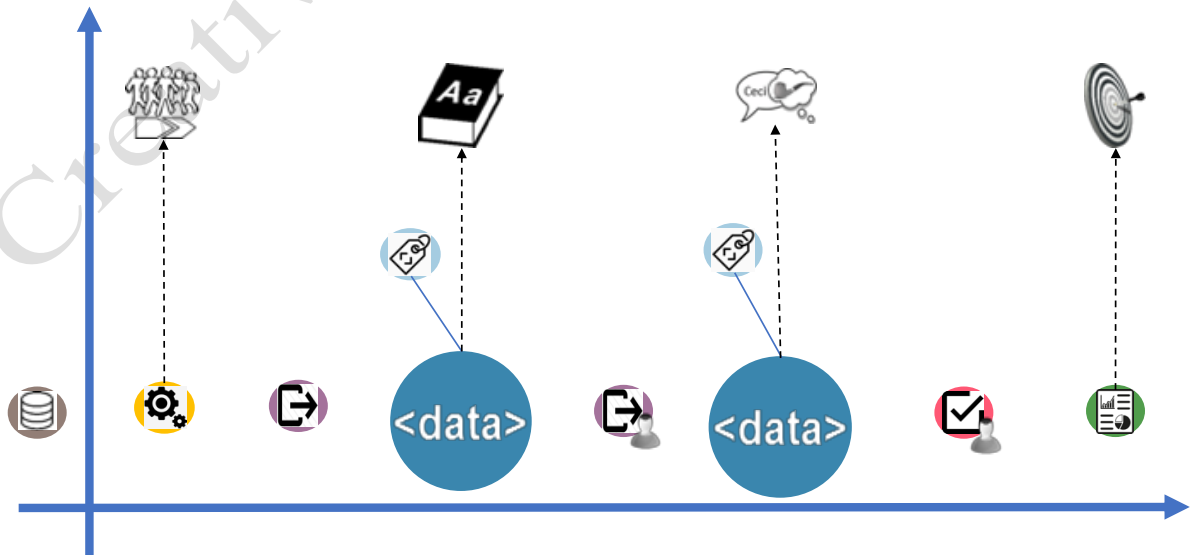
Assurer la traçabilité des données

Sujet	**Procédés de l'aspect logistique**
Objet du procédé	Répondre à des exigences d'analyse et d'optimisation en produisant le « lineage » de données
Mots clefs	*Data lineage, modélisation, exécution, données, Praxeme, méthode, procédé*

Référence	**PxPCD-64**	*État*	Validé
Version	1.1.0	*Date*	30 avril 2019
Auteurs, contributeurs	Contribution du cabinet CONIX		
Relecteurs	Dominique VAUQUIER		

Sommaire



Rappels méthodologiques

Dans le contexte de la méthode Praxeme, un *procédé* est « une façon de faire, un mode opératoire pour exécuter une tâche »[1]. Il s'agit donc d'une prescription à un niveau individuel, par opposition au *processus* qui est une réponse méthodologique au niveau collectif.

Les fiches de procédés ne font pas référence à d'éventuels processus dans lesquels ces procédés pourraient intervenir, ceci afin de faciliter leur réemploi dans plusieurs contextes.

Pour suivre l'actualité de la méthode publique

- Mailing list
- Groupe LinkedIn
- Twitter
- le wiki

Pour participer aux travaux du Praxeme Institute

- Adhésion au *Praxeme Institute*

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

Protection du document

L'initiative pour une méthode publique repose sur le bénévolat et la mutualisation des investissements entre ses contributeurs. Elle vise à élaborer et à diffuser une méthode ouverte et libre de droits. Sa dynamique n'est possible que si cet esprit est maintenu à travers les utilisations des documents qu'elle met à la disposition du public. C'est pourquoi les documents sont protégés par une licence « *creative commons* »[2] qui autorise l'usage et la réutilisation de tout ou partie d'un document du fonds Praxeme, sous seule condition que l'origine en soit citée. Les éventuels documents dérivés, qui reprennent du contenu de Praxeme, doivent s'appliquer à eux-mêmes les mêmes conditions, faire référence à la « *creative commons* » et porter les symboles idoines :

Actualisation de ce document

Pour obtenir la dernière version de ce document, se rendre sur le site du *Praxeme Institute*, à la page du catalogue : http://www.praxeme.org/telechargements/catalogue/.

L'historique du document

Indice	Date	Rédacteur	Contenu
0.0.0	26/11/2018	DVAU	Création de la fiche de procédé
0.0.x			Échanges JB / DVAU
1.0.0	26/02/2019	JB, DVAU	Première publication
1.1.0	27/04/2019	J. TOWARD	Relecture lors de la traduction et complément de JB (§ 5.4)
1.1.0	30/04/2019		Version actuelle du document

[1] Cf. rubrique Thesaurus sur le site du *Praxeme Institute* : http://wiki.praxeme.org/index.php?n=Thesaurus.Procedure.

[2] Voir la philosophie et le détail des licences sur : http://creativecommons.org/.

Remarque liminaire

Ce procédé s'insère dans un ensemble plus vaste : la méthode publique Praxeme, dont l'objet est la transformation maîtrisée des entreprises et des systèmes complexes. D'un côté, le procédé bénéficie de son articulation avec d'autres procédés, couvrant un large spectre d'intervention sur tous les aspects de l'entreprise. D'un autre côté, chaque fiche de procédé est conçue pour une utilisation autonome, au prix de quelques rappels dans les premières sections.

Le terme « aspect » renvoie à une notion centrale de la méthode. Cette notion est liée au cadre de représentation, la Topologie du Système Entreprise, formant le socle de la méthode et ordonnant une approche interdisciplinaire de l'entreprise[3].

1. Contexte d'application du procédé

1.1 Objet

Ce procédé « Assurer la traçabilité des données » constitue un guide pratique pour établir un *data lineage*, c'est-à-dire retracer le cheminement d'une donnée.

Le résultat produit (le *data lineage*) est un moyen qui sera ensuite utilisé pour répondre à des exigences d'analyse, d'optimisation, de réglementation mettant en jeu la production de données via la qualification de traces (voir §1.2 situations d'usage).

Pour qualifier un *data lineage*, on distingue des traces de niveau exécution (telle donnée a été saisie à telle date par telle personne) et des traces de niveau modélisation (telle nature de donnée a fait l'objet de tel traitement ou transformation dans le cadre de tel processus qui met en jeu tel objet métier).

La méthode Praxeme offre un cadre rigoureux qui permet de formaliser un *data lineage* (le qualifier et le décrire) et de l'inscrire dans une vision Entreprise plus large.

Ce procédé est un guide pratique qui bénéficie du cadre Praxeme pour décrire les opérations nécessaires permettant de recueillir et de représenter un *data lineage*. Cela passe par le recueil de données sur les données (ou traces sur les données), qu'on appellera par la suite méta-données (exemples : la nature d'un traitement sur une donnée, le nom de la source d'une donnée sont des données sur cette donnée). On cherchera à restituer ces données sur les données, dans la meilleure représentation possible, de façon à exploiter efficacement le résultat produit. Le cadre Praxeme permet de poser, de façon structurée, toutes les questions nécessaires au travail de collecte et de restitution des méta-données. On verra, par la suite, comment ces méta-données ou traces s'organisent naturellement dans le cadre Praxeme.

L'effort pour établir un *data lineage* est coûteux et fastidieux. Ce procédé a pour vocation d'améliorer l'efficacité, la portée et la productivité de cet exercice. Il fixe un cadre de représentation d'un *data lineage* ainsi que le mode opératoire associé permettant d'industrialiser la démarche.

Préoccupation centrale : Réaliser un data lineage *– documenter le cheminement d'une donnée – répond aux besoins de traçabilité sur une donnée ou un jeu de données.*

Par nature les données circulent *dans* et *entre* les systèmes d'information. Lorsqu'elles circulent, elles subissent des changements de valeur, des transformations numériques, des changements d'état et des événements. Recueillir les informations sur ces changements et événements permet de disposer d'une traçabilité des données de bout en bout, de l'acquisition à l'utilisation.

[3] Voir, en guise d'introduction, le Guide général, référence PxMDS-01.

Maîtriser cette traçabilité devient de plus en plus important, du fait de la valeur prise par certaines données, à l'exemple de données réglementaires ou encore de données liées à la connaissance des clients. Les données deviennent, à la fois, des produits et des actifs de l'entreprise. Assurer leur traçabilité revient à maîtriser :

- la façon selon laquelle elles sont produites : la chaîne de valeur des données (la *supply chain* de la donnée, associée à des traces de production) ;
- la valorisation des données en tant qu'actifs : la caractérisation de la donnée comme actif, en fonction de son origine et de sa qualité ;
- les risques associés : non conformité à une politique de la donnée, risque d'intégrité sur une donnée à une étape du cycle de traitement.

La production d'un *data lineage* doit permettre de répondre à ces enjeux.

L'effort que représente la reconstitution d'un *data lineage*, est souvent nécessaire pour pallier le déficit de construction d'un S.I. (défaut de traçabilité au moment de la construction).

Cet effort est également nécessaire du fait de la complexité croissante des chaînes de traitement (« empilage » historique de briques issues de systèmes différents – fusions/rapprochements de S.I., ouverture des S.I., nouvelles couches technologiques, absence d'urbanisation…).

1.2 Situations d'usage

Dans l'absolu, un *data lineage* répond au besoin qu'a l'organisation de garantir le bien-fondé des données manipulées et nécessaires à la conduite de son activité. Il a pour vocation de justifier l'exactitude des données : un *data lineage* contribue à la confiance dans les données.

Un *data lineage* et les métadonnées qui lui sont associées constituent un moyen pour répondre à différentes problématiques où les données sont un enjeu (stratégique, tactique, de performance).

Exemple de problématiques :

- Réglementation et pilotage (*Business Intelligence*) : apporter la preuve (trace) de la bonne constitution d'un indicateur, d'une donnée clé (par exemple, un chiffre d'affaires), dans le respect d'une réglementation ou pour satisfaire des besoins de pilotage (exemples : BCBS RDARR, BCBS 239 qui impose aux banques de prouver l'origine et le mode de construction d'indicateurs liés à la réglementation bâloise. Ou encore dans un projet BI s'assurer de la bonne fiabilité d'un tableau de bord).
- Analyse d'écarts constatés sur les valeurs d'une même variable (mesure, agrégat, indicateur), fournies par différentes sources.
- Analyse d'impact : lorsqu'une évolution a lieu sur une chaîne de traitement, quels impacts sur la production de données, de tableaux de bord… ?
- Compréhension de l'origine et de la nature des données dans les exercices de *data science* (charges de préparation des données).
- *Data Quality Management* (DQM)[4]: comment améliorer la qualité des données en posant des actions d'amélioration tout au long d'une chaîne de traitement (à la capture des données, par des contrôles au moment de l'intégration des données, etc.).
- Optimisation, simplification de la production de données, d'indicateurs : au fil du temps les circuits de production des données se complexifient. Leur représentation par un *data lineage* permet d'avoir une vision d'ensemble et de les optimiser (exemple : rationalisation des règles de gestion, optimisation d'achat de données externes).
- Contrôle du respect de règles de sécurité et d'accès aux données tout au long du circuit (répondre à une politique de diffusion, de protection des données).

[4] Gestion de la qualité des données en Français

Parmi ces usages, certains relèvent d'une approche *top-down* : on cherche à vérifier que le parcours d'une donnée respecte des règles et des exigences qualité imposées, par exemple, par une réglementation (du type RGPD ou BCBS 239 pour les banques).

D'autres usages relèvent d'une approche *bottom-up* : à partir du parcours d'une donnée, on cherche les dysfonctionnements, les optimisations possibles, par exemple dans le cadre d'un diagnostic qualité.

Plusieurs facteurs conditionnent la démarche de réalisation d'un *data lineage* :

- la complexité de l'organisation (nationale, internationale, multinationale, …) dans laquelle le *data lineage* s'inscrit,
- la multiplicité des sources sur un ensemble de données volumineux,
- la multiplicité des modes d'acquisition et de stockage (au sein d'ERP, dans le *cloud*, dans des plates-formes big data…),
- la globalisation des parcours d'acquisition (situation d'omnicanal),
- la multiplicité et la redondance de l'information,
- l'ouverture et les échanges avec des parties prenantes, partenaires, clients…

D'une façon générale, un *data lineage* apporte une vision globale sur le traitement d'une donnée, manipulée par de multiples briques issues de systèmes et environnements différents. Cette vision globale est le premier apport d'un *data lineage* avant toute résolution de problématique. Elle permet le dialogue entre les acteurs, à partir d'une représentation commune.

1.3 Cheminement nominal et écarts

L'exercice de *data lineage* vise à représenter la situation nominale, c'est-à-dire le cheminement normal, prévu à la conception et quand la chaîne de traitement s'exécute normalement.

Cette représentation permet d'analyser les écarts, lors de situations exceptionnelles. Par exemple, pendant la période d'été, le transfert manuel d'un flux de données s'est retrouvé dégradé.

La représentation d'un *data lineage* (situation nominale) sert de fond de carte pour mettre en évidence, par deltas, les situations exceptionnelles ou problématiques (voir le chapitre 4 mode opératoire pour une illustration de l'idée de positionnement d'un calque d'analyse des problématiques sur le fond de carte).

1.4 Positionnement dans la méthode

a. Place dans le cadre de référence

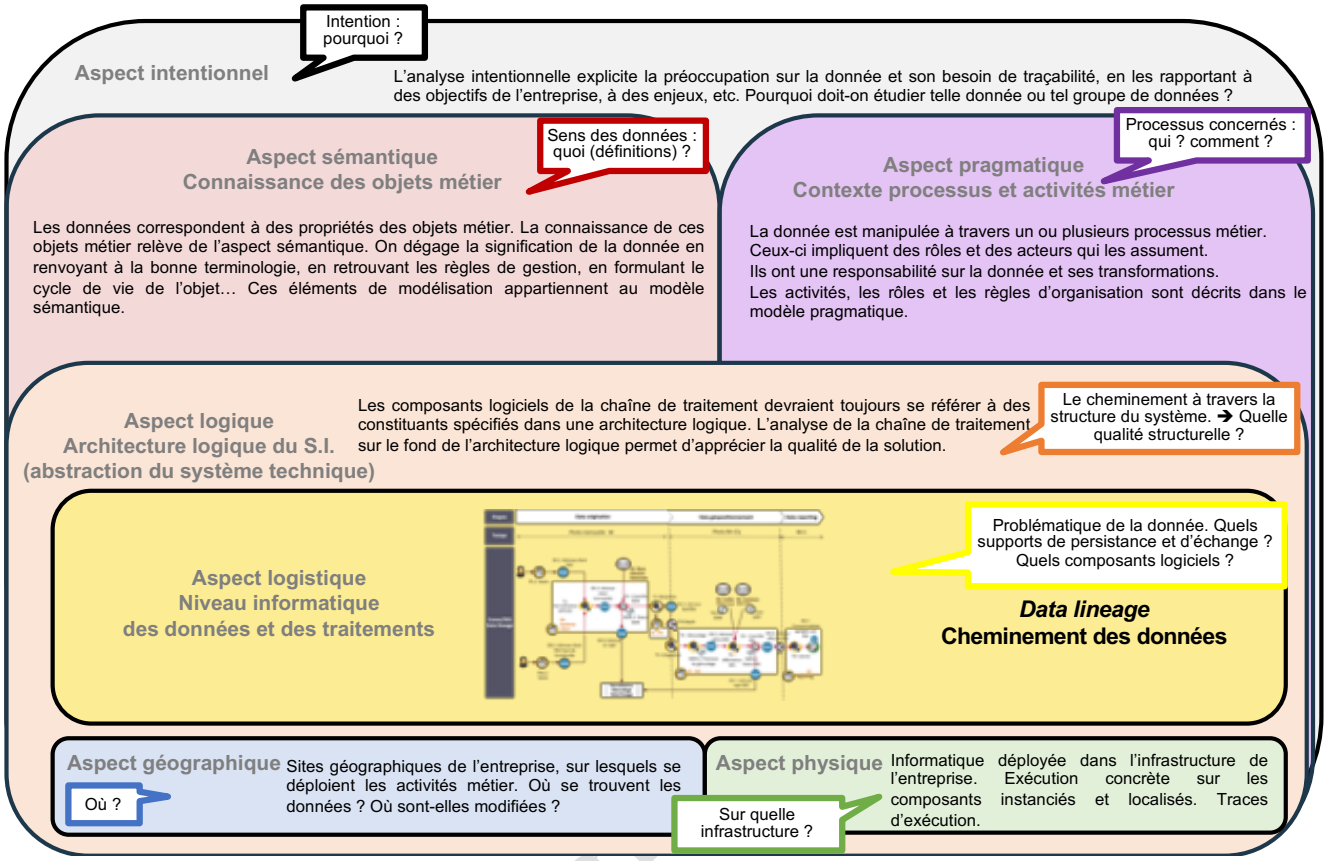
Établir la filiation ou le cheminement d'une donnée est un acte qui s'inscrit dans l'aspect logistique, tel que défini dans la Topologie du Système Entreprise. Cet aspect couvre les systèmes techniques au service des activités de l'entreprise, tout particulièrement les solutions informatiques. La demande qui déclenche l'exercice de ce procédé révèle toujours une préoccupation quant au système informatique.

En conséquence, l'essentiel du mode opératoire proposé ici porte sur le logiciel et le modèle logistique. Cependant, le procédé s'enrichit d'actions complémentaires qui permettront de contextualiser les données étudiées et de mieux aborder leurs usages. En cela, le procédé trouve place dans une véritable démarche d'architecture de l'entreprise, et oblige à fréquenter les autres aspects du système entreprise. La motivation de cet élargissement réside dans le souci de tirer le maximum de retombées de l'effort consenti. Répondre à la demande de traçabilité suppose un investissement non négligeable, dans l'état habituel d'un système mal documenté[5]. Nous recommandons d'utiliser cet effort pour jeter les bases d'une bonne documentation et alimenter le référentiel de description de l'entreprise. Cet effort supplémentaire entraînera un surcoût minime, tandis qu'il préviendra, à l'avenir, de nouvelles dépenses.

La figure suivante résume les contributions à la documentation de la traçabilité, que l'on peut tirer des différents aspects du système entreprise. Elles font l'objet des actions présentées dans le mode opératoire (chapitre 4 de cette fiche).

[5] Voir les indications de charges dans la section 3.

Figure PCD-64_1. Contextualisation d'un cheminement de données, par rapport aux aspects du Système Entreprise



b. Relations avec d'autres procédés

Ce procédé « Assurer la traçabilité des données » peut bénéficier des livrables produits par les procédés suivants :

1. **procédés terminologiques**, principalement « Définir un terme » (réf. PxPCD-14a) et « Élaborer un thesaurus » (PxPCD-14f), puisqu'une des premières choses à faire est de vérifier la bonne compréhension des données attendues en résultat ;
2. **procédés de modélisation sémantique** (principalement PxPCD-22a), pour donner une expression formelle aux notions impliquées dans la demande de traçabilité ;
3. **procédés de modélisation pragmatique**, quand le procédé en arrive à associer les données aux activités qui les manipulent (procédé PxPCD-32, s'intéressant à la perception « métier », en amont des solutions informatiques) ;
4. **procédés de l'aspect logique**, incluant l'architecture logique et l'urbanisation des SI, dont les produits peuvent guider l'intervenant pour retrouver les composants en jeu ;
5. **modélisation géographique** de l'entreprise, la distribution spatiale des activités et la définition des sites pouvant avoir des incidences sur la circulation des données ;
6. **modélisation de l'aspect logistique**, produisant le catalogue des composants logiciels (si ce catalogue existe et s'il respecte les règles de l'art, le travail d'enquête s'en trouvera grandement facilité) ;
7. **modélisation physique**, dont les diagrammes de déploiement éclaireront les problématiques liées à la duplication des sources de données et des traitements (domaine de la « production »).

Dans l'idéal, ces procédés ont été appliqués tout au long de la construction du système, ce qui permet, alors, de bénéficier des descriptions nécessaires pour assurer la traçabilité. En pratique, le plus souvent, l'intervenant devra lui-même produire une partie de ces informations.

c. Posture

Praxeme distingue les deux postures d'analyse et de conception, qui s'appliquent à tous les aspects de l'entreprise[6].

Établir la traçabilité des données relève de la posture d'analyse, puisqu'il s'agit de décrire – *a posteriori* – l'état d'un système existant.

La plupart du temps, l'application du procédé répond à la nécessité de se mettre en conformité avec une réglementation, ou au besoin de traiter une problématique qualité ou encore traiter un impact non prévu. On peut dire qu'il s'agit d'un geste défensif, réclamant une analyse pour répondre au plus vite. L'action documente l'existant, sans rien ajouter. C'est un travail d'enquête.

Cependant, au-delà de cette posture minimale et indispensable, le travail peut prolonger l'analyse et ses constats, en proposant des pistes d'amélioration. Celles-ci pourront, plus tard, déboucher sur un effort de conception (à l'exemple de la recherche d'optimisations dans une chaîne de traitement).

Ainsi, bien que ce procédé ressortisse à l'analyse, une partie de ses résultats peut aiguillonner une réflexion de conception.

2. Terminologie employée

2.1 Notions générales

La terminologie associée à la méthodologie est rassemblée sur le wiki de Praxeme[7].

Le tableau ci-dessous donne les définitions des principales notions utilisées dans cette fiche, et les commente dans le cadre de ce procédé.

Figure PCD-64_2. Notions générales et leur définition

Notion	Définition	Commentaire
Aspect	« Portion de la réalité, isolée pour en faciliter l'étude, en respectant sa logique interne »	L'approche multi-aspect enrichit les retombées de ce procédé[8].
Aspect logistique	« Aspect d'un système composé de ses moyens logistiques »	Ici, il ne s'agit que des moyens informatiques.
Référentiel	« Ensemble d'éléments partagés par une communauté »	
Référentiel de description de l'entreprise	« Référentiel qui contient tous les éléments accumulés au fil des travaux pour décrire le Système Entreprise »	Voir le chapitre 6, sur l'outillage. Dispositif central qui permet de travailler plus efficacement.
Objet métier	« Objet concret ou abstrait, essentiel à la mission du Système Entreprise »	Les objets métier sont les termes dans lesquels s'exprime la connaissance fondamentale du métier.
Processus	« Ensemble ordonnancé d'activités »	
Procédé	« Manière prescrite pour faire quelque chose »	
Donnée	Représentation informatique d'une information	
Méta-donnée	Donnée portant sur une donnée	Nous détaillons cette notion par la suite.

[6] Voir le Livre blanc, réf. « SLB-02 » et ibid.

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

[8] Pour la définition de tous les aspects, voir le thesaurus sur le wiki. Pour leur justification et l'explication du cadre de représentation, voir le guide méthodologique PxPRD-01.

2.2 Données et Méta-données

Dans la suite, nous parlerons de « donnée utile » pour désigner la donnée qui fait l'objet de la reconstitution pour répondre à la demande de traçabilité.

Le terme « méta-donnée » revêt deux significations, également impliquées dans le procédé de lignage :

1. D'une part, le radical « méta » évoque la représentation obtenue par un effort d'abstraction. Il s'agit alors du modèle de la donnée, non plus la valeur concrète, mais la variable positionnée au sein d'un modèle, accompagnée de sa description sous la forme de type, de libellé, de commentaires, de règles. Par exemple, la donnée s'exprime comme valeur d'un attribut dans une classe, celle-ci représentant une table dans une base de données, ou bien un concept au niveau sémantique.
2. D'autre part, on appelle « méta-données » les données qui accompagnent une donnée utile et qui sont, le plus souvent, liées à un usage ou à une exécution. Une autre expression pour les désigner est celle de « données d'enveloppe ». L'image est explicite : la donnée utile est la lettre, glissée dans l'enveloppe, le contenu utile ; les méta-données sont les informations inscrites sur l'enveloppe : destinataire, date, émetteur, exigences sur l'acheminement (tarif, recommandé, acquittement), etc.

Ces deux notions de méta-données sont également nécessaires dans la maîtrise des données d'un système. Nous les retrouverons dans la suite du procédé. Le tableau suivant les illustre.

Figure PCD-64_3. Exemples de méta-données et de questions relatives aux données

Exemple	Méta-données d'exécution	Méta-données de représentation
Statut de l'information (validée, confirmée, douteuse…) **Qualité de géo codage** **Date de la mise à jour** **Niveau de connaissance** **Campagne de capture**	La valeur change, potentiellement, pour chaque donnée.	Des variables (attributs) s'ajoutent à la donnée utile, et l'accompagnent. Le modèle doit donc s'enrichir. Une caractéristique des propriétés de ce type est qu'elles peuvent se définir à un niveau générique, par exemple sur la racine de tous les concepts métier (aspect sémantique) ou sur la racine des structures d'échange (aspect logique).
Source de l'information. **D'où vient la donnée ?** **Où est-elle utilisée ?**	Pour désigner une source particulière : individu qui a saisi la donnée ; organisation auprès de laquelle la donnée a été achetée… Ou un usage particulier.	La réponse peut aussi être produite au niveau du modèle : dans le cas où la donnée provient toujours de la même source (même composant logiciel, même fournisseur…).
Qui a créé cette donnée ? Qui l'utilise ? À qui appartient-elle ? Qui en assure le traitement et la maintenance ?	Réponse au niveau individuel (individu ou entité), potentiellement différente pour deux données d'un même type (par exemple : le commercial A a enregistré le client X).	Réponse au niveau du modèle (c'est toujours un commercial qui enregistre les données client ; « commercial » est un rôle qui apparaît dans le modèle pragmatique).
Quelle en est la définition métier ? **Quelles sont les règles métier ?** **Quel est son degré de sécurité ?** **Contraintes réglementaires…**		Réponses à travers la modélisation (les questions ne portent pas sur une donnée particulière, mais sur toutes les données d'un même type, d'une même signification). Aspect sémantique.
Où est stockée la donnée ? Quelles en sont les dénominations standard au sein des bases de données ?		Modèle de l'aspect logique, d'abord (modèle logique de données) et de l'aspect physique : le même schéma peut être instancié plusieurs fois dans l'architecture physique.

Exemple	Méta-données d'exécution	Méta-données de représentation
Pourquoi stocke-t-on cette donnée ? Quel est son usage et sa finalité ? Quel est le levier métier pour l'utiliser ?	Exceptionnellement, les réponses peuvent être individualisées (taux particulier pour une entité, pour des raisons spéciales).	Renvoi à des éléments d'intention (aspect intentionnel).
Quand cette donnée a-t-elle été créée, actualisée ? Quand doit-elle être effacée ?	Réponse possible au niveau de la donnée (par exemple, en lien avec des préférences du client).	Réponse possible au niveau de l'ensemble des données d'un même type (par exemple, renvoi à des règles d'archivage).
Comment cette donnée est-elle formatée ? Dans combien de base de données ou sources est-elle présente ?	Pour des acquisitions ponctuelles (le format peut varier).	Modélisation logistique (formats et supports) ; modélisation physique (duplication et phénomènes dynamiques résultant de la redondance physique).

2.3 Champ lexical de la traçabilité

Dans l'usage courant comme dans les contextes industriels, la traçabilité se définit comme :

> *Capacité à reconstituer une chaîne de détermination*[9].

La traçabilité contribue à établir la confiance en un système : processus de fabrication ou de livraison, organisation, système technique…

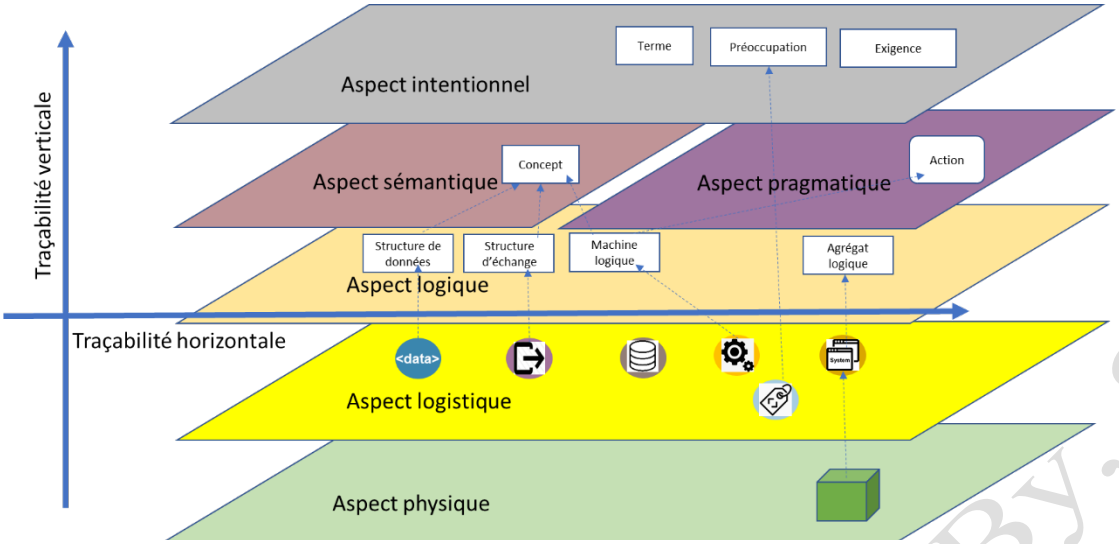
En ce qui concerne la traçabilité des données, nous devons distinguer deux types de traçabilité :

1. La *traçabilité horizontale*, qui se déploie sur le plan de l'exécution : elle s'établit sous la forme de chaînes de production, menant d'une ou plusieurs sources à un résultat.
2. La *traçabilité verticale*, développée sur le plan de la construction : elle s'inscrit dans les modèles, en reliant un élément d'un aspect aval à un élément d'un aspect amont.

Le schéma ci-dessous illustre la traçabilité verticale avec une sélection de catégories de représentation. Dans ce sens, une chaîne de traçabilité relie des éléments appartenant à des aspects différents. Ce schéma illustre la traçabilité horizontale uniquement à travers l'aspect logistique, mais on la trouve aussi dans d'autres aspects, par exemple sous la forme d'un processus dans l'aspect pragmatique.

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

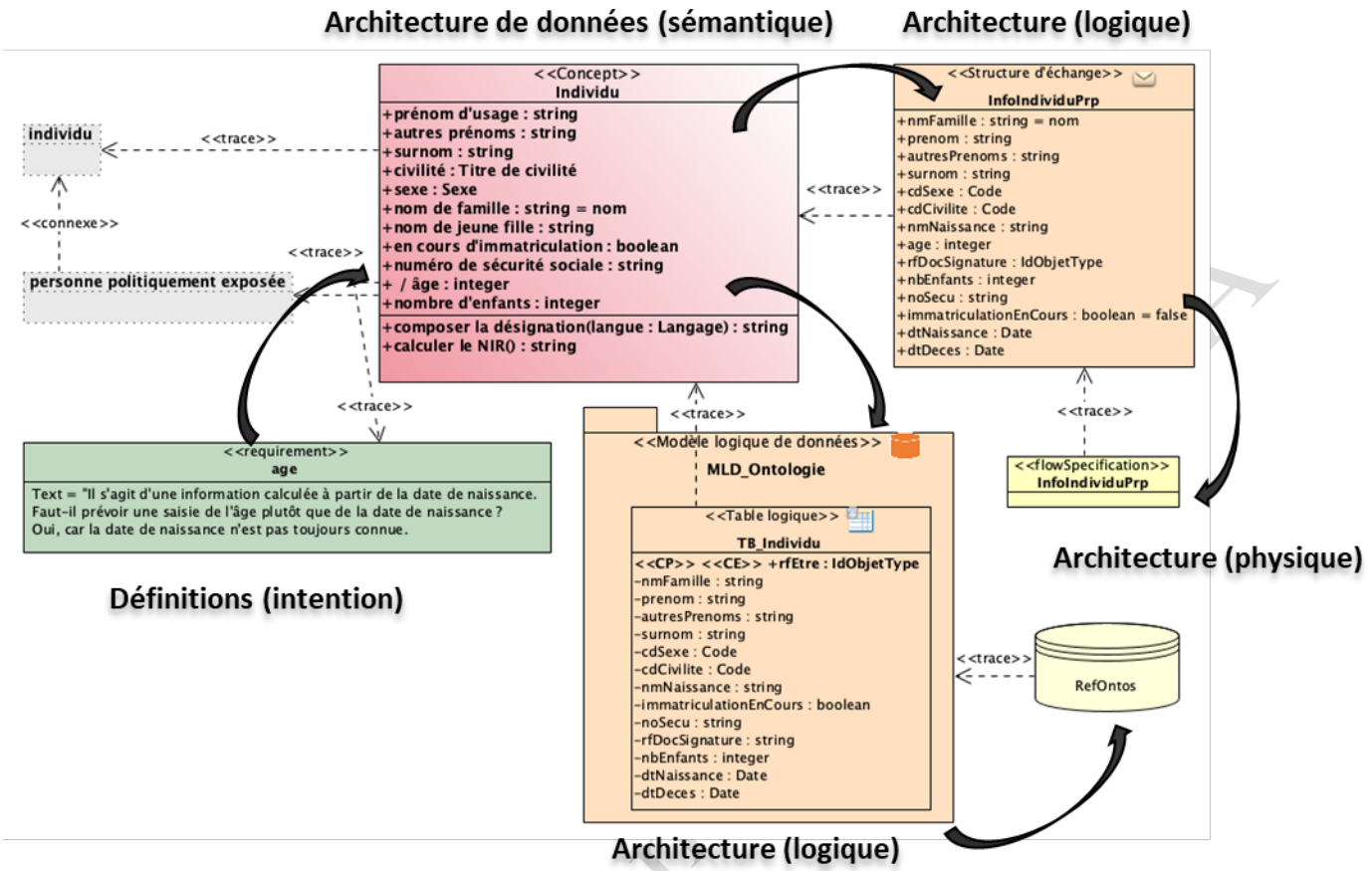
Figure PCD-64_4. Les deux dimensions de la traçabilité



La traçabilité horizontale d'une donnée s'exprime, d'abord, comme la succession des étapes nécessaires à sa production. Cette description est de l'ordre du modèle de l'aspect logistique, et répond pour un type de résultat – une propriété, une variable, une classe d'objets. Si la traçabilité demandée porte non pas sur un type de résultat mais sur un résultat particulier – une donnée, une valeur, une instance –, alors il convient d'ajouter, à la chaîne de production, des méta-données de type enveloppe, que nous appellerons des « traces d'exécution ».

La traçabilité verticale se compose de « traces de construction », c'est-à-dire de liens entre des éléments de modélisation d'aspects différents. Une chaîne de traçabilité complète permet de rattacher la propriété étudiée, de nature logistique, à sa spécification logique, et de remonter de cette dernière à un élément d'un aspect métier, et, de là, à un terme, une exigence ou un objectif dans l'aspect intentionnel. Ces traces de construction correspondent aux méta-données au sens de la modélisation. Elles se manifestent par des dépendances stéréotypées « *trace* », selon le mot réservé en notation UML. Praxeme impose de poser ces traces en respectant les dépendances entre les aspects de la Topologie du Système Entreprise. Cette règle réduit le couplage au sein du référentiel de description, et simplifie son exploitation.

Figure PCD-64_5. Illustration de la traçabilité verticale : la notion d'individu



Commentaire du diagramme

Les éléments intentionnels représentés sur cette figure sont de deux natures : des termes extraits du dictionnaire ; une exigence portant sur la propriété « âge ». En rouge, la classe sémantique « Individu » formalise le concept métier. Elle fournit le point de départ pour plusieurs projections dans l'aspect logique, dont deux sont montrées : la structure d'échange ; la structure de données (en orange, couleur de l'aspect logique). Ces deux éléments logiques se traduisent, au niveau logiciel (couleur jaune), en un schéma XSD, par exemple, et une table d'une base de données.

Quand ces chaînes de traçabilité sont parfaitement documentées dans le référentiel de description de l'entreprise, l'analyse d'impact, en cas de changement ou d'audit, peut être traitée en quelques minutes, au lieu de plusieurs jours.

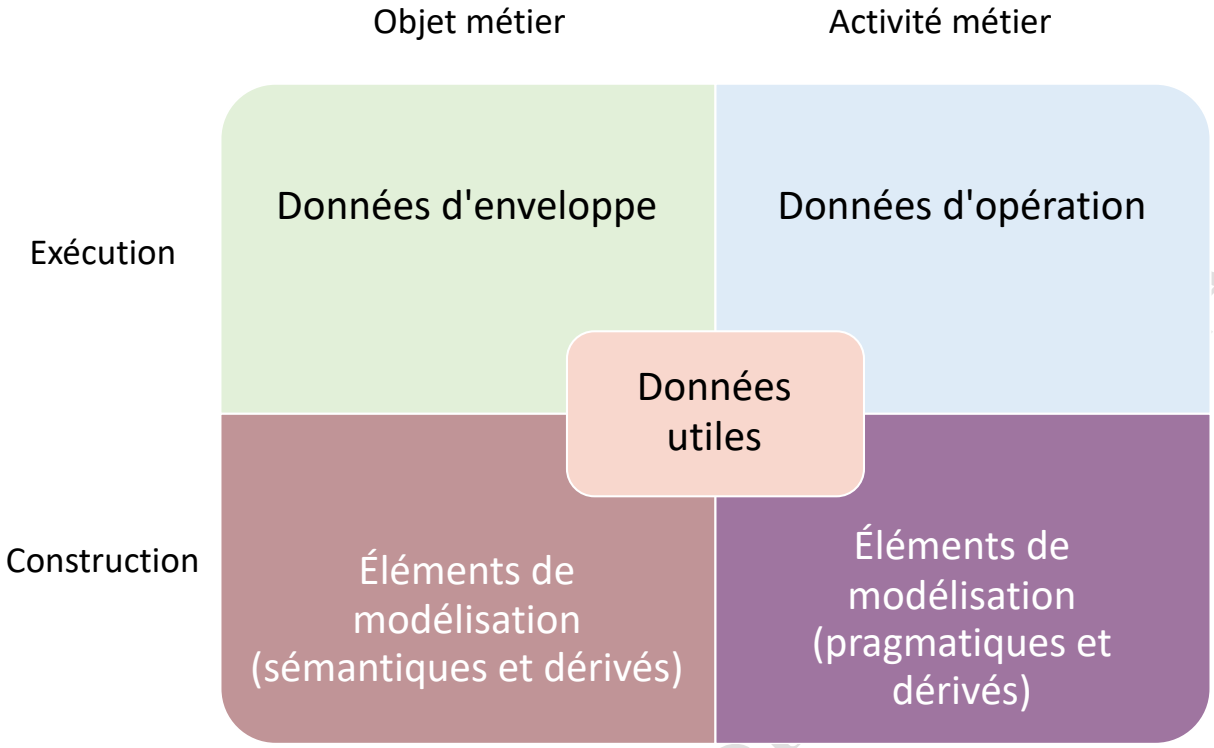
2.4 Catégories de méta-données

La donnée utile – celle qui fait l'objet de la demande de traçabilité – est, presque toujours, une information essentielle pour le métier : elle porte sur un « objet métier », c'est-à-dire une notion fondamentale. Retracer son histoire oblige à dégager les activités qui la manipulent. Apparaissent alors les « activités métier », inscrites dans les processus de l'entreprise. Ici aussi, nous trouvons les deux interprétations du terme « méta-donnée » :

1. L'activité se représente à travers un modèle (modèle de processus, cas d'utilisation…), et la chaîne de production de la donnée devra faire référence à l'activité et au processus dans laquelle elle se trouve plongée (exemple un processus comptable).
2. Lors de l'exécution, l'activité s'instancie, et il pourra s'avérer nécessaire de conserver des informations contextuelles, telles que : la date de l'action, l'identifiant de l'acteur, le contexte d'exécution de l'action, c'est-à-dire des informations « d'enveloppe ».

Ainsi, en combinant le couple objet-activité et la distinction enveloppe-modèle, on obtient quatre catégories de méta-données accompagnant la donnée étudiée, comme le montre la figure suivante.

Figure PCD-64_6. Les quatre catégories de méta-données accompagnant les données utiles



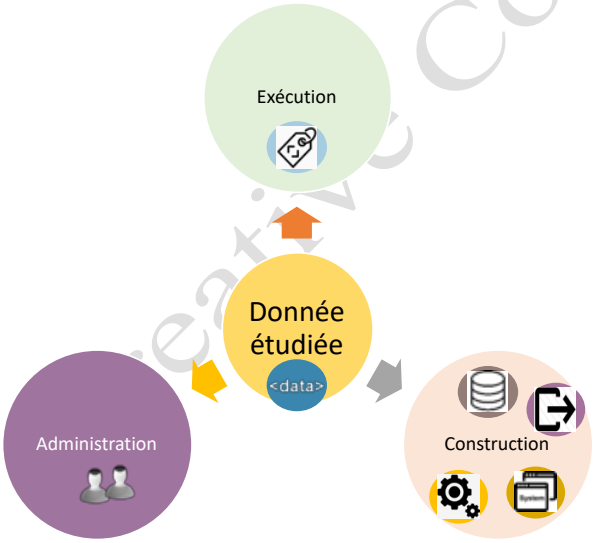
Commentaire du schéma

1. Par analogie avec un courrier, dont le contenu serait la donnée étudiée, les données d'enveloppe précisent les conditions d'exécution qui ont produit ou modifié cette donnée. Ces traces d'exécution permettent une analyse fine du destin de la donnée.
2. Ce destin mobilise des acteurs, dont il peut être nécessaire de mémoriser les actions. On obtient, alors, des méta-données opérationnelles qui nous parlent des manipulations de la donnée et de son enveloppe. Ce sont aussi des traces d'exécution ou de production. Le terme « production » ne se limite pas au sens qu'il a pris en informatique ; il désigne aussi les interventions humaines, outillées ou manuelles, sur la donnée.
3. La donnée trouve sa définition et sa signification dans un modèle. Au plus haut niveau, assurant la compréhension en termes métier, il s'agit du modèle sémantique. L'élément de modélisation qui définit la donnée sémantiquement est, le plus souvent, un attribut d'une classe. La classe représente elle-même le concept métier (ou « objet métier »). À partir de cet élément de modélisation, s'accrochent plusieurs filières de dérivation qui passent par l'aspect logique, et aboutissent à l'aspect physique. Ainsi, les éléments de modélisation auxquels se rattache la donnée étudiée, comprennent : des structures d'échange et de persistance (modèles logiques), leurs traductions en termes logiciels (aspect logistique), puis leurs instances dans l'architecture physique. Les traces de construction qui relient ces éléments sont précieuses pour la maîtrise du système.
4. De même, les activités font l'objet d'une modélisation. Elles sont décrites, en termes métier et organisationnels, dans le modèle pragmatique. Les éléments de modélisation sont les modèles de processus ou d'activités, les rôles et règles d'organisation, ainsi que les contextes d'exécution. Ces éléments se dérivent également, et ils s'imposent aux traitements portant sur les données.
5. Cette analyse nous permet de recenser les catégories d'information qu'il faudra collecter pour répondre à une demande de traçabilité sur une donnée. En élargissant le champ des préoccupations, nous découvrirons que d'autres natures de méta-données peuvent entrer en jeu, distribuées sur les aspects du système entreprise (voir l'action « Mettre la donnée en perspective »).

Les méta-données sont d'autant plus importantes que la donnée peut être produite de plusieurs façons et suivre plusieurs chemins. La traçabilité consiste alors à identifier le chemin pris par une donnée particulière.

Figure PCD-64_7. Notions liées à la traçabilité

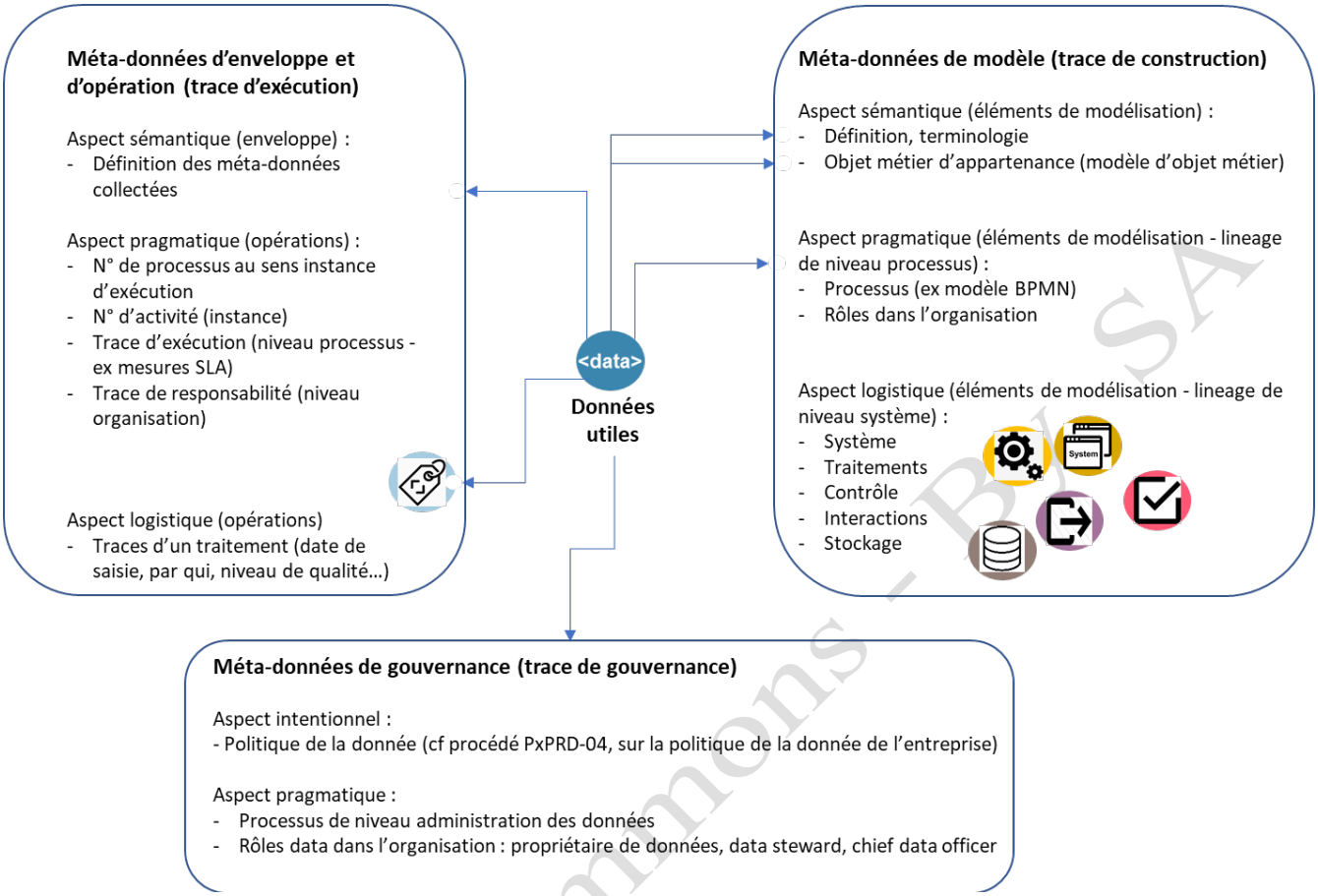
Notion	Définition	Commentaire
Traçabilité	**Capacité à reconstituer une chaîne de détermination**	Le sujet central de ce procédé.
Trace	**Marque laissée par un événement**	La trace est un signe ; elle porte de l'information.
Trace d'exécution	**Trace que quelque chose s'est produit, à un moment donné, dans un traitement**	L'événement que la trace révèle est lié à une exécution (quelque chose s'est produit dans le système, en production).
Trace de construction	**Trace reliant le choix de construction, dans un certain aspect, au choix de construction d'un aspect adjacent** Exemple : dérivation du modèle sémantique en un modèle logique de données, puis en un modèle physique.	Le choix de construction s'exprime, formellement, à travers un élément de modélisation. L'événement est, ici, un acte de construction du système. La trace relie deux éléments de modélisation, chacun pris dans un aspect différent.
Méta-donnée	**Donnée portant sur une donnée** Sens 1 : élément de modèle définissant une donnée ou une variable. Sens 2 : information sur la vie d'une donnée.	Le sens 1 renvoie au modèle. Le terme « variable » est plus approprié. La donnée est une valeur prise par une variable. Le sens 2, à l'enveloppe.
Enveloppe	**Ensemble des données accompagnant une donnée utile**	Équivalent de « trace d'exécution ».



Pour compléter notre typologie des informations à collecter, nous devons prendre en compte l'organisation et la définition des responsabilités en matière de données, ce qu'il est convenu d'appeler la gouvernance (couvrant majoritairement l'administration des données). À la construction et à l'exécution, nous ajoutons donc la dimension d'administration (figure ci-contre). Dans des cas extrêmes, la documentation de la donnée et de sa traçabilité devra comporter, en effet, la désignation des responsables et des procédures, associés à la donnée.

Figure PCD-64_8. Les trois dimensions de méta-données

Figure PCD-64_9. Lien entre données et méta-données (traces) – segmentation des méta-données selon les aspects



2.5 Chemin, cheminement de la donnée, chaîne de production de la donnée

Par « *data lineage* » (terme retenu dans le cadre général des processus de *data management*), on entend la reconstitution du chemin au sein d'un système d'information que va parcourir une donnée ou un ensemble de données, des sources d'acquisition jusqu'à une restitution pour un usage donné.

« Cheminement de la donnée » ou « chaîne de production de la donnée » sont des expressions équivalentes :

- « Cheminement » évoque plutôt la circulation d'une donnée, à peu près inchangée, entre un point initial, la source, et un point final, la fourniture.
- « Chaîne de production » évoque plutôt le calcul d'un résultat à partir de manipulations de données, par agrégation, consolidation, transformation, etc.

Dans tous les cas, il s'agit de montrer les étapes que traversent une donnée ou un ensemble de données, pour arriver à un résultat attendu. Ces étapes se précisent en termes de supports de stockage, de flots et de composants logiciels.

Établir un data lineage *est, d'abord, un travail d'enquête et de reconstitution de chaînes de traitement mettant en jeu les données. Il s'agit de cartographier les chaînes de traitement, couvrant tous les éléments de gestion ou d'exploitation successifs de la donnée. Dans cet exercice, l'intervenant détaille les règles de transformation de la donnée (codification, notation, normalisation, agrégation, etc. …), de la source de la donnée (sélection, filtrage) jusqu'à son usage.*

3. Compétences requises

Ce procédé spécialisé s'insère dans un corpus plus vaste, traitant de toutes les dimensions de l'entreprise. Il présente aussi l'intérêt de mettre en relation différentes compétences et de coordonner plusieurs perspectives sur l'entreprise (également reflet des aspects de la méthode évoqué précédemment). Outre la perception purement technique, nécessaire pour répondre à la préoccupation de traçabilité, le procédé convoque des disciplines s'intéressant aux processus et aux connaissances du métier.

Selon le profil de l'intervenant qui applique le procédé, deux cas de figure se présentent :

- soit il possède ces différentes compétences ou cette sensibilité, dans ce cas (rôle d'architecte d'entreprise), il pourra dérouler complètement le mode opératoire ;
- soit il complète son action en interagissant avec d'autres personnes : architecte métier, analyste métier, propriétaire de processus, responsable applicatif, urbaniste et architecte S.I., *data steward*, « consommateurs » des données – *data scientist*, statisticien…

Majoritairement l'exercice de *data lineage* est un exercice collectif et collaboratif, à la fois dans son élaboration et dans utilisation.

Praxeme coordonne ces disciplines en les assignant aux aspects fixés par le cadre de représentation. En cela, la méthode et ce procédé constituent un outil de dialogue.

Compétences impliquées :

- juridique (pour faire le point sur les exigences réglementaires),
- modélisation : de la sémantique au physique (objets métier, modélisation logique des données, modélisation physique),
- architecture : du métier à l'infrastructure (spécialisation au minimum en trois ensembles : architecture métier, architecture logique, architecture informatique),
- gouvernance : maîtrise de l'organisation, des rôles et des règles de gouvernance des données (politique des données[10], administration des données).

4. Mode opératoire

4.1 Analyser le besoin de traçabilité

Un *data lineage* s'effectue rarement dans l'absolu ; il répond toujours à une intention précise. La problématique à traiter détermine le travail d'enquête et de collecte des données. Pour une problématique de qualité de données, les défauts de qualité identifiés orientent la recherche des traitements, actions dans une chaîne de traitement pouvant être à l'origine des causes de non qualité. Pour une problématique de preuves, on recherchera plus spécifiquement, les contrôles, les risques de rupture, d'écart entre deux points (écart de valorisation – tel montant en entrée pour un autre montant en sortie, écart de couverture – une part du périmètre attendu n'est pas couvert – i.e. traces d'exécution). Il importe donc de clarifier l'intention qui préside à l'effort. C'est pourquoi le procédé commence par l'analyse de la demande de traçabilité.

Il s'agit de répondre aux questions :

> *Quelle intention réelle révèle la demande de traçabilité ? À quoi servira le résultat produit – le* data lineage *? Quelles sont les exigences de traçabilité ? Y-a-t-il une dimension réglementaire qui impose de fournir des preuves ou des traces ?*

Cette première action consiste en une analyse intentionnelle, au sens de l'aspect intentionnel de l'entreprise. Élucider l'intention conduit à s'intéresser aux points suivants :

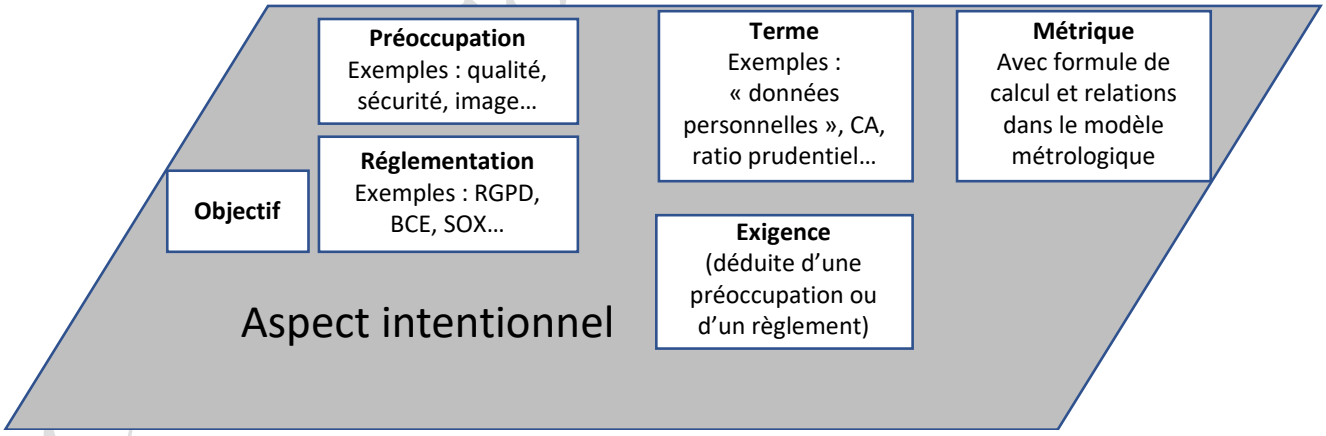
[10] Voir le formulaire et son mode d'emploi PxPRD-04, sur la politique de la donnée de l'entreprise (contribution du cabinet CONIX).

- **La nature de la problématique** et les enjeux pour l'entreprise : s'agit-il de répondre à une demande réglementaire de prouver le bon mode de production de données, ou de se conformer à une réglementation – comme le RGDP[11] – ou d'analyser une chaîne de traitement de données personnelles – dans la logique d'une EIVP (Étude d'Impact sur la Vie Privée), ou de traiter des dysfonctionnements qui se reflètent par une mauvaise qualité des données avec les risques inhérents, ou encore d'optimiser une chaîne de traitement pour être plus performant ?
- **Les priorités à traiter**, en lien à un programme de transformation ou à une politique informatique : des travaux prévus fournissent-ils l'occasion d'une simplification (par exemple, la réalisation d'un référentiel des personnes, couvrant des données sensibles) ? L'entreprise porte-t-elle une attention particulière à certaines données critiques ? La transformation numérique impose-t-elle de mieux maîtriser le patrimoine informationnel ? Etc.
- **Les acteurs et les responsabilités** : de quels acteurs émanent la demande de traçabilité ? Quels autres acteurs implique-t-elle ? Quels systèmes ou sous-systèmes concerne-t-elle ? Cette analyse délimite le périmètre de travail et le contenu des livrables (des systèmes peuvent en être exclus).
- **Les critères** permettant de vérifier que l'on a bien répondu à l'intention (voir chapitre « Critères d'acceptation »).

Ces attentes entrent dans le référentiel de description de l'entreprise, comme des éléments d'intention dont il faudra montrer la satisfaction. La première action consiste à les enregistrer et à les analyser. Les actions suivantes du mode opératoire satisferont aux critères dérivés de ces éléments (vérification de la satisfaction des intentions).

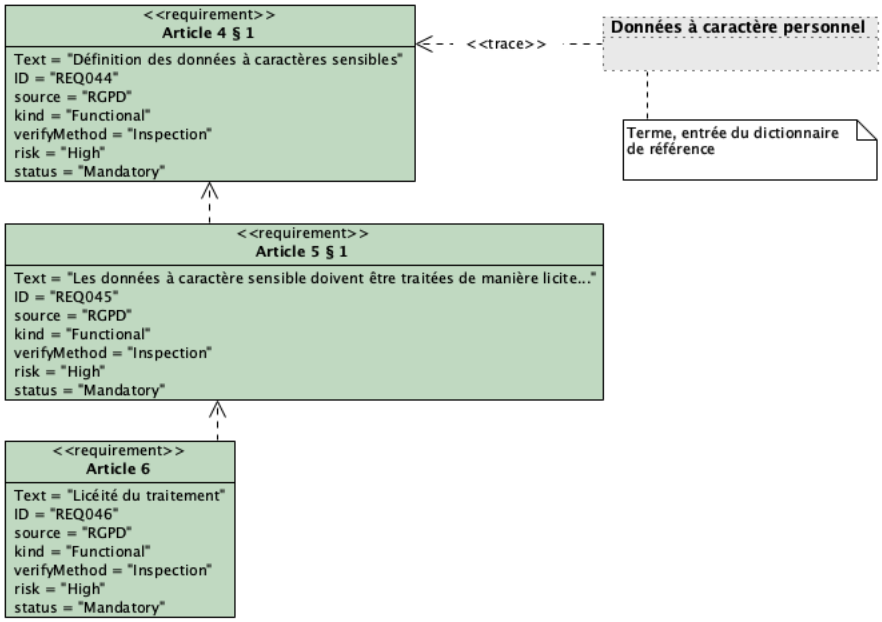
L'analyse intentionnelle se mène dans les termes propres à l'aspect intentionnel. Les éléments d'intention sont les formulations que l'on conserve dans le référentiel de description de l'entreprise (RDE). Par exemple, la réglementation est décomposée en formulations élémentaires (paragraphes, articles…), comme on le ferait pour des exigences. Au besoin, on attache, à ces formulations, des commentaires qui sont les résultats de l'analyse. Dans ce travail, il peut être nécessaire de renvoyer vers les entrées du dictionnaire de référence, lui aussi stocké dans le RDE. L'intervenant peut ajouter des diagrammes qui associent les formulations aux termes (voir figure PCD-64_11). Si l'analyse conduit à ajouter de nouveaux de termes, alors on applique les procédés terminologiques.

Figure PCD-64_10. Les types d'éléments manipulés dans l'action « Analyser le besoin de traçabilité »



[11] Règlement général pour la protection des données, issu de la Commission européenne.

Figure PCD-64_11. Illustration : résultat de l'action « Analyser le besoin de traçabilité »



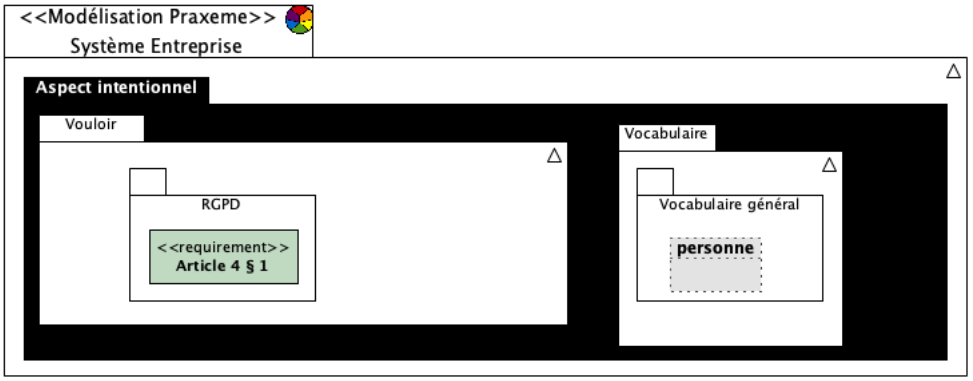
Commentaire du diagramme

Le texte réglementaire est décortiqué en formulations élémentaires, représentées ici sous la forme d'exigences.

Le diagramme montre également un terme, entrée du dictionnaire de référence, qui fournit la définition en référence au règlement.

La figure suivante montre comment se distribuent ces éléments dans la structure du référentiel de description de l'entreprise.

Figure PCD-64_12. La répartition des éléments d'intention dans le RDE (illustration)



4.2 Repérer les données

a. Identifier les données en partant de l'intention

L'action précédente a permis de clarifier la demande et d'en délimiter le périmètre. Il nous faut, ensuite, retrouver les données, sous la forme qu'elles prennent dans le système informatique.

La demande de traçabilité peut préciser la donnée visée (par exemple, le chiffre d'affaires) ou un type de données (les données personnelles) ; elle peut aussi exprimer une préoccupation plus générale (la sécurité des données, la pertinence des coordonnées…). À partir de ces expressions, nous devons repérer, dans le système, les données correspondantes.

À quel endroit, dans le système, sous quelle désignation et dans quelle forme les données étudiées existent-elles ? À quel contenu sémantique, ces données se réfèrent-elles ? Quels sont les objets métier concernés ? Quelle la définition métier des données analysées ? Quelles sont les règles métier qui les contraignent ? Quelles en sont les dénominations au sein des bases de données, des moyens de stockage et d'échange ?

b. Reconstituer la chaîne de construction liée aux choix de définition et de modélisation de données

Repérer les données intervient particulièrement sur l'aspect sémantique, puisque c'est là que doit se trouver la formalisation de l'information et sa meilleure expression, du point de vue de la connaissance du métier. En amont, la terminologie peut être un bon point de départ. Comme illustré précédemment, la demande de traçabilité absorbée dans l'aspect intentionnel s'analyse en termes et expressions, lesquels constituent des entrées du dictionnaire de référence. Chaque terme est « projeté » dans l'aspect qui lui convient, c'est-à-dire relié à un élément de modélisation appartenant à un autre des aspects. Le plus souvent, l'exigence de traçabilité portant sur une information « cœur de métier », le terme est repris par un élément sémantique. À titre d'exemple, citons : le chiffre d'affaires (attribut calculé, de portée ensembliste), les données personnelles (propriétés du concept d'individu), consommation d'énergie d'une zone géographique… Toutefois, il peut arriver que l'objet de la demande renvoie à un élément d'un autre aspect. Un exemple est la performance d'un processus (aspect pragmatique) ou d'une unité organisationnelle (aspect géographique).

Figure PCD-64_13. Les types d'éléments manipulés dans l'action « Repérer les données »

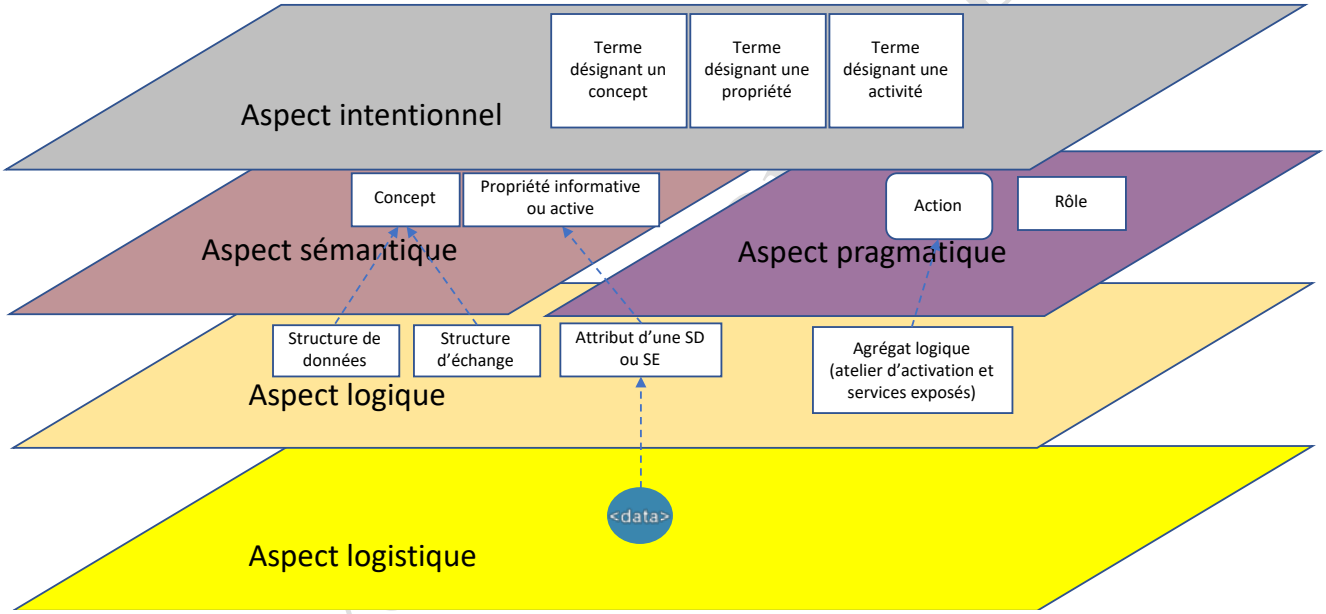
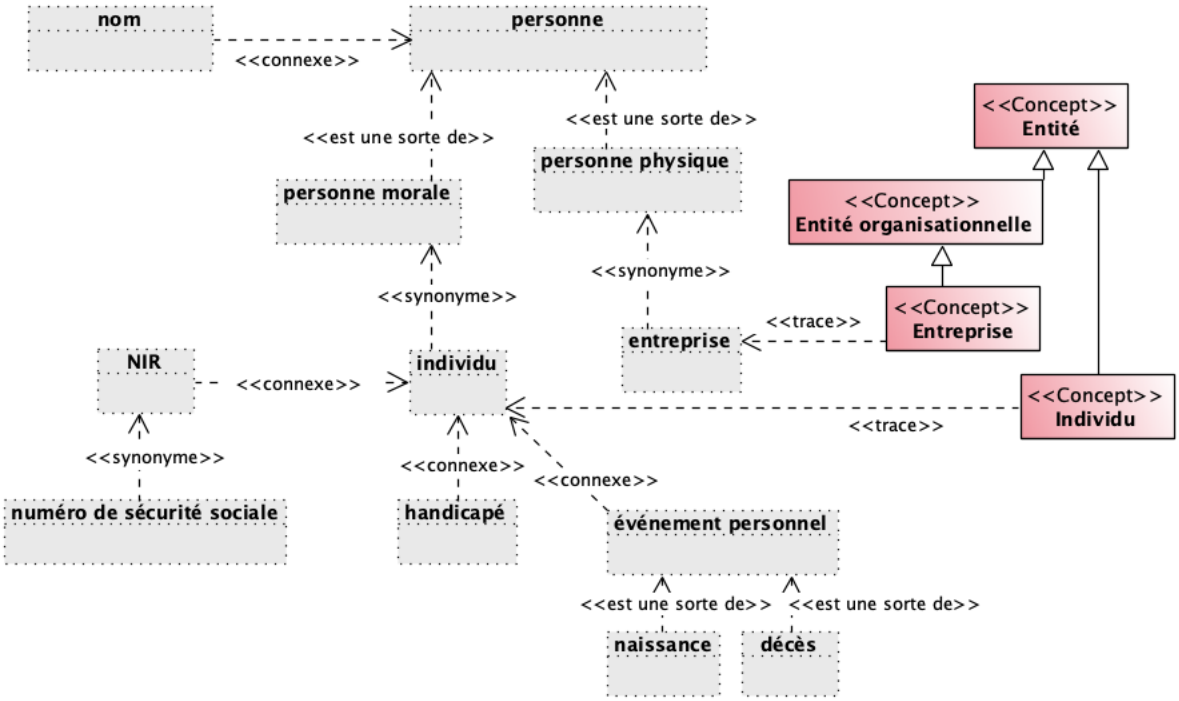


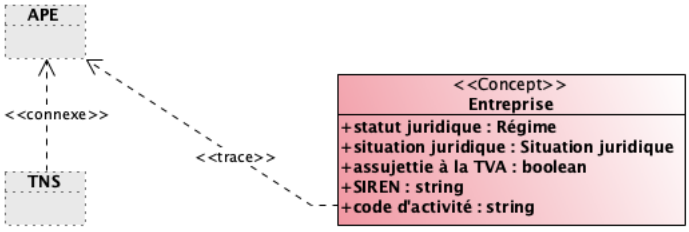
Figure PCD-64_14. Illustration de la projection de termes vers des éléments de l'aspect sémantique



Commentaire du diagramme

Ce diagramme illustre les résultats qui peuvent être produits par cette action (repérer les données). À partir du vocabulaire trouvé dans la demande de traçabilité (analyse intentionnelle menée lors de la première action), il s'agit de formaliser les notions, ici en renvoyant vers les concepts de l'aspect sémantique. Cette illustration reste au niveau des classes. Assez souvent, il sera nécessaire de pointer aux niveaux des propriétés des classes, comme montré dans la figure suivante.

Figure PCD-64_15. Illustration d'une projection vers un attribut d'une classe sémantique



Les traces de construction mises au jour par cette action de repérage ne s'arrêtent pas à l'aspect sémantique. Elles suivent les filières de dérivation pour mener aux structures de persistance et d'échange, comme le montre la figure ci-dessous.

Figure PCD-64_16. Les chaînes de traçabilité de construction (termes ➔ modèle logique ➔ implémentation)

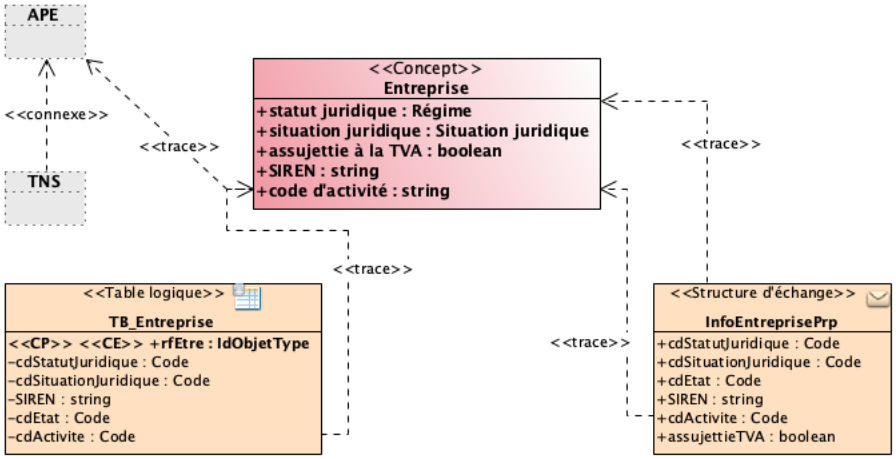


Illustration de défaut d'alignement sémantique : La qualité des données n'est plus vue comme une problématique de stock* mais de communication (changement de perspective). Toute acquisition de données est vue comme un message à destination d'autres acteurs (humain ou système). La qualité des données est alors vue comme un élément de la qualité de la communication. Et cette qualité de la communication dépend de la qualité sémantique des messages – signification des messages (lisibilité – encodage limité de façon à être lu sans décodage, compréhensible et formel – aligné sur les termes standard métier et les modèles métier, juste et complet – exempt de défaut – respect du sens et des règles métier au moment de l'alimentation…).

Cette qualité joue un rôle clé dans : les normes d'échange, les modèles de stockage et d'interprétation des données (à l'exemple des couches sémantiques des outils de *data visualization*) et par extension sur le cheminement de bout en bout des données.

Lorsqu'on qualifie une donnée objet du *data lineage* (tel indicateur), on est en bout de chaîne. Mais c'est la définition sémantique de cette donnée qui va fixer le besoin en alignement sémantique de toute la chaîne. Par exemple, dans le cadre de maintenance de canalisations ou d'un opérateur de transport, si on cherche à effectuer le *data lineage* d'un indicateur qui veut rendre la mesure de kms de réseau surveillés sur une période donnée, le respect de la définition tout ou long de la chaîne de traitement qui aboutira à la valeur mesurée est clé. Dans ce cas, soit on cherche à mesurer le nombre de kms en absolu surveillés, soit on cherche à mesurer le nombre de kms cumulés surveillés (ce qui n'est pas la même chose : dans le 1er cas, on va mesurer la part du réseau qui a fait l'objet d'une ou plusieurs surveillance – même si une partie de réseau a fait l'objet de plusieurs surveillances, on ne compte que les kms réseaux concernés ; dans le 2ème cas, on va mesurer le nombre de kms effectués en surveillance du réseau – si une partie du réseau a fait l'objet de plusieurs surveillance, on cumule autant de fois qu'il y a eu de surveillance les kms réseaux concernés. En fonction de cette différence de définition, on imagine l'impact d'une rupture de définition dans le *data lineage* de cette mesure).

* La qualité du stock ne peut pas être meilleure que la qualité issue des « communications ».

c. Dérouler le repérage des données au niveau de l'aspect logistique

En partant de la qualification de l'intention, on s'attache donc à qualifier les données qui feront l'objet du *data lineage* : un indicateur calculé, un tableau de bord, un jeu de données, une donnée élémentaire (comme une adresse mail).

Cette qualification couvre :

- **l'identification des données objet du *data lineage***. Par identification on entend une qualification permettant de reconnaître formellement la donnée (un identifiant, un titre de tableau de bord, un numéro de label d'un indicateur, un code de référence à l'exemple de la codification BCE de données dans le cadre de BCBS 239…) ;

- **la définition la plus formelle possible des données**. Nota : cet exercice de définition formelle est parfois la 1ère expression d'un dysfonctionnement. L'écart d'alignement sémantique (voir encart page précédente) est une cause fréquente de défaut que l'on identifie par un *data lineage*. Cet exercice de définition est à aligner avec l'aspect sémantique de la méthode, en particulier en faisant le lien avec les objets métier et leur cycle de vie (comme présenté dans les paragraphes précédents de cette étape) ;
- dans le cas de données calculées ou composées, **la formule de calcul ou de composition** avec les constituants élémentaires qui devront être tracés (objets du *data lineage*) ;
- **les règles de gouvernance** qui ont pu être associées à la donnée objet du *data lineage* (responsable de la donnée, politique à appliquer sur cette donnée – exemple réglementaire, confidentialité, sensibilité métier…). Cela constitue la récolte des premières méta-données de gouvernance.

Granularité des données objet du *data lineage* (« données utiles » dont on recherche les traces)

Le *data lineage* peut s'effectuer soit :

- au niveau de la donnée élémentaire ou un ensemble cohérent de données (une adresse, l'identité d'une personne, une transaction, un attribut-propriété d'un objet métier),
- au niveau d'un produit de pilotage : indicateurs, tableau de bord, mesures. In fine, l'exercice se déclinera au niveau des données élémentaires constituantes du produit de pilotage (les « données utiles » sont à la fois le produit de pilotage et ses composants).
- au niveau d'un ensemble de données (*datasets*), à l'exemple de publications de jeux de données pour des partenaires, pour un data lab, en open data,
- et parfois à un niveau flux de données, lorsqu'on cherche à retracer le parcours d'un flux de son émission à sa réception pour traitement, tout au long d'un processus métier (à l'image des CRE et CRI[12] en comptabilité, des flux de données client dans le processus d'entrée en relation et de KYC[13]).

4.3 Reconstituer la chaîne de traitement

Cette action consiste à récolter les méta-données de modèle, essentiellement les traces de construction de l'aspect logistique.

Comment sont construites les chaînes de traitement ? Comment les données sont valorisées, transformées, enrichies, filtrées, fusionnées, mappées, échangées … ?

Cette action porte sur l'aspect logistique du système entreprise. Elle reconstitue le modèle de la chaîne de traitement, et identifie les différents composants du système informatique qui vont se trouver sur le chemin d'une donnée.

Ces composants sont formés de briques logicielles ou d'activités manuelles :

- de transformation des données (calculs),
- de sélection de données (contrôles, filtres),
- de stockage des données (persistance, consolidation),
- d'échange de données (extractions, transferts, services d'exposition, mise en format d'échange).

On identifie également les systèmes (ou applications) supports à ces composants.

[12] Compte rendu d'événements (CRE) et Compte rendu d'inventaire (CRI)

[13] Know Your Customer – réglementaire bancaire

Figure PCD-64_17. Les types d'éléments manipulés dans l'action « Reconstituer la chaîne de traitement »

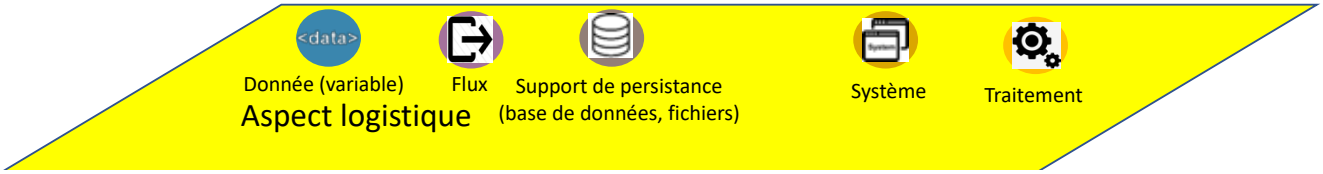
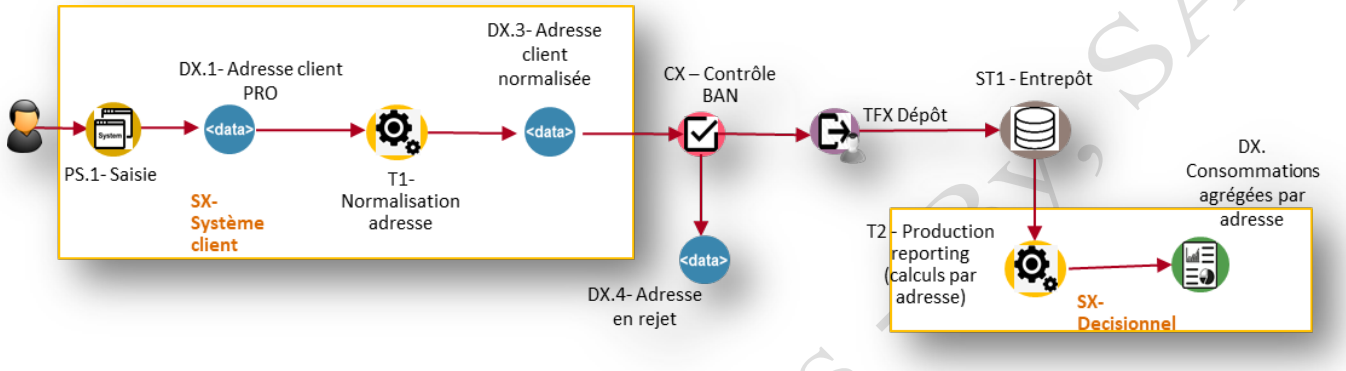


Figure PCD-64_18. Un exemple de chaîne de traitement



Cette chaîne décrit la succession de traitements, de la saisie d'une adresse à son utilisation, pour fournir des mesures agrégées par adresse.

Pour chaque composant, on recherche : sa description, les règles de transformation des données objet du *data lineage* (« données utiles »), l'événement déclencheur du traitement associé au composant.

Dans le cas d'un *data lineage*, par définition, on va rechercher tous les traitements (au sens générique) susceptibles d'influer sur la valeur des données :

- briques logicielles, applicatives qui portent des traitements de modification des données : calcul, filtre, enrichissement, transcodification, application de règles de gestion, contrôle amenant des rejets, valeurs par défaut, forçage de valeur, redressement de données …
- briques logicielles d'intégration : bus d'échange de données, plate-forme de service/d'API d'exposition de données (ces briques peuvent incorporer des règles de transformations de contenu ou de format, avec un impact sur la valeur ou l'interprétation de certaines données).

En parallèle, pour chaque bloc de traitement, on s'attache à suivre le cheminement des données visées par le *data lineage*, en identifiant les données en entrée et les données en sortie afin de suivre l'évolution des données et de recenser leurs transformations, tout au long de la chaîne.

Un exercice de *data lineage* s'effectue soit en partant d'un résultat et en reconstituant les étapes successives qui ont permis d'atteindre ce résultat (comment telle donnée a été constituée ?), soit en partant d'une source de données et en recensant les étapes successives de traitements et de transformation de ces données (comment telle donnée a été traitée ?). Au final, l'objectif est de collecter, puis de représenter les chemins qu'une donnée va parcourir dans le S.I.

De façon pratique, on alterne les deux approches :

1. en cherchant à reconstituer d'où provient la donnée (on remonte dans la chaîne de traitement en partant de la fin) ;
2. en vérifiant que la production de telle donnée entrant normalement dans la constitution de l'objet du *data lineage* (un indicateur) se retrouve bien consommée par les composants de la chaîne de traitement (on parcourt la chaîne de traitement en partant des sources).

L'identification des données doit faire le lien avec l'aspect sémantique (lien avec les objets métier) et respecter les définitions qu'un catalogue de données peut avoir recensées (voir la section outillage, chapitre 6 : les solutions de *data catalogage* offrent des capacités de représentation des *data lineage* s'appuyant sur les données de catalogage).

Dans l'idéal, toutes ces informations peuvent être retrouvées dans les référentiels de description du S.I. support aux architectes et urbanistes du S.I.

4.4 Retracer l'exécution

Par rapport aux précédentes, cette action change de registre : elle n'est plus sur le plan de la construction, mais sur celui de l'exécution. Elle récolte les méta-données d'enveloppe ou traces d'exécution. On parle, ici, d'instances : valeurs des données, objets au sens des instances de classes, instances de processus et d'activités, individus ayant mené les actions.

Comment s'exécutent les chaînes de traitement et quelles traces peut-on récupérer ? Quels événements se sont produits ? Quand ? Comment s'est déroulée l'exécution de la chaîne de traitement ?

Selon la problématique traitée et le contexte, on peut avoir besoin d'identifier les méta-données d'enveloppe (traces d'exécution).

Ces méta-données sont à rechercher dans les activités de production du S.I., dans l'exécution des activités attachées à un processus :

- délai d'exécution d'une suite de traitements,
- niveaux de qualité et de services constatés,
- couverture des données traitées, écarts de valorisation détectés,
- auteur d'une mise à jour,
- historique des valeurs successives d'une donnée (exemple : changements de n° de téléphone),
- identification de l'instance d'un flux de données, d'un *data set* ou d'une production, objet du *data lineage* (exemple tel tableau de bord à telle date).

Figure PCD-64_19. Les types d'éléments manipulés dans l'action « Retracer l'exécution »

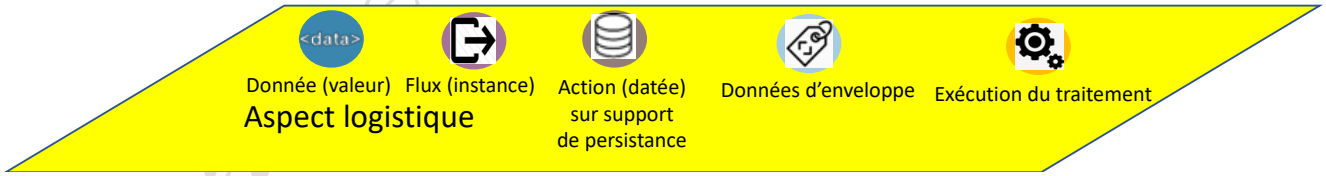
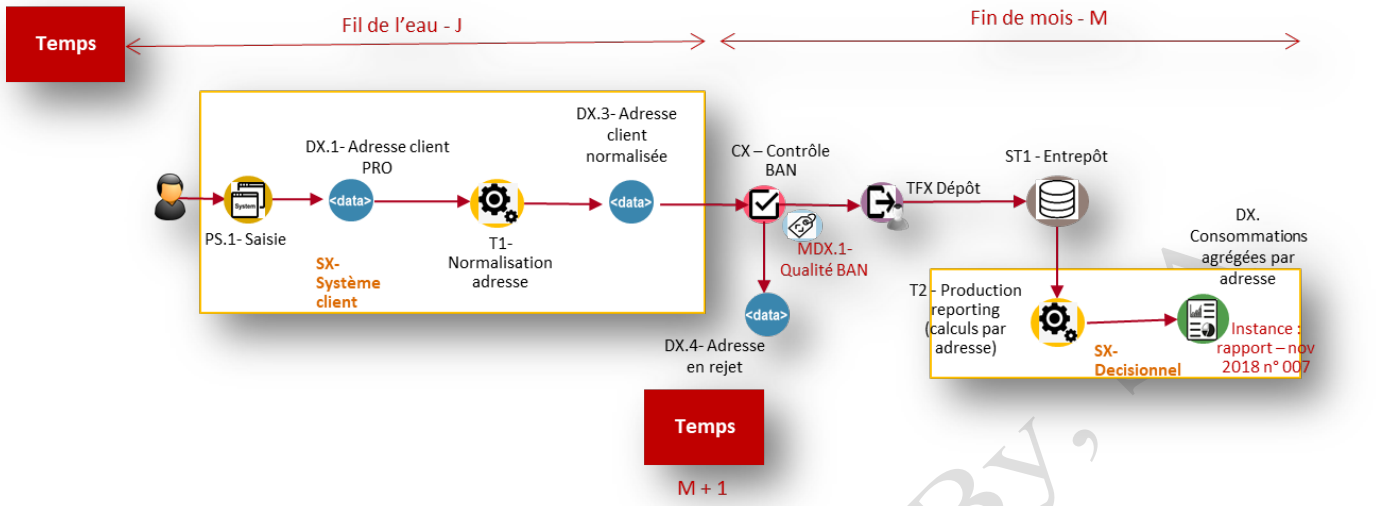


Figure PCD-64_20. Une exécution – datée – de la chaîne de traitement



Sur ce schéma, les méta-données d'exécution figurent en rouge : temps d'exécution, qualité de donnée obtenue – exemple qualité de l'adresse, instance et version d'un rapport objet du *data lineage*…

Certaines méta-données peuvent être fournies directement par la chaîne de traitement, lorsque, au sein de celle-ci, il a été prévu explicitement de collecter ces méta-données (exemples : *timestamp*, niveau de qualité d'une donnée résultant d'un traitement, auteur d'une mise à jour, conservation de la valeur précédant la mise à jour… plus généralement *logs*[14] d'exécution).

4.5 Apprécier les conditions de production de la donnée

La question de la qualité des données est quasiment intrinsèque à l'exercice de *data lineage* (la matière traitée étant la donnée, sa qualité est une caractéristique clé).

Apprécier la qualité de données relève d'un procédé à part entière. Les *data lineages* produits sont une connaissance en entrée, incontournable pour le traitement de la qualité des données (comprendre où peuvent avoir lieu les défaillances, remonter aux causes de non qualité…).

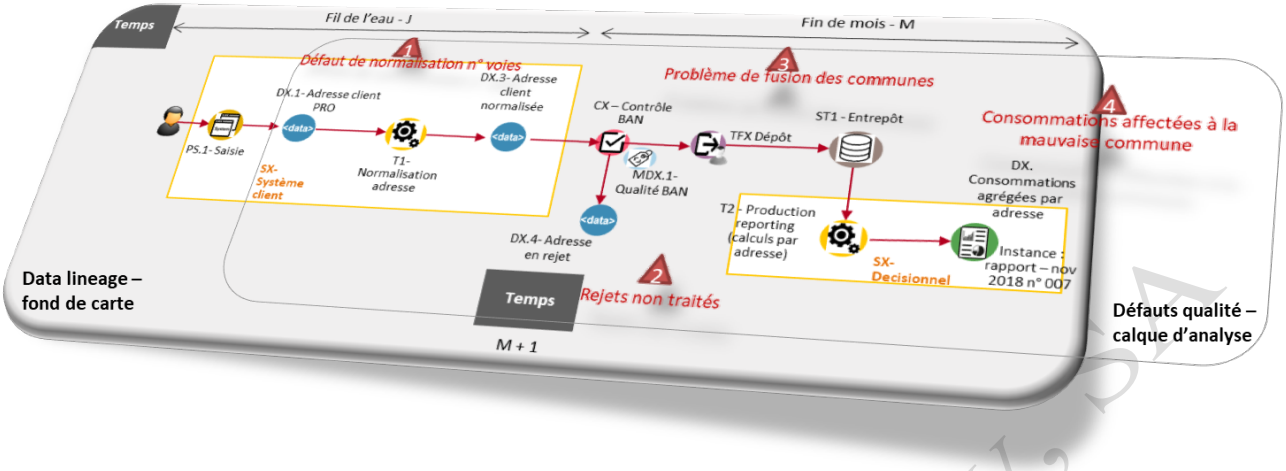
À titre d'illustration dans le cadre de l'exercice de *data lineage*, on peut identifier :

- les défauts de conception (traçabilité verticale -méta-données de modèle) à l'exemple des écarts sémantiques dans les implémentations (l'implémentation ne respecte pas la définition sémantique) ;
- les étapes pouvant générer une défaillance dans le traitement de la donnée et donc générer des défauts de qualité (et avec, pourquoi pas, un chemin de traitement spécifique dans le cas de rejets) ;
- les points dans la chaîne de traitement, pour mesurer le niveau de qualité des données (mesures qui peuvent être contrôlées automatiquement et stockées dans des méta-données dédiées, permettant à la suite de la chaîne de traitement de tenir compte des niveaux de qualité), ou pour vérifier l'absence d'écart de valorisation, d'occurrences entre le début d'un traitement et sa sortie (à l'exemple des contrôles comptables) ;
- d'une façon générale, toute situation (traitement, contexte) mettant en défaut la qualité des données en termes de : exhaustivité, exactitude, complétude, intégrité, cohérence, fraîcheur.

Toutes ces informations peuvent être retranscrites sur un calque d'analyse que l'on vient superposer à la représentation du *data lineage*.

[14] Journaux.

Figure PCD-64_21. La détection des dysfonctionnements le long de la chaîne de traitement

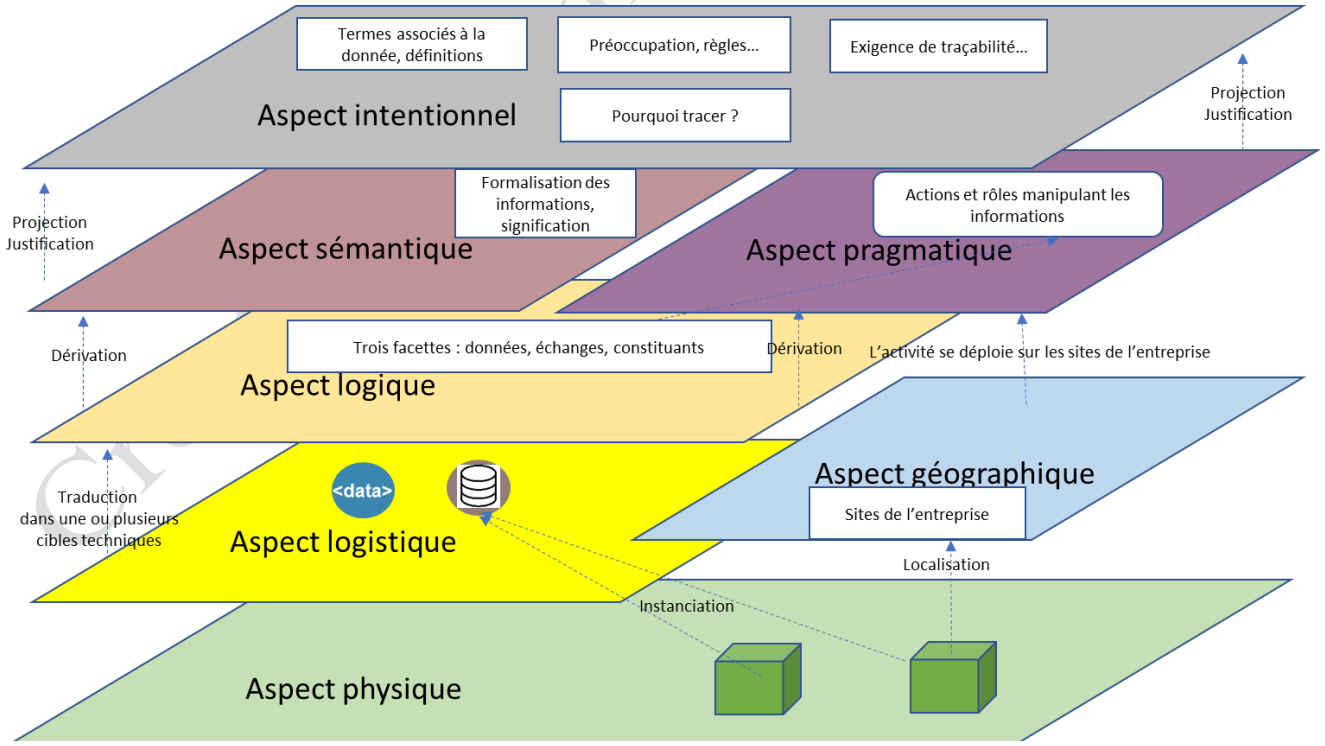


Ce schéma illustre l'idée de calque d'analyse sur le fond de carte de représentation du *data lineage*. Le calque permet de mettre en relief et de positionner les problématiques à traiter.

4.6 Mettre la donnée en perspective

Les actions précédentes suffisent pour répondre, strictement, à la demande de traçabilité. La mise en perspective demande un petit effort supplémentaire, dans le but de tirer le meilleur parti du procédé et de préparer l'avenir. Cette action consiste à documenter les éléments de construction relatifs aux données étudiées. Idéalement, elle reconstitue une chaîne de traçabilité verticale, de bout en bout : de l'aspect intentionnel jusqu'à l'aspect physique, en empruntant toutes les filières de dérivation d'un aspect à l'autre.

Figure PCD-64_22. Mise en perspective de la donnée : les types d'éléments selon les aspects



Mettre la donnée en perspective consiste à la relier à des éléments de modélisation appartenant aux aspects amont, c'est-à-dire aux univers « métier » :

- Sous l'aspect pragmatique : actions potentielles sur les données étudiées ; par exemple : qui peut modifier l'adresse (rôle et responsabilité dans les processus et l'organisation) ? Dans quel cadre ou contexte (pouvoirs, contexte) ?
- Sous l'aspect logique : si le modèle logique reflète l'existant, alors il montre la redondance ; d'où l'idée de simplifier le système.
- Sous les aspects logistique et physique : prendre en compte les phénomènes qui augmentent les risques sur les données, notamment le recours aux caches, les solutions d'informatique qui relèvent du « *shadow IT* », le *cloud computing*…
- Sous l'aspect géographique : la localisation des activités dans la géographie de l'entreprise est un facteur de multiplication puisqu'elle conduit à déployer les mêmes processus et solutions informatiques sur plusieurs sites.
- Sous l'aspect physique : le déploiement s'obtient par instanciation des composants logiciels et localisation sur des nœuds distribués dans la géographie de l'entreprise.

a. Contexte métier : aspect sémantique

Par nature et comme décrit dans la première action (repérer les données), l'aspect sémantique intervient dès le départ de la reconstitution d'un *data lineage*. Il permet de fixer la nature des données visées par le *data lineage*, puis la nature des données identifiées tout au long de la chaîne de traitement et qui concourent au résultat final (un rapport produit par exemple).

Les éléments que l'on considère dans l'aspect sémantique comprennent :

- la relation des données avec les objets métier de l'entreprise (exemple : telle donnée fait référence à l'objet métier Personne dans son rôle de Client),
- le domaine d'objets ou le sous-domaine d'objets auquel appartient l'objet considéré,
- les relations entre les objets métier,
- les règles de gestion contraignant les objets métier,
- les définitions formelles, obtenues en positionnant la notion (l'objet métier) dans le réseau des concepts.

Dans le cadre d'un *data lineage*, travailler sur l'aspect sémantique offre l'intérêt :

- de contextualiser les données (savoir de quoi on parle au sens métier),
- de s'assurer de la bonne traduction de la connaissance métier, à travers les données visées par le *data lineage* (traçabilité verticale).

b. Contexte métier : aspect pragmatique

Quels sont les activités, processus et les rôles qui amènent des transformations sur les données ? Dans les processus métier, quelle attention est portée aux données ?

Un *data lineage* par définition formalise les traitements sur les données. Ces traitements sont le reflet et le support d'activités des processus de l'organisation.

Tout *data lineage* s'inscrit donc dans l'exécution d'un ou plusieurs processus.

La représentation de ce niveau processus (dans l'aspect pragmatique) garantit la bonne interprétation et la compréhension d'un *data lineage*.

Dans le cadre d'un *data lineage*, le travail sur l'aspect pragmatique permet :

- de contextualiser les traitements sur les données (savoir dans quelle activité de quel processus ces traitements ont lieu),
- s'assurer de la pertinence des traitements sur les données visées par le *data lineage* par rapport aux activités et aux règles d'organisation (traçabilité verticale).

Certes, le *data lineage* se situe essentiellement dans l'aspect logistique. Il est d'abord à la main des équipes MOA / MOE, et permet de produire une représentation applicative du système et de la chaîne de traitement (souvent par

phase de traitement – du fait du silotage du S.I., plus rarement de bout en bout). Cette partie du travail intervient au plus près de la réalité de la réalisation informatique et de l'exécution. Toutefois, le travail sur les aspects métier – sémantique et pragmatique – rattache la donnée et son cheminement au contexte métier (processus, activité concernés). Il aide à dépasser les silos de traitements, en reconstituant une vision de bout en bout (de l'origine des données aux données produites). On parle parfois, à ce propos, de *data lineage* fonctionnel.

c. Contexte d'implémentation : aspect géographique et aspect physique

Aspect géographique : pour certaines chaînes de traitement, il peut être important de distinguer les sites d'exécution qui génèrent des comportements différents (par exemple, des sites industriels ou des entités juridiques dans des pays différents).

Comment la localisation des traitements sur les données intervient dans la compréhension de la chaîne de traitements, des traces ?

Où est stockée la donnée ? D'où vient-elle – de quel site ? Où est-elle utilisée, partagée ? À quelle norme réglementaire ou juridique locale / pays répond-t-elle ? Quelles sont les variations locales des règles de gestion, de la politique de la donnée ?

Dans certaines situations, le *data lineage* au niveau des traitements et des systèmes ne suffit pas pour analyser un dysfonctionnement. Il est nécessaire de descendre au niveau du déploiement physique (instance de stockage dans telle base de données, localisée, en passant par l'utilisation d'un proxy notamment). On complète alors l'exercice de *data lineage* d'un système, d'un traitement par l'étude de son déploiement dans l'aspect physique de l'entreprise.

En quoi le déploiement physique peut influer sur la façon dont la chaîne de traitement va se comporter (traces) ?

Où cette donnée est-elle physiquement localisée (base relationnelle, base NOSQL, système BI…) ? Dans combien de base de données ou sources est-elle présente ?
Avec quel niveau de service ?

4.7 Préciser l'administration des données et les responsabilités afférentes

Cette action consiste à recueillir ce que l'on peut nommer les méta-données (traces) d'administration. Ces données se trouvent à deux niveaux :

1. celui des modèles, qui désignent les rôles et responsabilités prévus, en tant que types ;
2. celui de l'exécution, avec les « traces d'administration » mentionnant les individus impliqués dans des occurrences particulières des traitements.

Comment est gérée l'administration des données objet du data lineage ?
Qui a créé cette donnée ? Qui la consomme ? À qui appartient-elle ? Qui en assure l'administration ? Quelle est la politique associée ? Quels sont les risques associés ? Quand la donnée doit-elle être effacée ? Quand cette donnée a-t-elle été créée, actualisée ?

Pour toutes les données entrant dans la chaîne de production étudiée par le *data lineage* (données résultant, données entrant dans la composition du résultat, données d'origine/source, données transformées), il s'agit d'identifier les règles d'administration associées :

- propriétaire de la donnée ou plus largement RACI[15] support au cycle de vie de la donnée ;
- politique de la donnée (sensibilité réglementaire, risques associés – exemple de publication - confidentialité) ;
- méthodologie de production de la donnée (particulièrement utile pour des indicateurs, par exemple objet d'une labélisation, ou encore dans le cadre de production de données statistiques – méthode statistique employée),
- règles de dépendance avec les systèmes de référence (respect de nomenclatures /codifications, cohérence avec des référentiels de données – exemple référentiel client)[16] ;
- reporting et mesure de qualité de données.

Ces méta-données sont utiles :

1. au niveau des actions effectuées dans le cadre de l'administration de données (normalement déjà identifiées au moment de la reconstitution de la chaîne de traitement, au travers des activités d'administration en rapport avec le cycle de vie des données – exemple de la suppression de données),
2. au niveau de la connaissance des données administrées pour une meilleure gestion de celles-ci.

5. Résultats produits

5.1 Exigences de représentation

La représentation d'un *data lineage* passe par l'association de descriptions textuelles (inventaires de méta-données) et de descriptions visuelles (adoption de formes symboliques de représentation des méta-données et des articulations entre elles).

La lecture directe de l'inventaire des méta-données collectées n'est pas immédiate. Il est difficile d'avoir une vision d'ensemble. Une représentation graphique complémentaire est indispensable. Cette représentation a pour vocation :

- de rendre visible la chaîne de traitement de bout en bout (vision d'ensemble),
- d'être un référentiel cartographique pour positionner les méta-données collectées et reproduire le cheminement au sein des traitements, les règles identifiées (rôle d'une carte),
- au moment de traiter la problématique, de pouvoir positionner les points d'analyse (point de dysfonctionnement, d'optimisation, d'approfondissement identifiés : idée de calque d'analyse évoqué dans les exemples précédents).

L'expérience montre que, dans beaucoup d'endroits de l'entreprise, des *data lineages* peuvent être réalisés. Il devient alors intéressant de définir une symbolique commune (un langage commun) à l'échelle de l'entreprise.

L'ensemble des métadonnées collectées lors de l'analyse du processus de traitement sont consignées et serviront de support à une représentation du *data lineage*. La façon de consigner les méta-données relève de l'outillage du procédé. Les solutions s'étagent de l'utilisation d'un fichier excel à la mise en place de solutions de data gouvernance à l'échelle de l'entreprise, en passant par le recours à des outils d'architecture d'entreprise. Ce point est détaillé dans le chapitre 6.

Avant de décider de l'outil, l'organisation qui se lance dans ce travail doit se fixer un niveau d'exigence quant aux représentations :

a) Soit elle adopte des représentations ad hoc, spécialement pour l'exercice de lignage : *flowcharts* intuitifs avec une symbolique évocatrice ad-hoc ou issue d'une autorité (à l'exemple de la BCE), établis à l'aide des outils ordinaires de dessin ou de communication ; tableaux ou formulaires pour présenter les méta-données collectées.

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

[16] Cette dépendance doit également apparaître au niveau des méta-données de modèles (traces de construction) – flux / interconnexion avec les référentiels de l'entreprise.

b) Soit elle met en avant les exigences globales de maîtrise de la connaissance du système, et elle opte pour une démarche de type « référentiel », imposant le recours aux notations standard.

La première option convient pour répondre rapidement à la demande de traçabilité et pour faciliter la communication entre les acteurs impliqués. Elle trouve ses limites dans la mise en perspective (action 7) et le travail sur les traces de construction (premières actions).

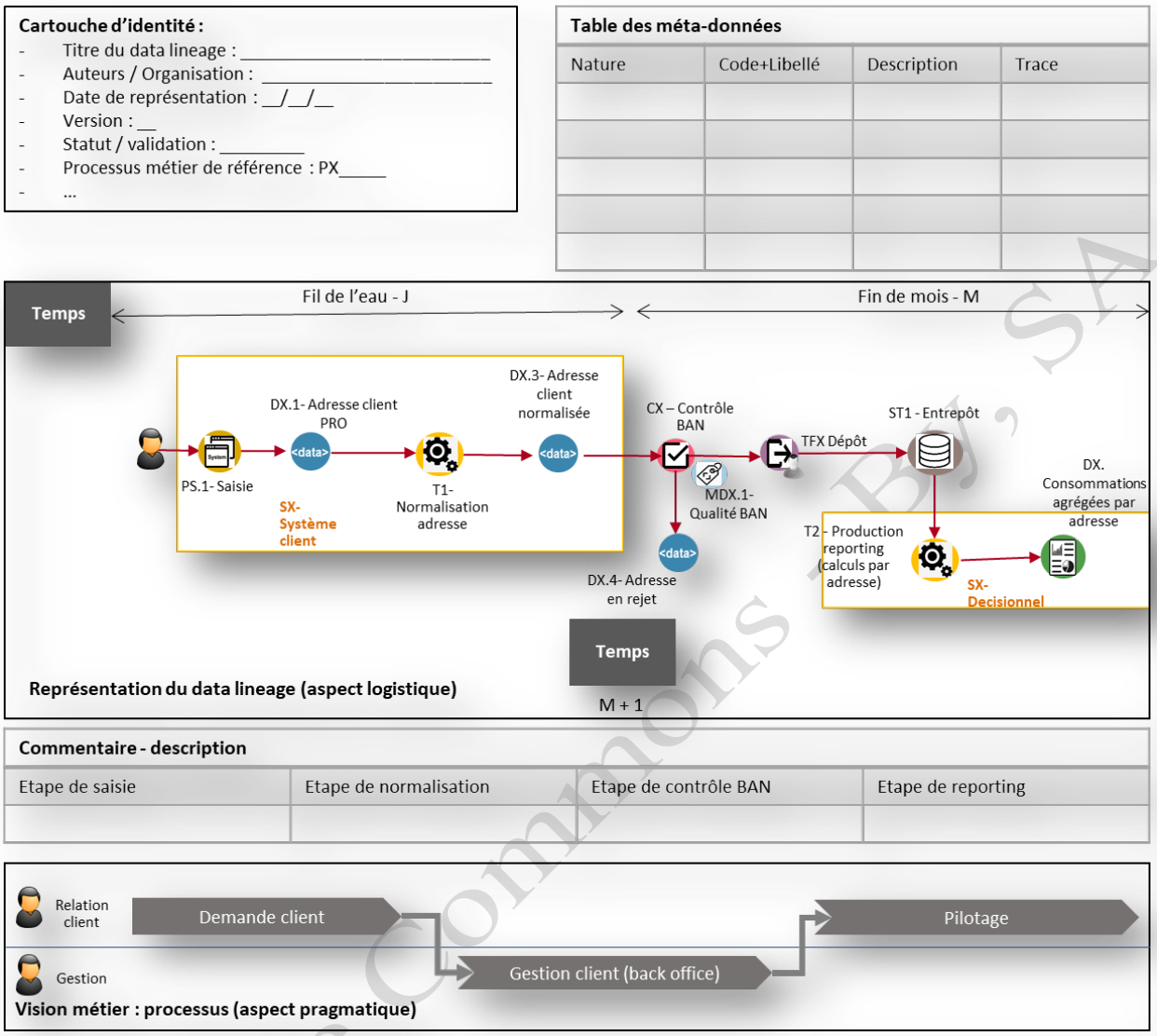
La deuxième option tire parti du référentiel de description de l'entreprise, avec un effet d'accélération à moyen terme grâce à la capitalisation de la connaissance sur le système entreprise, dans tous ses aspects. Elle exige de se familiariser avec les notations, tout particulièrement UML qui s'applique à la plupart des aspects définis par la Topologie du Système Entreprise. D'autres notations complètent la boîte à outils, particulièrement BPMN pour la modélisation des processus. L'aspect intentionnel peut, éventuellement, se formuler à l'aide de notations intuitives, moins formelles, telle qu'Archimate.

Le choix de l'une ou l'autre option dépend de plusieurs facteurs :

- Les demandes de traçabilité sont-elles ponctuelles ou, au contraire, le travail sera-t-il récurrent ?
- L'entreprise bénéficie-t-elle d'un bon niveau de maturité quant aux pratiques d'architecture d'entreprise ? Met-elle en œuvre une approche de type « référentiel » pour gérer efficacement son capital de connaissances ?
- Les compétences mobilisées couvrent-elles la pratique des notations et les techniques de modélisation ?
- L'enjeu lié à la traçabilité justifie-t-il un effort d'apprentissage ?
- Les résultats devront-ils être partagés entre plusieurs organisations (filiales, partenaires, prestataires…) ? Le cas échéant, on a intérêt à s'appuyer sur les standards, à condition de ne pas risquer un rejet.

La figure ci-dessous illustre une forme de représentation d'un *data lineage*. D'autres exemples viennent illustrer les résultats produits dans la suite de ce chapitre.

Figure PCD-64_23. Exemple de formulaire pour la publication d'un data lineage



5.2 Exemple de produit – cas n°1 : qualité des données

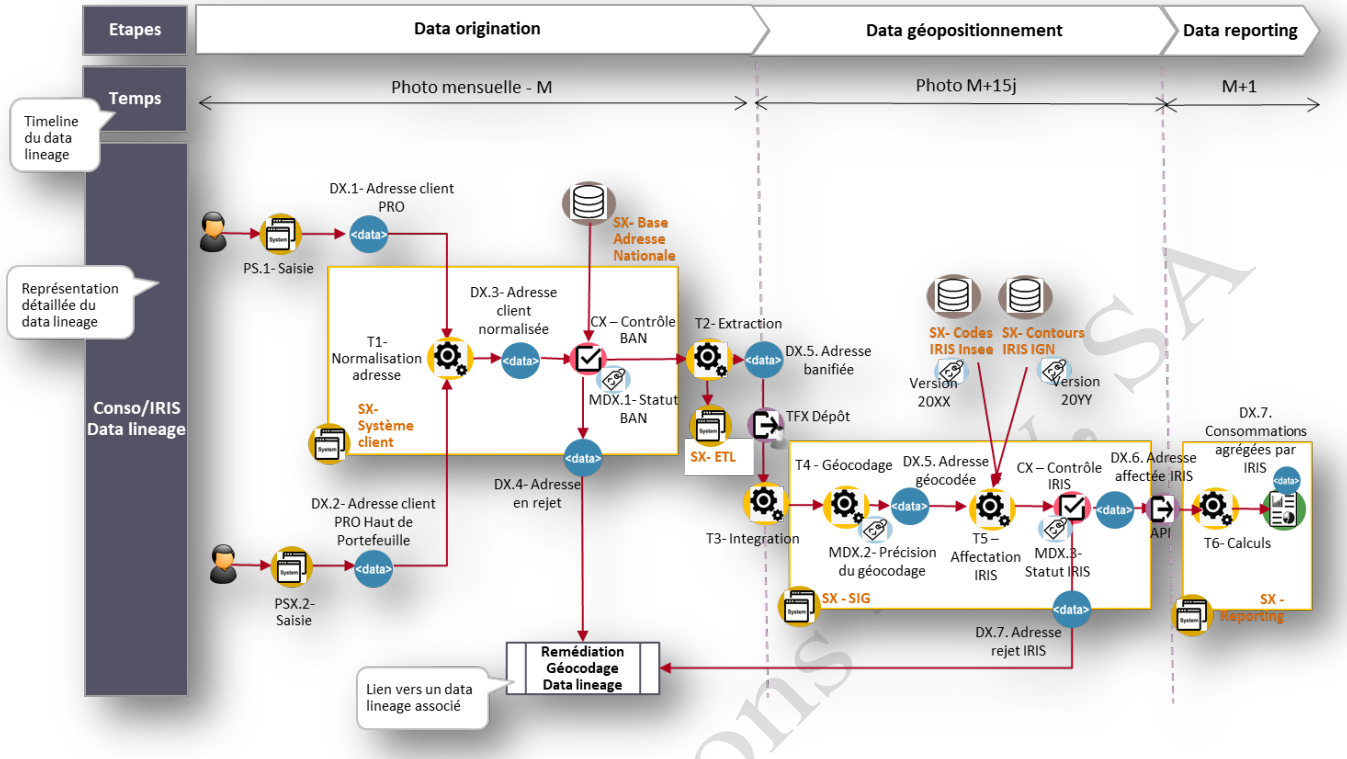
Données concernées :

- Résultat objet du *data lineage* : production de statistiques de consommation d'énergie par zones IRIS (découpage Insee/IGN).
- Données à tracer : données de géopositionnement (adresses, rattachement à une zone IRIS) permettant d'affecter les données de consommation de chaque point à la bonne zone IRIS de rattachement.

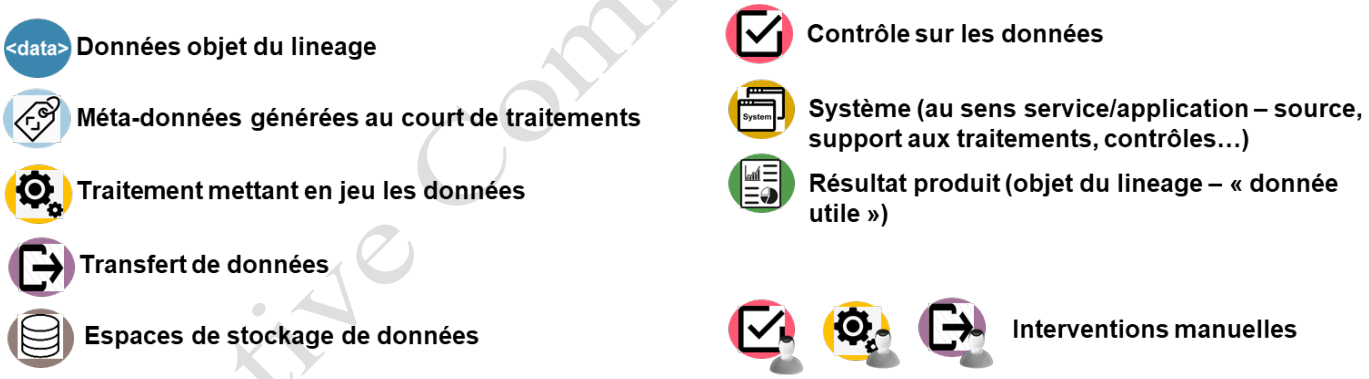
Problématiques :

- S'assurer (traçabilité) de l'exhaustivité de la couverture territoriale (tous les points de consommation ont pu être correctement géopositionnés à leur bonne zone IRIS).
- Identifier les pistes d'optimisation de la chaîne de production de statistiques : amélioration de la qualité du géopositionnement des données, optimisation des délais de traitement.

Figure PCD-64_24. Une forme de représentation d'un data lineage



Légende :



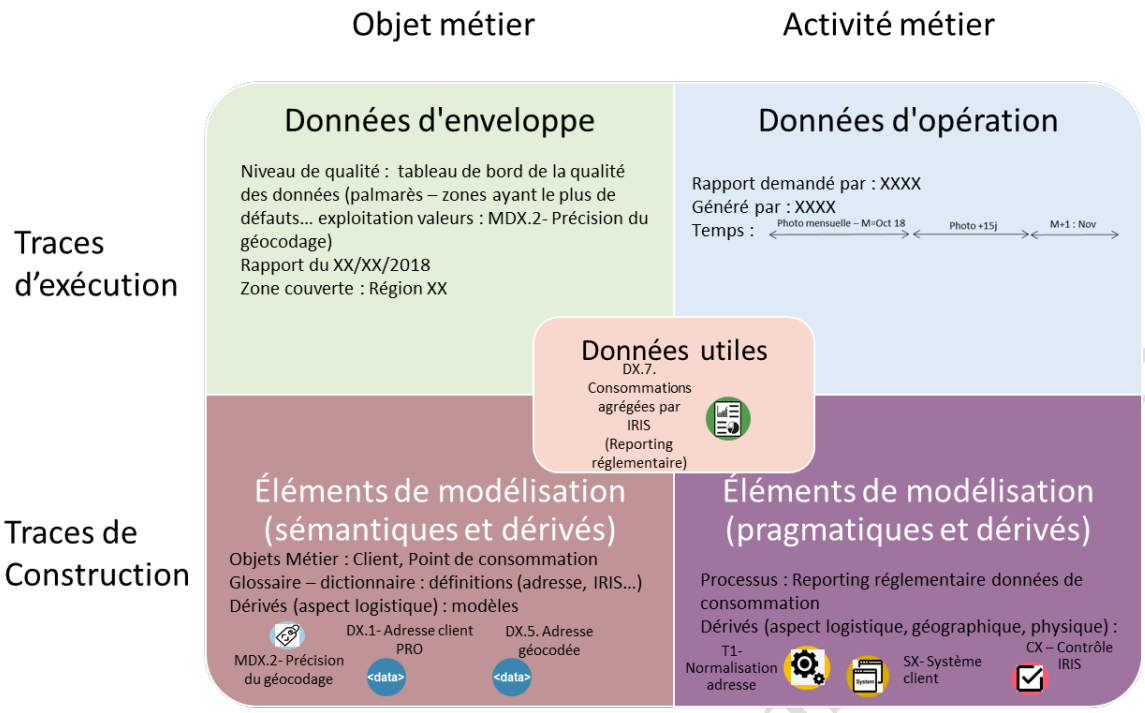
Commentaire de la figure

Cet exemple présente le *data lineage* des données concourant à la production d'un rapport réglementaire.

L'intention est double :

1. apporter la preuve de la bonne réalisation du rapport attendu (homologation du rapport),
2. analyser des dysfonctionnements apparus lors de l'exécution des premiers rapports (défauts qualité).

Figure PCD-64_25. Typologie des méta-données collectées



Exemple de table de recueil des méta-données :

Nature (exécution, construction)	Code+Libellé (DX1__, T1__, SX1___…)	Description _____	Aspect	Traces (….)	Commentaire

5.3 Exemple de produit – cas n°2 : réglementaire (RGPD, BCBS)

a. Domaine bancaire – réglementation BCBS 239

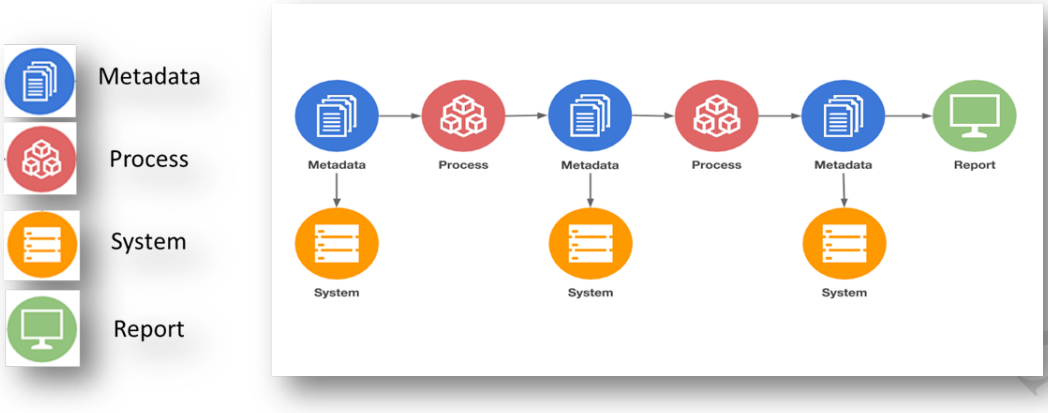
Les établissements bancaires font face depuis 2014 à un échéancier contraignant pour être en conformité avec le Comité de Bâle sur le contrôle bancaire (BCBS 239), qui a propulsé le *data lineage* (mise en place d'une traçabilité) dans le top 5 des chantiers incontournables à mettre en œuvre pour améliorer leurs capacités à produire et fiabiliser les rapports réglementaires. Rapports qui font l'objet d'exercices réguliers de tests par la Banque Centrale Européenne (BCE).

La BCE a ainsi émis plusieurs cadres de description de *data lineages* concernant la production de rapports réglementaires (type indicateurs liés au risque de crédit ou au risque de liquidité).

Ces cadres de description fixent une nomenclature des données de *reporting* à produire par les banques (équivalent d'un dictionnaire de données).

Ils fixent également une symbolique minimale à des fin de comparaison. Les banques doivent se concentrer sur quatre types d'éléments : les métadonnées, les systèmes, les processus et les rapports.

Figure PCD-64_26. La symbolique spécifique BCE[17]



Le *data lineage* consiste à fournir la chaîne de traitement permettant de produire un rapport réglementaire, en identifiant les systèmes (exemple une base de données) et processus (activités : traitement sur les données - exemples un script ETL, un algorithme) concourant à la production du rapport et en faisant ressortir les informations sur les données (méta-données : nom de colonne dans une table par exemple).

Instructions de la BCE pour la production du *data lineage* (extrait) :

« *Le data lineage* est défini comme le cycle de vie des données qui inclut les origines des données et leur déplacement dans le temps. Il décrit ce qu'il advient des données lorsqu'elles passent par divers processus. Il aide à fournir une représentation support à l'analyse et simplifie le traçage des erreurs jusqu'à leurs sources. En d'autres termes, un *data lineage* est une carte de processus, instrument utile pour évaluer la manière dont les banques mettent en œuvre les principes du BCBS239 sur les aspects liés à l'agrégation et au *reporting* des données de risque.

Le *data lineage* doit permettre aux banques d'assurer leur responsabilité vis-à-vis de leurs processus de consolidation des données de risque.

À des fins de comparaison, dans cette analyse en profondeur, les banques doivent se concentrer sur quatre types d'éléments : les métadonnées, les systèmes, les processus et les rapports.

Instructions fournies par la BCE au regard des éléments du présent procédé :

1. qualifier l'objet du *data lineage* (exemple le montant de dépôts et ses composants) -> *Procédé – « données utiles » - repérer les données ;*
2. établir le *data lineage* : i) comment les chiffres correspondants sont complétés ; ii) par quels systèmes informatiques ; et iii) s'il existe une ou plusieurs sources de données. Pour chaque base de données incluse dans le cheminement, indiquez : 1) la solution technique (SGBD) et 2) la taxonomie des données. -> *Procédé – « méta-données de construction » - reconstituer la chaîne de traitement ;*
3. identifier et expliquer les étapes du processus de traitement des données qui sont entièrement automatisées, partiellement automatisées ou manuelles ;
4. documenter le type de contrôles, y compris pour les processus partiellement automatisés et manuels, mis en place à chaque étape, les fonctions responsables de ces contrôles et les incitations du personnel à tous les niveaux du groupe bancaire (siège, succursales, filiales). Les personnes ou fonctions responsables doivent être nommés. En particulier, expliquer à quel niveau du processus les contrôles sont effectués, quels sont leurs résultats, comment les erreurs sont identifiées, signalées et corrigées (escalade), comment ces chiffres sont réconciliés avec d'autres sources, si les données sont sans ambiguïté définies dans un dictionnaire de données. *-> Procédé – méta-données de construction (aspects sémantique et pragmatique), de gouvernance – mettre la donnée en perspective ;*

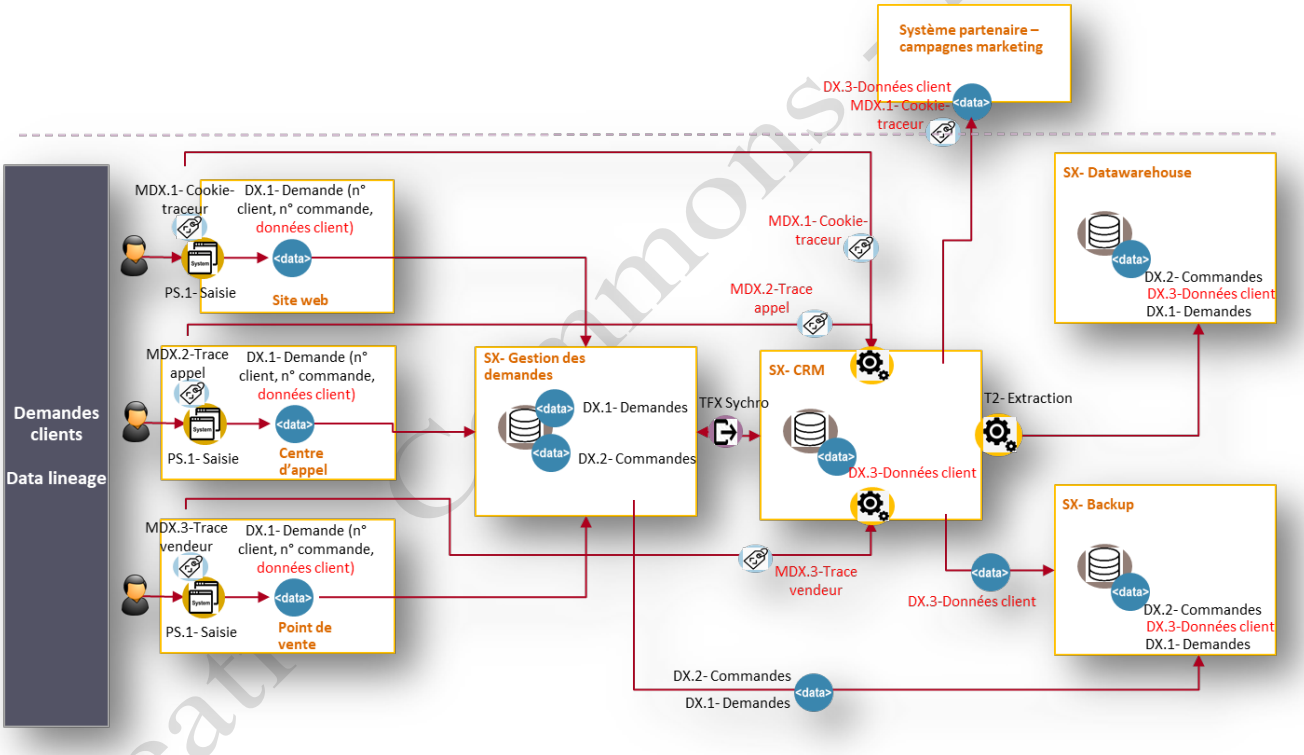
[17] Il n'y pas de référentiel officiel de cette symbolique. La BCE y fait référence. Elle se serait inspirée d'un éditeur logiciel.

5. expliquer quels indicateurs de qualité des données sont en place, y compris les niveaux de tolérance, et expliquez le processus en cas de violation. -> *Procédé – Apprécier les conditions de production de la donnée…*
6. décrire si des modifications éventuelles dans le traitement des données pourraient améliorer l'efficacité du processus et la fiabilité des chiffres -> *Procédé – usages.*

Remarque : Via BCBS, la BCE a fixé ses exigences de traçabilité. Ces exigences ont porté d'abord sur des données agrégées. L'évolution actuelle, au travers de BIRD - *Banks' Integrated Reporting Dictionary* et de l'exemple Anacredit, va amener les banques à fournir non plus des données agrégées mais des données granulaires (élémentaires) normalisées (alignées sur le dictionnaire de la BCE). La traçabilité attendue devra être au niveau de ces données élémentaires avec toutes les conséquences et les nouvelles exigences à respecter en termes de maîtrise de cette traçabilité à une maille élémentaire au niveau des systèmes bancaires.

b. Gestion des données personnelles - Le RGPD

Figure PCD-64_27. Un cheminement de données personnelles, au sens du RGPD

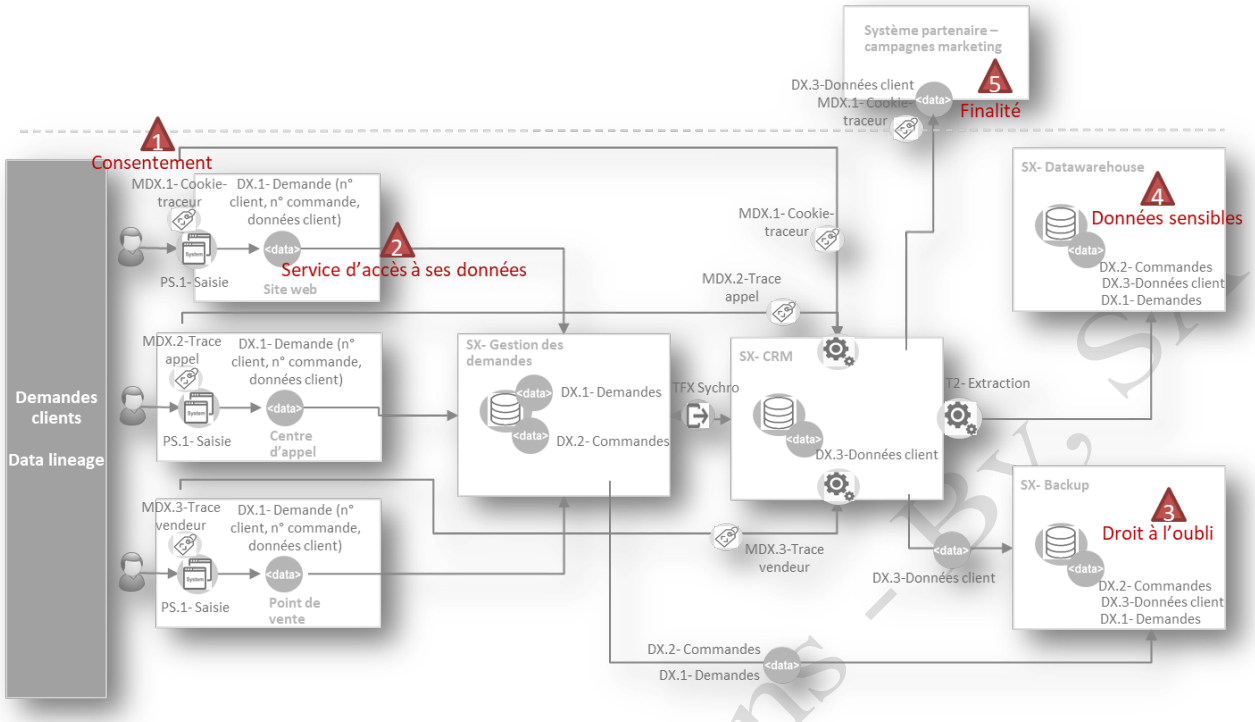


Cet exemple illustre le cheminement de données personnelles à partir de canaux de communication, jusqu'aux systèmes *back office* en couvrant également les systèmes de *backup*[18].

À partir de cette représentation, des points de vigilance vis-à-vis du RGPD sont identifiés (voir schéma suivant) et devront faire l'objet d'un traitement.

[18] Relève de l'exercice de « data flow mapping »

Figure PCD-64_28. Détection de points de vigilance sur le parcours des données sensibles



5.4 Exemple de produit – cas n°3 : Framework LIME (*Lineage in Malicious Environment*)

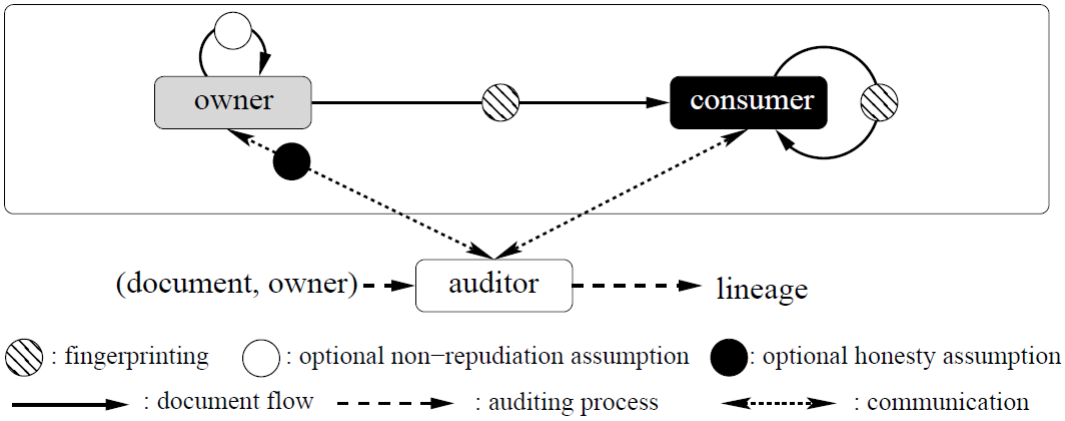
LIME est un cadre de *data lineage* en situation de flux de données entre entités qui assument deux rôles principaux :

- fournisseur et propriétaire des données ;
- consommateur des données.

Le cadre de *data lineage* permet de mettre en évidence les dispositifs ou non de sécurité des données : identification, non répudiation, préservation de l'intégrité, risques de fuites de données (divulgation), traçage des flux de transfert (marquage/trace du fournisseur et du destinataire dans les données par le fournisseur), preuves post fuites…

LIME met en évidence un troisième rôle d'auditeur, en relation avec les deux autres rôles. L'auditeur est le gardien du *framework* et, par-delà, de la confiance attendue.

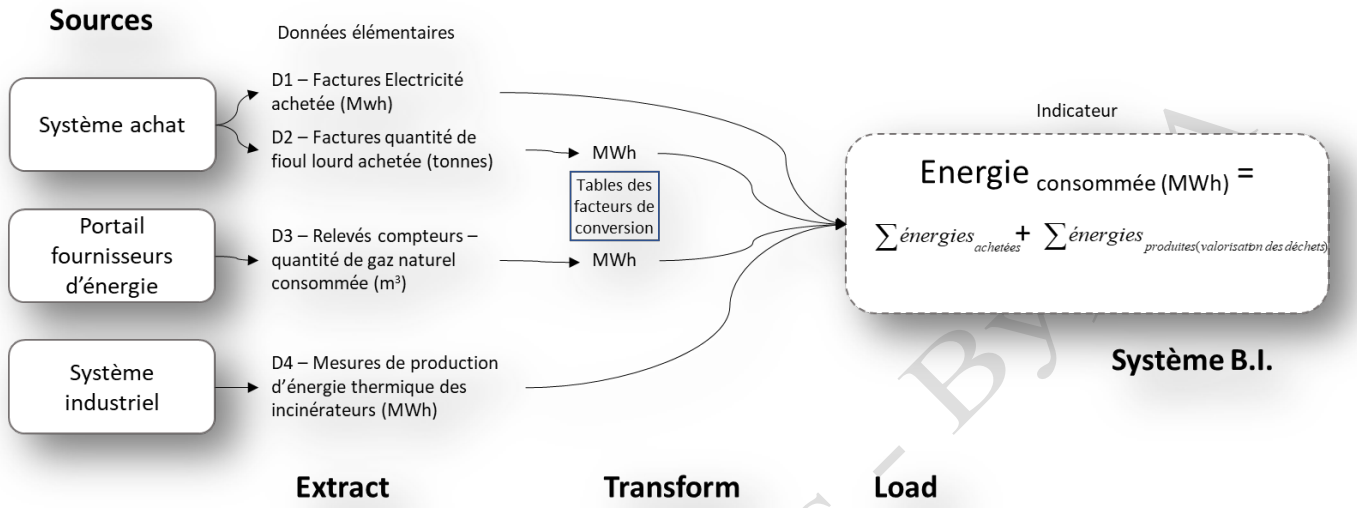
Figure PCD-64_29. Data Lineage in the Malicious Environment



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

5.5 Exemple de produit – cas n°4 : Composition d'un indicateur

Figure PCD-64_30. Data Lineage en environnement B.I. (Business Intelligence) – composition d'un indicateur

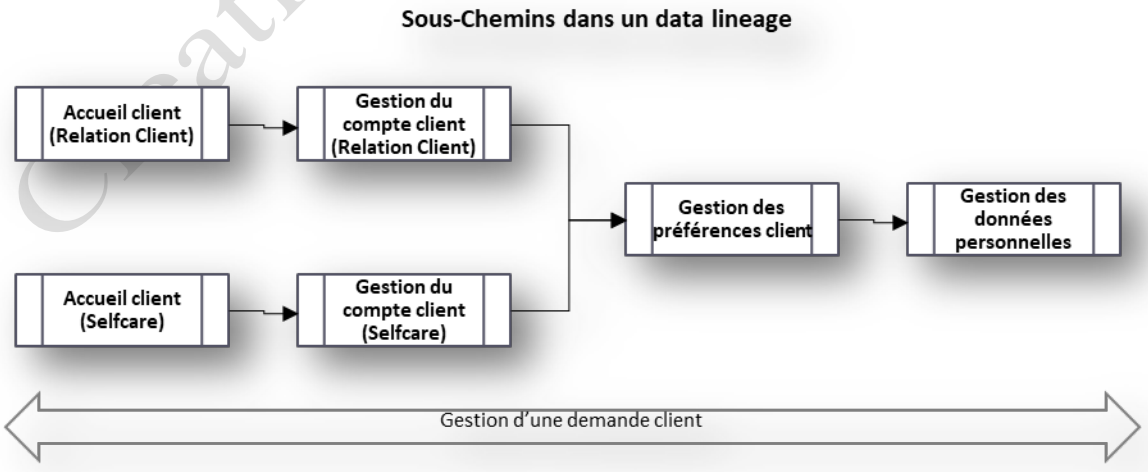


5.6 Décomposition des chemins

Parfois, le *data lineage* d'une donnée demande un effort conséquent pour mettre en lumière une multitude d'étapes dont la vue d'ensemble devient difficile à dégager. Il est, alors, de bonne pratique de s'appuyer sur une approche par sous-chemins (ensemble cohérent d'étapes autour d'une sous-finalité de traitement). Exemples : *data lineage* pour la production des données en dénominateur et *data lineage* pour la production des données en numérateur d'un indicateur, *data lineage* des étapes de préparation de données avant le *data lineage* des étapes de calculs. Le bon découpage d'un *data lineage* concourt à son efficacité et à sa lisibilité (vue macro d'ensemble, réutilisation de briques de parcours, décomposition de l'effort, etc.).

La figure ci-dessous, présente une vision macro de *data lineage* concernant la gestion des données personnelles du client. La décomposition en sous-chemins facilite la gestion, la représentation et la lecture d'un *data lineage*.

Figure PCD-64_31. Décomposition d'un chemin



Cet exemple, a priori simple, est pourtant représentatif. Plusieurs canaux d'acquisition de la donnée portent des composants applicatifs différents avec des interactions en propre (représentés par des sous-chemins différents). Et une même suite de traitements est commune à l'ensemble des canaux.

L'exercice rejoint les démarches de conception d'architecture S.I. À ce titre, les architectes S.I. peuvent aider au bon découpage de la vue d'ensemble en sous-chemins (vues détaillées) d'un *data lineage*.

5.7 Point de cohérence

Dans un système complexe ou très réparti, il est important de disposer de points de cohérence. Garantir des points de cohérence de la donnée est un des moyens pour découper le *data lineage* sur des territoires organisationnels plus restreints (voir le point précédent sur l'idée de sous-chemins).

Un point de cohérence apporte une information complémentaire. On considère qu'à une étape de la chaîne de traitement, la donnée est fiable, et est conforme à l'attente (par exemple à un niveau de qualité). L'effort d'analyse du *data lineage* va alors porter sur la suite de la chaîne de traitement.

Conserver les points de cohérence pour d'autres analyses peut être un facteur de productivité, en évitant de réanalyser une partie de la chaîne.

5.8 Critères d'acceptation

Le résultat produit doit répondre aux exigences qualité suivantes :

- lisibilité de la représentation (c'est un élément clé, le *data lineage* produit va servir de carte commune entre acteur face à une analyse, à une problématique à traiter) ;
- fidélité de la représentation (les architectes et responsables applicatifs peuvent être amenés à valider le résultat produit) ;
- identification des zones d'incertitudes ;
- couverture du périmètre de données (pour un jeu de données, s'assurer que l'on capte l'ensemble du périmètre : par exemple, vérifier qu'un site ou une nature de données ne sont pas oubliés) ;
- conservation des éléments de preuves recueillis au cours du travail de reconstitution du data lineage (traces) ;
- actualisation de la représentation (le résultat produit a fait l'objet d'une mise à jour récente) ;
- responsabilité pour assurer le partage des résultats (il existe une entité en charge des résultats et de leur publication dans un référentiel que l'on peut consulter avec le niveau de fiabilité attendu).

Au final, ces critères doivent permettre de disposer de la bonne représentation (la bonne carte par analogie), avec les bonnes informations (les bons composants de la carte).

La réglementation BCBS 239 de la BCE, dans le domaine bancaire, illustre ces critères d'acceptation. Elle précise que le *data lineage* doit permettre de comprendre :

- comment les données attendues sont valorisées dans le sens affectation d'un résultat ;
- par quels systèmes informatiques ;
- s'il existe une ou plusieurs sources de données ;
- quelle est la part automatisée ou manuelle...

Pour chaque base de données incluse dans le *data lineage*, la réglementation demande d'indiquer : 1) la solution technique (SGBD) ; 2) la taxonomie des données et la mise en correspondance des différentes taxonomies. La BCE indique attendre également, si d'éventuelles modifications des activités et des chemins identifiés par l'exercice de *data lineage* pourraient améliorer l'efficacité du processus et la fiabilité des chiffres fournis.

5.9 Capitalisation et gestion des résultats produits

Un *data lineage* représente un effort continu et conséquent. Les résultats de cet effort doivent être capitalisés.

Comme tout capital de connaissance, la façon dont il sera formalisé (sous forme de document bureautique ou dans un système logiciel dédié – type référentiel d'entreprise) va influer sur sa gestion : maintenance, correction, capacité de partage.

Cette gestion est incontournable. La matière représentée dans un *data lineage* est extrêmement vivante.

Une fois le *data lineage* représenté, le risque fort est de se retrouver, après un certain temps, avec des représentations non à jour, « fantômes » (dont on a presque oublié l'existence), difficilement exploitables sans la présence de ceux qui ont pu en produire la représentation.

C'est pourquoi, l'exercice de *data lineage* est un des composants importants de la *data governance*. Celle-ci doit définir la politique de *data lineage*, ainsi que les moyens nécessaires à sa gestion (organisation, processus dédié de niveau gouvernance des données, outillage).

6. Outillage du procédé

6.1 Collecte des méta-données

La base de l'outillage repose sur les moyens de collecte des méta-données.

Soit la collecte est assurée à l'échelle de l'entreprise par des dispositifs de type RDE (Référentiel de Description de l'Entreprise), et le *data lineage* n'a plus qu'à exploiter ces dispositifs.

Soit il est nécessaire de mener une démarche de collecte ad hoc pouvant aller jusqu'à effectuer le rétro-engineering de parties des systèmes étudier. Cette démarche de collecte passe par un travail d'enquête, de recueil des méta-données, de qualification de ces méta-données.

Définir explicitement la stratégie de collecte des méta-données est clé pour la réalisation de *data lineage*.

Soit on considère qu'il s'agit d'un effort ponctuel et la stratégie de collecte va mobiliser des moyens temporaires (*task force*, ressources d'un projet, mobilisation d'une équipe de *data quality management*, *data scientists* en phase de préparation de données…).

Soit on considère qu'il s'agit d'un effort récurrent avec des enjeux dans le temps et la stratégie de collecte doit s'attacher à définir :

- une organisation pérenne (activités et rôles dédiés au sein d'une équipe de data management),
- des dispositifs de collecte prévus à cet effet : automatisation du recueil des méta-données via des connecteurs/sondes sur les systèmes existants, via un « *design by trace* » prévu dès la conception des systèmes – en réponse à des exigences de traçabilité (à l'exemple de logs que l'on peut positionner).

6.2 Solutions logicielles support

La collecte des traces d'exécution (traçabilité horizontale, au niveau de la chaîne de traitement), la connaissance des traces de construction (traçabilité verticale) et les représentations que l'on peut en retirer forment un corpus riche d'informations.

Il existe plusieurs types de solutions logicielles pour gérer et exploiter un tel corpus :

- les solutions de type référentiel d'entreprise qui donnent la meilleure vision d'ensemble des différents niveaux de représentations (par aspects) d'un *data lineage* et surtout, qui peuvent gérer la conservation des traces de construction entre aspects (par exemple lien entre les choix au niveau de l'aspect pragmatique – processus et les choix au niveau de l'aspect logique puis logistique). Le résultat d'un data lineage par nature vient compléter des vues/modèles référencés dans les outils support aux démarches d'AE (modélisation des processus, des objets métier, du S.I…)[19] ;

[19] Ces solutions intègrent également la possibilité de représentations utilisant UML (choix d'une symbolique UML de représentation d'un *data lineage*), permettant l'interopérabilité et la navigation entre les modèles de représentation (on peut alors facilement identifier que telle donnée objet du *data lineage*, appartient à tel objet métier qui est opéré par au niveau de tel processus via tel services).

- les solutions de *data management*/*data governance* qui, en partant de la connaissance des données (glossaire métier, dictionnaire de données, catalogue de données), l'étendent aux cycles de vie des données et aux *data lineages* associés[20] ;
- des solutions spécifiques et dédiées au *data lineage*, solutions qui, au-delà de dimension représentations des *data lineages*, visent à collecter le plus automatiquement possible les méta-données nécessaires à ces représentations[21] ;
- les solutions de type « ETL » ou « *data preparation tool*s », dans le cas du domaine d'application des systèmes de *business intelligence*, *Big data*/*data lake*[22] ;
- solutions reposant sur des moyens bureautiques avec toutes les limites afférentes (intégrité, utilisabilité, maintenabilité…)[23].

Nous recommandons de rechercher la simplicité et la meilleure intégration possible de l'outillage dans la chaîne de production du SI, plutôt que d'ajouter des solutions dédiées, qui risquent de mal communiquer avec le reste.

Quelques critères de sélection de l'outillage :

- couverture des aspects et axes de traçabilité ;
- respect des standards de représentation (UML, BPMN) ;
- en prise avec le S.I. pour collecter des méta-données, directement ou non ;
- passage à l'échelle de la gouvernance des données ;
- lisibilité des résultats produits pour des exercices collectifs d'analyse (intégration de fonctions collaboratives).

6.3 Gouvernance de la production de *data lineage*

L'effort de *data lineage* est toujours conséquent et mobilise des moyens importants (le cumul, à l'échelle d'une entreprise, des efforts liés directement à la production de *data lineages* peut être surprenant).

L'outillage doit comprendre des dispositifs de gestion, de production et d'administration des *data lineages*.

Un tel dispositif peut reposer sur la mobilisation de « *data designers* » au sein d'une équipe centralisée. La constitution d'une équipe de data designers qui tournent sur les activités de *data lineage* et capitalisent est un facteur de productivité et d'efficacité (plutôt que de tout réinventer à chaque besoin de *data lineage*).

Les « *data designers* » ont en charge :

- la réalisation des *data lineages ;*
- la relecture croisée des *data lineages* afin d'en garantir la bonne représentation (respect de la norme de représentation, homogénéité de représentation) et la bonne lisibilité ;
- la capitalisation des produits au travers d'un portefeuille de données (les « données utiles ») qui ont fait l'objet d'un *data lineage,* et sur lesquelles l'équipe de *data designers* peut être interrogée à des fins de réutilisation (j'ai un problème sur telle donnée, avez-vous le *data lineage* correspondant *?*) ;
- la relation avec les projets, conseil aux projets et suivi des projets sur l'angle traitements des données.

Cette relation permet d'initier un cercle vertueux avec l'exercice de représentation des *data lineages* (conformité des projets à un dictionnaire de données et ce même dictionnaire référent pour la représentation des *data lineages*, retours d'information sur ce qui est mis en œuvre par les projets en termes de traitement sur les données et représenté dans les *data lineages*, partage des retours d'expérience et expertise sur certains traitements sur les

[20] Ces solutions intègrent souvent une dimension collaborative (exercice de reconstitution d'un data lineage par « production participative »).

[21] Ces solutions relèvent de la famille des *metadata systems*.

[22] Ces solutions disposent de moyens de représentation des différentes étapes de traitements qui ont conduit à l'alimentation de ces systèmes (de la source à l'intégration dans un entrepôt, une plate-forme Big Data).

[23] Ces solutions aboutissent, au bout d'un moment, au risque d'être confronté à un référentiel de data lineage « fantôme » (on sait que cela a été fait, on ne sait plus où est la bonne version, ni si elle est à jour ; les personnes qui géraient cela sont parties...).

données (par exemple, distinction entre *default value* et forçage de valeur, règles de normalisation des noms de données…).

7. Approfondissements

7.1 Correspondances avec d'autres référentiels

Il existe peu de cadres pratiques explicitant une démarche de *data lineage* (et aucun avec une approche d'ensemble, dans une dimension d'architecture d'entreprise).

Le plus approchant en étant orienté data management – dans le cadre de l'initiative DMBOK (Data Management Body Of Knowledge) – La version 2[24] intègre un chapitre « - Chapitre 12 Méta-data management 12.4 Techniques 12.4.1 *Data lineage and Impact Analysis* ».

Cela couvre :

- Relation entre *data lineage* et *data flow* (nota a priori les deux termes sont interchangeables) ;
- Exigences relatives au format et à l'outillage de la documentation relative au *data lineage* (nota un lien avec les processus métier et éléments d'organisation-rôles est fait. On distingue deux niveaux de représentation élevé et détaillé) ;
- Relation entre le *data lineage* et le cycle de vie des données (nota : les données ont non seulement un cycle de vie, mais également un *data lineage* (c'est-à-dire un chemin dans lequel elles se déplacent de leur point d'origine à leur point d'utilisation, parfois appelée « chaîne de données ») ;
- Concept de *data lineage* dans différents domaines de connaissances de la gestion des données (nota : un cas exposé – pour traiter les problématiques de qualité des données, avec l'accent mis sur les méta-données, les exigences sur les modèles – architectures - architecture de données, modélisation et conception de données, intégration et interopérabilité des données, données de référence, implémentation de type DW & BI…).

La BCE dans le cadre de BCBS 239 propose des guides (réglementaire[25]) spécifiant son attente en termes de restitution du *data lineage* des données qu'elle cherche à contrôler (indicateurs Bâlois).

Les solutions de data gouvernance incluent des modules de *data lineage*. Les éditeurs de ces solutions proposent pour certain des guides de production des *data lineages* en ligne avec les fonctions de leur solution.

7.2 Autres démarches proches

Il existe des démarches proches voire synonymes :

a. *Data supply chain*

Contexte : la production de données est la vocation du processus métier. Exemple pour un data lab cela peut correspondre à la phase de préparation de données avant leur exploitation par des algorithmes de data science. Autre exemple pour une place de marché de la donnée, cela correspond à la gestion du cycle de vie des données de leur identification à leur « commercialisation » par la place de marché.

La donnée devient une matière, un produit fini et distribué comme dans une *supply chain* traditionnelle (chaîne d'approvisionnement).

On doit maîtriser cette chaîne d'approvisionnement et cela commence par la décrire.

Cette description est équivalente à la production d'un *data lineage*. La différence porte essentiellement sur le fait que les étapes d'une *supply chain* sont caractéristiques (la chaîne d'approvisionnement de données se compose de trois parties. D'abord, du côté de l'offre, les données sont créées, capturées et collectées. Ensuite, les données sont

[24] La version 1 ne proposait rien.

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

enrichies, contrôlées et améliorées. Enfin, du côté de la demande, les données sont utilisées, consommées et exploitées).

b. Data biography

Certains acteurs parlent de biographie des données[26].

L'approche est complémentaire et va dans le sens des questions que l'on peut se poser dans le cadre d'une activité et des objectifs de data science. Elle met le focus sur la production des *data sets* utilisés par les *data scientists*.

L'approche distingue des questions amont qui rejoignent une approche de type *data lineage* :

- Comment les données ont été collectée ? Dans quel but ?
- Le processus de collecte a-t-il évolué entre deux campagnes de collecte (recherche des variations de collecte) ?
- Quelle représentativité des données collectées, de l'échantillon obtenu ?
- D'où viennent les données, d'une autorité, de l'agrégation de différentes sources ? Qui a financé la collecte ?
- Comment les données ont été nettoyée ? Qu'est ce qui a été adopté pour les valeurs aberrantes (supprimées, conservées) ?

On recherche principalement à caractériser les *data sets* obtenus en termes de qualité et de biais possibles (ne pas prendre les données utilisées pour argent comptant).

Et des questions d'intention : Quel but poursuivi : prouver un modèle d'analyse, interpréter des comportements pour en déduire un modèle ?

L'objectif final est de conserver l'histoire (biographie) des *data sets* utilisés par les *data scientists* (autrement dit le *data lineage* de production des *data sets*).

Plus largement cette approche est familière aux statisticiens, qui associent à leur publication la méthodologie statistique utilisée (à l'instar de l'Insee : https://www.insee.fr/fr/information/2838097).

Cela rejoint également, les efforts de formalisation des étapes de préparation des données pour les travaux de *data science* et leurs déclinaisons dans des solutions de type « *data preparation tool* », dans lesquels on retrouve la capacité de manipuler et représenter des formes de *data lineage* des *data sets* utilisés.

De façon plus anecdotique, on trouve des démarches de type « *data genealogy* » (Quelles sont les données mères de telles données ? Quelles données ont été générées à partir de telles données ?).

7.3 Ouvertures

La traçabilité est au cœur de ce procédé. Elle devrait être une exigence initiale à la construction des systèmes.

Le contenu de ce procédé peut être, ainsi, vu comme un premier guide vers :

- la reconception d'une chaîne de traitement : profiter de l'exercice de *data lineage* pour revenir à une démarche d'Architecture d'Entreprise, remonter aux aspects amont (logique, pragmatique et sémantique) que propose la méthode en exploitant le niveau logistique, pour optimiser localement et de bout en bout la chaîne de traitement ;
- la conception de nouvelles chaînes de traitement dans une logique « *trace by design* », où les exigences de traçabilité sont partie intégrante des exigences générales et où des moyens explicites de gestion des traces sont mis en œuvre ;

Il existe des cas de construction de systèmes qui font appel à une logique « Métadata System Centric » – où le cœur du système repose sur un référentiel des méta-données. À l'exemple d'organisations dont la donnée est le métier – market place data / broker data, instituts de statistiques et qui bâtissent leur S.I. à partir d'une vision des méta-données (la donnée étant le produit qu'ils vendent). La vision des *data lineages* est alors automatique et au cœur du *business model* (cela rejoint aussi l'idée de *supply chain* de la donnée vu précédemment). Cela permet

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

de gérer des dispositifs de type « *automatic change data management* », c'est-à-dire la capacité à détecter les changements sur la construction – évolution de modèles, voire les changements de traitements sur les données.

Avec l'idée de repenser les systèmes d'information dans une logique *data centric*, l'obligation de penser « *trace by design* » est indispensable.

7.4 Bibliographie pratique

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

DAMA – Data Management Association : Volet *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 des illustrations

Table des matières analytique