ESCP BUSINESS SCHOOL

IT ALL STARTS HERE BERLIN | LONDON | MADRID | PARIS | TURIN | WARSAW

ECOLE DOCTORALE DE MANAGEMENT PANTHÉON-SORBONNE n°559

Exploration du processus de construction des données RH – Qualification, Capitalisation et Requalification

Thèse de Doctorat présentée en vue de l'obtention
du grade de docteur en sciences de gestion et du Management
par

Catherine LESPERANCE

dirigée par

Géraldine GALINDO, Professeure – ESCP Business School

Soutenance le 20 décembre 2024

Devant un jury composé de :

Rapporteurs :	Clotilde CORON Professeure des universités - Université Paris-Saclay
	Ewan OIRY Professeur des universités - Université de Poitiers
Suffragants :	Michel COSSETTE Professeur agrégé - HEC Montréal
	Roxana OLOGEANU-TADDEI Professeure associée - TBS Education
	Laurent TASKIN Professeur ordinaire - Université catholique de Louvain

L’Université n’entend donner aucune approbation ou improbation aux opinions émises dans les thèses. Ces opinions doivent être considérées comme propres à leurs auteurs.

Remerciements

Cette thèse est le fruit d'un effort collectif. D'un intérêt balbutiant à Montréal à un engagement pleinement assumé en France, je tiens à remercier chaleureusement toutes celles et ceux que j'ai eu le privilège de côtoyer. Merci pour votre contribution à ce parcours.

Tout d'abord, je tiens à exprimer ma profonde gratitude envers ma directrice de thèse, Géraldine Galindo. Géraldine, merci infiniment pour la confiance que tu m'as accordée dès nos premières rencontres. Ta présence et ton soutien ont été une boussole, guidant mes pas avec assurance à travers les périodes de doute, les moments d'euphorie et mon désir insatiable de tout lire. Tu as su canaliser mon énergie et m'orienter habilement jusqu'à l'achèvement de cette thèse. Je te suis extrêmement reconnaissante pour ta bienveillance et tes conseils toujours si précieux. Enfin, je tiens à te remercier tout particulièrement pour le modèle féminin que tu représentes pour moi et pour notre communauté académique en GRH. Voir comment tu gères d'une main de maître, les publications, l'encadrement de doctorants, l'enseignement et la vie familiale m'inspire et me motive à exceller dans tous ces domaines. J'espère de tout cœur que nos échanges et notre collaboration se poursuivront bien au-delà de cette thèse.

Je souhaite exprimer ma sincère reconnaissance à tous les membres de mon jury. Je vous remercie profondément pour votre disponibilité à évaluer ce travail, auquel vous avez tant apporté. Un merci particulier à Clotilde Coron et Ewan Oiry pour vos retours constructifs et votre bienveillance lors de la pré-soutenance. Clotilde, vos recherches ont été le catalyseur de ma curiosité pour les données quantifiées en GRH. Vous êtes une grande source d'inspiration et je suis honorée que vous ayez accepté d'être ma rapporteure. Ewan, tes conseils et ton dynamisme inépuisable ont été des phares dans mes moments de doute, me permettant de me surpasser et de gagner en confiance. Je te suis également très reconnaissante pour ce séjour de recherche à Montréal qui a renoué et renforcé mon affection pour cette ville. Merci à Michel Cossette pour nos

conversations enrichissantes au fil des années, pour ton soutien et ton intérêt pour mes travaux. Chaque retour à Montréal est un moment que j'attends avec impatience, grâce à la perspective de nos échanges. Un merci tout aussi spécial à Roxana Ologeanu-Taddei. Vous avez confirmé la possibilité d'une belle synergie entre les SI et la GRH, et pour cela, je vous suis infiniment reconnaissante. Merci également à Laurent Taskin d'avoir accepté d'être suffragant. Je suis impatiente à l'idée de nos futurs échanges. Je vous remercie toutes et tous sincèrement d'avoir accepté de faire partie de ce jury.

Je tiens à remercier Claire Dambrin pour la confiance qu'elle m'a témoignée en m'accueillant au sein du programme doctoral de l'ESCP. Claire, je te suis particulièrement reconnaissante pour tes encouragements à retourner vers Bruno Latour, l'auteur qui m'a fait tomber amoureuse de la recherche pour la première fois. Un grand merci également à Régis Coeurderoy, qui a ensuite pris la relève et m'a aidée à me remettre sur les rails lorsque je ne m'en sentais plus capable. Merci aussi à Valentina Carbone, aujourd'hui directrice du programme, qui continue d'enrichir cet environnement si stimulant. Je suis convaincue que c'est grâce à votre dévouement que l'atmosphère au sein de ce programme est si dynamique et chaleureuse, contribuant à forger une communauté soudée. Enfin, et sans conteste, la pierre angulaire du programme : Christine Rocque. Ta bonne humeur, ton esprit festif, ta porte toujours ouverte et ton soutien sans faille sont inestimables. Pour tout cela, et bien plus encore, merci.

Je souhaite exprimer ma reconnaissance envers le département de Management de l'ESCP et tous les professeurs pour leur encadrement et leur soutien constants. Un remerciement particulier est adressé aux professeures de GRH – Almudena Cañibano, Emmanuelle Léon et Maral Muratbekova - pour votre passion contagieuse pour la recherche en GRH. Merci également à mes camarades doctorants, Arthur, Bianca, Chris, Domenico, Janice, Katya, Mariann, Olivier S., Sara, Sofia, Yaëlle, Trang, et à tous les autres. A bientôt, je l'espère.

Un merci tout spécial à mes équipes de co-auteur.es : Maral, Anna, Diana, Matilde, Sophie et Tristan, merci pour cette formidable première expérience de rédaction collective sur un sujet aussi passionnant ! Nous l'avons fait ! Je tiens également à remercier Sophie R. pour sa façon si singulière de voir le monde.

Cette thèse n'aurait évidemment jamais vu le jour sans la richesse du terrain auquel j'ai eu la chance d'accéder. Merci de tout cœur aux *data scientists* : Artus, Augustin, Guillaume, Julien, Kevin, Paul, Simon, Sophie, Stefan, Thomas, Tiphaine et à tous les autres avec qui j'ai eu le plaisir de collaborer. Un merci tout particulier à Anna et Rami, qui m'ont ouvert les portes de leur organisation. Vous avez pris le temps de me guider dans la découverte de votre métier et je vous en suis profondément reconnaissante. Votre bienveillance et votre passion ont été une source de motivation inestimable pour mener à bien cette thèse.

Mes remerciements vont bien sûr à mes ami.es doctorant.es et docteur.es : Éléonore, Sophie et Valentin. Les mots me manquent pour exprimer tout ce que vous m'avez apporté. À Tristan, merci pour nos discussions passionnées et ton soutien indéfectible. J'espère te retrouver très bientôt à Montréal ! Un immense merci au groupe des « livreuses » : Claire, Éléonore, Justine, Sophie et Sophie G. Ce beau prétexte pour nous retrouver et discuter de tout, sauf du livre. Vous avez illuminé ce parcours.

Merci à tous mes ami.es en dehors du domaine académique pour votre soutien constant. Je vous en suis profondément reconnaissante. Un merci particulier à Adriana, Aref, Caroline, Gab, Gabe, Katheline, Marine et Mitra pour les parenthèses que vous m'avez offertes.

Merci à mes familles Amar - De Wulf - Lespérance. Je me sens incroyablement chanceuse de vous avoir. Si ce parcours a parfois été ardu et que sa durée a mis ma patience à l'épreuve, votre affection et vos encouragements ont fait toute la différence. Merci du fond du cœur.

Merci à toi, Luc, mon amour, mon roc. Tu m'as soutenue tout au long de cette épreuve sans jamais laisser le poids de cette thèse peser sur nous. Aucun mot ne pourrait exprimer tout ce que tu représentes pour moi. Tant de choses se sont passées - la France a été notre premier grand projet commun. D'autres projets nous attendent maintenant.

Les remerciements de cette thèse sur l'ANT ne seraient pas complets sans souligner l'apport discret mais essentiel de la technologie : un merci spécial à mon casque à

réduction de bruit Sony WH, fidèle compagnon du quotidien, qui a été indispensable à ma concentration tout au long de ces années.

Table des matières

Index des figures

Index des tableaux

Introduction

Le mythe de la donnée « naturelle » (Power, 2023, p. vii) [1] perpétue l'idée selon laquelle les données existent à l'état brut, indépendamment de toute intervention humaine. Il est renforcé par l'étymologie même du terme « donnée », comme le rappelle Jensen (1950 : ix, cité par Kitchin, 2022d, p. 5) :

« *C'est un accident malheureux de l'histoire que le terme « datum »* [donné] *plutôt que « captum »* [capturé] *soit venu symboliser l'unité-phénomène en science. Car la science ne traite pas de « ce qui a été donné » par la nature au scientifique, mais de « ce qui a été pris » ou sélectionné de la nature par le scientifique en fonction de son objectif*. ».

La définition de ce qui constitue une donnée est influencée non seulement par des conventions établies, mais aussi par des pratiques - calculatoires ou non - qui façonnent et segmentent le monde en éléments susceptibles d'être comparés, agrégés et analysés (Alaimo & Kallinikos, 2024 ; Passi & Jackson, 2018 ; Passi & Sengers, 2020). L'expression « données brutes » peut être considérée comme un oxymore ; les données sont toujours « cuisinées » et donc jamais totalement « brutes » (Gitelman & Jackson, 2013, p. 2). Ainsi, bien loin de l'idée reçue des données « naturelles » (Power, 2023, p. vii) ; celles-ci sont le fruit d'une démarche humaine et organisée (Alaimo & Kallinikos, 2024; Kitchin, 2022d).

Historiquement, la production, l'analyse et l'interprétation des données représentaient des processus coûteux et chronophages. Elles fournissaient le plus souvent des informations statiques et partielles. En raison de leur rareté, les données de bonne qualité étaient donc perçues comme une marchandise précieuse, jalousement protégée ou échangée à prix élevé (Kitchin, 2022d).

[1] J'utilise l'expression « mythe de la donnée naturelle » pour traduire le terme anglais « *myth of the given* », que l'auteur emploie pour remettre en question l'objectivité des données.

Ces dernières années ont été marquées par des changements radicaux. Si la valeur des données reste, les transformations induites par la « révolution numérique » ont profondément modifié leur mode de production et de valorisation (ibid.).

1. La révolution numérique et l'avènement des organisations *data-driven*

L'importance de saisir la nature « cuisinée » des données est intrinsèquement liée à la révolution numérique. En effet, la numérisation croissante de nos vies crée de nouvelles dynamiques où les données ne se contentent pas de refléter la « réalité », mais participent activement à sa construction. Elles redéfinissent ainsi les limites entre les sphères privée et professionnelle, modifient nos manières de travailler, de communiquer et de consommer (Alaimo & Kallinikos, 2022, 2024).

La révolution numérique est souvent comparée à l'adoption de l'imprimerie, soutenant l'idée que le numérique pourrait, lui aussi, constituer une culture (Cardon, 2019). Cette analogie repose sur des racines historiques similaires, à la fois profondes et diversifiées. Comme le démontre Eisenstein (1980), l'adoption de l'imprimerie a favorisé l'émergence de la Réforme protestante, du libre arbitre et du développement des marchés, entraînant des bouleversements intellectuels, religieux, sociaux, économiques et politiques.

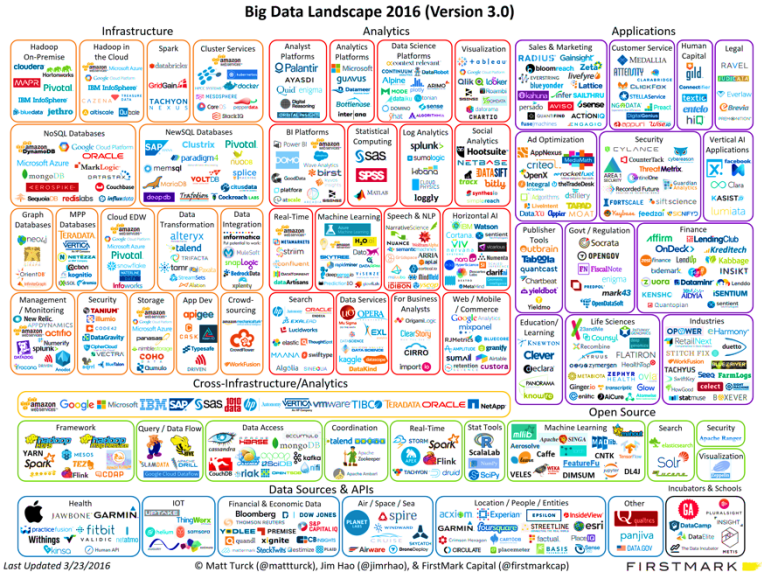
De manière comparable, la culture des données transforme aujourd'hui l'ensemble de nos systèmes (van Dijck, 2014). Alaimo & Kallinikos (2017) illustrent cette transformation en examinant la multifonctionnalité des « *likes* » sur les réseaux sociaux. Un *like* sur ces plateformes agit non seulement comme une unité de mesure, mais également comme un moyen de communiquer une approbation ou un accord, souvent interprété comme un signe de préférence et de construction identitaire.

Dans ce contexte, chaque *like* acquiert une nouvelle dimension : il sert non seulement de baromètre pour mesurer l'engagement des utilisateurs, mais devient aussi un agrégat de données essentiel pour les algorithmes qui modélisent les comportements de ces mêmes utilisateurs, également considérés comme des consommateurs. Chaque *like* est donc une contribution à une base de données, influençant la diffusion de contenu, les

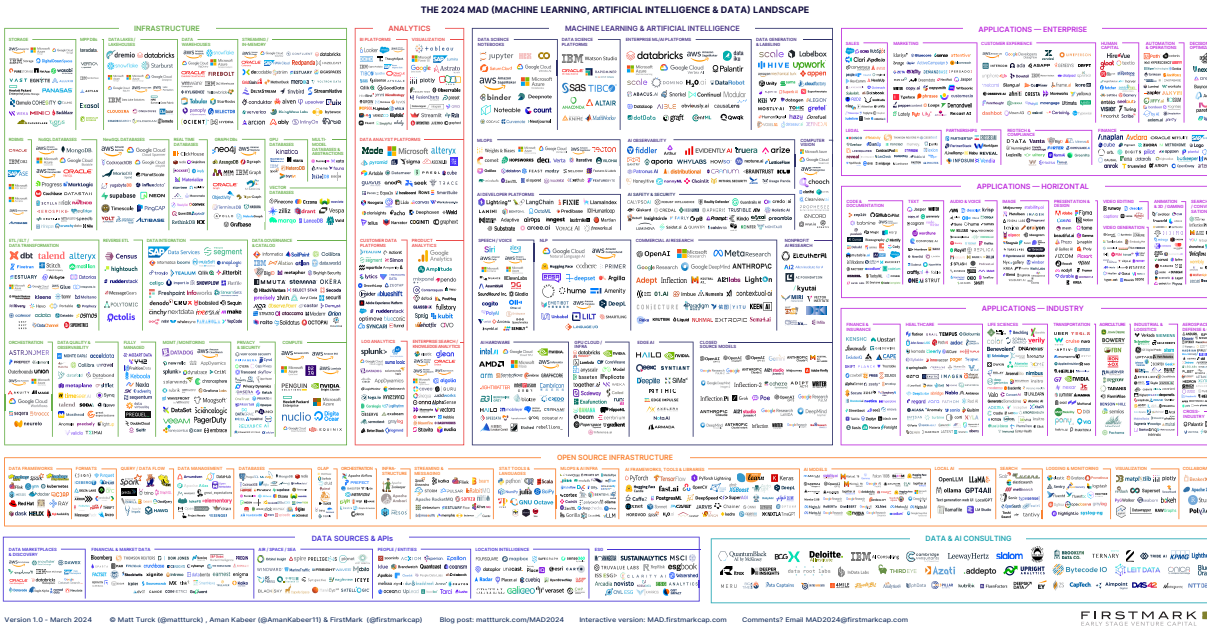
stratégies marketing et la dynamique commerciale des plateformes. Cet exemple montre ainsi comment une action apparemment simple comme un *like* peut avoir de profondes répercussions. Il influence les modèles économiques des plateformes et modifie la manière dont les contenus et les préférences sont évalués et valorisés dans l'espace numérique.

Plutôt que d'être rares et en accès limité, les données sont désormais produites en abondance et à moindre coût, devenant de plus en plus ouvertes et accessibles. En conséquence, ces dernières se transforment en une ressource clé dans le monde moderne, une marchandise importante activement achetée et vendue sur le marché mondial, qui se chiffre en plusieurs milliards de dollars (Kitchin, 2022d).

La Figure 1 ci-dessous compare le paysage des données entre 2016 et 2024. Elle révèle non seulement une augmentation quantifiable des technologies numériques disponibles, mais aussi une sophistication croissante de ces dernières, stimulée par l'accumulation des données. Cela témoigne d'une spécialisation accrue dans leur gestion et leur analyse, notamment par le biais des techniques d'apprentissage automatique et d'intelligence artificielle (IA). Cette évolution montre ainsi que le paysage des données, autrefois considéré comme un secteur technique de niche, est devenu un pilier essentiel et omniprésent de notre monde moderne.



a. Le paysage des données en 2016



b. Le paysage des données en 2024

Figure 1 : Évolution du paysage des données entre 2016 et 2024 (Turck, 2024)

Cette évolution souligne également l'importance croissante des données en tant que catalyseurs de la dynamique économique. Leur valeur est telle que sans elles, les technologies numériques et les modèles d'affaires de nombreuses entreprises, en particulier celles « *digital natives* » - telles que les plateformes ou certains cabinets de conseil spécialisés -, seraient incapables de fonctionner ou de générer un chiffre

d'affaires (Sadowski, 2019). En tant qu'actifs et marchandises achetés et vendus de manière intensive (Alaimo & Kallinikos, 2022; Sadowski, 2019), les données occupent un rôle central dans la médiation de la production et de la circulation du capital (Dalton et al., 2016). Ce rôle constitue la force motrice derrière l'essor rapide des organisations qualifiées de « *data-driven* ».

Les organisations *data-driven* produisent, exploitent et interprètent les données pour justifier leurs prises de décision, piloter leurs processus et stimuler l'innovation (Kitchin & Lauriault, 2018). En outre, les données ne se contentent pas de soutenir les opérations ; elles transforment également de manière fondamentale ces entreprises. Elles redéfinissent les types d'activités économiques qu'elles entreprennent (Sadowski, 2019), tout en modifiant profondément le travail et les expériences professionnelles de leurs salariés (Woodcock, 2021). Avec l'adoption croissante d'une culture orientée vers les données, il devient également impératif de s'interroger sur le rôle de la gestion des ressources humaines (GRH) au sein de ces organisations.

2. Des organisations à une GRH *data-driven*

Avec l'expansion de la culture *data-driven*, la GRH ne fait pas exception à la règle : ses pratiques s'alignent progressivement vers une exploitation toujours plus intensive de ses données.

Selon un rapport de Kowu (2024) pour *Markess by Exaegis*, le marché numérique RH se divise en deux grandes catégories : (1) les logiciels qui incluent les technologies *SaaS* (logiciels hébergés sur le cloud) et *On-premise* (logiciels installés localement) ; et (2) les services numériques, qui se répartissent entre le conseil et l'*outsourcing*. Ces deux catégories se concentrent autant sur des activités RH variées telles que le recrutement, la gestion des talents et la gestion administrative, que sur des phénomènes RH tels que la gestion de l'absentéisme.

Les prévisions du rapport indiquent que le marché RH français dépassera les 4 milliards d'euros en 2023, avec une croissance projetée jusqu'en 2027, sans signe d'essoufflement. La Figure 2 illustre cette croissance avec une augmentation annuelle régulière de +11,6% parmi les quatre segments du marché. Notamment, la part des solutions *SaaS* connaît une croissance particulièrement significative, révélant un engouement pour les solutions *cloud*, réputées pour leur flexibilité et leur scalabilité.

Parallèlement, le segment de l'*outsourcing* affiche également une progression notable, ce qui reflète une tendance à la sous-traitance de certaines activités de la fonction RH. Si cette tendance se maintient, le marché est en bonne voie pour presque doubler de taille, atteignant ainsi 6,7 milliards d'euros d'ici 2027, soulignant l'importance croissante de la GRH *data-driven*.

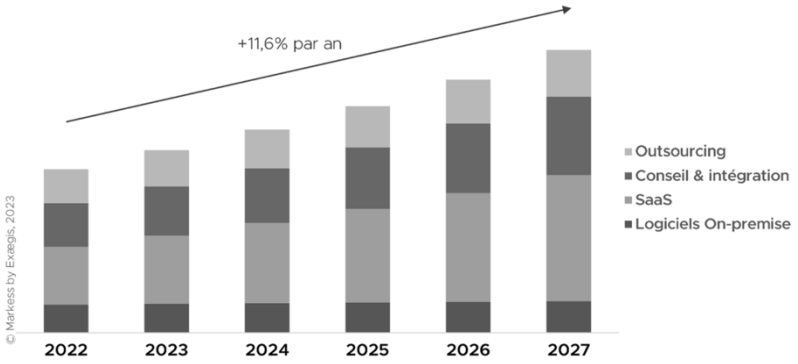


Figure 2 : Prévisions des quatre segments du marché numérique RH entre 2022-2027 (Kowu, 2024)

L'expansion du marché RH - en cohérence avec l'évolution globale du marché numérique - démontre que les principaux leviers des technologies RH ne sont pas uniquement techniques, mais profondément sociaux et économiques. Comme l'illustre la Figure 2, les données RH sont produites pour soutenir le développement de nouvelles technologies et de nouveaux services, afin d'assurer une croissance continue des profits qu'ils génèrent (Ruppert et al., 2017). Compte tenu de l'interdépendance entre les données RH et les technologies numériques qui les exploitent (Borgman, 2017a), il est nécessaire de s'interroger sur les types de données perçus comme ayant une valeur actuelle ou un potentiel de valorisation future pour la GRH (Alaimo et al., 2020).

3. La valorisation par la construction des données RH

Un corpus important de travaux en GRH s'intéresse aux données. Ces recherches couvrent des thématiques allant de la définition d'indicateurs jusqu'à l'adoption de méthodes analytiques avancées, telles que l'analytique RH ou bien l'IA (Angrave et al., 2016; Davenport et al., 2010; Madsen & Slåtten, 2019; Rasmussen & Ulrich, 2015).

Toutefois, bien que la sélection des données soit un sujet fréquemment débattu dans l'analyse des activités et phénomènes RH, ce thème demeure largement sous-représenté dans la littérature. La GRH, et plus particulièrement la e-GRH et l'instrumentation de GRH (Coron, 2022), tendent à privilégier les effets de l'exploitation des données RH déjà intégrées aux technologies numériques (Tambe et al., 2019). Cette sous-représentation est d'autant plus préoccupante que, dans le contexte empirique, environ 80 % du temps de travail d'un *data scientist* est consacré à cette étape[2].

En tenant compte des dynamiques engendrées par la révolution numérique et de l'importance croissante des données dans la GRH, cette thèse se concentre sur une question centrale :

« ***Comment les données RH sont-elles construites dans un contexte de marchandisation ?*** »

Cette question vise à révéler la complexité des données RH en tant qu'objets socio-technico-économiques. Elle met en lumière les tensions entre l'objectivité et la subjectivité de ces données, façonnées par les interventions humaines et les dynamiques de valorisation économique.

L'approche adoptée pour explorer le processus de construction des données RH et répondre à la question de recherche se traduit par la formulation de quatre objectifs relatifs à autant de thématiques :

- Les « données RH » : examiner l'écosystème sous-jacent à ce que la littérature associe aux « données RH ».
- La « marchandisation » : investiguer le contexte empirique de cette recherche en analysant l'offre de construction des données RH en relation avec la demande du marché.
- Le « comment » : cartographier le processus de construction des données RH et cerner les mécanismes sous-jacents à son développement.

[2] Notes issues du journal de bord (18/07/2023).

- La « construction » : approfondir les étapes constitutives du processus de construction des données RH afin d'identifier les éléments qui permettent de les définir.

4. Structure générale de la thèse

Partie I : Perspectives théoriques et méthodologiques

Cette première partie est consacrée à la conceptualisation des données RH à travers une exploration des littératures issues des systèmes d'information (SI) et de la GRH. Cette revue de littérature vise à définir et analyser l'écosystème sous-jacent, essentiel pour appréhender la nature des données RH ainsi que les contextes et infrastructures dans lesquels elles s'inscrivent.

Un cadre théorique est ensuite développé, combinant la production de connaissances (Latour, 2005a, 2007) et l'économie des qualités (Callon et al., 2000, 2002). Cette synergie conceptuelle, désignée sous le terme de processus *QCR* - Qualification, Capitalisation, Requalification, décrit la manière dont les données RH sont construites pour acquérir le statut de biens économiques. Elle s'articule autour de trois séquences clés :

1. La Qualification : concerne l'évaluation des qualités des données RH. Il s'agit de déterminer les critères initiaux qui définissent leur singularité potentielle.
2. La Capitalisation : examine les projets qui s'appuient sur la qualification pour développer des connaissances. L'objectif est de concevoir la fonction épistémique des données RH, permettant leur singularisation en tant que biens économiques.
3. La Requalification : consiste à requalifier les données RH à partir des projets de capitalisation.

La méthodologie de cette étude, fondée sur une approche qualitative et séquentielle, est également présentée. Elle s'inscrit dans un continuum entre recherche-action et ethnographie. Elle est rendue possible par une immersion de trois ans et sept mois au sein d'un cabinet de conseil en *data science*, dans le cadre d'une convention *CIFRE*. Le processus de construction des données RH est analysé à travers la conception de *DSN Analytics*, un outil d'IA destiné à l'analyse de l'absentéisme.

Partie II : Le processus de construction des données RH

Cette deuxième partie est dédiée à l'application du processus de construction des données RH, tel que théorisé dans la première partie. Ce processus est structuré autour des trois séquences clés préalablement définies, chacune constituant un chapitre de résultats distinct :

1. La Qualification : qui révèle la qualité *normative* des données RH.
2. La Capitalisation : qui met en évidence trois projets de connaissances distincts, chacun caractérisé par un type spécifique de connaissances sur l'absentéisme :
    1. L'absentéisme *réel univarié* : connaissances *descriptives* ;
    2. L'absentéisme *réel multivarié* : connaissances *explicatives* ;
    3. L'absentéisme *latent* : connaissances *prédictives*.
3. La Requalification : qui met en lumière la nouvelle qualité *extensive* des données RH, se manifestant sous deux formes : *externe* et *interne*.

Partie III : Discussion et conclusion

Enfin, dans cette troisième partie, une réflexion est engagée sur les résultats obtenus, en abordant les contributions théoriques et empiriques de ma recherche, ainsi que ses limites et perspectives.

Sur la base des résultats et, plus généralement, de l'ambition de cette thèse de déconstruire le mythe des données « naturelles », trois contributions théoriques sont mises en lumière :

1. La conceptualisation des données RH en tant qu'objets d'étude en GRH ;
2. La conceptualisation des dispositifs socio-numériques des données RH à travers la spirale *data-driven* ;
3. La théorisation du processus de construction des données RH articulée en trois séquences (*QCR*).

Trois contributions empiriques sont ensuite soulignées, offrant des perspectives variées sur le processus de construction des données RH :

1. Pour la gestion de projet : en identifiant les facteurs limitants et les facteurs de succès du processus de construction des données RH ;
2. Pour les *data scientists* : en renforçant la sensibilisation et la formation en GRH afin de garantir une contextualisation adéquate des données RH ;

3. Pour la fonction RH : en garantissant une sensibilisation et une formation en *data science* pour assurer un contrôle adéquat sur la construction des données RH.

Afin de bien contextualiser l'ensemble de ces contributions, j'identifie également les limites liées à cette recherche. Elles sont classées en deux grandes catégories : la première limite est de nature théorique, résultant de la dilution du pouvoir entre les agents économiques impliqués dans le réseau de construction des données RH. La seconde est empirique et concerne l'influence de mon implication active dans l'étude.

Enfin, deux perspectives de recherche sont envisagées, réparties en deux catégories principales : la première, théorique, interroge la qualification des « bonnes » données RH, tandis que la seconde, empirique, examine la fétichisation des données RH à travers leur accumulation.

En conclusion, cette thèse aspire à démontrer que les données RH ne sont pas simplement « données ». Loin d'être des entités brutes, elles émergent de pratiques - calculatoires ou non - qui sont à la fois discrétionnaires et situées. Positionnées à la croisée des dimensions sociales, techniques et économiques, elles révèlent des tensions entre objectivité et subjectivité, façonnées par des logiques de marchandisation. Ainsi, cette étude cherche à enrichir la définition des données RH, en mettant en lumière les enjeux qui sous-tendent leur construction et, plus largement, leur portée en GRH.

La Figure 3 ci-dessous offre une vue d'ensemble de la structure générale de la thèse et illustre son articulation avec les différents composants de la problématique.

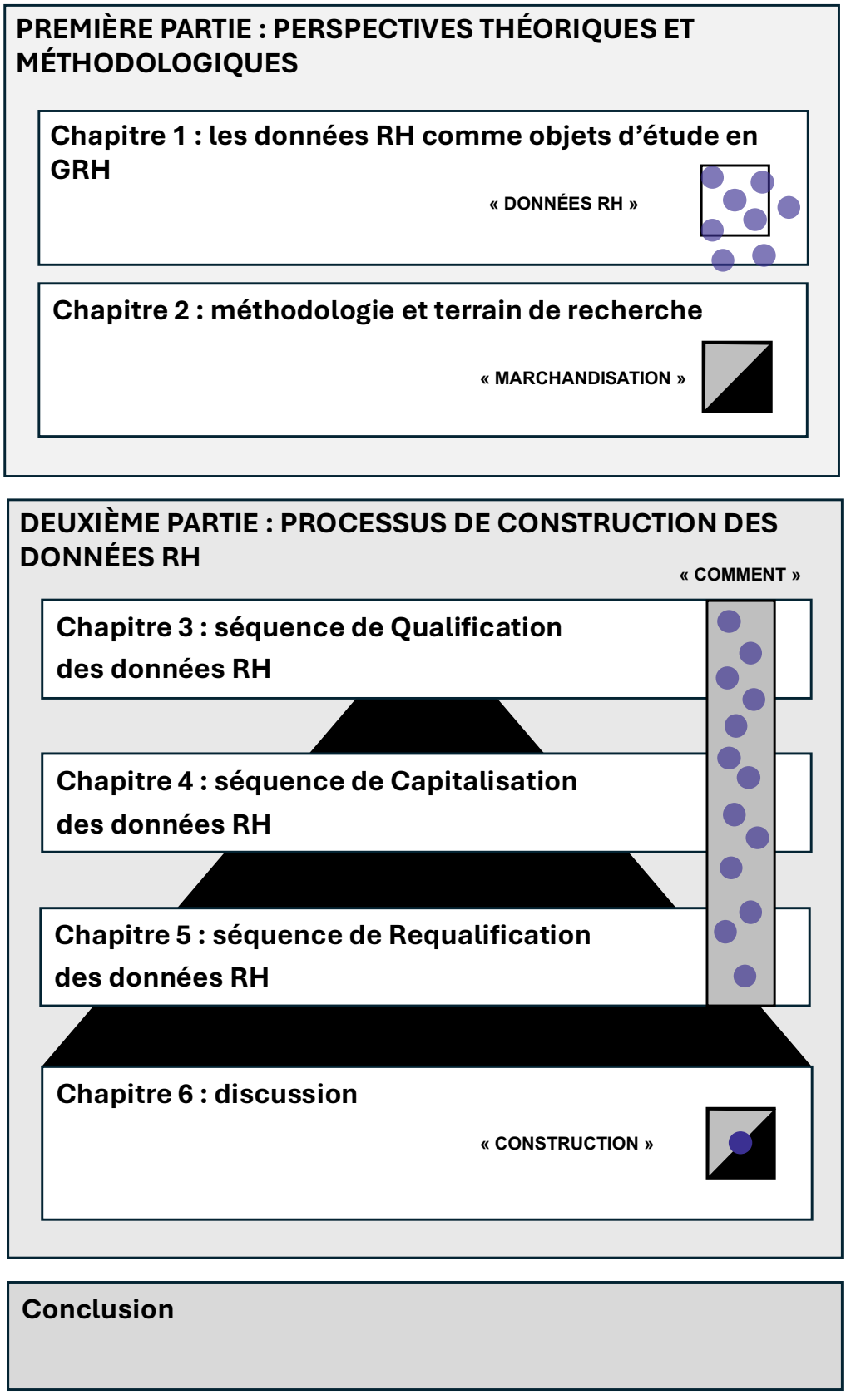


Figure 3 : Structure générale de la thèse

Partie I. PERSPECTIVES THEORIQUES ET METHODOLOGIQUES

Chapitre 1. Les données RH comme objets d'étude en GRH

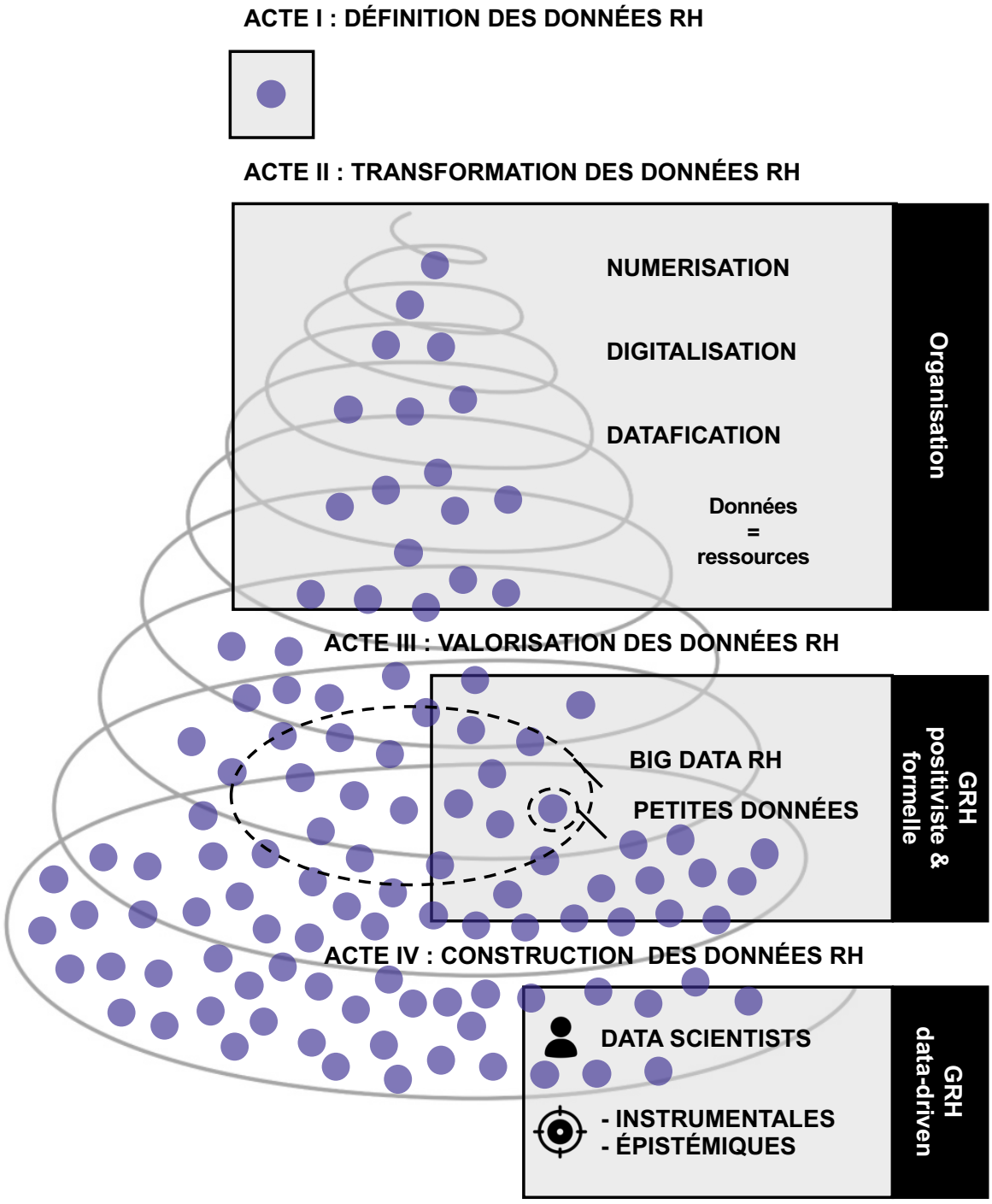


Figure 4 : Conceptualisation des données RH à travers quatre grands actes

1. Introduction

L'histoire des organisations est intrinsèquement liée à celle des données, qui servent non seulement d'enregistrements mais aussi de fondements aux systèmes chargés de représenter, suivre et réguler les activités organisationnelles (Alaimo & Kallinikos, 2022). Elles fournissent la preuve de la génération de connaissances utilisée par les organisations pour se comprendre et pour comprendre leur environnement (Kitchin, 2022a). Toutes les fonctions organisationnelles sont impactées : le contrôle de gestion, la finance, le marketing, les ressources humaines. Aucune exception à la règle. L'augmentation des capacités de calculs des ordinateurs et l'essor de la quantité de données disponibles et relativement peu coûteuses ne cessent de renforcer la puissance de cette pratique.

Face à ce contexte, ce premier chapitre structure l'exploration de la problématique de recherche en deux sections principales. La première section, centrée sur la conceptualisation des données RH, vise à définir et investiguer l'écosystème sous-jacent à ce que la littérature associe aux « données RH ». Cette étape est essentielle pour approfondir la compréhension de leur nature ainsi que les contextes et infrastructures dans lesquels elles opèrent.

La seconde section se rapporte, quant à elle, à la sélection d'un cadre théorique adapté pour analyser cette dynamique. Il s'agit de fournir les concepts nécessaires pour développer une approche à la fois structurée et ancrée dans un fondement théorique solide.

2. Quatre actes de conceptualisation des données RH : définition, transformation, valorisation et construction

Les organisations jouent un rôle crucial dans le développement des conventions, des techniques et des systèmes de notation qui définissent les types de données produites. Ces développements influencent directement les stratégies de gestion, démontrant l'importance des données dans la structuration des pratiques organisationnelles (Alaimo & Kallinikos, 2022; Chandler, 1990). En transcendant leur rôle d'actifs, les données deviennent les éléments centraux des stratégies de gestion organisationnelle (Kellogg et al., 2019; Leonardi & Treem, 2020). Cette centralité se

manifeste par l'adoption de démarches intégrées dans l'exploitation et la gouvernance des données, révélant ainsi leur potentiel transformateur (Alaimo & Kallinikos, 2022).

Cette première section se consacre donc à une exploration approfondie des fondations et dynamiques transformationnelles associées aux données RH. Elle est structurée en quatre actes distincts, chacun apportant une couche de complexité supplémentaire à la conceptualisation des données RH :

1. Définition des données RH ;
2. Transformation des données RH ;
3. Valorisation des données RH ;
4. Construction des données RH.

Chaque acte vise à décomposer et analyser les différentes facettes des données RH. Ensemble, ils contribuent à brosser un portrait exhaustif de leur écosystème, conceptualisé ici sous la forme de la spirale *data-driven* (voir Figure 4). Ainsi, cette section est conçue pour offrir une perspective holistique, illustrant comment les données RH influencent les paradigmes organisationnels contemporains et transforment la GRH.

2.1. Acte I : définition des données RH

En général, les données sont définies comme : « [...] *représentatives, capturant le monde sous forme de nombres, de caractères, de symboles, d'images, de sons, d'ondes électromagnétiques, de bits, etc. et détenant les préceptes d'être abstraites, discrètes, agrégées, non-variantes et significatives* » (Kitchin & Lauriault, 2018, p. 5).

D'un point de vue étymologique, le terme « donnée » trouve son origine dans le latin *dare*, qui se traduit par « donner ». En ce sens, les données sont envisagées comme des éléments bruts extraits des phénomènes mesurés et enregistrés selon diverses méthodes analytiques. Dans le langage courant, le terme « donnée » fait référence aux éléments qui sont « pris » ; extraits par le biais d'observations, de calculs, d'expériences et d'enregistrements (Kitchin, 2021b).

Techniquement, au regard de ces caractéristiques, le concept de données, couramment utilisé, correspondrait donc à des « capta » (du latin *capere*, signifiant « prendre ») ; il s'agit en effet de données qui sont choisies et collectées parmi

l'ensemble de toutes les données possibles. Comme le dit Jensen (1950 : ix, cité par (Kitchin, 2022d, p. 5) : « *C'est un accident malheureux de l'histoire que le terme « datum »* [...] *plutôt que « captum »* [...] *soit venu symboliser l'unité-phénomène en science. Car la science ne traite pas de « ce qui a été donné » par la nature au scientifique, mais de « ce qui a été pris » ou sélectionné de la nature par le scientifique en fonction de son objectif*. ».

En d'autres termes, les données sont envisagées comme une manifestation transparente du monde, se présentant sous diverses formes. Cette représentation se rapporte au statut de la « réalité » des activités et phénomènes RH, tel que retranscrit par les données, qui sont perçues comme le reflet neutre et objectif de cette « réalité » (Coron, 2019g; Desrosières, 2013b). Les expressions couramment employées pour décrire le traitement des données, telles que « collectées », « saisies », « compilées », « stockées », « traitées » et « exploitées », contribuent également à soutenir ces notions de neutralité et d'objectivité (Gitelman & Jackson, 2013; Kitchin & Lauriault, 2018).

Soumises uniquement à des contraintes techniques, les données sont donc perçues comme des mesures transmettant une vérité intrinsèque sur le monde, considérées comme exemptes de tout biais individuel et indépendantes des coutumes locales, de la culture, des connaissances et du contexte (Porter, 1995). Cependant, cette posture a largement été remise en question ces dernières années, notamment par les approches critiques sur les données (Alaimo & Kallinikos, 2022; Passi & Jackson, 2018; Passi & Sengers, 2020). Ces dernières avancent que les données sont intrinsèquement liées aux idées, techniques, technologies, personnes, systèmes et contextes qui les conçoivent, les produisent, les traitent, les gèrent et les analysent (Kitchin & Lauriault, 2018). La sociologie de la quantification s'attache notamment à déconstruire ce mythe en s'intéressant aux pratiques de quantification et en démontrant en quoi elles sont socialement construites (Coron, 2019d; Desrosières, 2013a; Diaz-Bone & Didier, 2016).

Les études sur la définition des données révèlent leur nature instable, mettant en évidence les incertitudes quant à ce qu'elles sont et comment elles sont produites. Cette instabilité crée également des doutes sur leur signification, rendant difficile la compréhension de ce que les données doivent représenter ou transmettre (Alaimo et

al., 2020). Parallèlement, Kitchin (2022d) définit les caractéristiques techniques qui influencent leur gestion et leurs applications potentielles :

1. Les formes et les structures des données ;
2. Les sources de données ;
3. Leurs producteurs de données ;
4. Les types de données.

Les définitions subséquentes de ces caractéristiques sont enrichies par des exemples spécifiques aux données RH.

2.1.1. Définition par les formes et structures des données

Les données sont généralement divisées en deux grandes catégories :

1. Les données quantitatives ;
2. Les données qualitatives (e.g. texte et images).

Plus spécifiquement, les données quantitatives se distinguent principalement en deux sous-catégories :

1. Les données numériques : qui sont mesurables et quantifiables (e.g. salaire mensuel, nombre d'heures travaillées).
2. Les données catégorielles : qui établissent des classifications distinctes pour attribuer, par exemple, les salariés à des catégories spécifiques (e.g. métier, entité géographique).

Les données quantitatives se divisent en quatre niveaux de mesure différents (voir Tableau 1) qui définissent la manière dont elles doivent être traitées et analysées numériquement. Ces niveaux déterminent les types de calculs et d'opérations statistiques applicables à ces données.

Tableau 1 : Niveaux de mesure des données RH

Niveaux de mesure	Caractéristiques	Exemples
Données nominales	Catégorisation sans ordre spécifique ou hiérarchie.	Salaire, nombre d'enfant, niveau d'éducation, etc.

Niveaux de mesure	Caractéristiques	Exemples
Données ordinales	Classement avec un ordre mais sans valeur numérique.	Statut dans l'entreprise, échelle de performance, etc.
Données d'intervalle	Mesure avec des valeurs numériques et un intervalle fixe, mais sans point de départ absolu.	Âge, ancienneté, nombre d'heures de formation, nombre de jours d'absence, etc.
Données de ratio	Mesure avec des valeurs numériques, un intervalle fixe et un point de départ absolu.	Ratio de rotation du personnel (nombre de départs / nombre total de salariés), ratio d'embauche (nombre d'embauches / nombre de candidatures), etc.

Les données peuvent ensuite être classées selon trois types de structures :

1. Les données structurées : se conforment à un modèle de données explicite qui définit la structuration, le stockage et la manipulation des données au sein des systèmes d'information (e.g. SIRH). En GRH, elles englobent des données telles que celles de la rémunération (salaire mensuel, primes) ou bien l'historique des évaluations de performance.
2. Les données semi-structurées : ne se conforment pas entièrement à un modèle de données prédéfini, bien qu'elles présentent une certaine organisation, elles ne possèdent pas une structure rigide. Parmi ces types de données, on retrouve, par exemple, les fiches de poste des différents métiers, les rapports d'entretiens d'évaluation et les documents de conformité réglementaire.
3. Les données non structurées : ne sont pas régies par un modèle de données préétabli et peuvent afficher une grande diversité. En GRH, elles comprennent une gamme variée d'éléments tels que les CV des candidats, les retours d'expérience des salariés lors de sondages, etc.

2.1.2. Définition par les sources de données

Les données RH peuvent être collectées selon deux méthodes principales.

1. La méthode directe : qui comprend des techniques telles que les observations ou les enquêtes ciblées, spécifiquement conçues pour générer des données applicables à des objectifs RH définis.
2. La méthode indirecte : qui implique l'utilisation de données secondaires, considérées comme des « sous-produits » du SIRH, plutôt que comme des résultats principaux (Manyika et al., 2011). À titre d'exemple, les systèmes de planification du temps, bien qu'utilisés essentiellement pour le suivi des heures de travail, fournissent également des données qui facilitent l'analyse des tendances de présence/absence des salariés.

2.1.3. Définition par les producteurs de données

La variation de la nature des données RH, révélée par leur défintion de formes, de structures et de sources, souligne également l'importance de considérer les différents types de producteurs, qui sont essentiels à la production de connaissances et au développement de technologies numériques RH.

Trois catégories de producteurs de données RH peuvent être identifiées :

1. Les producteurs de données primaires : collectent directement des données auprès des salariés ou via la fonction RH pour des objectifs spécifiques, tels que les évaluations de performance ou les enquêtes de satisfaction.
2. Les producteurs de données secondaires : fournissent des données initialement collectées pour d'autres finalités, mais qui sont réutilisées en GRH pour des analyses comparatives ou des évaluations. Ces données peuvent inclure, par exemple, les résultats d'enquêtes démographiques ou de rapports sectoriels, comme ceux publiés par l'Insee[3].
3. Les producteurs de données tertiaires : génèrent des données à partir de données secondaires, les transformant en analyses statistiques ou en synthèses d'études.

[3] L'Insee, ou Institut national de la statistique et des études économiques, est chargé de produire, d'analyser et de publier les statistiques officielles en France.

Ces producteurs offrent des perspectives approfondies et répondent à des questions analytiques spécifiques.

2.1.4. Définition par les types de données

Les types de données se réfèrent au rôle et à l'utilisation des données dans le SIRH ou pour des analyses spécifiques. Il s'agit d'une classification plus générale qui englobe un large éventail de données, y compris des données qualitatives. On peut distinguer trois types de données :

1. Les données indexicales : caractérisées par leur capacité à permettre l'identification et la connexion entre les informations, comprennent des identifiants uniques tels que les numéros de matricule et les codes postaux. Elles jouent un rôle crucial dans la (ré)agrégation et le suivi des volumes importants de données non indexicales, facilitant ainsi divers traitements et analyses.
2. Les données attributaires : représentent des caractéristiques qualitatives ou quantitatives d'un phénomène sans fonction d'identification directe, telles que l'âge, le genre ou la catégorie socio-professionnelle. Ces données constituent la majorité des données stockées dans le SIRH.
3. Les métadonnées : ou appelées « données sur les données », concernent soit le contenu des données, soit l'ensemble des données elles-mêmes. Elles incluent les noms et descriptions de champs spécifiques, comme les en-têtes de colonnes dans une feuille de calculs ou la définition des données utilisées. Les métadonnées sont essentielles pour aider les utilisateurs (e.g. fonction RH) à comprendre la structure d'un jeu de données, sa méthode d'utilisation et son interprétation.

Considérant ces caractéristiques techniques, les données RH sont ainsi intrinsèquement modifiables. En tant que médiums de signification et de représentation des activités et phénomènes RH, elles sont définies par leur capacité à être constamment révisées, adaptées, renouvelées et étendues (Kallinikos et al., 2013). De surcroît, la portabilité entre les producteurs - primaires, secondaires et tertiaires - les rend décontextualisables, permettant ainsi de véhiculer des récits qui transcendent leur origine et leur usage initial (Alaimo et al., 2020; Alaimo & Kallinikos, 2017; Monteiro & Parmiggiani, 2019). Ces caractéristiques confèrent aux données RH une fluidité qui peut influencer considérablement leur utilisation au sein des

organisations, lesquelles sont elles-mêmes productrices de ces données (Alaimo et al., 2020).

2.2. Acte II : transformation des données RH

Au cours des dernières décennies, les chercheurs ont examiné comment la numérisation, la digitalisation et la datafication facilitent la diversification des formes organisationnelles et, par conséquent, favorisent la création de valeur à partir des données (Alaimo & Kallinikos, 2022; Leonardi & Treem, 2020; Østerlie & Monteiro, 2020; Zuboff, 2019). Ces trois phénomènes permettent aux organisations de créer de nouvelles opportunités (Kitchin & Lauriault, 2018) et de renforcer leur capacité à générer toujours plus de valeur par la transformation de leurs structures et modèles économiques.

Cette valeur peut être étudiée à travers trois concepts (Lycett, 2013; Normann, 2001a) :

1. La dématérialisation ;
2. La liquéfaction ;
3. La densité.

Ces concepts, détaillés dans le Tableau 2 ci-dessous, permettent de repenser le processus de création de valeur à partir des données numériques.

Tableau 2 : Processus de création de valeur à partir des données numériques

Concepts	Phénomènes d'origine	Descriptions
Dématérialisation	Numérisation	Capacité à dissocier les données analogiques de leur utilisation dans le domaine physique, grâce à leur conversion au format numérique.
Liquéfaction	Digitalisation	Aptitude des données, une fois dématérialisées, à être aisément manipulées et transférées via une infrastructure adéquate, facilitant ainsi leur dissociation et reconfiguration.

Concepts	Phénomènes d'origine	Descriptions
Densité	Datafication	Capacité à recombiner de façon optimale les données mobilisées en fonction d'un contexte particulier, à un moment et en un lieu spécifiques, pour une finalité donnée.

Toutefois, pour comprendre comment ces trois phénomènes transformationnels contribuent à la création de valeur à partir des données RH, il est essentiel de les définir clairement.

2.2.1. La numérisation : transformation par la « dématérialisation » des données

La numérisation selon Leonardi & Treem (2020), est définie comme le phénomène de transformation d'actions ou de représentations d'actions en données binaires (zéros et uns). Cette conversion des données analogiques au format numérique constitue un processus technique fondamental introduit par les informaticiens dès l'apparition des premiers ordinateurs. Ce format permet aux données codées d'être lues, traitées, transmises et stockées par des technologies numériques (Flyverbom, 2019; Legner et al., 2017; Tilson et al., 2010).

Les ordinateurs ont ainsi marqué un changement radical par la numérisation, ébranlant les modèles de services dominants des industries de l'information traditionnelles telles que la radio, le téléphone et la télévision (Tilson et al., 2010). Les premières vagues de numérisation ont toutefois été progressives dans la dissolution des réseaux socio-techniques existants à cette époque, en reproduisant initialement les fonctionnalités analogiques sous un format numérique (ibid.).

La numérisation possède donc de nombreux avantages. Contrairement aux données analogiques qui nécessitent des déplacements physiques et des infrastructures spécifiques, elle permet un accès, un partage et un stockage des données de manière instantanée et depuis n'importe quel endroit (Atasoy & Morewedge, 2017; Leung et al., 2022). En effet, lorsque les ordinateurs sont devenus moins chers, plus petits et plus puissants, leur utilisation s'est étendue au-delà des

coulisses des organisations. Parallèlement, l'émergence de réseaux numériques généraux, tels qu'Internet, a permis aux technologies numériques de communiquer, de stocker et de traiter une grande variété de données (cf. convergence des technologies). Ces réseaux ont alors été en mesure de prendre en charge presque tous les types de services d'informations (cf. convergence des réseaux). La flexibilité inhérente à la numérisation a ainsi pu être pleinement exploitée, les données numériques devenant véritablement infrastructurelles (Tilson et al., 2010).

Cette dématérialisation vers les données numériques a fondamentalement modifié les processus organisationnels en diminuant ou en supprimant leurs contraintes matérielles traditionnelles, incluant les restrictions de temps, d'espace, de localisation et de capital nécessaire (Leonardi & Treem, 2020). En effet, la réduction continue des coûts associés à la production et au stockage offre aux organisations l'opportunité d'élargir de manière significative leur accès aux données numériques.

Si la numérisation des données est un concept utile, il apparaît plus pertinent de l'analyser à travers le prisme de la digitalisation. En effet, la numérisation se limite essentiellement à sa dimension technique, consistant à convertir des données analogiques au format numérique. La digitalisation, quant à elle, ne se contente pas de cette conversion technique. Elle intègre les données numériques dans toutes les couches organisationnelles et sociales, ouvrant ainsi la voie à des transformations plus profondes et systémiques. En exploitant la nature numérisée des données, la digitalisation permet d'instaurer de nouvelles configurations organisationnelles et de redéfinir les pratiques traditionnelles (Leonardi & Treem, 2020). Ce processus marque l'entrée dans un phénomène transformationnel d'une portée beaucoup plus vaste, où les structures, processus et pratiques évoluent de manière drastique.

2.2.2. La digitalisation : transformation par la « liquéfaction » des données

La digitalisation se définit notamment comme le phénomène par lequel les organisations sont modelées et influencées par les données numériques (Flyverbom, 2019; Leonardi & Treem, 2020). Alors que la numérisation se concentre principalement sur la conversion des données analogiques au format numérique, la digitalisation adopte une perspective beaucoup plus large. Elle renvoie aux dispositifs socio-

techniques découlant des processus d'adoption et d'utilisation des données numériques dans des contextes individuels, organisationnels et sociétaux (Legner et al., 2017).

Les études sur la digitalisation, bien que diverses dans leurs perspectives, convergent sur plusieurs points clés. Par exemple, Yoo et al. (2010) caractérisent les technologies numériques par leur programmabilité algorithmique et leur structure en couches. De manière complémentaire, Normann (2001) souligne la capacité de ces technologies à engendrer la « liquéfaction » des données numériques, caractérisée par leur dématérialisation et leur détachement de tout support physique. Cette idée est également soutenue par Lusch & Nambisan (2015) ainsi que par Monteiro & Parmiggiani (2019). La liquéfaction, en rendant les données numériques plus malléables et indépendantes de leur support d'origine, peut également amener les organisations à se « liquéfier » elles-mêmes. Orlikowski & Scott (2015) décrivent ce phénomène comme un « phénomène algorithmique », tandis que Kallinikos (2009) le qualifie de « représentation computationnelle » de la réalité organisationnelle.

La capacité transformative de la digitalisation est également mise en avant par Østerlie & Monteiro (2020). Selon ces auteurs, la précision avec laquelle les données numériques reflètent le domaine physique oscille entre une représentation fidèle, une simple ressemblance, ou une complète dissociation (p.3). En d'autres termes, les données numériques possèdent des propriétés uniques[4] que l'on ne retrouve pas dans le domaine physique (Kallinikos et al., 2013). Cette spécificité se manifeste de manière concrète à travers diverses applications. Par exemple, les vidéos sur *YouTube* sont régies par des normes de codage vidéo communes. Ces normes ne se contentent pas de faciliter le visionnage des vidéos, mais permettent également des pratiques socio-techniques complémentaires telles que l'annotation, la réappropriation, le remixage et la modification des codages originaux, effectuées tant par des humains que par des machines. Cela engendre un espace en constante expansion de création, de négociation et de standardisation du sens (Kallinikos et al., 2013; Tilson et al., 2010).

Ainsi, les données numériques, bien qu'initialement basées sur des éléments physiques, acquièrent progressivement une autonomie et se détachent de leurs

[4] Les propriétés des données numériques sont développées de manière exhaustive dans la section 2.4.2.1, intitulée « Les données RH comme instruments du capital ».

origines matérielles (Østerlie & Monteiro, 2020). La focalisation sur la digitalisation en tant que phénomène transformationnel permet ainsi d'examiner comment les rationalités collectives liées aux propriétés des données numériques se manifestent à l'intersection des pratiques socio-techniques (Kallinikos et al., 2013).

Cette capacité transformative est également étroitement liée à ce que Zittrain, (2008) décrit comme la générativité des données numériques, c'est-à-dire leur capacité d'extension illimitée. Avec l'intensification de la numérisation, la digitalisation s'appuie sur cette générativité, où les données numériques sont conçues sur : « […] *la notion qu'elles ne sont jamais complètement achevées, qu'elles ont de nombreux usages encore à imaginer, et que le public et les membres ordinaires des organisations peuvent être dignes de confiance pour inventer et partager de bonnes utilisations* » (ibid. p. 43).

À titre d'illustration de ces propriétés génératives, Tilson et al. (2010) expliquent que le langage de balisage étendu (XML) possède la capacité d'intégrer de nouvelles pratiques socio-techniques, fondées sur des normes de données convenues. Désormais, les définitions de type de document (DTD) basées sur le langage XML peuvent être définies localement tout en se rapportant simultanément à des normes émergentes au sein et entre les organisations. Cette générativité ajoute des propriétés imprévues aux données numériques, telles que la capacité des producteurs – primaires, secondaires et tertiaires - à recombiner leurs caractéristiques (Kitchin, 2022d) et à générer, assembler et redistribuer du nouveau contenu.

Cette générativité, induite par la digitalisation, se manifeste à travers diverses évolutions organisationnelles. Ces dernières englobent, par exemple, l'adoption de modalités de travail distribuées et flexibles (Hinds & Kiesler, 2002), l'automatisation des processus administratifs, l'implémentation de systèmes de gestion des connaissances (Alavi & Leidner, 2001), ou bien l'utilisation des réseaux sociaux d'entreprise comme principales plateformes de communication (Treem & Leonardi, 2012). Ces évolutions, comme mentionné précédemment, sont rendues possibles par la numérisation, qui réduit significativement le coût marginal de production et de stockage, facilitant ainsi l'accès des organisations à un vaste volume de données numériques (Hinings et al., 2018).

Cette générativité a également permis le développement de nouvelles dynamiques de marché. Cela a notamment été rendu possible par l'émergence de nouveaux acteurs, tels que les plateformes et les cabinet de conseil spécialisés (Flyverbom, 2019; Kallinikos et al., 2013; Zuboff, 2019). La générativité des données numériques, résultant directement des phénomènes de numérisation et de digitalisation, a considérablement intensifié la nécessité de créer de la valeur à partir de ces données (Tilson et al., 2010). Cette tendance conduit à l'émergence d'un troisième processus transformationnel au sein des organisations, connu sous le nom de : datafication.

2.2.3. La datafication : transformation par la « densité » des données

La datafication se définit comme le phénomène par lequel les activités sociales et organisationnelles sont transformées en données numériques. Ces données sont jugées d'« intérêt » en raison de leur capacité à être accumulées, cette accumulation possédant le potentiel de générer de la valeur (Mayer-Schönberger & Cukier, 2013). Introduit par Cukier et Mayer-Schoenberger (2013), ce concept a attiré l'attention de nombreux chercheurs, notamment dans le domaine croissant des études critiques sur les données (Alaimo et al., 2020; Alaimo & Kallinikos, 2017, 2022; Monteiro & Parmiggiani, 2019; Østerlie & Monteiro, 2020).

Le terme « datafication » avait toutefois déjà été utilisé auparavant. Par exemple, Brown & Duguid (2000) l'emploient dans un sens qui résonne avec les discours critiques actuels sur la datafication : « *Il n'est donc pas surprenant que les enthousiastes de l'information* [données] *exultent devant le simple volume d'informations [données] que la technologie rend maintenant disponible. Ils comptent les bits, octets et paquets avec enthousiasme. Ils applaudissent la désagrégation de la connaissance en données (et fournissent un nouveau mot – datafication - pour la décrire)* » (p.12).

Dans le contexte actuel, marqué par l'interaction entre trois phénomènes transformationnels (la numérisation, la digitalisation et la datafication), une attention particulière est accordée à la maximisation de la valeur des données numériques. Ainsi, les données peuvent être soumises à des processus d'analyse, de classification et, de plus en plus, de marchandisation (Davenport, 2014; Zuboff, 2019).

En effet, en s'appuyant sur les deux premiers phénomènes, la datafication est étroitement liée à la marchandisation des données numériques. Elle met effectivement en lumière le rôle croissant des données numériques dans les dynamiques du capitalisme moderne (van Dijck, 2018). Dans cette perspective, Morozov (2015) conceptualise le « capitalisme des données » comme un système économique visant à capturer notre comportement en temps réel afin de stocker et utiliser ces données de manière personnalisée. Ce système permet une variété d'activités allant de la réservation de taxis (par exemple, Uber) à la commande de nourriture (par exemple, Deliveroo), en passant par le réseautage professionnel (par exemple, LinkedIn), la socialisation (par exemple, Instagram), l'écoute de musique (par exemple, Spotify), ou encore l'achat et la vente de produits de mode (par exemple, Shopify). Il trouve, dans le même temps, une résonance particulière dans le mouvement *Quantified Self*, initié en 2007 (Nafus & Sherman, 2014). Ce mouvement, souvent décrit comme la « mesure de soi », tire ses origines du *self-tracking* et a favorisé l'intégration de capteurs de plus en plus miniaturisés et performants dans une multitude d'objets du quotidien (e.g. montres connectées). Ces derniers permettent une automatisation de la mesure de paramètres physiologiques et cognitifs en collectant une variété de données (poids, nombre de pas, distances parcourues, calories brûlées) qui sont ensuite transformées en métriques personnelles. Une fois analysées, ces données peuvent, dans certains cas, être partagées et commercialisées sous forme d'abonnements (par exemple, Strava), transformant ainsi les données personnelles en biens économiques (Ranck, 2012).

La datafication redéfinit également les processus organisationnels en augmentant le volume et la diversité des données disponibles pour la prise de décision. Par exemple, les salariés sont de plus en plus surveillés et évalués par des algorithmes qui analysent leurs comportements passés. Cette tendance, liée à l'intensification de la numérisation et de la digitalisation conduit inévitablement à une accélération de la datafication (Sadowski, 2019). En conséquence, cette transformation de la main-d'œuvre en données numériques permet également à ces dernières de devenir un capital exploitable.

La valeur des données ne réside pas uniquement dans leur visibilité et leur disponibilité (Leonardi & Treem, 2020), mais aussi dans leur capacité à être intégrées dans des infrastructure et des technologies numériques. Ces données jouent un rôle

essentiel dans la création de valeur en éliminant les contraintes de temps, de lieu, d'acteurs et de configurations nécessaires pour réaliser des analyses (Normann, 2001b). Elles permettent de dépasser la « connaissance figée ». En effet, l'auteur explique que les données analogiques sont utiles parce qu'elles sont reproductibles et prévisibles, mais elles immobilisent toutefois les connaissances au moment de leur production.

En résumé, les trois phénomènes transformationnels ont conjointement amplifié l'importance des données numériques au sein des organisations et, plus largement, dans la société contemporaine. En raison de leur interconnexion, ils forment ainsi une spirale de renforcement mutuel avec les données numériques que je désignerai par « *data-driven* ».

L'interconnexion de ces trois phénomènes accroît en retour la dématérialisation, la liquéfaction et la densité des données, stimulant ainsi le développement des technologies numériques et des méthodes analytiques. Par conséquent, chaque avancée dans la numérisation, la digitalisation et la datafication, renforce la capacité de traitement des données et augmente de ce fait leur potentiel à générer de la valeur. Cette augmentation de la valeur incite à des investissements supplémentaires dans les technologies et les méthodes analytiques, ce qui, à son tour, renforce encore davantage la croissance des trois phénomènes transformationnels. En somme, la spirale de renforcement *data-driven* à l'œuvre, illustre comment les phénomènes technologiques et la prolifération croissante des données numériques se nourrissent réciproquement, redéfinissant ainsi les dynamiques socio-économiques et organisationnelles.

La tendance au renforcement par et pour les données numériques a inévitablement influencé la recherche en GRH, et plus particulièrement en e-GRH et en instrumentation de la GRH. Cette évolution révèle effectivement une montée en puissance de l'intérêt de la fonction RH pour l'analyse de ses données (Angrave et al., 2016). Cet intérêt accru est également à l'origine de trois phénomènes transformationnels dans les travaux en GRH.

2.3. Acte III : valorisation des données RH

De nombreux travaux en GRH portent une attention particulière aux données numériques. Ces études mettent en valeur les différentes pratiques d'exploitation des données, allant de la définition d'indicateurs jusqu'à l'adoption de méthodes analytiques avancées telles que les IA. Ces études s'inscrivent majoritairement dans trois phénomènes transformationnels à l'échelle de la GRH :

1. Les *big data* RH ;
2. L'analytique et les métriques RH ;
3. L'épistémologie positiviste et formelle.

Plusieurs revues de littérature ont mis en lumière les différentes perspectives d'exploitation des données RH (Chalutz Ben-Gal, 2019; Coron, 2022; Garcia-Arroyo & Osca, 2019; Margherita, 2022). Comme le souligne Coron (2022), bien que chaque revue apporte une contribution spécifique à la conceptualisation des données RH, leur compilation offre cependant une vision fragmentée de leur cycle de vie (de leur construction à leur exploitation). Malgré cette disparité, ces discussions ont en commun l'ambition de (re)positionner la fonction RH comme un acteur stratégique au sein de l'organisation (Hermans & Ulrich, 2021; Marler & Boudreau, 2017 ; Rasmussen & Ulrich, 2015).

2.3.1. Les *big data* RH : valorisation par les « grandes » données

L'interconnexion des trois phénomènes transformationnels à l'échelle de l'organisation, à savoir la numérisation (cf. dématérialisation), la digitalisation (cf. liquéfaction) et la datafication (cf. densité) des données numériques, se traduit notamment dans la GRH à travers le concept des *big data* RH[5]. Les *big data* RH illustrent comment l'interconnexion de ces trois phénomènes peut influencer et enrichir les pratiques et politiques de GRH en fournissant un accès à un vaste volume de

[5] Étant donné que le terme « *big data* » est pluriel en anglais puisqu'il désigne la (re)combinaison de multiples ensembles de données, il est plus précis de parler « des » *big data* en français. Il en est de même pour le terme « intelligences artificielles » qui renvoie également à une disparité de techniques.

données diversifiées (Coron, 2019a, 2019b, 2019e; Strohmeier, 2020; Zhang et al., 2021).

La principale innovation des *big data* RH réside dans l'utilisation du terme « *big* », qui formalise une distinction avec les « petits » ensembles de données issus du SIRH (Kitchin & Lauriault, 2015). Cette distinction reste toutefois relativement récente. Avant 2008, les données numériques étaient rarement considérées comme « petites » ou « grandes » ; ces dernières étant simplement vues comme des données, indépendamment de leur volume. En raison de facteurs tels que les coûts, les ressources nécessaires et les défis liés à la gestion du cycle de vie des données, seuls des volumes limités de données de « bonne » qualité étaient produits, visant à assurer la représentativité des activités et phénomènes RH à l'étude (ibid.).

Avec l'accélération exponentielle de trois phénomènes transformationnels à l'échelle de l'organisation, ainsi que le développement des technologies numériques et des méthodes analytiques, les « petites » données RH évoluent rapidement. Elles peuvent désormais être complétées par des ensembles de données beaucoup plus volumineux et diversifiés. Cette transition vers des volumes plus importants offre ainsi une nouvelle dimension à l'analyse des données RH.

Cependant, ces « grandes » données présentent des caractéristiques ontologiques très différentes (Kitchin & Lauriault, 2015). Le terme « *big* » apparaît alors comme réducteur, les *big data* étant caractérisés par bien plus que leur volume (Boyd & Crawford, 2012; Mayer-Schönberger & Cukier, 2013).

La définition des *big data*, qui trouve son origine dans le rapport Gartner (Laney, 2001), est couramment articulée autour du modèle des « 3V ». Ce modèle met en lumière trois dimensions essentielles des données : (1) le volume, (2) la variété et (3) la vélocité (McAfee & Brynjolfsson, 2012). D'autres dimensions, telles que la véracité et la valeur des données, sont souvent ajoutées pour enrichir cette définition. Cependant, comme le soulignent Kitchin & McArdle (2016), il est rare que toutes ces dimensions soient simultanément présentes. Ainsi, les *big data* ne constituent pas une catégorie homogène et il est possible d'identifier différentes « typologies » de *big data*. En outre, bien que ces dimensions soient principalement mises en avant dans les travaux sur le sujet, cette focalisation tend à occulter d'autres enjeux essentiels pour comprendre pleinement la complexité de leur exploitation. Parmi ceux-ci, les méthodes

analytiques employées pour le traitement des données numériques, ainsi que les finalités visées par ces traitements, méritent également une attention particulière (Coron, 2019c).

Afin de clarifier les différences entre les « petites » et les « grandes » données, Kitchin & Lauriault (2015) ont élaboré un tableau comparatif mettant en lumière leurs principales distinctions (voir Tableau 3).

Tableau 3 : Comparaison entre les « petites » et les « grandes » données (Kitchin & Lauriault, 2015)

Caractéristiques	**Petites données**	**Grandes données**
Volume : quantité totale de données stockées.	Limité à important	Très important
Exhaustivité : étendue de la couverture des données dans le contexte étudié.	Échantillons	Populations entières
Résolution et précision : détail et exactitude des données.	Variables (de faibles à élevées)	Élevées
Relationnalité : capacité des données à être liées et analysées en relation avec d'autres données.	Variable (de faible à élevée)	Forte
Vélocité : vitesse à laquelle les données sont générées et traitées.	Lente (arrêt sur image)	Rapide
Variété : diversité des types de données et de leurs sources.	Limitée à variée	Très importante
Flexibilité et évolutivité : facilité d'adaptation des structures de données et capacité à augmenter en volume.	Faibles à moyennes	Très élevées

Ainsi, selon Kitchin & McArdle (2016), les *big data* se définissent par les caractéristiques ontologiques présentées ci-dessous :

- Un volume très important : les données sont mesurées en téraoctets ou pétaoctets, reflétant la quantité massive de données numériques accumulées. Par exemple, le

groupe *Meta* (*Facebook*) capte quotidiennement l'attention de 36,2 millions d'internautes à travers ses plateformes, couvrant ainsi plus de la moitié de la population française (57,4%) (Billon, 2024).

- Une portée exhaustive : les données utilisées visent à englober des populations entières ou des systèmes complets.
- Une résolution et une précision élevées : les données doivent être détaillées au maximum, avec des identifications uniques et une indexation pour chaque élément de donnée.
- Une relationnalité forte : les données incluent des champs communs qui permettent l'association de différents ensembles, facilitant ainsi leur interconnexion.
- Une vélocité rapide : les données sont générées et traitées en temps réel ou quasi-réel, soulignant la rapidité de leur flux.
- Une variété très importante : les données présentent une variété importante, incluant à la fois des données structurées, semi-structurée et non structurées, souvent accompagnées de références temporelles et spatiales.
- Une flexibilité et une évolutivité très élevées : les données se caractérisent par leur extensibilité (facilité d'ajout de nouveaux champs de données) et leur scalabilité (capacité à augmenter rapidement en taille), adaptant ainsi les structures de données aux besoins évolutifs.

Les caractéristiques ontologiques des *big data*, telles que conceptualisées par Kitchin & McArdle (2016), fournissent une base théorique importante pour analyser leur influence dans le domaine de la GRH. À cet égard, et pour illustrer concrètement cette influence, l'étude menée par Coron (2019e, 2019c) examine l'impact potentiel des *big data* sur trois pratiques spécifiques de GRH :

1. La gestion de l'absentéisme ;
2. La présélection des candidatures ;
3. L'orientation des programmes de formation.

Cette étude révèle que les *big data* ont le potentiel de transformer radicalement les pratiques de GRH en favorisant des objectifs d'automatisation, de personnalisation et de prédiction. Ces objectifs représentent effectivement un changement substantiel par rapport à l'usage traditionnel des technologies numériques et méthodes analytiques en GRH, principalement centrés sur de larges groupes de salariés et sur l'analyse

d'activités et phénomènes passés. Ils suggèrent ainsi une modification structurelle au sein de la GRH, en mettant en lumière des objectifs auparavant non explorés dans ce domaine.

Bien que le débat sur les *big data* RH gagne progressivement en visibilité, l'état des connaissances actuelles sur le sujet reste embryonnaire. La majorité des contributions sont principalement de nature conceptuelle, soulignant ainsi un écart notable entre l'élaboration théorique et l'application pratique en GRH (Coron, 2019a, 2019b, 2019e; Strohmeier, 2020; Zhang et al., 2021). L'adoption des *big data* RH est notamment entravée par des défis tels que les contraintes de protection des données personnelles et les coûts associés aux technologies et infrastructures numériques nécessaires (Kitchin, 2022e; Kitchin & Lauriault, 2015). À ce jour, les données structurées issues du SIRH restent les plus exploitées, utilisées principalement pour la génération de rapports et de tableaux de bord (Angrave et al., 2016).

En outre, le scepticisme quant à l'utilité des *big data* RH est renforcé par des chercheurs tels que Cappelli (2017), qui soulignent également les défis pratiques liés à un contexte comme celui de la GRH où les avantages escomptés semblent restreints. Cappelli (2007) soutient que, pour la majorité des entreprises, qui comptent seulement quelques milliers de salariés et réalisent principalement des évaluations annuelles, l'investissement dans de vastes ensembles de données numériques pourrait être disproportionné. Cela remet en question la pertinence de leur utilisation compte tenu des bénéfices potentiellement limités.

Cette préoccupation mène également à une autre interrogation concernant l'intégration des *big data* RH : la qualité des données collectées et analysées. Dans un domaine où les décisions ont une portée sociale significative (Coron, 2019g), il est essentiel de s'assurer de la qualité des données en tenant compte de différents paramètres (Kitchin & Lauriault, 2015) :

- L'exactitude des données : garantie par l'absence d'erreurs et de lacunes, assurant ainsi la fiabilité des données.
- La fidélité des données : démontrée par la capacité à reproduire les mêmes résultats dans des mesures répétées, affirmant la constance des données.
- La cohérence des données : observée à travers l'uniformité des informations recueillies via différentes sources et contextes.

- L'objectivité des données : exigée pour éliminer tout biais affectant les données[6].
- La véracité des données : mesure la précision avec laquelle les données représentent la réalité qu'elles sont censées refléter.
- La traçabilité des données : faculté de retracer l'origine des données, importante pour vérifier leur authenticité et leur adéquation à l'usage envisagé.

Ces paramètres sont ainsi essentiels pour garantir la qualité des ensembles de données numériques employés dans le processus décisionnel en GRH, dans la mesure où ils influencent directement la qualité des décisions prises (Coron, 2019c). Comme le dicton le souligne : « *garbage in, garbage out* »[7].

2.3.2. L'analytique et les métriques RH : valorisation par les « petites » données

La progression rapide et l'influence des *big data* ont amené certains chercheurs à se demander si ce phénomène pourrait entraîner l'obsolescence des « petites » données, ou si la pertinence des études basées sur ces dernières pourrait se voir diminuée en raison de leurs contraintes en termes de volume, de vélocité et de pertinence relative (Kitchin & Lauriault, 2015).

Bien que les « petites » données numériques, telles que les données RH, puissent être limitées, elles jouissent néanmoins d'une longue tradition de développement dans le domaine de la GRH, ainsi qu'au sein d'organisations gouvernementales, non gouvernementales et d'entreprises. Elles reposent sur des méthodes analytiques bien établies et largement reconnues, qui ont historiquement produit, de manière générale, des résultats probants (ibid.).

A titre d'illustration, considérons un système de pointage du temps de travail. Ces données RH structurées, bien que limitées à des séries temporelles, enregistrent précisément les heures d'arrivée et de départ des salariés. Elles documentent aussi

---

[6] Il est plus approprié de viser la réduction ou la régulation des biais, ce qui évite l'illusion d'une élimination totale, souvent associée au « mythe de la quantification objective » selon Coron (2019g, p. 58). En complément, l'approfondissement de la notion d'« objectivité limitée » peut s'avérer pertinent (ibid., p.64).

[7] Le dicton « *garbage in, garbage out* » est issu de l'informatique et signifie que des données d'entrée de mauvaise qualité - qui comportent des erreurs ou des incohérences - entraînent inévitablement de mauvais résultats.

les pauses, les heures supplémentaires et d'autres variations du temps de travail quotidien, offrant ainsi une haute résolution. Chaque donnée est donc précise et granulaire, facilitant des analyses approfondies des comportements de présence et d'absence des salariés. En outre, enregistrées de manière continue, ces données permettent de suivre les habitudes de présence des salariés sur une base quotidienne ou hebdomadaire. Bien que limitées en variété, la précision de ces données est importante pour évaluer l'efficacité des politiques de gestion du temps de travail, notamment pour la planification des effectifs. Elle permet également d'identifier des tendances ou de détecter des anomalies dans les habitudes de travail des salariés, jouant ainsi un rôle significatif dans la gestion de l'absentéisme.

Ainsi, le deuxième phénomène transformationnel à l'échelle de la GRH réside dans les approches par les métriques et, plus largement, par l'analytique RH. Ces deux approches, basées sur les « petites » données, constituent les typologies d'études les plus exploitées en GRH.

2.3.2.1. L'analytique RH : les « petites » données structurées et internes à l'organisation

L'analytique RH mobilise des méthodes analytiques avancées pour favoriser une meilleure prise en compte de la complexité des activités et phénomènes RH (Madsen & Slåtten, 2019; Marler et al., 2017; Marler & Boudreau, 2017). Ceux-ci sont fréquemment multifactoriels, et de ce fait, difficilement interprétables par le biais de simples croisements de données (Coron, 2019c). L'analytique RH vise ainsi à approfondir la compréhension des liens de causalité ou de corrélation entre diverses données, qu'elles soient liées aux RH ou non (ibid.).

Appelé successivement « *People Analytics* », « *Talent Analytics* » ou « *Workforce Analytics* » (Madsen & Slåtten, 2019; Marler & Boudreau, 2017), cette approche, par la valorisation des « petites » données numériques, se définit comme : « *rendue possible par les technologies de l'information, qui utilisent une analyse descriptive, visuelle et statistique des données relatives aux processus RH, au capital humain, à la performance organisationnelle et aux repères économiques externes, afin d'établir l'impact commercial et soutenir la prise de décision basée sur les données* » (Marler & Boudreau, 2017, p. 13). Cette définition met ainsi en évidence que l'analytique RH ne

se concentre pas exclusivement sur les données RH mais implique l'intégration d'autres données fonctionnelles et internes à l'organisation. Ces données sont collectées, manipulées et analysées pour soutenir les décisions à destination des salariés et établir un lien entre décisions RH, résultats commerciaux et performance organisationnelle.

La prégnance de la valeur présumée de l'analytique RH dans les discours des praticiens reste, toutefois, fortement contrastée par le nombre de travaux académiques sur le sujet (Angrave et al., 2016; Greasley & Thomas, 2020). Selon Rasmussen & Ulrich (2015), les preuves empiriques des bénéfices engendrés par l'exploitation de l'analytique RH sont limitées, davantage fondées sur des convictions et le plus souvent publiées par des consultants possédant un intérêt marchand.

Les résultats de deux enquêtes menées par Lismont et al. (2017) révèlent que l'analytique RH, utilisée par seulement 40% des répondants, est moins répandue et donc moins avancée que dans d'autres fonctions organisationnelles. Les auteurs précisent qu'au sein de ce sous-groupe, 21% des répondants abordent l'analytique RH pour la gestion de la performance des salariés, contre 15% pour la rétention et 13% pour le recrutement. Dans ce contexte, Angrave et al. (2016) éclairent cette situation en pointant le manque de capacités analytiques au sein de la fonction RH, un déficit qui peut être attribué à ses racines historiques dans le domaine légi-social.

Cette lacune devient d'autant plus problématique avec l'expansion rapide de l'industrie de l'analytique RH. Selon Tambe et al. (2019), les technologies numériques dédiées à la GRH tendent à fournir des réponses trop générales, qui ne parviennent pas à adresser adéquatement les enjeux éthiques, techniques et sociaux auxquels la GRH est confrontée.

2.3.2.2. Les métriques RH : les « petites » données structurées et internes à la GRH

Les métriques RH renvoient, quant à elles, plus précisément aux tableaux de bord et reporting RH. Les données utilisées sont le plus souvent des données agrégées au niveau de la GRH et s'effectue souvent sous la contrainte d'obligations légales ou de mise en conformité (Coron, 2019c). Lawler III et al. (2004), mettent en évidence trois

différents types de métriques pour comprendre et évaluer les activités et phénomènes de GRH :

1. L'efficience : s'axe sur le développement d'indicateurs de productivité et de coût issus des tâches administratives (e.g. coût par embauche). Comme en font état les auteurs, l'intérêt pour l'exploitation des données administratives est renforcé par le fait qu'il s'agisse de données *normative*s, et donc similaires et comparables entre les organisations.
2. L'efficacité : favorise la production d'indicateurs qui mesure l'effet des politiques et pratiques de GRH sur les populations de salariés cibles (e.g. retour sur investissement de la formation).
3. L'impact : renvoie au développement de métriques opérationnelles relatives au développement des compétences des salariés (e.g. productivité par salarié).

Les recherches concernant les métriques RH tendent souvent à adopter une approche univariée ou bivariée pour évaluer les activités ou phénomènes RH. Autrement dit, elles présentent fréquemment leurs mesures et résultats à travers l'analyse croisée de deux données RH, telles que l'âge et le statut des salariés, pour calculer des indicateurs comme le taux d'absentéisme (Coron, 2019c). Cependant, cette approche analytique « basique » s'avère souvent insuffisante pour saisir pleinement la complexité des activités et phénomènes RH, tels que l'absentéisme, nécessitant une analyse plus nuancée et multifactorielle.

En réponse à cette limitation, certains chercheurs et défenseurs d'une GRH positiviste et formelle estiment néanmoins que les reportings et tableaux de bord permettent de réduire l'incertitude dans la prise de décision et le pilotage des activités et phénomènes RH. Cette perspective est notamment illustrée par Tootell et al. (2009) dans leur article : « *Metrics : HRM's holy grail ? A New Zealand case study* ».

Compte tenu de l'évolution actuelle et des avantages offerts par l'analyse des « petites » données, il est évident que celles-ci continueront de constituer un des piliers dans le domaine de la GRH. Sur le plan théorique, on suppose que l'utilisation de ces données s'intensifiera avec l'intégration des nouvelles technologies et infrastructures numériques. L'intégration de ces données devrait non seulement faciliter leur conservation, mais également permettre leur réutilisation et leur combinaison avec d'autres ensembles de données, qu'ils soient petits ou grands. En

conséquence, cela aura pour effet d'augmenter leur densité et, par extension, leur valeur, comme le souligne le concept des *big data* RH (Kitchin & Lauriault, 2015). Cependant, sur le plan empirique, les études existantes sur l'analytique et les métriques RH restent insuffisantes pour confirmer de manière fiable ces affirmations ou mesurer les progrès réalisés dans ce domaine (Van den Heuvel & Bondarouk, 2017).

Ainsi, bien qu'il soit possible d'observer une spirale *data-driven* à l'échelle organisationnelle, largement documentée par la littérature en SI, cette dernière n'est pas aussi clairement établie en GRH. Même si la pertinence des données est un sujet important et régulièrement débattu dans l'analyse des activités et phénomènes RH, ce thème demeure largement sous-représenté dans cette littérature, et plus particulièrement en e-GRH et en instrumentation de la GRH (Coron, 2022). En effet, ces dernières privilégient davantage les effets de l'exploitation des données RH déjà intégrées aux technologies numériques (Tambe et al., 2019). Cette orientation rend ainsi difficile l'étude du renforcement mutuel entre la GRH et les données, si tant est qu'il en existe un.

Toutefois, ce renforcement demeure un sujet central dans les débats en GRH, notamment en raison de la pression croissante exercée par deux forces convergentes :

1. Les discours technophiles des acteurs *digital natives* qui ont un intérêt économique à soutenir l'adoption des technologies numériques en GRH (Marler et al., 2017).
2. L'épistémologie positiviste et formelle des chercheurs, notamment d'Amérique du Nord, qui privilégient une approche scientifique fondée sur des données quantitatives et des preuves (Harley, 2015).

Ces forces convergent ainsi pour renforcer la spirale *data-driven* en GRH, soulignant la visibilité croissante des données dans ce domaine.

2.3.3. L'épistémologie positiviste et formelle : valorisation par l'institutionnalisation d'un « *One Best Way* » en GRH

L'utilisation des données en GRH suscite de nombreux débats épistémologiques et méthodologiques, divisés principalement en deux grandes perspectives (Coron, 2019d). D'une part, l'approche positiviste, adoptée par des disciplines telles que la psychologie et l'économie, repose sur la conviction que les phénomènes humains peuvent être observés et mesurés objectivement, offrant ainsi une compréhension fidèle et prédictive des comportements humains au travail. D'autre part, les courants constructivistes et interprétativistes en GRH remettent en question cette objectivité, soulignant que la complexité multidimensionnelle de l'expérience humaine échappe à une capture complète par les seules données numériques. Ces courants plaident pour une approche plus holistique qui intègre les contextes sociaux, culturels et subjectifs.

Dans ce cadre, la spirale *data-driven* tire profit de l'approche positiviste, employant les données comme un levier pour transformer la fonction RH en un pilier stratégique axé sur la performance (Fitz-enz, 2010). Cette influence de l'approche positiviste se manifeste à travers deux tendances actuelles et interconnectées (Greasley & Thomas, 2020; Harley, 2015) :

1. Une focalisation croissante de la recherche en GRH autour du paradigme positiviste ;
2. La montée en puissance de la pratique fondée sur les preuves (*Evidence-Based Management* - EBM).

La première tendance en GRH provient donc principalement de la psychologie et de l'économie, deux disciplines aspirant à la rigueur des sciences physiques. En mettant un fort accent sur une épistémologie positiviste, cet héritage disciplinaire a consolidé le positivisme comme le paradigme le plus légitime, sinon le plus dominant, dans la recherche en GRH (Greasley & Thomas, 2020). À titre d'exemple, Harley (2015) cite la théorie psychologique qui examine principalement l'impact des pratiques de GRH et des caractéristiques du lieu de travail sur les attitudes et comportements individuels, tels que l'engagement, la satisfaction et la motivation. Cette théorie, souvent décrite comme « en boîtes et flèches » ou « corrélationnelle » (Alvesson &

Gabriel, 2013; Delbridge & Fiss, 2013), se concentre sur l'effet global des variables indépendantes en analysant le monde social en termes de relations linéaires et de corrélations. Cette approche tend cependant à occulter d'autres variables significatives telles que les structures de pouvoir, les mécanismes réglementaires et les acteurs collectifs comme les syndicats, qui jouent un rôle crucial dans l'explication des activités et phénomènes RH.

La deuxième tendance en GRH réside dans l'institutionnalisation d'une approche basée sur les preuves, perçue comme le moyen d'éliminer « *les préceptes discrédités, les remèdes partiels ou les cures miracles de gestion non testées »* (Pfeffer & Sutton, 2006, p. 63). Cette approche met en avant la supériorité des preuves « scientifiques » dérivées de ses méthodologies, visant à réduire les irrationalités et les biais fréquemment observés chez la fonction RH (Agrawal et al., 2019). À titre illustratif, cette dernière peut prendre des décisions d'embauche défavorables en se laissant influencer par des facteurs « non pertinents » pour la performance au travail, tels que les traits de personnalité, l'apparence ou le comportement des candidats (Broek et al., 2021). Bien que l'approche basée sur les preuves valorise une diversité de types de preuves et prône une pluralité théorique, elle est souvent critiquée pour sa conception restrictive de ce qui constitue une preuve « scientifique » valide (Greasley & Thomas, 2020).

Avec la progression croissante d'une GRH à dominante positiviste et formelle, Sandberg & Tsoukas (2011) se sont intéressés à l'écart entre les logiques théoriques et pratiques qui sous-tendent le travail académique et le travail managérial. Ils mettent en évidence que, tandis que les chercheurs adoptent une approche scientifique rationnelle d'observation et d'analyse détachées, les managers privilégient au contraire une rationalité pratique ancrée dans leurs expériences concrètes. Ils suggèrent également que cette différence de logiques rend les idées basées sur la rationalité scientifique d'une valeur discutable pour les praticiens de la gestion.

Malgré ces critiques, la prévalence du paradigme positiviste et formel continue néanmoins de s'affirmer. Conjuguée à la prolifération des données numériques et à l'accélération des avancées technologiques, cette prévalence des approches scientifiques formelles converge pour inaugurer un nouveau paradigme dans la recherche scientifique : celui de la science pilotée par les données, ou *data-driven* (Kitchin, 2014, 2022f). Ce nouveau paradigme peut être interprété comme une ré-

émergence ou un renouveau de l'empirisme. Il se caractérise par une capacité accrue des données numériques, ainsi que des technologies et méthodes analytiques qui leur sont associées, à produire des connaissances dépassant les cadres théoriques traditionnellement appliqués dans la recherche. Dans ce contexte, Prenski (2009), cité par Kitchin (2022f, p. 115), souligne :

« […] *les scientifiques n'ont plus besoin* […] *de construire des hypothèses et des modèles et de les tester par des expériences basées sur des bases de données. Au lieu de cela, ils peuvent exploiter l'ensemble complet des données pour découvrir des motifs révélant des effets, produisant des conclusions scientifiques sans expérimentation supplémentaire. En d'autres termes, plutôt que de tester si certains motifs ou relations hypothétiques existent dans l'ensemble de données, les algorithmes travaillent sur les big data pour découvrir des associations significatives entre les données sans être guidés par des hypothèses préalables.* ».

Cependant, les « petits » et « grands » ensembles de données numériques sur lesquels se fonde ce nouveau paradigme ne constituent pas de simples éléments extraits de manière neutre et objective des activités et phénomènes RH (Ribes & Jackson, 2013). Les données numériques sont le produit de processus complexes influencées par trois catégories de facteurs :

1. Facteurs techniques ;
2. Facteurs contextuels ;
3. Facteurs réglementaires.

Chacun de ces facteurs joue un rôle déterminant dans la forme finale des données numériques collectées (Kitchin & Lauriault, 2015). Le Tableau 4 fournit une synthèse détaillée de ces facteurs.

Tableau 4 : Facteurs influant sur les données (Kitchin & Lauriault, 2015)

Catégories	Facteurs	Descriptions
Techniques	Champ de vision ou cadre d'échantillonnage	Détermine quelles données sont collectées, en fonction de l'emplacement et des paramètres des dispositifs de collecte.

Catégories	Facteurs	Descriptions
	Technologies et plateformes numériques utilisées	Affecte la nature des données recueillies en raison des caractéristiques spécifiques des technologies et plateformes numériques utilisés.
	Ontologie des données	Influence la manière dont les données sont définies, mesurées et catégorisées, impactant ainsi leur interprétation.
Contextuels	Contexte de génération des données	Conditionne les données selon les événements et les circonstances prévalant au moment de leur collecte.
Réglementaires	Environnement réglementaire	Régit la manipulation et l'utilisation des données en fonction des lois sur la confidentialité, la protection et la sécurité des données (variables selon les juridictions).

Ainsi, l'émergence du paradigme *data-driven* soulève de nouvelles questions, tant ontologiques qu'épistémologiques, car la signification et la représentation des données numériques font l'objet de débats, de contestations et de négociations (Saifer & Dacin, 2022). En effet, cette idée suppose que les données puissent « parler d'elles-mêmes ». Par conséquent, elle implique que quiconque maîtrisant les technologies et méthodes analytiques appropriées serait en mesure de les construire et de les interpréter sans tenir compte du contexte ou des connaissances spécifiques au domaine dont elles proviennent (Kitchin, 2014, 2022f).

2.4. Acte IV : construction des données RH

Les données RH doivent être considérées comme le produit de cadres discursifs et de médiations socio-techniques (Kitchin & Lauriault, 2018; Saifer & Dacin, 2022). Elles sont construites à partir de catégories et d'échelles de mesure, de protocoles,

ainsi que de normes et de conventions spécifiques. Les choix des acteurs intervenant dans le processus de construction des données s'inscrivent également dans un environnement opérationnel marqué par des dispositifs socio-numériques[8] qui comprend des éléments tels que : des contextes culturels, des cadres épistémiques dominants et des systèmes et des pratiques déjà établis (Kitchin & Lauriault, 2018).

Ainsi, les données ne représentent pas de manière neutre et objective les composantes de la GRH ; elles constituent une construction partielle de celle-ci. Comme l'affirme Borgman (2017b, p. 17) : « *Les données ne sont ni la vérité ni la réalité, elles peuvent être des faits, des sources de preuves ou des principes d'argumentation utilisés pour affirmer la vérité ou la réalité* ». Cette citation souligne que, pour appréhender la « vérité » et la « réalité » du passage vers une GRH *data-driven*, il est essentiel de (re)connaître à la fois les acteurs impliqués et les finalités plurielles de l'instrumentation sous-jacente à la construction des données RH.

2.4.1. Les *data scientists* : travailleurs d'une construction « partielle » des données

Il y a dix ans, Davenport & Patil (2012) publiaient l'article intitulé « *Data scientist : le métier le plus sexy du 21ème siècle* ». Dans cet article, ils définissaient les *data scientists* comme : des « *professionnels de haut niveau avec la formation et la curiosité nécessaires pour faire des découvertes dans le monde des big data* » (p.2). Aujourd'hui, la demande pour ces travailleurs des données atteint des niveaux sans précédent avec la croissance de l'intégration des IA dans les organisations. Entre 2012 et 2021 au Canada, au Royaume-Uni et aux Etats-Unis, le nombre d'offres d'emploi en ligne visant les *data scientists* a été multiplié par 40 (OECD, 2022).

Pour soutenir cette dynamique, le 29 mars 2018, le Président de la République française a introduit une politique ambitieuse, annonçant le lancement d'une Stratégie Nationale pour les IA (SNIA). Cette stratégie, bénéficiant d'un budget de 1,5 milliard d'euros sur une période de cinq ans (2018-2022), visait trois objectifs (Inria, 2018)[9] :

[8] Compte tenu du sujet de cette thèse, je privilégie la notion de « socio-numérique » plutôt que « socio-technique » lorsque j'aborde les dispositifs selon la Théorie de l'Acteur-Réseau (ANT).

[9] Inria (2024). *French national artificial intelligence research program*. https://www.inria.fr/en/french-national-artificial-intelligence-research-program (consulté le 17/01/2024).

1. Atteindre l'excellence scientifique dans le domaine en formant et en attirant les meilleurs talents mondiaux ;
2. Promouvoir largement l'intégration des IA dans l'économie et la société, notamment par le biais de *start-ups*, de partenariats public-privé et du partage de données ;
3. Établir un cadre éthique pour régir son utilisation.

Parallèlement à la montée en popularité du travail en *data science*, un nombre croissant de chercheurs en SI (Aaltonen & Tempini, 2014; Østerlie & Monteiro, 2020; Parmiggiani et al., 2022) s'attellent aujourd'hui à investiguer les coulisses du travail des données. En effet, ce dernier dépasse la simple exécution d'algorithmes pour la création de modèles analytiques. Le monde « réel » se manifeste que très rarement sous une forme préstructurée, compliquant sa représentation à travers une approche basée sur les données (Passi & Jackson, 2018). De fait, un investissement substantiel est requis de la part de ces travailleurs pour filtrer, compléter et organiser les données en vue de leur exploitation.

Le domaine de la *data science* repose sur l'utilisation de systèmes largement distribués pour la production de connaissances (Parmiggiani et al., 2022). Pour Leonelli (2016), ces systèmes se caractérisent généralement par quatre aspects distincts.

1. Une forte dépendance aux technologies et infrastructures numériques : qui facilitent le stockage, l'intégration, l'exploration et l'analyse des données.
2. La mobilisation d'une diversité d'expertises dans le traitement des données : qui englobe les connaissances spécifiques au domaine pour l'interprétation des données, ainsi que des compétences en informatique, en programmation, en statistiques et en visualisation.
3. Un fonctionnement dans divers contextes : depuis la collecte de données par des non-spécialistes dans le cadre d'initiatives citoyennes jusqu'aux activités dans des établissements publics et privés avec des missions et spécialisations variées.
4. Des interdépendances complexes : entre institutions, gouvernements, industries et réseaux engagés dans le développement et/ou l'utilisation d'infrastructures numériques interopérables, de données et d'algorithmes réutilisables, ainsi que de normes communes.

Du fait de sa nature distribuée, qui s'appuie sur des réseaux étendus de travailleurs et de technologies fonctionnant souvent de manière indépendante, la *data science* peut être assimilée à un artisanat, mélangeant planification, essais et erreurs (Suchman & Trigg, 1993). Comme tout artisanat, elle nécessite un travail collaboratif, discrétionnaire et situé (Passi & Jackson, 2018; Passi & Sengers, 2020). La nature adaptative et expérimentale de la *data science* est précisément ce qui confère aux pratiques de ses travailleurs une grande flexibilité. Ces pratiques ne sont pas contraintes par des règles, mais se basent sur elles (Passi & Jackson, 2017).

Selon Miller, (2010), le travail des données transcende la simple automatisation activée par un presse-bouton ; elle requiert un ensemble de compétences spécialisées, une expertise approfondie et des capacités de réflexion critique ; des qualités que les machines ne possèdent pas encore. Ce travail complexe englobe plusieurs étapes essentielles dans le cadre de la « découverte » de connaissances dans les bases de données, un domaine également désigné par l'acronyme « KDD » pour « *Knowledge Discovery from Database*s » (ibid.). Ce dernier est illustré dans le Tableau 5.

Tableau 5 : Étapes du travail des données d'après le KDD (Miller, 2010)

Étapes du travail des données	**Définitions**
Sélection des données	La première étape consiste à identifier un sous-ensemble spécifique de données dans une base. L'objectif est de faciliter la découverte de connaissances pertinentes.
Prétraitement des données	La deuxième étape inclut la mise en qualité des données sélectionnées pour éliminer les interférences, supprimer les données dupliquées et résoudre les problèmes liés aux données manquantes ou aux violations de normes. Cette phase peut également inclure l'enrichissement des données par l'intégration de sources externes telles que les données de recensement ou de marché.

Étapes du travail des données	Définitions
Réduction et projection des données	La troisième étape réduit la dimensionnalité des données à travers des transformations qui produisent des représentations plus compactes et efficaces de l'ensemble des données.
Fouille ou analyse des données	La quatrième étape implique l'application d'algorithmes pour détecter des motifs dans les données.
Interprétation des données et rapport	La cinquième et dernière étape englobe l'évaluation, la compréhension et la communication des informations extraites des données, afin de les rendre utiles dans des contextes pratiques.

Ces cinq étapes soulignent l'importance d'identifier le « signal » au sein des données. Ce signal doit être compréhensible par les humains, applicable à des situations générales, potentiellement utile et innovant (Miller, 2010). Cette perspective met ainsi en évidence le caractère incomplet et conditionnel du travail des données. En effet, elle souligne l'arbitrage dans les choix opérés par les *data scientists* élaborés à partir d'hypothèses et de méthodes analytiques pour concilier le monde « réel » avec ses représentations abstraites (Passi & Jackson, 2017).

2.4.1.1. La partialité inhérente au travail des données : marginalisation des dimensions humaines et organisationnelles et ambiguïté des frontières professionnelles des *data scientists*

Malgré une prise de conscience croissante parmi les *data scientists* de l'importance des dimensions humaines et organisationnelles dans leurs pratiques (Zook et al., 2017), ces dernières demeurent souvent sous-valorisées et invisibles lors de l'évaluation de leur travail (Passi & Sengers, 2020). L'étude récente de Smaldone et al. (2022) révèle une tendance persistante à survaloriser les compétences techniques, telles que le traitement des données, les IA et les algorithmes, reléguant ainsi les compétences humaines et organisationnelles à un rôle secondaire.

Cette omission limite non seulement une compréhension holistique du profil des *data scientists* mais souligne également un décalage par rapport aux compétences interdisciplinaires essentielles à cette profession. Cette situation est confirmée par l'analyse de cinq études de cas réalisées par Saltz & Grady (2017), révélant une grande variété d'approches dans l'adoption et l'application du rôle des *data scientists*. En somme, la diversité observée, couplée au déséquilibre dans la valorisation des compétences, rend ainsi difficile l'élaboration d'une définition précise de leur rôle (Smaldone et al., 2022).

L'absence d'une définition uniforme prend une importance particulière dans un contexte où les frontières professionnelles s'estompent, notamment lorsque ces travailleurs opèrent dans des cabinets de conseil et endossent le rôle de consultants en *data science*. En tant que consultants, ils occupent une position centrale au sein des organisations. Certaines études soulignent la capacité des consultants à performer les fonctions de divers rôles (Sturdy et al., 2009). D'autres les considèrent comme des agents catalyseurs dans la création, la diffusion et l'adoption de nouveaux concepts, en grande partie en raison des implications commerciales qu'ils engendrent (Cabantous & Gond, 2012). De ce fait, ils participent à la popularisation de nouvelles pratiques professionnelles (Madsen & Slåtten, 2019; Marler et al., 2017) ce qui pose des défis supplémentaires quant à la définition et à la (re)connaissance de leurs compétences.

En conséquence, une redéfinition du rôle des *data scientists* s'avère nécessaire pour promouvoir une approche plus intégrative. Cette démarche est particulièrement cruciale face aux défis croissants liés au traitement et à la gestion des données, notamment en GRH, où une manipulation éthique et responsable des données personnelles des salariés est non seulement attendue mais également réglementée (Coron, 2019f).

2.4.1.2. La partialité inhérente au travail des données : conscientisation aux enjeux éthiques des *data scientists*

L'éthique, pilier essentiel de la philosophie, vise à systématiser, défendre et promouvoir des conceptions de ce qui constitue un comportement juste et injuste (Kitchin, 2022c; Zook et al., 2017). Selon Fieser (2003) et cité par Kitchin (2022c, p.

268), les théories éthiques peuvent être classées en trois catégories interdépendantes :

1. La *métaéthique* : qui examine l'origine et la signification des principes éthiques ;
2. L'éthique *normative* : qui établit les normes morales et définit comment les choses devraient être et ;
3. L'éthique *appliquée* : où les concepts de la *métaéthique* et de l'éthique *normative* sont opérationnaliser pour aborder et résoudre des problèmes spécifiques et controversés.

Dans le prolongement de la classification de Feiser (2003), l'éthique des données, telle que décrite par Boyd & Crawford (2012), illustre une application contemporaine des principes éthiques qui élargit les frontières traditionnelles de cette discipline. Initialement concentrée sur des préjudices tangibles comme la douleur physique ou la réduction de l'espérance de vie, la recherche éthique s'étend désormais à des domaines moins palpables mais tout aussi significatifs, tels que l'impact sur la vie privée et la discrimination par les données (Metcalf & Crawford, 2016). Ce déplacement souligne l'importance de l'éthique appliquée dans le traitement des enjeux propres à la *data science*, notamment en ce qui concerne les pratiques de collecte, de gestion et d'utilisation des données numériques par ses travailleurs.

Dans leur étude sur la recherche biomédicale, Metcalf & Crawford (2016) soutiennent que l'éthique des données redéfinit notre perception de ce qui constitue une « donnée ». En étendant la notion traditionnelle de l'individu à des groupes ou des catégories de personnes, cette nouvelle approche transforme radicalement notre compréhension des données. Celles-ci sont désormais théoriquement considérées comme : « *infiniment connectables, indéfiniment réutilisables, continuellement actualisables, et aisément dissociables de leur contexte initial de collecte* » (p.2). Ce bouleversement appelle à une réévaluation des normes éthiques traditionnelles, auparavant limitées par des contraintes temporelles, contextuelles, techniques et financières.

Selon Manroop et al. (2024), le travail des données introduit également des dilemmes éthiques pour la GRH. Les auteurs citent, par exemple, les controverses liées à *Facebook* qui a échangé des données avec des entreprises technologiques sans l'accord des utilisateurs (LaForgia et al., 2019) ou bien les discriminations raciales

observées dans les systèmes de reconnaissance faciale (National Institute of Standards and Technology, 2019). Ces pratiques entraînent potentiellement un profilage des salariés puisqu'elles se basent sur des données RH sensibles tels que l'âge, la nationalité, le genre ou le statut socio-économique (Manroop et al., 2024).

En tant que travailleurs des données, les *data scientists* ont ainsi accès à ces dernières sans jamais intervenir directement dans la vie des salariés pour les obtenir. Ils peuvent les prédire, les déduire, ou les collecter à partir de bases de données non reliées entre elles (Metcalf & Crawford, 2016). Par conséquent, l'exploitation de ces données soulève un ensemble de questions éthiques, notamment celles sur la propriété des données, le risque de violation de la vie privée des salariés, ainsi que les limites éthiques concernant les inférences que ces derniers peuvent légitimement faire concernant les salariés et les candidats potentiels (Mittelstadt et al., 2016; Resseguier & Ufert, 2024; Zwitter, 2014).

Les préoccupations éthiques, liées aux pratiques des *data scientists*, découlent notamment du contexte historique et disciplinaire de la *data science* (Metcalf & Crawford, 2016). Cela s'explique par le fait que ses disciplines fondatrices, comme l'informatique, les mathématiques appliquées et les statistiques, n'ont historiquement pas pris en compte qu'elles conduisaient des recherches impliquant des sujets humains. Bien que les statistiques représentent *in fine* des individus, les recherches en mathématiques n'ont que rarement soulevé les préoccupations relatives aux sujets humains intégrées dans les réglementations éthiques de la recherche. Ces définitions réglementaires, qui reposent sur des suppositions éthiques et épistémiques, sont désormais remises en question par les méthodes associées aux *big data* (ibid.).

Dans ce contexte, les chercheurs en études critiques plaident de plus en plus en faveur d'une approche proactive qui va au-delà de la simple critique du travail des données. Ils mettent en avant la nécessité de développer de nouvelles « réponses - habilités[10] » qui incluent un engagement plus profond et des relations de soin dans la recherche et le travail des données (Zakharova & Jarke, 2024). Loukissas (2019) décrit son propre engagement avec les données comme étant guidé par une « éthique du

[10] L'expression « réponses - habiletés » est utilisée pour enrichir la compréhension du concept de « responsabilité », en soulignant que les responsabilités sont plus facilement assumées lorsque les compétences adéquates soutiennent les réponses apportées au travail des données.

soin », soulignant ainsi l'importance croissante d'une approche éthique et responsable dans la manipulation des données :

« *Contrairement à la réflexion critique, le soin embrasse l'affect, l'engagement matériel et une multitude de préoccupations parfois invisibles dans le travail conventionnel avec la technologie. Le soin est critique en ce qu'il attire l'attention sur des choses négligées. Mais il est plus que la réflexion critique ; c'est une pratique en action. En poursuivant les opportunités non seulement pour la réflexion critique sur les données, mais également en soutien au soin, j'espère intégrer des sensibilités locales largement méconnues et non récompensées dans les efforts pour comprendre les données* » (p. 9).

Ce qui distingue spécifiquement l'éthique du soin dans le contexte des données, c'est la compréhension hautement située, influencée par la pratique et l'affect, qui s'inscrit dans la construction des données. L'acte de construire de « bonnes » données est souvent abordé sous l'angle de la qualité des données et du travail de nettoyage nécessaire. Cependant, ces données incorporent également des éléments affectifs et des pratiques techno-politiques, soulevant ainsi la question de la valeur morale émanant des « bonnes » données. Pour ces auteurs, les émotions et les expériences personnelles contribuent à la signification et à la représentation de ces données (Lindén, 2021; Zakharova & Jarke, 2024).

Dans le prolongement de cette réflexion, Lupton (2020) propose une approche épistémologique de « penser avec soin » qui vise à sensibiliser aux dimensions affectives, sociales, culturelles et politiques du travail des données. Elle soulève des questions telles que : « *Quelles sont les données considérées comme pertinentes et dans quel but sont-elles utilisées ?* » Ces questions déplacent ainsi l'attention vers les expériences subjectives et affectives des pratiques des acteurs (travailleurs et utilisateurs) des données.

Le « bon soin » des données n'est donc pas neutre puisqu'il s'inscrit dans un environnement marqué par des contextes individuels et culturels, des cadres épistémiques dominants et des systèmes et pratiques déjà établis (Kitchin & Lauriault, 2018). Il est donc essentiel de prendre en compte les dispositifs socio-numériques dans l'étude de la construction des données RH. En effet, ces dernières sont

intrinsèquement façonnées par les dispositifs qui les produisent, agissant comme des instruments de ces derniers.

2.4.2. La fonction instrumentale des données : construction aux finalités « plurielles »

L'instrument se définit comme un moyen ; conçu pour un usage spécifique et exempt de finalité propre (Gilbert, 2021). Alors que les données RH se distinguent par leur fluidité, manifestée par leur capacité à être constamment révisées, adaptées, renouvelées, étendues et recontextualisées, elles opèrent néanmoins comme des relais techniques des dispositifs socio-numériques qui président et encadrent leur construction.

Dans le domaine de la recherche critique sur les données, certains chercheurs, tels qu'Alaimo & Kallinikos (2022, 2024), mettent en lumière les diverses fonctions que les données peuvent performer. Ces derniers ont par ailleurs développé une typologie centrée sur trois fonctions principales :

1. La fonction sémiotique : qui se concentre sur la construction de sens ;
2. La fonction épistémique : qui implique la production de connaissances ;
3. La fonction communicative : qui vise à faciliter l'échange d'informations.

Mon étude se concentre principalement sur l'analyse de la fonction épistémique telle qu'explicitée par les chercheurs et explore sa connexion avec une quatrième fonction :

4. La fonction instrumentale : qui concerne l'utilisation pratique des données pour atteindre des objectifs spécifiques.

Ainsi, la mise en lumière de ces quatre fonctions, s'étendant de la construction de sens à l'application pratique des données, en passant par la production et l'échange de connaissances, illustre non seulement une progression itérative mais également une interdépendance. Ensemble, elles forment une spirale fonctionnelle, où chacune influence et conditionne mutuellement les autres, enrichissant ainsi notre compréhension de la nature intrinsèque des données.

En ce qui concerne la fonction instrumentale, cette dernière s'articule autour de trois finalités distinctes :

1. Le capital ;
2. La légitimité ;
3. La surveillance.

L'accent mis sur ces finalités met en évidence la diversité des effets de l'écosystème numérique RH.

2.4.2.1. Construction des données comme instruments du capital

L'utilisation des données considérées comme des instruments du capital, souvent décrites comme le « nouveau pétrole », met en évidence leur impact sur la croissance économique (Nolin, 2019). Bien que cette métaphore soit critiquée pour sa simplicité (Bolin, 2022), elle illustre néanmoins l'importance croissante des données RH qui permettent d'extraire de la valeur des informations concernant les salariés (Cousineau et al., 2023; Zuboff, 2015, 2019). Cette dynamique économique favorise ainsi l'expansion des technologies numériques dans le domaine de la GRH, rendant ainsi l'accès aux données cruciales pour la viabilité des acteurs présents sur le marché (Bolin, 2022; Sadowski, 2019).

Parallèlement, la croissance exponentielle et la diffusion accrue des données définissent une ère où leur importance est impérative et aussi sujette à une fétichisation, caractéristique de la spirale *data-driven* (Fourcade & Healy, 2016). En conséquence, la GRH est soumise à une pression constante de la part d'acteurs extérieurs à l'organisation, l'incitant à adopter des pratiques basées sur les données de plus en plus sophistiquées (Madsen & Slåtten, 2019; Marler et al., 2017).

La littérature reconnaît les dimensions sociales, politiques et économiques des données, qui sont conceptualisées comme des marchandises dans les transactions entre les différents acteurs (Parker et al., 2016). Cependant, Sadowski (2019) élargit cette perspective en soutenant que les données doivent être considérées non seulement comme des instruments du capital, mais aussi comme une forme de capital à part entière. Selon lui, bien que les données ne soient pas directement équivalentes aux profits, elles adhèrent à une logique similaire d'accumulation et de circulation, soulignant leur valeur grâce à leur potentiel d'utilisation répétée.

La logique de valorisation économique sous-jacente à la construction des données repose sur trois propriétés caractéristiques de toutes les ressources numériques (Floridi, 2010 cité par Kitchin, 2022d, p. 13) :

1. La non-rivalité des données : plusieurs utilisateurs peuvent exploiter simultanément les mêmes données, une caractéristique inexistante pour les biens matériels.
2. La non-exclusivité des données : le partage des données s'effectue aisément, nécessitant des mesures spécifiques pour restreindre ce partage, telles que l'application de droits de propriété intellectuelle ou la mise en place de barrières financières.
3. Le coût marginal nul des données : lorsque les données sont rendues accessibles, leur reproduction engendre généralement un coût marginal insignifiant.

Ainsi, les données, en tant qu'instruments du capital, exercent une influence croissante sur les différentes strates de l'organisation, y compris la GRH, en raison de leur valeur commerciale. Cette valeur, qui repose sur l'accès et l'accumulation des données, est indispensable pour assurer la viabilité des acteurs du marché numérique RH, en leur permettant de développer de nouveaux services.

2.4.2.2. Construction des données comme instruments de légitimation

La légitimité est conceptualisée comme le degré d'acceptation sociale d'un phénomène (Suchman, 1995). Elle revêt une importance particulière en GRH où elle sert à la fois de processus et de résultat (Belizón & Kieran, 2022).

La fonction RH est régulièrement confrontée à la nécessité de prendre des décisions importantes dans la gestion de ses activités, chaque décision ayant un impact significatif sur les salariés. Ces décisions doivent être solidement justifiées pour garantir leur légitimité, accroître leur acceptabilité et minimiser les contestations internes (Coron, 2019c). En outre, les obligations légales et éthiques notamment en matière de non-discrimination exigent que ces décisions soient de plus en plus étayées quantitativement, ce qui place les données au cœur de la légitimation des pratiques en GRH. Ces données : « *facilitent la production de rapports, contribuent à la constitution de dossiers probants en cas de litiges, et aident à atténuer les biais* » (ibid., p.28). Dans ce contexte, il devient impératif de projeter une image de « neutralité » et

d'« objectivité » dans la prise de décision en GRH (Coron, 2019g). Cette exigence est d'autant plus accentuée par les discours prédominants au sein de l'organisation, qui s'appuient de plus en plus sur une épistémologie positiviste et formelle (Greasley & Thomas, 2020).

Les données RH utilisées se classent en trois catégories principales et servent des objectifs spécifiques (Coron, 2019d) :

1. Les données sur les salariés : comprend essentiellement les données socio-démographiques et de performance, qui sont utilisées pour la classification et l'évaluation des salariés.
2. Les données sur le travail : comportent les données sur la classification des emplois pour permettre l'évaluation de la charge de travail.
3. Les données sur l'activité de la fonction RH : inclut les données qui permettent de mesurer la manière dont les politiques et les pratiques RH influencent la performance globale de l'organisation.

Bien que les données soient utilisées comme des instruments de légitimation de la fonction RH, la fragilité historique de ses revendications quant à sa contribution à la performance expose cette fonction à des enjeux de confiance auprès de ses parties prenantes (Caldwell, 2003). En effet, cette quête continue de légitimité a engendré de nombreux débats sur la nécessité de moderniser ce qui est souvent perçu comme le « parent pauvre » des professions de gestion (Wright, 2008, p. 1066).

Ainsi, les données RH ont ravivé le débat sur la légitimité de cette fonction au sein de l'organisation (Angrave et al., 2016). En réponse, des chercheurs tels que Belizón & Kieran (2022) et Rasmussen & Ulrich (2015) préconisent de décloisonner les données RH en adoptant une approche « *outside-in* ». Selon cette approche, les données atteignent leur plein potentiel lorsqu'elles sont utilisées comme instruments de légitimation et intégrées au-delà des frontières traditionnelles de la GRH. Toutefois, comme l'indiquent Belizón & Kieran (2022), à travers un exemple de décentralisation des données RH vers les unités commerciales, une telle intégration nécessite impérativement que la fonction RH ait préalablement consolidé sa propre légitimité.

2.4.2.3. Construction des données comme instruments de surveillance

L'évolution du capitalisme prend différentes formes, notamment celle du capitalisme de surveillance (Bolin, 2022). Selon la perspective développée dans la section 2.4.2.1, les données ne devraient pas être simplement perçues comme des éléments préexistants destinés à être traités. Plutôt, elles devraient être envisagées comme des instruments d'investissement au sens large (Ruppert et al., 2017).

Cette vision est incarnée dans l'utilisation des données RH comme instruments de surveillance, lesquelles ont historiquement contribué à structurer la gestion des salariés par leur profilage, leur classification et leur segmentation (Boyd & Crawford, 2012). Cet usage s'aligne sur les travaux d'Hacking (1986, 1995), qui a introduit le concept de « fabrication des individus » pour illustrer comment ces derniers se définissent et sont définis par les données.

L'utilisation des données RH en tant qu'instruments de surveillance, établit un cadre où l'accumulation et l'exploitation de ces données devient une logique dominante, influençant les comportements et notamment ceux des salariés. Ces données, principalement générées par les salariés eux-mêmes, facilitent leur contrôle en les rendant quantifiables à travers leur classification dans des catégories RH prédéfinies (Bowker & Star, 1999; Coron, 2019c; Sadowski, 2019).

La surveillance va ainsi de pair avec la notion de contrôle. La théorie du contrôle organisationnel se concentre sur les tentatives de l'organisation pour augmenter la probabilité que les salariés et les groupes de salariés se comportent de manière à atteindre les objectifs organisationnels (Cousineau et al., 2023).

Le contrôle exercé à travers les données RH vise à identifier, surveiller, suivre, réguler, prédire ou prescrire les comportements des salariés. Il a désormais transcendé les formes traditionnelles de contrôle technique (e.g. fréquence et durée des tâches) et bureaucratique (e.g. règles et descriptions de postes) pour intégrer le contrôle numérique (ibid.). En effet, le développement des technologies numériques de surveillance des salariés a notamment été amplifié par la généralisation du télétravail et des horaires flexibles induits par la pandémie de COVID-19. Les auteurs

ont systématiquement classé ces technologies en 21 formes, regroupées en trois catégories distinctes :

1. La surveillance de l'activité numérique ;
2. La surveillance par caméra et audio ;
3. La localisation et la surveillance bio-physique.

Ces chercheurs ont également évalué leur impact en termes d'invasivité personnelle et sociale.

En tant qu'instruments de surveillance, les données RH suscitent en effet de nombreuses interrogations concernant la vie privée ainsi que les droits à l'anonymat et à la confidentialité des salariés. Cette préoccupation découle notamment du potentiel de ces dernières à générer de la valeur et à devenir une classe d'actifs économiques (Ruppert et al., 2017).

En résumé, l'exploration des trois finalités des données révèle la complexité de l'écosystème numérique RH. Elle expose la diversité des interactions tant internes qu'externes qui marquent la transition vers une GRH *data-driven*. Cette dynamique souligne le rôle central des données, lesquelles transcendent désormais leurs frontières traditionnelles et offrent une perspective renouvelée sur le plan instrumental mais aussi épistémique.

2.4.3. La fonction épistémique des données : construction génératrice de connaissances

La fonction épistémique des données RH se manifeste par leur capacité à produire des connaissances. Par exemple, le classement des emplois ou les évaluations des salariés illustrent des types de connaissances qui sont largement diffusés et appliqués au sein des diverses activités RH.

Les données RH sont perçues comme des artefacts sémiotiques utilisés pour capturer, représenter, connaître et agir sur la GRH (Alaimo & Kallinikos, 2022; Østerlie & Monteiro, 2020). Elle sont en outre perçues comme une ressource omniprésente et un moyen que la fonction RH doit appréhender afin de répondre aux contingences internes et externes auxquelles elle est confrontée (Angrave et al., 2016; Rasmussen & Ulrich, 2015).

Les données RH acquièrent cette fonction épistémique du fait de leur incorporation au sein d'une infrastructure de connaissances RH, englobant systèmes, technologies, outils, professions, processus et pratiques (Alaimo & Kallinikos, 2022, 2024; Cadin et al., 2012). La partialité inhérente à ces données peut être attribuée à la complexité de cette infrastructure qui préside la génération de nouvelles connaissances RH. Cette partialité peut se manifester à travers les choix de conception des technologies, des dépendances avec celles existantes, des convictions établies ainsi que des objectifs en matière de GRH. Ces facteurs peuvent élargir ou restreindre les options quant à ce qui peut être enregistré sous forme numérique et, par conséquent, ce qui peut être élaboré en tant que connaissances RH (Alaimo & Kallinikos, 2022). Pour certains chercheurs (Carter, 2018; Kitchin & Lauriault, 2018), cette infrastructure est désignée par le terme d'assemblage de données, en résonance avec le concept de « dispositif » développé par Foucault (1977). Ils décrivent cet assemblage comme un :

« […] *système socio-technique* [numérique] *complexe composé de nombreux appareils et éléments étroitement imbriqués, dont le centre concerne la production, la gestion, l'analyse et la traduction de données et dont les produits d'information sont dérivés à* [diverses] *fins, commerciales, administratives, etc.* » (Kitchin & Lauriault, 2018, p. 8).

L'accélération exponentielle de la spiale *data-driven*[11] va favoriser l'homogénéisation des connaissances et leur production (Yoo et al., 2010). En effet, la conversion des activités et phénomènes RH en données numériques permet une manipulation de ces données avec une certaine indépendance vis-à-vis du domaine de la GRH et de ses contextes spécifiques. Cela les rend plus facilement transportables et moins enclines à dépendre des modalités de connaissances RH existantes.

Pour Alaimo & Kallinikos (2022), ces avancées desserrent l'étau des connaissances du domaine d'origine et rééquilibrent l'importance relative entre les références internes et externes à ce domaine. Ils qualifient cette évolution généralisée de « décentrement des organisations » (p.6). De plus en plus significative, cette

[11] Il convient de rappeler que la spirale *data-driven* fait référence à l'interconnexion des trois phénomènes transformationnels à l'échelle de l'organisation : (1) la numérisation, (2) la digitalisation et (3) la datafication.

évolution affaiblit l'emprise traditionnelle des types de connaissances établis et des données générées par le domaine (de GRH, par exemple), redéfinissant ainsi le processus de production de connaissances (Monteiro & Parmiggiani, 2019; Pachidi et al., 2021).

Le Tableau 6, élaboré par Alaimo & Kallinikos (2022), offre une synthèse éclairante des trois caractéristiques des données qui influent sur la production des connaissances.

Tableau 6: Caractéristiques épistémiques des données numériques (Alaimo & Kallinikos, 2022)

Caractéristiques	Définitions	Implications
Agnostique quant au contenu	Indifférence au contenu et au contexte de l'enregistrement.	La production de données se fait sans référence précise aux connaissances du domaine et aux méthodes spécifiques.
Partiale	Les prédispositions sont intégrées dans les décisions de conception des dispositifs.	Les technologies et les objectifs organisationnels imposent des biais, limitant ce qui peut être encodé et élaboré comme données.
Homogénéisante	Traduction des conventions culturelles et des systèmes d'information dans le langage des machines.	Toutes les données peuvent être traitées de la même manière, réduisant les différences entre divers domaines de connaissances et d'industrie.

L'analyse de la spirale fonctionnelle des données RH met en lumière la nécessité de dépasser la perspective traditionnelle dans la production de connaissances (Weinberger, 2010)[12]. En effet, les données sont désormais reconnues comme des produits manufacturés porteurs de valeurs, d'influences et de rationalités (Ruppert et al., 2017).

[12] La hiérarchie de la production de connaissances est souvent représentée sous forme pyramidale, où les données se trouvent à la base et les connaissances accumulées au sommet.

Les données RH ne sont toutefois pas intrinsèquement utiles. Leur potentiel réside dans leur exploitabilité, c'est-à-dire leur capacité à être manipulées de manière pertinente (Greasley & Thomas, 2020). Pour agir en tant que médiums de représentation et de signification des activités et phénomènes RH, elles requièrent l'intervention d'algorithmes (par exemple, Faraj et al., 2018; Kellogg et al., 2019; Orlikowski & Scott, 2015). Ces derniers jouent un rôle crucial dans la transformation des données RH en connaissances.

La notion d'algorithme se réfère essentiellement à une séquence d'instructions finie qui nécessite des données d'entrée (Coron, 2019c). Les algorithmes privés de données ne sont rien d'autre que des exercices mathématiques (Gillespie, 2014). Ce sont les données qui, agissant comme des « capteurs », leur permettent de transcender leur fonctionnement purement calculatoire et de se connecter à la « réalité » des activités ou phénomènes à l'étude (Alaimo & Kallinikos, 2022).

Cependant, l'émergence des algorithmes d'apprentissage automatique, qui s'ajustent en fonction des données d'entrée, représente une avancée notable dans le domaine. L'apprentissage automatique se subdivise en deux catégories principales (Kitchin, 2022b) :

1. L'apprentissage « supervisé » : nécessite que l'algorithme associe les données d'entrée à des résultats connus, s'appuyant sur un ensemble de données préétablies pour l'entraînement.
2. L'apprentissage « non supervisée » : permet à l'algorithme de découvrir indépendamment des motifs et des structures dans les données sans recourir à des données d'entraînement spécifiques.

Alaimo & Kallinikos (2022) décrivent ainsi une relation symbiotique où données et algorithmes sont indissociables, représentant les deux facettes d'une même pièce. Cette symbiose permet l'application d'une variété de méthodes analytiques organisées autour de quatre types de connaissances :

1. Connaissances *descriptives* ;
2. Connaissances *explicatives* ;
3. Connaissances *prédictives* ;
4. Connaissances *prescriptives*.

Chaque type de connaissances se définit également par une question distincte (Kapoor & Kabra, 2016; Kitchin, 2022b).

1. Les connaissances RH *descriptives* : *quels événements ou phénomènes RH ont été observés et à quelles périodes ? Quelle est la fréquence de leurs occurrences ?*

Ce premier type de connaissances implique la description des événements ou phénomènes RH observés et leur périodicité sans investiguer leurs causes sous-jacentes. Son développement est principalement motivé par les exigences de reporting social en France, qui imposent la collecte et la publication de données RH quantifiées. Ces données RH issues de tableaux de bord découlent généralement d'une approche univariée ou bivariée et se présentent sous la forme de croisement entre deux données tels que l'absentéisme par âge et ancienneté (cf. section 2.3.2.2) (Coron, 2019c; Kapoor & Kabra, 2016).

2. Les connaissances RH *explicatives* : *quels sont les causes et les effets des événements ou des phénomènes RH étudiés ?*

Ce deuxième type de connaissances vise à approfondir la compréhension des événements ou phénomènes RH en analysant leurs causes et effets. Il offre une perspective analytique et argumentative pour appuyer la prise de décision. Cette approche favorise une analyse multivariée permettant d'examiner plusieurs données simultanément afin de comprendre leurs relations. Un exemple pertinent est l'étude des facteurs influençant les accidents du travail, où les données pertinentes sont sélectionnées pour leur potentiel explicatif (cf : section 2.3.2.1) (Coron, 2019c).

3. Les connaissances RH *prédictives* : *quels événements ou phénomènes RH sont susceptibles de se produire dans un futur proche ? Quels effets pourraient découler des différentes actions entreprises ?*

Ce troisième type de connaissances s'appuie sur l'utilisation d'algorithmes prédictifs pour anticiper les événements ou phénomènes RH à venir. À la différence des approches *descriptives* et *explicatives*, les connaissances *prédictives* utilisent un vaste ensemble de données historiques (cf. 2.3.1) pour projeter les occurrences futures. Un exemple illustratif serait un algorithme conçu pour prédire les risques d'absentéisme. Bien que les méthodes analytiques prédictives, telles que les

régressions linéaires ou logistiques, peuvent être communes aux approches descriptives et explicatives, elles sont utilisées dans le cas échéant pour identifier les facteurs susceptibles d'influencer l'absentéisme futur, facilitant ainsi la prise de décision (Coron, 2019c; Kapoor & Kabra, 2016).

Cependant, il reste souvent difficile de déterminer quel type d'algorithme ou quelle variante offrira a priori la meilleure performance pour un ensemble de données spécifique. Chaque algorithme présente des particularités, avec ses avantages et ses inconvénients, qui influencent la précision des prédictions et la minimisation des erreurs, en fonction de la nature spécifique du problème ou des données choisies. Cette incertitude est en partie attribuable au fait que les algorithmes sont conçus en fonction des connaissances préexistantes sur le fonctionnement d'un système (Kitchin, 2022b).

4. Les connaissances RH *prescriptives* : *quel est le résultat le plus optimal pour influencer les événements ou phénomènes RH observés ? Quels sont les moyens nécessaires pour y parvenir ?*

Ce quatrième et dernier type de connaissances s'emploie, à l'instar de la connaissance *prédictive*, à analyser les données antérieures afin d'identifier les occuurences futures et déterminer les scénarios optimaux permettant de maximiser les objectifs déterminés. Par exemple, une analyse prescriptive consisterait à recommander une action préventive spécifique pour atténuer les risques d'absentéisme au sein d'un sous-groupe de salariés (Kapoor & Kabra, 2016).

Ainsi, la fonction épistémique des données RH est structurée autour de quatre catégories principales de connaissances :

1. Connaissances *descriptives* ;
2. Connaissances *explicatives* ;
3. Connaissances *prédictives* ;
4. Connaissances *prescriptives*.

En plaçant les données au cœur de la transformation de la production de connaissances en GRH ; elle invite également à reconsidérer le rôle de la fonction RH dans le passage à une GRH *data-driven*. En effet, les trois caractéristiques des données : (1) l'agnosticisme, (3) la partialité et (3) la tendance à l'homogénéisation,

exacerbent la tension entre l'indépendance - soit la capacité à générer des connaissances sans dépendre directement de l'expertise situé de la fonction RH - et la pertinence - qui implique l'application directe de ces connaissances dans le domaine de la GRH (Broek et al., 2021).

Toutefois, la nature codifiable et prétendument indépendante du contexte des connaissances promue par la spirale *data-driven* est remise en question par des études critiques sur les pratiques qui révèlent le caractère essentiellement situé et social de la connaissance. Ces recherches mettent en exergue que les connaissances, bien que structurées pour guider l'action, nécessitent des ajustements continus et des jugements adaptés à chaque contexte. Par exemple, Tsoukas & Vladimirou (2001) observent que dans un centre d'appels en Grèce, les salariés ne se contentent pas de suivre des règles explicites ; ils ajustent constamment leur comportement en réponse à des circonstances particulières, modifiant ainsi l'intention originale des règles. Cette perspective critique suggère donc que la capacité à générer des connaissances de manière décontextualisée pourrait être largement surestimée, accentuant la nécessité d'engager un processus de production de connaissances qui soit actif et sujet à négociation.

Dans un contexte similaire en GRH, Tambe et al. (2019) remettent en question l'approche dite « indépendante » adoptée par les éditeurs de logiciels dans l'industrie de l'analytique RH. Les auteurs soulignent que ces éditeurs ne prennent pas suffisamment en compte les spécificités de la GRH, notamment ses contraintes sociales, techniques et réglementaires. En conséquence, les technologies numériques proposées échouent souvent à répondre aux besoins particuliers de la fonction RH, réduisant ainsi la pertinence des connaissances RH produites dans ce domaine.

De plus, la prudence exprimée par la fonction RH à l'égard de l'adoption de données quantitatives - potentiellement influencée par les déceptions liées aux pratiques des éditeurs - incite à revisiter les considérations ontologiques et épistémologiques en GRH. Ces considérations déterminent ce qui est reconnu comme des connaissances valides et des preuves « appropriées » dans ce domaine. Cette réflexion, soutenue par les travaux de Greasley & Thomas (2020) et Angrave et al. (2016), souligne l'importance de ces débats dans l'évaluation du passage à une GRH *data-driven*.

Pour conclure, cette première section procède à une exploration de l'écosystème des « données RH » tel que défini par la littérature. Elle met l'accent sur leur conceptualisation ainsi que sur les contextes et infrastructures dans lesquels elles évoluent. Structurée en quatre actes distincts, différentes facettes des données sont décomposées et analysées :

1. Définition : les données RH se distinguent par leur nature modifiable et décontextualisable, ce qui leur permet de véhiculer des narrations dépassant leur origine et usage initial. Cinq caractéristiques techniques ont été examinées, couvrant leurs formes, structures, sources, producteurs et types.
2. Transformation : l'accumulation croissante des données RH, à travers les phénomènes de numérisation, digitalisation et datafication, entraîne l'accélération exponentielle d'une spirale *data-driven* qui redéfinit les dynamiques organisationnelles et socio-économiques en modifiant les pratiques et structures au sein des organisations.
3. Valorisation : cette spirale *data-driven* entraîne l'intégration de données volumineuses et diversifiées aux « petites » données RH, créant ainsi un nouveau paradigme de GRH *data-driven*. Ce paradigme se caractérise par l'utilisation des *big data* RH, de l'analytique et des métriques RH, ainsi qu'une prédominance épistémologique positiviste et formelle.
4. Construction : enfin, les transformations successives et de plus en plus intensives des données RH conduisent à l'émergence de spécialistes, tels que les *data scientists*, qui travaillent au développement de diverses fonctions performées par les données RH. Ces fonctions instrumentale (capital, légitimation et surveillance) et épistémique (connaissances *descriptives*, *explicatives*, *prédictives* et *prescriptives*), illustrent ainsi l'impact des dispositifs socio-numériques sur la conceptualisation des données RH.

Dispersés entre les littératures en SI et GRH, ces quatre grands actes montrent qu'ils font partie intégrante d'un tout : la spirale *data-driven*. Cette spirale met en évidence la nécessité d'une approche théorique permettant de saisir pleinement la complexité des données RH et leurs dispositifs socio-numériques. Ainsi, l'utilisation de la Théorie de l'Acteur-Réseau (ANT) (Akrich et al., 2006) est privilégiée afin de théoriser la conceptualisation des données RH sans en réduire la complexité.

3. L'ANT pour appréhender les dispositifs socio-numériques sous-jacents à la construction des données RH

La première section visait à explorer la conceptualisation des données RH, en investiguant l'écosystème sous-jacent à ce que la littérature identifie comme des « données RH ». En mettant en lumière les quatre grands actes de conceptualisation de ces données, il a été démontré qu'ils font partie intégrante d'un tout : la spirale *data-driven*.

Cette observation souligne la nécessité d'adopter un cadre théorique qui permette de saisir pleinement les données RH et les dynamiques qui se jouent au sein de leurs dispositifs socio-numériques. Afin de théoriser leur conceptualisation sans en réduire la complexité, l'approche de l'ANT est privilégiée.

Pour ce faire, la richesse de l'ANT conduit à sélectionner les principes qui répondent précisément à ma problématique de recherche. Je me concentrerai principalement sur le rapprochement entre la production de connaissances, telle que définie par Latour (2005a, 2007), et la qualification des biens économiques, décrite sous le terme d'« économie des qualités » par Callon et al. (2000, 2002). Cette synergie conceptuelle, enrichie par l'exploration des controverses - concept central de l'ANT - structure de manière cohérente mon cadre théorique et fournit les fondations nécessaires à l'analyse de la construction des données RH.

3.1. L'ANT dite la sociologie de la traduction

La Théorie de l'Acteur-Réseau[13] est élaborée dans les années 80 par Bruno Latour et les chercheurs affiliés au Centre de Sociologie de l'Innovation (CSI) de l'École des Mines de Paris. Parmi ces chercheurs, Madeleine Akrich et Michel Callon ont également apporté des contributions significatives à cette théorie ainsi qu'à l'établissement et au développement du domaine des *Science and Technology Studies* (STS).

[13] Il convient de noter que les notions « théorie de l'acteur-réseau », « ANT », « sociologie de la traduction » et « modèle de la traduction » sont utilisés de manière interchangeable et sans distinction.

Le concept de « traduction », introduit par Callon et emprunté à Michel Serres dans les années 70 est : « *le mécanisme par lequel le monde social et naturel se met progressivement en forme et se stabilise pour aboutir, si elle* [la traduction] *réussit, à une situation dans laquelle certaines entités* […] *mettent en forme des aveux qui demeurent vrais aussi longtemps qu'ils demeurent incontestés* » (Callon, 1986, p. 205). L'introduction de ce concept vise ainsi à comprendre le processus de production et de stabilisation des connaissances scientifiques, en partant du principe que les espaces de production se constituent au cours de la construction des problématiques, indépendamment de leur contenu (Callon & Ferrary, 2006).

Une étude portant sur l'innovation des véhicules électriques est le point de départ de Callon pour introduire le néologisme d'« acteur-réseau ». Selon lui, afin de comprendre les succès ou les échecs de cette innovation, il est essentiel de reconnaître que le fonctionnement d'une voiture repose sur un réseau socio-technique. L'utilisation du terme « réseau » reflète la nécessité pour le véhicule électrique de progressivement s'articuler avec tous les éléments essentiels à sa survie et à son développement. Quant au terme « acteur », il souligne que cet environnement n'existe initialement pas et doit être imaginé et construit par le réseau lui-même. L'expression « acteur-réseau » est ainsi choisie pour sa capacité à rendre intelligible la combinaison de ces deux caractéristiques. Grâce à elle, les chercheurs du CSI s'inscrivaient encore dans un domaine qui demeurait familier à leurs pairs : « *Avec l'acteur-réseau, les chercheurs en sciences sociales étaient en pays familier : tout le monde sait ou croit savoir ce qu'est un acteur, tout le monde sait ou croit savoir ce qu'est un réseau. L'acteur-réseau était aux sciences sociales ce que le maïs hybride a été aux sciences agricoles : le changement dans la continuité.* » (Callon & Ferrary, 2006, p. 42).

Le modèle conceptuel de l'ANT, initialement élaboré par Callon (1986), qui utilise l'exemple emblématique de la raréfaction des coquilles Saint-Jacques, souligne quatre grandes étapes dans la traduction :

1. La problématisation : l'identification des acteurs et des enjeux et la définition des rôles assignés à chaque acteur.
2. L'intéressement : la conception de stratégies pour persuader les acteurs de la pertinence des rôles qui leur ont été assignés lors de la problématisation.
3. L'enrôlement : l'attribution de rôles spécifiques aux acteurs et l'officialisation de leur engagement en tant qu'alliés.

4. La mobilisation : le regroupement et la mise en action des alliés indispensables pour appuyer le projet.

Ces quatre étapes suivent une trajectoire non linéaire, de nature plutôt « tourbillonnaire », jalonnée de controverses qui permettent de mieux comprendre les facteurs de succès ou d'échec des innovations.

Dans la perspective des chercheurs en sciences de gestion, l'ANT comporte de nombreux bénéfices. Premièrement : « *La traduction est un processus avant d'être un résultat* » (Callon, 1986, p. 205) c'est-à-dire que cette théorie privilégie l'exploration du processus menant de la construction à la stabilisation des connaissances plutôt qu'à leur validité scientifique. Elle promeut également une symétrie entre les acteurs humains et non-humains, ce qui engendre un renoncement aux approches organisationnelles formelles et fonctionnelles. Cette approche privilégie davantage une prise en compte des facteurs contextuels, ainsi que des entités sociales et techniques - dans mon cas, numériques - incluant les connaissances, les individus, les organisations, les données RH, parmi d'autres. Elle valorise également la diversité des discours et des points de vue, intégrant tous ces éléments au sein d'un même espace conceptuel : le réseau (Akrich et al., 2006).

Considéré comme une configuration organisationnelle distincte, le réseau remet en question le concept traditionnel de frontières. Cette remise en question devient d'autant plus significative dans un contexte marqué par l'accélération exponentielle de la spirale *data-driven*. Dans cet environnement, l'abondance des données RH donne l'impression que la production de connaissances à leur sujet devient progressivement plus autonome, les détachant ainsi de leur contexte d'origine.

3.2. L'ANT et la connaissance des données RH

Latour s'est penché sur les conditions de production de la science (Latour, 2005a, 2007) et des techniques (Latour, 1992), articulant une vision où la connaissance est envisagée comme une trajectoire dynamique d'apprentissage. Selon lui, la connaissance fonctionne comme un vecteur qui confère rétroactivement son « *pouvoir de validation* » (Latour, 2007, p. 3), signalant ainsi que notre perception des connaissances acquises n'est pas statique, mais est, au contraire, en perpétuelle évolution.

Cette conceptualisation de la connaissance, comme processus évolutif, est particulièrement bien illustrée dans son analyse de l'exposition sur les fossiles de chevaux au Museum d'histoire naturelle (2007). Les conservateurs de cette exposition adoptent une approche novatrice en orchestrant deux séquences parallèles : la première traçant l'évolution physique des chevaux au fil du temps, et la seconde dévoilant comment notre compréhension de cette évolution a elle-même progressé.

Cette mise en scène ne révèle pas seulement la transformation physique des chevaux, mais également la dynamique de transformation de l'histoire des sciences à leur sujet, illustrant ainsi que la connaissance n'est pas un produit fini, mais un processus évolutif en constante révision (Latour, 2007). La conception de Latour résonne ainsi fortement avec la théorie de la « rectification » de Bachelard (1969), qui décrit les connaissances comme une série de corrections et de validations successives.

La notion de « connaissance » ne peut donc être définitivement circonscrite, car elle émane de la capitalisation d'éléments humains (e.g. paléontologues) et non-humains (e.g. découvertes fossiles et avancées scientifiques) accumulés au fil du temps, enrichissant et reformulant continuellement la construction des connaissances (e.g. évolution physique des chevaux). Platon résume ce principe par la maxime « *connaître, c'est reconnaître* », mettant en lumière l'importance du processus de « reconnaissance » dans la production de connaissances (cité par Latour, 2005a, p. 525).

Approfondissant cette idée, Latour (2005a) souligne que la connaissance n'émerge pas instantanément mais se développe à travers l'itération et la familiarité avec l'objet étudié : « *La première fois que nous sommes confrontés à un événement, nous ne le connaissons pas ; nous commençons à connaître quelque chose quand c'est au moins la deuxième fois que nous le rencontrons, c'est-à-dire quand il est devenu familier* » (Latour, 2005a, p. 525).

Dans la lignée de l'ANT, la construction des données RH, au centre de cette thèse, doit donc être perçue comme un processus de (re)connaissance, soulignant ainsi le caractère continu et itératif de la familiarisation avec les données. Cette familiarisation, renforcée par la capitalisation progressive d'éléments (humains et non-humains), accentue le caractère expansif de la (re)connaissance des données RH au fil de leur

construction. Ainsi, plutôt que de concevoir la construction comme un processus linéaire, je la décris comme un phénomène spiralé, amplifié par la capitalisation. Cette conceptualisation est représentée par la Figure 5.

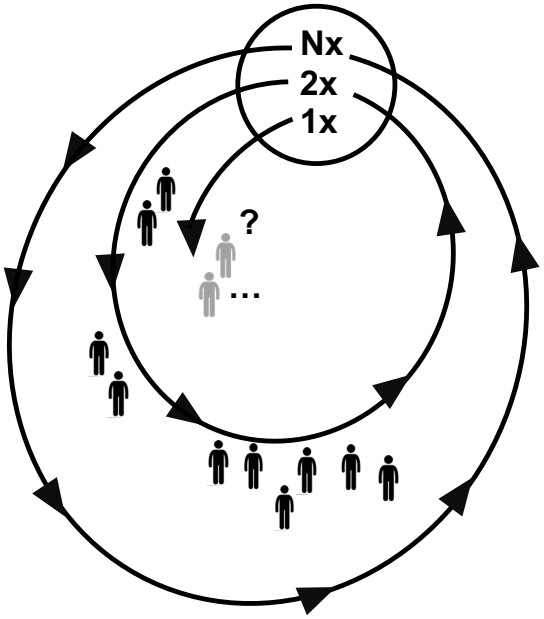


Figure 5 : Spirale de construction des données RH (adaptée de Latour, 2005a)

Au regard de cette figure, Latour (2007) met en garde contre une fixation rigide de la connaissance, argumentant que de telles pratiques scientifiques ne rendent pas justice à la nature dynamique et évolutive de la compréhension humaine. Il reproche plus spécifiquement aux épistémologues leur concentration excessive sur la lutte contre le scepticisme, souvent au détriment de l'adoption de méthodes qui permettraient des ajustements continus, l'amélioration des instruments et l'élargissement des processus de validation par les pairs et le grand public (p.10).

En préconisant une « *désépistémologisation* » et une « *réontologisation* » de l'activité scientifique, Latour (2007) plaide pour l'adoption d'une « *Providence Épistémologique* ». Cette approche, au lieu de prescrire un chemin linéaire dans la connaissance, encourage l'examen approfondi du chemin de ses controverses avant que la connaissance ne soit reconnue comme un fait établi. Les controverses sont considérées et sont vues dans cette thèse comme des lieux de négociation (Callon,

2006). Ce cadre critique ouvre ainsi la voie à une interrogation plus fondamentale sur la nature de la vérité des connaissances, un thème que Latour (2005a, 2007, 2012) a continué d'explorer en profondeur.

Il est effectivement ardu de statuer de manière absolue sur la véracité d'une connaissance. Latour (2007, 2012), dans ses réflexions philosophiques, souligne la profonde complexité de la vérité, réfutant l'idée d'une vérité monolithique. En faisant cela, il met en lumière la multiplicité des façons dont les vérités sont construites et validées, englobant les domaines juridique, moral, scientifique, religieux, politique ou technique. Cette diversité des vérités amène Latour (2012) à considérer la notion d'existence non seulement comme une manifestation particulière de l'être mais aussi comme un critère de vérification à travers l'action, intégrant ainsi la dynamique des modes d'existences dans le processus de validation des connaissances (Dambrin & Grall, 2021).

Ainsi, tout mode d'existence doit être validé. C'est pourquoi il est essentiel de mettre en évidence l'importance des « épreuves » pour Latour (2005a). Le chercheur est tenu de rendre compte de l'action des forces à l'œuvre en spécifiant la nature des épreuves et des preuves tangibles qu'elles ont laissées (documents, objets physiques, marques ou signes, etc.). En l'absence de compte rendu de transformation d'une situation donnée, de différence observable ou de cadre de référence détectable, il est impossible de déclarer qu'une entité (humaine ou non-humaine) est en action : « *Une forme d'existence invisible qui ne génère aucune existence, aucune transformation, qui ne laisse aucune trace et qui n'est mentionnée dans aucun rapport n'est pas une forme d'existence. C'est aussi simple que cela. Elle est agissante ou elle ne l'est pas* » (Latour, 2005a, p. 76).

Cette approche est donc particulièrement bien adaptée à l'analyse du processus de construction des données RH, lequel se trouve à l'intersection de plusieurs modes d'existence (Latour, 2007, 2012). Parmi ces modes, le mode économique occupe une place centrale (Callon et al., 2000, 2002).

3.3. De la (re)connaissance à la marchandisation des données RH : les marchés en tant que mode d'existence économique

En s'écartant des perspectives traditionnelles, l'ANT offre une relecture critique des modes d'existence, notamment celui économique. Elle montre comment les dispositifs peuvent influencer ou dicter les comportements d'agents économiques, qu'ils soient du côté de l'offre ou bien de la demande (Callon et al., 2000, 2002).

Cette vision s'intègre dans un cadre plus large du mode d'existence économique, où les actions, objets et processus sont rendus économiques par leur intégration dans des réseaux de valorisation, de tarification et de circuits commerciaux, les rendant ainsi économiquement commensurables et échangeables (Espeland & Stevens, 1998; Muniesa et al., 2007). C'est ce que l'on appelle le « processus d'économisation » (Callon & Muniesa, 2005; Muniesa et al., 2007).

Dans ce cadre, les actions des dispositifs économiques sont souvent décrites à l'aide de termes tels que la « *désimbrication* » et, plus spécifiquement, l'« *abstraction* », mettant en lumière comment ces actions favorisent le mode d'existence économique des réseaux (Callon & Muniesa, 2005). Toutefois, cette abstraction se manifeste principalement en interaction avec d'autres dispositifs auparavant moins dominés par une logique économique (Muniesa et al., 2007). Selon Callon & Muniesa (2005), l'abstraction devrait être interprétée non pas comme un simple adjectif, mais comme une action de transformation et de déplacement d'un dispositif vers le mode d'existence économique.

En GRH, l'abstraction se manifeste par la transformation des activités et phénomènes RH en données structurées et quantifiables. Cette transformation initie un processus d'économisation, où les données sont valorisées et tarifées par des agents économiques, tels que les éditeurs de logiciels et les plateformes numériques, intégrant ces données dans des circuits commerciaux.

Dans cette thèse, une attention particulière est accordée à une forme spécifique de dispositif au sein du mode d'existence économique : les marchés. Callon et al. (2000, 2002) mettent en lumière la transformation profonde des règles qui régissent ces derniers, qu'il caractérise comme des « forums hybrides ». Cette hybridité, issue

de la diversité et de l'hétérogénéité des agents impliqués, permet aux marchés d'interroger simultanément divers modes d'existence. Les marchés se trouvent donc à la croisée de multiples modes d'existence, intégrant la dynamique de ces modes dans leurs processus de (re)validation.

La relation entre les marchés et les technosciences est intrinsèquement liée à l'architecture du capitalisme. Cette interaction est particulièrement évidente dans le contexte actuel de la spirale *data-driven*, où les données ne sont plus seulement considérées comme des instruments du capital, mais suivent également un processus d'économisation, évoluant vers une forme de capital à part entière (cf. fonction instrumentale) (Ruppert et al., 2017).

Comme tout dispositif engagé dans un processus d'économisation, les marchés fonctionnent comme des collectifs visant à parvenir à des compromis (Callon & Muniesa, 2005). Ces compromis ne se limitent pas à la définition de la nature des biens produits et distribués ou à l'attribution de leur valeur ; ils émergent souvent dans des contextes de marché marqués par l'incertitude, où coexistent des modes d'existence divergents parmi les agents économiques. Dans ce contexte, l'analyse du processus de construction des données RH requiert une compréhension approfondie des espaces de négociation entre ces agents (Callon, 2006; Callon et al., 2000, 2002). Ces espaces, qualifiés de « controverses », constituent des lieux privilégiés où se forgent les compromis nécessaires à la valorisation des données RH sur le marché.

Les interactions qui se déroulent au sein de ces controverses révèlent non seulement la dimension économique des données RH en tant que biens, mais aussi la manière dont la notion même d'« économique » est définie en pratique (Callon & Muniesa, 2005). Ainsi, une attention particulière est accordée au processus de qualification de ces données en tant que biens économiques, un processus essentiel pour appréhender l'incertitude inhérente au marché numérique RH (Callon et al., 2000, 2002).

3.4. La marchandisation des données RH par leur qualification en tant que biens économiques

Parler des données RH en tant que biens économiques renvoie à l'étymologie du terme. Dans le cadre de la théorie économique, les termes « biens économiques » et

« produits » sont fréquemment utilisés de manière interchangeable. En raison de cette ambiguïté, Callon et al. (2000) proposent une distinction nuancée mais étymologiquement fondée entre ces deux concepts. Il définit le « bien » comme l'essence de l'activité économique, conçu pour satisfaire des besoins en offrant ce qui est bon, recherché et désiré. Ainsi, un bien économique requiert une stabilisation des qualités qui motivent la demande et facilitent sa marchandisation. En contraste, le « produit » est appréhendé à travers le prisme de sa production, distribution et consommation, envisagé comme une séquence organisée de transformations successives. Ce processus décrit les réseaux coordonnant les agents impliqués dans sa conception, fabrication, et commercialisation, soulignant ainsi que le produit est non seulement un processus mais aussi le résultat des interactions entre ces agents, qui par leurs ajustements et itérations, définissent ses qualités.

Au regard de cette distinction et dans le contexte de cette thèse qui examine le processus de construction des données RH, ces dernières sont considérées comme des produits engagés dans une trajectoire de stabilisation de leurs qualités. Cette stabilisation vise à transformer les données RH, initialement engagées en tant que produits, en biens économiques capables de satisfaire les besoins spécifiques de la fonction RH.

Callon et al. (2000, 2002) privilégient le terme « qualité » plutôt que « caractéristique », le dernier terme occultant les transformations progressives du produit ainsi que la nécessité d'investissements continus dans les épreuves de caractérisation. Ils affirment que les qualités ne se révèlent qu'à travers un processus de qualification. Chaque acte de qualification vise ainsi à établir une liste de qualités temporairement stabilisées, qui attribuent au bien ses propriétés marchandes.

Dans l'économie des qualités, la qualification des biens est centrale à la compétition économique (ibid.). L'établissement d'une liste de qualités d'un bien implique la mise en relation, voire la co-construction, d'une offre et d'une demande. Ce sont évidemment les agents économiques qui construisent les singularités et substituabilités des biens économiques. Le défi central pour ces derniers réside dans la conciliation délicate entre les attentes et les besoins du consommateur d'une part, et l'offre qui lui est proposée d'autre part. Dans ce contexte, le rôle principal des agents est alors de procéder à la qualification des biens, impliquant une classification, une évaluation et une appréciation fondées sur des comparaisons et des mises en relation.

Ainsi, les biens économiques se caractérisent par une nature dualiste : « *Différent mais semblable ; singulier et comparable, telle est la nature paradoxale du bien économique qui est constitutive de la dynamique des marchés.* » (Callon et al., 2002, p. 201).

Chaque bien économique se définit donc par une singularité. Cette singularité découle d'un ensemble de qualités spécifiques qui permettent de positionner le bien au sein d'un système de différences et de similitudes, créant ainsi des catégories à la fois interconnectées et distinctes. Ces qualités sont façonnées par les dispositifs de marchés utilisés pour les évaluer et les mesurer et leur interprétation varie selon les agents économiques impliqués (Callon et al., 2000, 2002). S'appuyant sur la théorie de la concurrence monopolistique de Chamberlin (1953), les auteurs mettent en évidence que les propriétés, distinguant un bien d'un autre, incluent non seulement des qualités intrinsèques, mais également extrinsèques tels que les marques, les emballages, ou les conditions spécifiques de vente. Ils citent Chamberlin pour souligner que toutes ces qualités sont intégrantes du bien, affirmant que le consommateur acquiert non seulement le produit « matériel » mais également l'ensemble des qualités périphériques tels que la réputation et l'honnêteté du vendeur. Cette approche suggère que toutes ces qualités partagent un statut ontologique identique, remettant en question la pertinence de distinguer les qualités primaires et secondaires, ou les qualités intrinsèques et extrinsèques. Selon cette perspective, toutes contribuent de manière égale à la définition et à la valorisation des biens économiques sur le marché.

Mon travail sur le processus de qualification-requalification des données RH en tant que biens économiques se concentre sur trois actes de (re)qualification clés :

1. (Re)définition des besoins des clients ;
2. (Re)rationalisation des coûts d'investissement ;
3. (Ré)enrôlement des agents économiques.

L'analyse approfondie de ces actes, notamment par la mise en évidence de leurs controverses, est essentielle pour comprendre la dynamique de construction des données RH dans un marché hautement compétitif.

Bien que le processus de qualification-requalification (Callon et al., 2000, 2002) soit intrinsèquement lié aux dynamiques des différents modes d'existence dans les dispositifs de marché, il apparaît qu'une séquence importante semble être négligée

pour l'analyse de la construction des données RH : la capitalisation des éléments humains et non-humains. Cette étape est essentielle, car elle nourrit directement la (re)connaissance des données RH, un processus qui suit une dynamique spiralée, reflétant la construction progressive de leurs singularités et substituabilités en tant que biens économiques (Callon et al., 2000, 2002; Latour, 2005a, 2007, 2012). En effet, la capitalisation permet d'attribuer une identité distincte aux données RH sur le marché. Ces nouvelles connaissances, obtenues lors de la qualification initiale, permettent ensuite de procéder à la requalification des données RH, renforçant leur position et leur valeur sur le marché.

Dans le cadre de la capitalisation sur les qualités des données RH, j'entreprends une analyse détaillée des divers projets de connaissances, structurée selon trois dimensions :

1. Territoire d'exploration épistémique des données RH ;
2. Savoir-faire des agents économiques ;
3. Epreuves d'exploration.

À l'instar des actes de (re)qualification, ceux de capitalisation des données RH révèlent également des controverses.

Ensemble, ces actes mobilisent divers éléments, humains et non-humains, et s'intègrent dans le processus de construction des données RH, que j'ai théorisé en trois séquences distinctes sous l'appellation « *QCR* » :

1. La Qualification des données RH ;
2. La Capitalisation des données RH ;
3. La Requalification des données RH.

La Figure 6 schématise le processus *QCR* et montre comment les données RH passent d'une première séquence de qualification à des cycles successifs entre capitalisation et requalification, formant ainsi une spirale dynamique qui conduit à leur transformation en biens économiques.

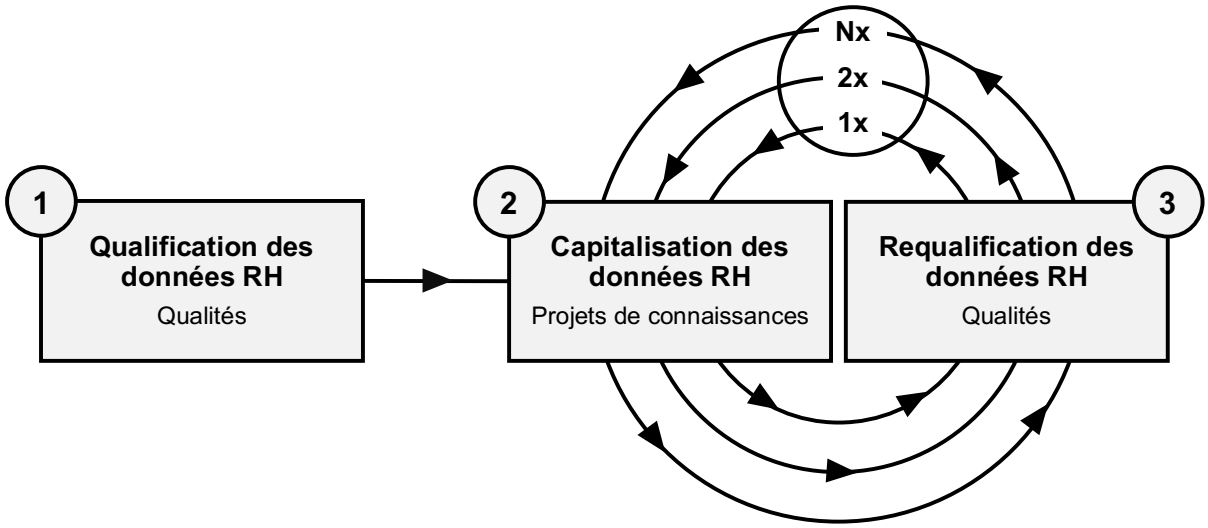


Figure 6 : Processus de construction des données RH

Cette deuxième section se consacre ainsi à l'exploration de l'ANT pour théoriser le processus de construction des données RH. Dans le contexte de la spirale *data-driven*, les données ne sont plus simplement des instruments du capital, mais suivent un processus d'économisation, évoluant vers une forme de capital à part entière (cf. fonction instrumentale). Cette évolution souligne l'importance de reconnaître la place centrale des données RH et leur rôle au sein des réseaux de valorisation, de tarification et de circuits commerciaux.

Le processus de construction des données RH implique leur considération en tant que produits engagés dans une trajectoire de stabilisation de leurs qualités. Cette stabilisation vise à transformer les données RH initialement considérées comme des produits en biens économiques.

La combinaison de la production de connaissances (Latour, 2005a, 2007) et de l'économie des qualités (Callon et al., 2000, 2002) s'avère particulièrement pertinente pour structurer mon cadre conceptuel. Cette synergie a non seulement facilité la théorisation du processus de construction des données RH (voir Figure 6), mais elle a également révélé la diversité des perspectives que les agents économiques adoptent à l'égard de ces données. L'analyse des controverses entre les différentes perspectives permet de comprendre comment ces agents négocient les qualités des données avant de les établir comme des « vérités incontestées » (Callon, 2006).

Les controverses mettent ainsi en lumière l'existence de nombreux compromis qui précèdent et encadrent les décisions liées à la construction des données RH (ibid.). En s'appuyant sur ces compromis, les agents économiques développent progressivement la fonction épistémique des données RH, permettant leur transformation en biens économiques.

4. Conclusion

Ce premier chapitre se divise en deux sections principales. La première section se concentre sur l'écosystème sous-jacent à ce que la littérature associe aux « données RH ». Cette étape est cruciale pour approfondir la compréhension de leur nature et des contextes et infrastructures au sein desquelles elles opèrent.

La revue de littérature est organisée en quatre actes distincts, chacun visant à décomposer et à examiner les différentes facettes des données :

1. Définition des données RH ;
2. Transformation des données RH ;
3. Valorisation des données RH ;
4. Construction des données RH ;

Bien que dispersés entre les domaines des SI et de la GRH, ces quatre grands actes constituent les éléments d'un ensemble cohérent : la spirale *data-driven*. Elle se caractérise par le renforcement mutuel des phénomènes de transformation et d'accumulation des données. Ce dépassement des frontières disciplinaires souligne la nécessité d'adopter une approche théorique capable d'appréhender pleinement la complexité des données RH et leurs dispositifs socio-numériques.

Dans cette optique, la seconde section de ce chapitre se consacre au choix d'un cadre théorique pertinent pour analyser cette dynamique. L'ANT est ainsi retenue pour son adéquation avec cet objectif. Plus spécifiquement, la combinaison de la production de connaissances (Latour, 2005a, 2007) et de l'économie des qualités (Callon et al., 2000, 2002) se révèle particulièrement pertinente.

Cette synergie conceptuelle permet de théoriser le processus de construction des données RH en trois séquences (*QCR*), mettant en lumière leur transformation progressive et controversée en biens économiques.

Chapitre 2. Méthodologie et terrain de recherche

1. Introduction

Le premier chapitre est consacré à la conceptualisation des données RH ainsi qu'à l'identification d'un cadre théorique pertinent. Au regard de la nature interdisciplinaire de cette étude, l'ANT est retenue. Elle permet la combinaison de la production de connaissances (Latour, 2005a, 2007) et de l'économie des qualités (Callon et al., 2000, 2002) et conduit à la théorisation du processus de construction des données RH (*QCR*).

En cohérence avec ce cadre théorique, l'approche méthodologique choisie permet d'explorer l'application du processus *QCR*.

Pour garantir une présentation transparente et cohérente de mon travail, ce chapitre méthodologique est structuré en plusieurs sections. La première section établit le contexte et les conditions de la recherche, en précisant les motivations et objectifs de cette thèse. La deuxième présente le cas d'étude de manière détaillée. La troisième section expose et justifie la démarche méthodologique adoptée. La quatrième fournit une description exhaustive des types de données recueillies, incluant les sources et les méthodes de collecte. Enfin, la cinquième section détaille les méthodes d'analyse des données, en démontrant leur adéquation pour répondre à la problématique de recherche.

2. Contexte et conditions de la recherche : la convention *CIFRE*

La thèse est réalisée dans le cadre d'une convention *CIFRE* (Convention Industrielle de Formation par la Recherche) entre le ministère de la Recherche, l'ESCP Business School et le cabinet de conseil *QIA*.

La signature de cette convention a été rendue possible grâce aux échanges que j'ai préalablement eus avec des *data scientists*, dont l'un avait occupé un poste au sein du cabinet à l'étude. Ma volonté d'être fortement ancrée dans le domaine empirique a

ultérieurement conduit à un processus de recrutement et à mon embauche. L'une des motivations sous-jacentes à la décision du cabinet était ma capacité, grâce à mon expertise théorique en GRH, à approfondir les données d'intérêt, qu'elles soient quantifiables ou non. Cela visait à établir des liens entre les phénomènes RH et la *data science*, en vue du développement d'une offre commerciale RH. En effet, les 90 salariés de *QIA* proviennent essentiellement des domaines de l'ingénierie, de l'informatique, de la *data science* et de la finance, et disposent de peu d'expérience dans le domaine de la GRH.

La direction de *QIA* n'anticipait pas bien, au départ, quels pouvaient être les résultats d'une recherche en GRH et la façon dont le cabinet pourrait en tirer parti. Cette collaboration avait, par conséquent, une dimension opportuniste de par l'appétence académique de ma responsable au sein du cabinet, elle-même docteure en sciences de gestion. Différents commentaires issus d'échanges ont, par ailleurs, permis de souligner l'écart de compréhension avec d'autres membres de la direction : « *Comment peux-tu faire une thèse en management alors que tu n'as jamais managé ?* » (Directeur de filiale, notes issues du journal de bord, 21/10/2020) ; « *C'est bien de chercher mais on veut que tu trouves…* » (Directeur général, notes issues du journal de bord, 11/09/2020).

Le Contrat à Durée Indéterminée (CDI) était structuré de manière à permettre à l'entreprise d'amortir son investissement indépendamment des résultats de mes travaux. Concrètement, cela signifiait que mon temps de travail était divisé entre des activités de conseil facturées aux clients et des activités de recherche (enseignement, formation, etc.) sur lesquelles l'entreprise pouvait générer des « actifs »[14]. Mes activités étaient donc définies comme suit :

- Recenser, analyser et documenter les besoins du marché, de la concurrence et du savoir-faire de *QIA* sur les sujets en GRH ;
- Proposer des méthodologies innovantes et des pistes de développement sur des sujets RH ;
- Contribuer proactivement au développement d'une offre RH « augmentée » ;

[14] Terme utilisé par ma responsable opérationnelle lors de nos discussions sur mes objectifs.

- Intervenir dans le processus itératif d'analyse de données et de conception des livrables au sein des équipes dans le cadre de prestations RH ;
- Réaliser et favoriser la capitalisation interne sur l'implantation des solutions d'IA et de conseil en RH.

Une certaine souplesse dans le partage du temps entre travail productif et de recherche était prévue dès le départ. Les activités de conseil nécessitant un investissement à plein-temps auprès des clients, le partage des temps a ainsi évolué afin de tenir compte des impératifs du cabinet et de ses besoins. Afin de répondre à ces impératifs, il était nécessaire d'adapter mes rôles en fonction des exigences qui se présentaient et des ressources mobilisables sur le projet. En tant que *HR business analyst*, mes responsabilités ne se limitaient pas seulement à ce rôle, mais englobait également d'autres charges. La fluctuation de ma fonction et du contexte des projets RH sur le terrain a ainsi entraîné l'évolution de mon approche méthodologique.

Cette flexibilité a cependant soulevé une question fondamentale qui constituera la trame de fond cette thèse : celle de mon expertise et plus particulièrement de ma légitimité en tant que référente RH dans les projets en *data science*. Un bref retour en arrière est nécessaire afin de pleinement comprendre le contexte *in situ*. Je considère également cette introspection comme une condition préalable au renforcement de la crédibilité de cette recherche.

Préalablement à ma thèse, j'ai occupé le poste de consultante en gestion des conditions de travail au sein d'un cabinet de conseil spécialisé en ergonomie. Au cours de cette fonction, j'ai interagi avec des DRH et les ai accompagnés, sans toutefois assumer directement leurs responsabilités. En conséquence, mes connaissances du domaine de la GRH étaient principalement de nature théorique et manquaient d'applications concrètes impactant, de ce fait, mon implication chez *QIA*. L'écart entre les dimensions théoriques et pratiques de mes connaissances a donc suscité des défis de positionnement, qui ont persisté tout au long de la phase empirique de cette recherche. Mon « incarnation » de la fonction RH a également soulevé des interrogations concernant la perception de cette fonction par les *data scientists*. Ainsi, en raison de mon manque d'expérience en GRH, un paradoxe a émergé, à savoir la coexistence simultanée et contradictoire de la présence et de l'absence de la fonction RH sur l'ensemble des projets auxquels j'ai participés.

Plusieurs études mettent en évidence la nature confidentielle du travail des *data scientists* et les difficultés rencontrées par les chercheurs pour accéder à des observations directes (Passi & Sengers, 2020). Dans mon cas, en tant que *HR business analyst*, ma position m'a accordé un accès exhaustif aux données empiriques nécessaires pour mener une recherche qualitative : participation à des réunions formelles et informelles, tant en internes qu'avec des clients potentiels, production et analyse de divers types de documents, échanges de courriels, conversations informelles et ainsi de suite (voir Tableau 9). De plus, dans le cadre de mes actions, j'ai pu orienter les discussions vers certains aspects de ma réflexion, parfois de manière ouverte, parfois en abordant plus directement les enjeux concrets des projets auxquels j'ai participés. Les *data scientists* étaient informés de mon travail de recherche, mais n'avaient initialement pas de connaissance précise de son contenu. Aucun d'entre eux n'a exprimé de préoccupations concernant la confidentialité ou l'utilisation des données au cours de mon intervention. Par ailleurs, il a été décidé de rendre anonymes tous les noms d'entreprises ou de personnes mentionnés dans la présente thèse.

Le matériau empirique utilisé dans cette étude provient de mes observations et de mon implication directe dans le travail de construction de données RH chez *QIA*. Depuis plus de 15 ans, *QIA* développe des outils d'IA personnalisés d'aide à la prise de décision pour des entreprises opérant dans les secteurs de la santé et de la finance. Le cabinet connait, depuis quelques années, une phase de croissance qui l'a incité à explorer de nouvelles opportunités. Comme l'explique Emilie lors d'un entretien : « *QIA se met en déséquilibre pour trouver une finalité. Nous, on n'est pas dans une phase de business stable. On est dans une phase de croissance. Il faut tester plein de choses…* » (Directrice associée, entretien no1, 31/05/2021). Cette situation de déséquilibre a motivé la direction de *QIA* à étendre ses activités, ouvrant ainsi la porte à de nouveaux projets et à de nouveaux investissements. En 2022, *QIA* a, d'ailleurs, réussi à lever plus de 10 millions d'euros auprès d'investisseurs pour financer le déploiement de nouveaux outils d'IA. Par conséquent, ma présence sur le terrain découle de l'intention du cabinet de conseil de se positionner également sur le marché des outils et services numériques RH.

Les données empiriques de cette étude ont été recueillies lors de ma participation à quatre projets sur une période de trois ans et sept mois. La sélection de ces projets,

réalisée en collaboration avec ma responsable opérationnelle, reposait sur trois critères principaux :

1. Des opportunités commerciales pour le cabinet ;
2. Des projets nécessitant la construction de données RH ;
3. Une diversité de missions.

En ce qui concerne le dernier critère, ma responsable accordait une grande importance à mon implication dans des contextes d'intervention variés afin de me familiariser avec la méthodologie des projets en *data science*.

Le Tableau 7 fournit un aperçu des projets de construction des données RH réalisés au cours de mon immersion sur le terrain, constituant ainsi la source de mon matériel empirique.

Tableau 7 : Synthèse des projets réalisés sur le terrain de recherche

Projets	Dimensions du projet	Enjeux principaux	Durée	Agents
2020-2021 **Dimensionnement d'effectif 1**	Budget : 230K€ Diagnostic ; Modélisation ; Déploiement ; Exploitation et amélioration continue Population ciblée : 30 salariés Client : contrôle de gestion	Prévoir les volumes d'activité et les effectifs nécessaires pour l'ensemble des équipes intervenants dans le processus.	7 mois	6
2021-2022 **Dimensionnement d'effectif 2**	Budget : 180K€ Diagnostic Population ciblée : 26 salariés Client : contrôle de gestion	Prévoir les volumes d'activité et les effectifs nécessaires pour l'ensemble des équipes intervenants dans le processus.	9 mois	3
2020-2022* **Offre RH « augmentée »**	Budget : N/A Développement commercial Client potentiel : fonction RH	Développer une offre commerciale data RH généraliste.	21 mois	3
2021-2022** **Outil *DSN Analytics***	Budget : N/A Développements techniques et commerciaux Clients potentiels : fonctions RH, financières et opérationnelles	Concevoir un outil d'IA pour la gestion de l'absentéisme.	21 mois	6

* Echéance de ce projet est fixée à mai 2022, suite à la décision de le placer en mode d'attente (« *stand-by* ») afin de focaliser les efforts sur l'outil *DSN Analytics*. / ** Echéance de ce projet est fixée à octobre 2022, à la suite de la décision de me retirer du terrain.

À la lumière de ce tableau de synthèse, deux typologies de projets de construction des données RH peuvent être distinguées :

1. La construction des données RH dans le cadre d'une demande du contrôle de gestion : englobe la prévision et l'adéquation des ressources pour la planification des effectifs, ainsi que la capacité à gérer l'activité dans le respect des contraintes de qualité.
2. La construction des données RH à destination de la fonction RH : couvre la quantification, la caractérisation et la modélisation des activités et phénomènes RH.

Face à cette situation, j'ai rapidement soulevé des interrogations quant à la cohérence des liens pouvant être établis entre ces deux types de projets. Bien que les deux explorent les pratiques - calculatoires ou non - des *data scientists* dans la construction des données RH, la fonction RH est toutefois absente des projets mandatés par le contrôle de gestion. Ces projets sont de nature opérationnelle et ne nécessitent pas son intervention. Compte tenu de ce constat et du volume de données recueillies sur le terrain, j'ai donc fait le choix de centrer mon attention sur l'étude de *DSN Analytics*, outil d'IA destiné à la gestion de l'absentéisme.

3. Présentation du cas : *DSN Analytics* en tant que nouvel outil d'IA pour la gestion de l'absentéisme

L'origine de cette étude réside dans la volonté du cabinet de conseil de s'implanter sur le marché des outils et services numériques liés aux RH et plus particulièrement dans la gestion de l'absentéisme :

« [L'absentéisme] *c'est un coût pour l'entreprise, c'est une perte de performance potentielle et une source de désorganisation opérationnelle. Dans les cas extrêmes, c'est une exposition juridique puisque ça créer des mauvaises conditions de travail* » (Tristan, chef de projet, entretien no2, 04/05/2022).

La convergence des enjeux liés à l'absentéisme conduit ainsi les *data scientists* à ressentir la nécessité de répondre à la demande opérationnelle de : « *comprendre et d'agir sur le sujet de l'absentéisme…* » (Xavier, directeur de projet, notes prises lors d'une réunion, 21/07/2021).

Cette thèse repose sur le travail des *data scientists* dans la construction des données RH lors de la conception de *DSN Analytics*. Sur une période d'un an et demi (de janvier 2021 à octobre 2022), j'ai étudié les dispositifs socio-numériques ainsi que les pratiques calculatoires et non calculatoires des *data scientists*. Au cours de cette période, l'équipe projet a officialisé une première version de l'outil d'IA, suffisamment aboutie pour être présentée sur le marché RH.

3.1. De l'idée au problème de l'absentéisme

« […] *je pense que c'est essentiel. Ça coûte de l'argent à tout le monde, l'absentéisme. Dans toutes les entreprises, il y a des gens qui sont payés pour le résoudre, a minima pour vérifier que ça ne devient pas un problème. Tout le monde a de l'absentéisme. C'est un phénomène humain.* […] *C'est un coût pour l'entreprise, à chaque fois que tu en as, c'est une perte de performance potentielle et une source de désorganisation opérationnelle dans une proportion plus ou moins importante.* » (Tristan, chef de projet, entretien no1, 15/12/2021).

L'idée de *DSN Analytics* est initiée par Benoît, un actuaire à la recherche d'un partenaire possédant une solide expertise en gestion des données. Son objectif est de concrétiser son outil RH - *Assur'act* - par la mise en place d'un moteur de calculs plus performant. *Assur'act* se voulait devenir un outil de pilotage des coûts de la prévoyance collective en entreprise reposant sur quatre grands piliers : (1) l'identification en temps réel de la dynamique des arrêts maladie, (2) l'anticipation des impacts financiers, (3) la cartographie des gisements de l'absentéisme par métier et par entité géographique et (4) la maîtrise de l'efficacité des investissements sociaux :

« Assur'act, c'est uniquement ça. […] *c'est un tableau de bord, ça ne doit pas être un dictionnaire de données* […]*, c'est le nombre minimal d'indicateurs pertinents qui contient toute l'information. »* (Benoît, partenaire en actuariat, entretien no1, 24/11/2021).

Ainsi, pour donner vie à *Assur'act*, Xavier, occupant le rôle de directeur de projet, et Tristan, en qualité de chef de projet, sont désignés pour la gestion des données RH. Néanmoins, en dépit de l'effort de réduction du nombre d'indicateurs, l'outil *Assur'act* demeure complexe et ne se prête pas aisément à une compréhension opérationnelle de l'absentéisme.

« On est vraiment sur des choses incompréhensibles, très compliquées pour rien en plus. On ajoute de la complexité là où ce sont des règles de trois. Et donc, je [lui] *ai assez vite dit qu'il fallait partir sur une offre RH, compréhensible par les RH. Ça ne suffit pas de rajouter un radar quelque part dans un coin pour dire : là, vous pouvez faire un zoom opérationnel. Il faut vraiment que l'outil soit construit pour ce public-là* […] *C'était ma conviction. »* (Xavier, directeur de projet, entretien no1, 31/10/2023).

La collaboration entre Xavier (directeur de projet) et Benoît (partenaire en actuariat) fait donc émerger l'idée d'une nouvelle offre RH, centrée sur une approche plus opérationnelle de l'absentéisme. Cette initiative, pensée comme un complément à *Assur'act*, vise à enrichir l'offre existante avec une perspective orientée vers une application plus pratique.

3.2. D'un problème opérationnel à la formulation d'un problème *data-driven*

C'est donc à travers la perspective de Xavier (directeur de projet) que *DSN Analytics* nait, émergeant comme un outil complémentaire à *Assur'act*. Ce dernier est, et reste pour le moment, une fiction. Il n'existe pas et ne saurait exister autrement que sous la forme d'une idée qui émane de l'esprit de Xavier (directeur de projet) avec le soutien de Benoît (partenaire en actuariat). Il n'y a donc pas de différence entre l'idée et l'objet *DSN Analytics* ; ce dernier étant en cours d'« objectivation ». Pour qu'il devienne officiellement un projet, Xavier (directeur de projet) dois recruter des agents économiques, mais aussi faire en sorte que leur enrôlement soit sûr afin de garantir la formation d'un réseau convergent. Dans l'objectif de réaliser cette ambition, l'équipe projet se compose initialement de Xavier, directeur de projet ; Tristan, chef de projet ; Olivier, *data scientist* ; Ariane, UX/UI *designer freelance* ; et moi-même, en tant que *HR business analyst*.

Il sera ensuite nécessaire d'intéresser d'autres agents économiques, de les séduire et de traduire leurs intérêts, tout en veillant à éliminer toute dissension potentielle. Avec le futur soutien des agents qu'il aura enrôlés, Xavier (directeur de projet) s'efforcera progressivement d'augmenter le degré de réalité de *DSN Analytics*. En effet, par le biais de ce nouvel outil d'IA, Xavier (directeur de projet) aspire à transformer la gestion de l'absentéisme. Constatant l'inefficacité des outils actuels sur le marché RH, il juge

indispensable de les repenser, ce qui passe préalablement par la sélection des données RH d'intérêt.

3.3. Les données *DSN* : une enthousiasmante réglementation

L'extrait du rapport d'information n°743, déposé le 15 juin 2023[15], célèbre le triomphe de la Déclaration Sociale Nominative (*DSN*) en tant que pilier de la conformité réglementaire des entreprises vis-à-vis de l'État français. Émanant d'une impulsion vers la simplification administrative, la *DSN* est applaudie comme une révolution technique, particulièrement pour son impact sur la qualité des données RH.

LE SUCCÈS D'UNE SIMPLIFICATION : LA DSN

Le premier constat lié à la mise en place de la DSN est tout d'abord celui ***d'une réussite technique*** *: un système d'information solide a été mis en place, garantissant une déclaration dématérialisée réalisée par les employeurs puis transmise de façon sécurisée à l'ensemble des acteurs de la protection sociale à une échelle industrielle. Ce succès est tel que la DSN a dépassé ses limites originelles : d'abord destinée à se substituer aux déclarations sociales obligatoires, elle a permis la mise en place du prélèvement à la source à partir de 2019, mais également la création du Dispositif de Ressources Mensuelles (DRM) qui constitue la pierre angulaire des ambitieux chantiers de modernisation des prestations sociales versées sous condition de ressources (et demain de la solidarité à la source).*

Pensée dès l'origine comme une émanation du « choc de simplification administrative », la DSN a été construite pour simplifier les procédures administratives au profit des employeurs, avec l'idée que c'est en simplifiant « à la source » que l'on ***améliore la qualité des données****, avec des effets en cascade de facilitation pour la gestion par les organismes et de sécurisation in fine des droits des individus. La DSN s'appuie donc aujourd'hui sur trois piliers de simplification administrative au bénéfice des employeurs :*

*Demander aux employeurs les données qu'ils maîtrisent. C'est pourquoi la DSN véhicule les données de la paie, ce qui permet d'****améliorer la qualité des données utilisées****. Par*

[15] Rapport d'information n°743 (2022-2023), déposé le 15 juin 2023. *La sobriété normative pour renforcer la compétitivité des entreprises.* Réponse écrite de la DGE. https://www.senat.fr/rap/r22-743/r22-7436.html (consulté le 14/05/2024).

*rapport aux anciennes déclarations, c'est une **véritable révolution**. Au lieu de demander à l'employeur des données définies d'abord par les consommateurs des données (organismes et administrations), on leur demande d'utiliser les **données présentes en paie*** […].

Source : réponse écrite de la DGE [Direction Générale des Entreprises]

(À noter que les passages mis en exergue sont des emphases personnelles)

La *DSN* est devenue obligatoire pour toutes les entreprises du secteur privé en janvier 2017, et a été étendue au secteur public en janvier 2022. La généralisation de son obligation réglementaire permet une unification des modalités de déclaration sociale pour l'ensemble des entreprises, quels que soient leur taille et leur secteur d'activité. Elle a ainsi remplacé toutes les déclarations antérieures, qu'elles soient périodiques ou événementielles, qui étaient auparavant adressées par les employeurs à une multitude d'acteurs tels que la CPAM, l'URSSAF, l'AGIRC ARRCO, les organismes complémentaires, Pôle emploi, le Centre des impôts, les caisses des régimes spéciaux, etc. Ainsi, plus de 40 procédures ont été substituées au profit de la *DSN*.

La *DSN* se décline en deux types :

1. *DSN* périodique ;
2. *DSN* évènementielle.

Dans le cas de la *DSN* évènementielle, celle-ci concerne principalement trois situations, notamment :

1. L'arrêt de travail : lorsqu'un salarié se trouve en arrêt de travail pour des raisons telles que la maladie, la maternité, la paternité, et autres circonstances similaires.
2. La reprise anticipée : cette situation survient lorsque le salarié reprend son activité professionnelle avant la date prévue de fin d'arrêt de travail initial.
3. La fin de contrat de travail : il s'agit de la situation où un salarié quitte l'entreprise, indépendamment du motif de cessation d'emploi.

Les données *DSN*, reflètent ainsi la situation mensuelle d'un salarié, mettant en évidence les évènements survenus au cours d'un mois donné, tels que des périodes

de maladie, de maternité, des changements dans les éléments du contrat de travail, qui ont un impact sur sa rémunération.

La structure de la *DSN* se compose de 55 blocs thématiques, qui sont subdivisés en 527 rubriques. La Figure 7 ci-dessous présente le bloc « contrat de travail » ainsi que les rubriques qui lui sont associées.

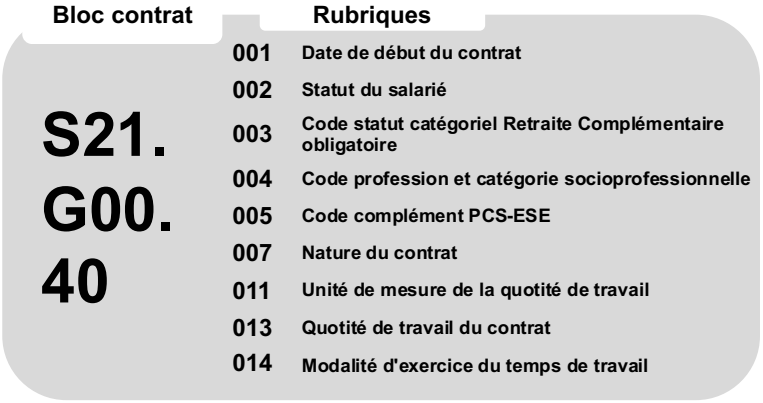


Figure 7 : Exemple de la structure du bloc « contrat » et ses rubriques spécifiques telles qu'elles apparaissent dans la *DSN*

Les données *DSN* peuvent être regroupées en différentes catégories distinctes :

- Les données d'identification : cette catégorie englobe les informations permettant d'identifier de manière unique les salariés, telles que leur nom, prénom, date de naissance, numéro de sécurité sociale et adresse postale.
- Les données contractuelles : les données contractuelles comprennent les informations relatives aux dispositions contractuelles, telles que la date d'embauche, la nature du contrat, la durée du travail, la rémunération et autres.
- Les données relatives à la paie : les données relatives à la paie concernent les différents éléments composant la rémunération, par exemple le salaire brut, les primes, les indemnités, les retenues (comme les cotisations sociales et l'impôt sur le revenu), les avantages en nature et autres.
- Les données d'absences et d'arrêts de travail : cette catégorie regroupe les informations relatives aux absences du salarié, telles que les congés payés, les congés maladie, les arrêts de travail, les périodes de maternité/paternité et autres.

- Les données liées à la protection sociale : ces données englobent les informations concernant l'affiliation du salarié à des régimes de protection sociale, tels que l'assurance maladie, l'assurance chômage, les régimes complémentaires de retraite, les régimes de prévoyance et autres.
- Les autres données spécifiques : selon les exigences spécifiques des organismes sociaux, d'autres données supplémentaires peuvent être incluses dans la *DSN*. Cela peut inclure des informations sur les accidents du travail, les maladies professionnelles, les attestations de l'employeur, etc.

Considérant le potentiel significatif de ces nouvelles données RH sur le marché, les *data scientists* officialisent donc l'exploitation de la *DSN* dans la conception de *DSN Analytics* :

« Ah ! Si vous passez à la DSN, c'est formidable. Moi, je vous extrais la DSN, c'est beaucoup plus simple. Je n'ai pas à reconstruire les données qu'il vous faut, c'est beaucoup mieux ! » (Tristan personnifiant un client d'*Assur'act*, chef de projet, entretien no1, 15/12/2021).

4. Exposition et justification du bricolage méthodologique : une démarche qualitative abductive et séquentielle

Dans cette étude, l'intérêt empirique précède l'intérêt théorique. Elle débute par une attention particulière portée aux données RH, un sujet qui attire largement l'attention malgré sa nature élusive. Cette focalisation oriente les thématiques explorées lors des entretiens et des observations subséquentes avec les *data scientists*. En optant pour une méthode abductive qui oscille entre empirie et théorie, je cherche également à tempérer les préconceptions théoriques évitant ainsi un risque de circularité.

En ce qui concerne la méthodologie employée, je m'inscris dans une approche qualitative, visant à suivre les acteurs (Latour, 2005b, 2005a) lors du processus de construction des données RH. Afin de garantir la pertinence de l'étude, cette démarche méthodologique a cependant nécessité une adaptation continue au contexte spécifique du terrain.

Mon rôle initial de *HR business analyst* a ainsi évolué pour inclure la fonction de cheffe de projet pendant une période déterminée, nécessitant des ajustements méthodologiques. De ce fait, le matériel empirique de cette thèse repose sur une combinaison de recherche-action (Eden & Huxham, 1996) et d'ethnographie (Akrich et al., 2006; Callon, 1986; Latour, 2005a), choisies en fonction des rôles que j'ai assumés au sein de l'équipe. Ces approches représentent pour moi les pôles opposés d'un continuum méthodologique, transitant entre implication complète et observation distanciée. La plupart du temps, je me situais à différents stades intermédiaires entre ces deux extrêmes, chaque approche servant de boussole intellectuelle pour m'orienter dans la complexité du terrain étudié. Toutefois, il est indéniable que cette pratique peut susciter des questionnements, tant parmi les partisans de la recherche-action que parmi les « puristes » de l'ethnographie. Ces réflexions méthodologiques, bien qu'elles soulèvent des débats, enrichissent selon moi la rigueur et la profondeur de l'analyse menée dans cette thèse.

Cette section se consacre d'abord au bricolage méthodologique comme approche pragmatiste de ma recherche. Ensuite, elle examine la combinaison de la recherche-action et de l'ethnographie utilisée lors de mes transitions vers une autre fonction. L'objectif principal est d'analyser et de mettre en évidence leurs similarités et leurs différences, ainsi que les ajustements apportés au cours du processus de collecte des données.

4.1. Le bricolage méthodologique comme approche pragmatiste de la recherche

Dans son essai intitulé « La pensée sauvage », Lévi-Strauss (1962), philosophe et anthropologue, remet en question la dépréciation du bricoleur en le juxtaposant à l'ingénieur. Le but n'est pas de les hiérarchiser, mais d'illustrer différentes approches au travail : d'un côté, une logique rationnelle et structurée, et de l'autre, une approche plus pratique ; toute deux étant fondamentalement empiriques :

« *Le bricoleur est apte à exécuter un grand nombre de tâches diversifiées ; mais, à la différence de l'ingénieur, il ne subordonne pas chacune d'elles à l'obtention de matières premières et d'outils, conçus et procurés à la mesure de son projet : son univers instrumental est clos, et la règle de son jeu est de toujours s'arranger avec les*

« moyens du bord », c'est-à-dire un ensemble à chaque instant fini d'outils et de matériaux, hétéroclites au surplus, parce que la composition de l'ensemble n'est pas en rapport avec le projet du moment, ni d'ailleurs avec aucun projet particulier, mais est le résultat contingent de toutes les occasions qui se sont présentées de renouveler ou d'enrichir le stock, ou de l'entretenir avec les résidus de constructions et de destructions antérieures.» (P. 31).

La compréhension de l'esprit du bricoleur, tel que défini par Lévi-Strauss (1962), pourrait susciter une réflexion méthodologique en sciences de gestion, visant à mettre en lumière les détours complexes qui caractérisent cette discipline. Dans cette optique, Bell & Willmott (2020) examinent l'évolution des pratiques de recherche, marquée par une prédominance croissante d'approches standardisées. Les auteurs soulignent l'importance accordée à la rigueur méthodologique, généralement associée à la conformité aux normes de la science « dure » et à l'utilisation de techniques analytiques traçables. Ils ne sont pas les seuls à observer cette tendance (voir également Eisenhardt, 2021; Eisenhardt et al., 2016; Langley & Abdallah, 2011; Pratt et al., 2022).

Langley & Abdallah (2011) ont identifié deux paradigmes méthodologiques prédominants dans la recherche qualitative : la perspective des méthodes de cas d'Eisenhardt (1989) et l'approche de la théorie ancrée de Gioia et al. (2013). Elles ont analysé les structures logiques et rhétoriques sous-jacentes à ces paradigmes et démontré leur efficacité. Toutefois, ces autrices ont également souligné les limites d'une dépendance exclusive à ces deux paradigmes et ont recommandé l'exploration d'autres options méthodologiques, telles que les approches discursives et pratiques. Même les créateurs de ces deux modèles méthodologiques émettent des avertissements et des mises en garde contre l'appropriation sans réserve de leurs « orientations » méthodologiques (Eisenhardt, 2021; Pratt et al., 2022). Par exemple, Gioia et al. (2013, p. 25) avertissent :

« *Les chercheurs en organisation semblent appliquer la méthodologie comme un modèle... d'autres semblent la traiter comme une « formule », reproduisant essentiellement le format exact de la structure des données des études récemment publiées. Même plusieurs sections méthodologiques semblent maintenant adopter des formats et des descriptions procédurales presque identiques à ceux des travaux publiés. Cette tendance est quelque peu préoccupante... nous la voyons comme une*

orientation flexible vers la recherche qualitative et inductive, ouverte à l'innovation, plutôt qu'un « livre de recettes. » ».

Par la remise en question des paradigmes méthodologiques dominants dans la recherche qualitative, les auteurs préconisent une approche plus contextuelle et flexible, prenant en compte les spécificités et les enjeux propres à chaque étude en gestion. Cette vision est renforcée par l'analogie du bricoleur de Lévi-Strauss (1962), qui se trouve constamment confronté à la nécessité de « *s'arranger avec les moyens du bord* ». Cette expression illustre de manière pertinente la capacité du chercheur en sciences de gestion à tirer parti des ressources disponibles et à s'adapter aux contraintes imposées, en appelant à une approche plus réflexive sur la recherche « *en train de se faire* » (Latour, 2005a, p. 29).

Dans cette optique, Pratt et al. (2022) introduisent le concept de bricolage comme une métaphore pour repenser les choix méthodologiques en recherche, mettant en avant l'importance de la réflexion sur ces choix. Le bricolage méthodologique est envisagé comme une approche plutôt qu'une méthode, valorisant ainsi la flexibilité et l'adaptabilité du chercheur. La métaphore du bricolage met ainsi en lumière l'agentivité, la créativité et l'habileté des chercheurs, qualifiés de bricoleurs. Elle se structure autour de trois éléments centraux (Baker & Nelson, 2005; Pratt et al., 2022) :

1. Se débrouiller ;
2. Utiliser les ressources disponibles ;
3. Combiner ces ressources pour de nouveaux usages.

Cette approche vise à renforcer la crédibilité de la recherche en nécessitant des choix méthodologiques actifs et réfléchis, visibles dans la section méthodologique et dans la rédaction des résultats. Le bricolage méthodologique permet ainsi de relever de manière créative les défis propres à chaque projet, en favorisant l'expérimentation et les découvertes inattendues.

S'inspirant de Lévi-Strauss (1962) et de Pratt et al. (2022), je souhaite mettre en avant l'assemblage et la diversité inhérente à mon activité de recherche. Cette approche permet de considérer la « recherche en tant qu'artisanat » (*research-as-craft*) plutôt que simplement comme une technique (*research-as-technique*) (Bell &

Willmott, 2020). Pour ce faire, j'ai décidé de m'appuyer sur les trois orientations méthodologiques suivantes (Olivier De Sardan, 2018, p. 205) :

1. Insister sur une approche des phénomènes sociaux qui privilégie la proximité avec le terrain, mon engagement actif, ainsi que la discussion des interprétations locales des situations.
2. Promouvoir une compréhension dynamique des événements, évitant ainsi de se restreindre à des découpages disciplinaires rigides.
3. Donner la priorité aux « micro-réglages » plutôt qu'aux choix théoriques ou épistémologiques majeurs.

Ces trois orientations méthodologiques m'ont ainsi conduit à entreprendre une recherche-action couplée à une ethnographie. Les sections suivantes présentent ces méthodes de recherche et mon positionnement par rapport à ces dernières.

4.2. L'engagement actif à travers la recherche-action

La recherche-action est une méthodologie fondée sur la collaboration entre chercheurs et praticiens selon laquelle la production de connaissances découle de la pratique (Eden & Huxham, 1996) : « *Contrairement aux sciences sociales conventionnelles, son objectif n'est pas principalement ou uniquement de comprendre les arrangements sociaux, mais aussi d'apporter les changements souhaités afin de générer des connaissances et d'habiliter les parties prenantes* » (Bradbury-Huang, 2010, p. 93). Cette méthode est donc le fruit d'une action conjointe et d'une réalité négociée dans la pratique, triangulés grâce aux points de vue des participants (Lüscher & Lewis, 2008).

Ma collaboration avec les *data scientists* dans la conduite de cette recherche a également été essentielle pour garantir sa pertinence (Eden & Huxham, 1996). Les deux axes de la recherche-action, à savoir la génération de connaissances pratiques (locales) et scientifiques, étaient particulièrement bien adaptés à mes rôles de *HR business analyst* ou de cheffe de projet. Lorsque Tristan (chef de projet) a annoncé son départ en janvier 2022, j'ai été désignée pour reprendre temporairement une part de ses responsabilités, notamment dans l'élaboration et le suivi de la feuille de route technique et commerciale (jusqu'en mars 2022). Cela a impliqué une collaboration

étroite avec Xavier (directeur de projet), qu'il décrit comme une dynamique de « ping-pong », illustrant la proximité et la réciprocité de ses interactions avec le chef de projet.

Mon objectif était double :

1. Théoriser le travail de construction des données RH par l'analyse des pratiques calculatoires et non calculatoires des *data scientists* ;
2. Favoriser le bon développement technique et commercial de *DSN Analytics*[16]

4.2.1. L'exploration ethnographique par le prisme de l'ANT

Au-delà de la recherche-action, mon terrain m'a également conduit à entreprendre une étude ethnographique, principalement par le biais de phases d'observation non-participante du travail des *data scientists*. En immersion pendant trois ans et sept mois, j'ai pu suivre les acteurs de manière quasi quotidienne en consignant mes observations dans un journal de bord.

L'ANT est utilisée à la fois comme cadre conceptuel et méthodologique dans cette thèse. Les quatre phases de la traduction (Callon, 1986) : (1) la problématisation, (2) l'intéressement, (3) la mobilisation et (4) l'enrôlement, m'ont accompagnées intuitivement, m'orientant de manière non systématique dans l'exploration du processus de construction des données RH. En mettant l'accent sur la « science en action » (Latour, 2005a), l'ANT adopte une approche ethnographique pour examiner les dynamiques en jeu par la description des interactions entre acteurs, humains et non-humains. Comme le souligne Latour (2005b, p. 146), l'explication découle de la description elle-même : « *Nous nous engageons dans des descriptions... De bonnes enquêtes produisent toujours beaucoup de nouvelles descriptions*. ».

L'approche de l'ANT repose ainsi sur la réhabilitation de la description en tant qu'élément au centre de l'approche scientifique (Dumez, 2011). Elle préconise de débuter les descriptions au cœur même des phénomènes et de suivre les actions, sans présumer ou minimisant l'existence d'un contexte constitué de groupes sociaux, d'intérêts construits, de classes, d'habitus, et autres. En d'autres termes, l'ANT peut

[16] Le processus de construction technique des données RH est schématisé en annexe I, p. 280

être considérée comme une « technologie de la description » (ibid.). Elle ne se présente pas comme une théorie « de » quelque chose, mais plutôt comme une démarche opérationnelle ou un « guide d'utilisation » méthodologique. Ce que Latour (2005a) cherche à comprendre, décrire et expliquer réside dans l'abandon de l'idée selon laquelle le « social » est une propriété essentielle qui peut être découverte et mesurée (Czarniawska, 2006). Ainsi, la question n'est pas de déterminer dans quelle mesure quelque chose est social, mais plutôt de comprendre comment les objets, les individus et les idées se connectent et s'assemblent pour former des entités plus complexes.

En qualité de *HR business analyst*, j'ai donc été à même d'observer les interactions entre les *data scientists* et avec les données RH, étudiant attentivement leurs discours, leurs perceptions et la mise en pratique de leur travail (Akrich et al., 2006; Callon, 1986; Latour, 2005a).

4.2.2. Le bricolage méthodologique séquentiel : continuum entre recherche-action et ethnographie

Le corpus de recherche actuel révèle un manque de conclusions claires concernant les opportunités et les défis de la combinaison méthodologique de la recherche-action et de l'ethnographie (Piovesan, 2022). En effet, ces deux approches divergent sur plusieurs points. Tout d'abord, les chercheurs en recherche-action acquièrent une compréhension des phénomènes sociaux en intégrant action et réflexion, créant ainsi un lien direct avec les sujets étudiés. En revanche, l'ethnographie selon l'ANT (Akrich et al., 2006; Callon, 1986; Latour, 2005a), cherche à approfondir la compréhension des phénomènes sociaux par une description des interactions, en suivant les acteurs humains et non-humains sans intervenir directement. Par la suite, concernant la production de connaissances, la recherche-action co-construit activement ces dernières avec les participants, alors que l'ANT les extrait des interactions observées. Enfin, la recherche-action se caractérise par une posture engagée, soucieuse du type de changement qu'elle promeut et intervient activement dans le processus de transformation sociale (Saija, 2014). À l'opposé, l'ANT adopte une position d'observateur détaché, se limitant à documenter les interactions sans y participer.

Malgré ces distinctions, trois similarités peuvent être soulignées (Piovesan, 2022). Tout d'abord, ces approches partagent la particularité de générer des connaissances sensibles au contexte. Cela se traduit par une méthodologie basée sur des études de cas qualitatives, qui rejettent les généralisations en faveur de réflexions « relatives ». Ensuite, les deux approches s'efforcent de mettre en lumière des représentations de la réalité qui se veulent pluralistes. Par exemple, la recherche-action intègre la perspective des praticiens en cherchant à les impliquer activement, reconnaissant ainsi leur expertise et leur connaissance pratique (Lüscher & Lewis, 2008). De son côté, l'ANT prend en compte la perspective des acteurs non-humains, soulignant leur rôle et leur capacité d'agir dans la configuration des réseaux socio-numériques (Akrich et al., 2006; Callon, 1986; Latour, 2005a). Cette inclusivité permet d'élargir la compréhension des réalités étudiées en prenant en compte une diversité d'acteurs et de points de vue. Enfin, tant la recherche-action que l'ANT reconnaissent l'importance de la transparence pour garantir la validité de la recherche. La recherche-action s'efforce d'assurer la transparence des processus de création et de manipulation des objets auprès des praticiens (Lüscher & Lewis, 2008), tandis que l'ANT met en avant la nécessité de rendre compte de manière transparente des relations et des influences multiples qui façonnent un réseau donné.

Le Tableau 8 résume les similarités et différences entre la recherche-action et l'ethnographie selon l'ANT.

Tableau 8 : Similarités et différences entre la recherche-action et l'ethnographie selon l'ANT

Similarités	**Différences**	
	Recherche-action	**Ethnographie selon l'ANT**
Sensibilité au contexte	Combinaison d'action et de réflexion	Combinaison de description et de réflexion
Représentations pluralistes	Co-création des connaissances avec les participants	Extraction des connaissances des interactions des participants
Transparence	Engagement éthique et intentionnalité	Observations détachées

Par ailleurs, afin de justifier mes choix méthodologiques, il convient de préciser que je n'avais initialement pas envisagé la transition séquentielle entre ces deux approches. Toutefois, ma présence sur le terrain, notamment en raison des rôles qui m'ont été attribués, a suivi une trajectoire non linéaire, davantage façonnée par les circonstances et les contingences fortuites que par un processus méthodologique préétabli. Compte tenu de ce changement de perspective, il m'était impératif d'explorer les éventuels parallèles permettant de saisir le potentiel succès de cette transition séquentielle.

Ainsi, la combinaison entre recherche-action et ethnographie selon l'ANT a joué un rôle essentiel dans ma capacité à m'adapter à l'évolution de mon positionnement. Impliquée dans la recherche-action, j'ai pu maintenir une perspective orientée vers les objectifs du projet RH. Parallèlement, en tant qu'observatrice, j'ai pu réfléchir au travail de construction des données RH en utilisant un langage conceptuel adapté, ce qui a renforcé ma compréhension des acteurs (humains et non-humains) impliqués dans le réseau à l'étude. Par ailleurs, en suivant les recommandations de Lewis (2007), cette combinaison a également permis de tenir compte de la progression de mes actions et de mes intérêts en fonction des rôles que j'ai assumés. Cela m'a permis d'appréhender de manière réflexive les interactions formelles et informelles - souvent imprévues - auxquelles j'ai été confrontée.

5. Présentation des données

Dans l'élaboration de ma démarche de recherche, j'ai opté pour un découpage temporel (Langley, 1999) selon les principes de la recherche-action (Eden & Huxham, 1996; Greenwood & Levin, 2007; Herr & Anderson, 2015). Ce découpage a permis de structurer ma démarche en trois phases distinctes :

1. La phase exploratoire ;
2. La phase de collecte ;
3. La phase d'analyse.

Elle sont décrites ci-dessous, détaillées selon leur calendrier et leurs objectifs. Le Tableau 9 donne un aperçu des principales sources de données.

La collecte des données empiriques est réalisée de manière flexible tout en étant systématique. En tant que responsable des comptes-rendus au sein du projet à

l'étude, j'ai intégré sans difficulté mes observations (expressions, échanges, etc.), mes réflexions et mes moments d'étonnement au fur et à mesure de mes activités. Pour ce faire, j'ai utilisé l'outil de prise de notes numérique *Microsoft OneNote* afin de systématiser mon suivi. Ce bloc-notes multifonctionnel a rendu plus aisée l'intégration de mes notes avec les détails de mes réunions, y compris les sujets abordés et les participants, en utilisant les informations de mon calendrier *Microsoft Outlook*, ainsi que les courriels correspondants. De plus, les multiples versions des supports et documents de travail produits par les *data scientists*, qui ont été une source de données empiriques essentielle pour mon étude, ont été systématiquement consignées et classées chronologiquement. Cette organisation a permis d'intégrer efficacement ces documents aux notes recueillies sur le terrain. Grâce à l'interopérabilité des outils de la suite *Microsoft*, la centralisation de ces informations a considérablement été simplifiée.

Tableau 9 : Synthèse des données collectées

Sources de données	Données collectées	Caractéristiques
190 entrées de journal de bord	155 entrées de journal de bord pour 155 réunions comme membre active du projet (réunions internes et externes). 35 entrées de journal de bord pour retracer les interactions et les réflexions lors de ma présence sur le terrain.	Les réunions duraient en moyenne une heure. Les entrées du journal de bord varient considérablement en longueur selon qu'elles capturent ma réflexion, une réunion avec les membres du projet ou un échange prolongé avec ma directrice de thèse. Elles vont de 1 à 5 pages.
16 entretiens* * Certaines personnes sont interviewées à plusieurs reprises.	- **Phase exploratoire** : 1 entretien avec le chef de projet, 1 entretien avec un *data scientist* et 1 entretien avec une directrice associée. - **Phase de collecte** : 1 entretien avec le chef de projet, 1 entretien avec le *data scientist* junior, 1 entretien avec le *data scientist* senior, 2 entretiens avec des partenaires externes et 2 avec des praticiens RH. **- Phase d'analyse*** : 2 entretiens avec le directeur de projet, 2 entretiens avec les chefs de projet, 1 entretien avec le *data scientist* senior et 1 entretien avec un partenaire externe. * La phase d'analyse désigne la période marquant l'arrêt de ma participation active au projet, bien que je sois toujours employée par le cabinet de conseil.	Les entretiens ont duré entre 20 et 90 minutes. La plupart des entretiens ont été enregistrés. Les entretiens sont principalement semi-directifs, offrant ainsi une certaine liberté aux personnes interrogées. La grille d'entretien a été régulièrement mise à jour en fonction de la spécificité du rôle de chaque participant au sein du projet.

Sources de données	Données collectées	Caractéristiques
Descriptions exhaustives	3 descriptions détaillées du processus de construction des données RH articulées autour des phases principales du projet : (1) la structuration, (2) la modélisation et (3) la commercialisation.	Ces descriptions, dont la longueur varie de 10 à 25 pages, ont été révisées à plusieurs reprises, d'une à cinq versions, et ont été élaborées en coopération avec les *data scientists* afin de garantir une perspective pluraliste de l'étude.
400 documents d'archives	Plus de 323 mails et plus de 100 documents, dont un nombre significatif de descriptions des postes, de rapports et de présentations.	La longueur des documents varie de 1 à 10 pages. Les documents analysés comprennent principalement des présentations de projets destinées aux secteurs bancaire et assurantiel, des offres commerciales, ainsi qu'un modèle de compétences du cabinet qui décrit les parcours d'évolution des salariés vers des rôles techniques ou de gestion de projet. Le corpus inclut également des articles de conférence et des synthèses de veille de marché.

5.1. Phase exploratoire

La phase initiale de mon étude a débuté en janvier 2021, sous la coordination de Xavier, le directeur de projet. Une série de réunions préliminaires a été organisée pour lancer les travaux de conception de *DSN Analytics*, un nouvel outil d'IA destiné à la gestion de l'absentéisme. Lors de ces réunions, nous avons analysé les résultats d'une étude sur l'absentéisme menée en 2019 pour une entreprise de nettoyage industriel, ce qui a permis l'élaboration d'un premier inventaire des données RH pertinentes pour comprendre le phénomène. En parallèle du développement technique de l'outil, j'ai également contribué à la création de sa première offre commerciale.

Conformément aux méthodologies recommandées par Herr & Anderson (2015), j'ai par la suite établi une liste de données importantes pour l'étude. Cette liste comprenait des entretiens semi-directifs, des documents internes, des supports de présentation, des courriels, et des observations sur le terrain. Cette phase préparatoire, et en particulier la première série de réunions, a été essentielle pour développer une compréhension initiale des acteurs et des dynamiques au sein du projet.

5.2. Phase de collecte

La deuxième phase de mon étude, s'étendant de février 2021 à octobre 2022, est caractérisée par ma transition vers le rôle de cheffe de projet de janvier à mars 2022. Pendant cette période, mon intégration complète dans l'équipe projet m'a permis de contribuer activement à l'élaboration et au suivi de la feuille de route pour le développement de *DSN Analytics*. En outre, j'ai participé à plus de 155 réunions à la fois internes et externes et mené plus de sept entretiens approfondis. La grille d'entretien est détaillée en annexe III, p. 289.

Les sessions de travail ont été essentielles non seulement pour faciliter la collecte de données contextuelles mais aussi pour stimuler l'émergence des diverses unités de sens, qui seront examinées plus en détail dans la section suivante. Intégrées aux développements commerciaux et techniques de *DSN Analytics*, ces sessions ont facilité mes observations, tant participantes que non-participantes, auprès d'une diversité d'acteurs exprimant leurs perspectives. Cette interaction a substantiellement enrichi l'analyse du processus de construction des données RH.

De plus, mon engagement soutenu dans le partage de connaissances en GRH s'est concrétisé par plusieurs présentations destinées à l'équipe. Ces exposés avaient pour but d'approfondir la compréhension, par les *data scientists*, des enjeux liés aux activités et phénomènes RH, contribuant ainsi à enrichir l'approche globale du projet.

5.3. Phase d'analyse

La phase post-terrain de mon étude a été structurée en deux étapes distinctes. La première étape s'est concentrée sur la rédaction de descriptions détaillées du processus de construction des données RH articulées autour des trois phases principales du projet :

1. La structuration ;
2. La modélisation ;
3. La commercialisation.

Ces descriptions, variant de 10 à 25 pages, ont subi plusieurs révisions - allant d'une à cinq versions - et ont été élaborées en collaboration avec les *data scientists* pour assurer une représentation pluraliste de l'étude. Un exemple de description est présenté en annexe II, p. 282. Cette collaboration a ainsi permis une meilleure compréhension du travail de construction des données RH et a aussi contribué à une reconnaissance plus précise de ma présence au sein du cabinet. Elle a également facilité la prise de recul des *data scientists* vis-à-vis leurs décisions et leurs arbitrages.

La seconde étape a été motivée par la nécessité de traiter et de donner sens à la vaste quantité de données recueillies sur une période de 21 mois. Pour ce faire, j'ai réalisé des analyses quantitatives simples afin d'identifier les moments clés du projet, en développant trois indicateurs appliqués à deux niveaux de granularité : le projet dans son ensemble et ses différentes phases (structuration, modélisation et commercialisation) :

1. L'intensité mensuelle : il s'agit du nombre total de réunions par mois. Cet indicateur permet d'examiner la dynamique globale de charge de travail lors du projet, en identifiant les pics et les chutes d'activité.
2. L'intensité cumulée : cet indicateur classe les acteurs en fonction de leur présence cumulée lors des réunions. Ce classement permet notamment de filtrer les acteurs

du projet selon différents critères tels que leur importance relative (primaires/secondaires) et leur affiliation organisationnelle (internes/externes).

3. La complexité : cet indicateur représente le nombre moyen d'acteurs différents présents lors des réunions. Il permet d'identifier les moments d'inclusion et d'exclusion des participants dans le processus de construction des données RH.

Ces analyses avaient ainsi pour objectif d'examiner la structure et la dynamique du réseau d'acteurs, en quantifiant les réunions pour comprendre le découpage temporel du processus de construction des données RH (Langley, 1999). Il est toutefois important de préciser que cette « étape quantitative » n'apporte qu'une valeur ajoutée limitée à cette étude, dont la démarche repose essentiellement sur une approche qualitative. Je ne la considère donc pas comme indispensable à la réussite de cette recherche. En effet, les résultats sont affectés par plusieurs biais, notamment l'absence d'accès au plan de charge de chaque acteur et le fait que seules les réunions auxquelles j'ai participé ont été comptabilisées. Les résultats de ces analyses secondaires ont principalement servi à préparer mon analyse qualitative principale, notamment en identifiant les acteurs clés à mettre en évidence dans les résultats.

6. Méthodes d'analyse des données

Étant donné que cette étude repose sur une analyse approfondie de la nature des données RH, le choix des *data scientists* d'utiliser la *DSN* comme base pour la conception de leur outil d'IA a particulièrement retenu mon attention. Je me suis notamment interrogée sur les raisons pour lesquelles ces données RH étaient jugées plus pertinentes que d'autres en matière de gestion de l'absentéisme.

Pour répondre à cette question, je me suis orientée vers des cadres conceptuels qui intègrent les principes de l'ANT et une perspective socio-économique. En combinant les concepts de « connaissances » de Latour (2005a, 2007) et des « qualités » des biens économiques de Callon et al. (2000, 2002), j'ai pu effectuer un découpage temporel (Langley, 1999) fondé sur une théorisation du processus de construction des données RH :

1. La Qualification : concerne l'évaluation des qualités des données RH. Il s'agit de déterminer les critères initiaux qui définissent la singularité potentielle des données RH.

2. La Capitalisation : examine les projets qui s'appuient sur la qualification pour développer des connaissances. L'objectif est de concevoir la fonction épistémique des données RH, permettant leur singularisation en tant que biens économiques.
3. La Requalification : consiste à requalifier les données RH à partir des projets de capitalisation.

La Figure 8 schématise ce découpage en illustrant comment les données RH passent par une première séquence de qualification, suivie de boucles successives entre capitalisation et requalification, formant ainsi une spirale dynamique de construction des données RH jusqu'à leur transformation en biens économiques[17].

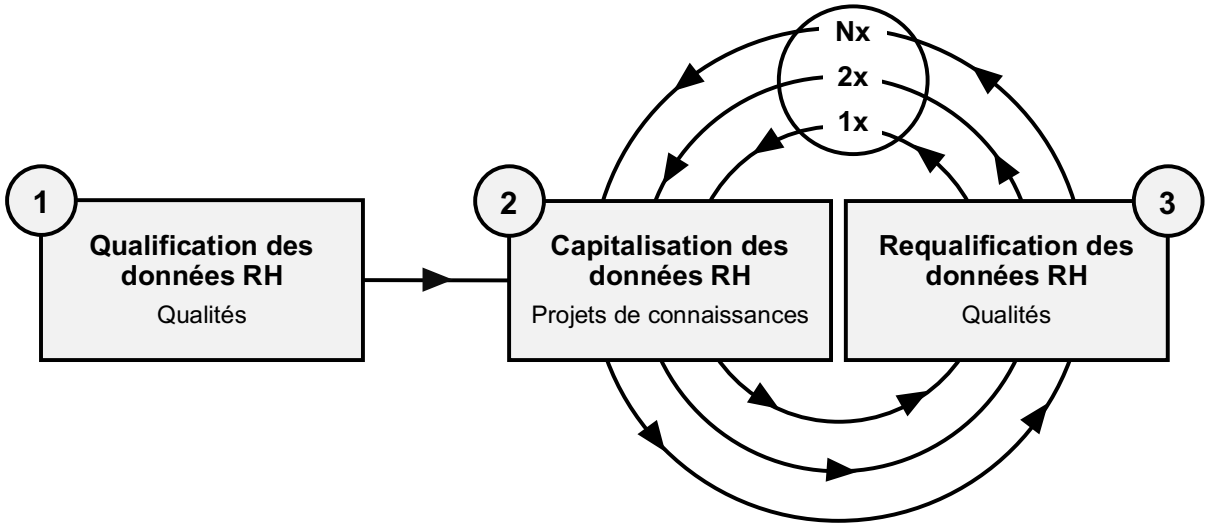


Figure 8 : Découpage temporel des trois séquences du processus de construction des données RH

Sur la base de ce découpage, j'ai procédé au codage des différentes unités de sens à partir de l'importante quantité de données collectées. Guidée par une approche abductive, j'ai d'abord conceptualisé les données RH en tant que biens économiques. Cette démarche a nécessité d'explorer les interactions entre les concepts d'offre et de demande dans une logique de marchandisation, établissant ainsi des bases solides pour les phases ultérieures de codage.

[17] La Figure 8, présentée dans le cadre théorique (voir Figure 6), mérite néanmoins d'être rappelée, car elle illustre le processus de construction des données RH, élément central de cette thèse.

Partant de cette conceptualisation, mon approche m'a conduit à une exploration détaillée des composantes d'une demande de marché, cette dernière étant intégrée dans les séquences de (re)qualification du processus. Celle-ci a souligné l'importance de clarifier les besoins des clients et de fixer un prix de vente, ce qui implique d'évaluer les ressources financières et humaines nécessaires pour le projet. À la suite de cette exploration, trois principales unités de sens ont été identifiées :

1. (Re)définition des besoins des clients ;
2. (Re)rationalisation des coûts d'investissement ;
3. (Ré)enrôlement d'agents économiques clés.

Ensemble, ces trois unités de sens ont mis en évidence des controverses de (re)qualification des données RH.

Une fois ces unités de sens identifiées, il a ensuite été essentiel de les intégrer dans la formulation de l'offre. Cette dernière, développée dans le cadre de la séquence de capitalisation, repose sur l'accumulation des connaissances issues de la demande et tire parti du savoir-faire des agents impliqués. En s'appuyant sur ces éléments, trois unités de sens ont été distinguées pour cette séquence :

1. Territoire d'exploration épistémique des données RH ;
2. Savoir-faire des agents économiques ;
3. Epreuves d'exploration.

À l'instar des séquences de (re)qualification, ces trois unités de sens ont également mis en lumière des controverses entourant la capitalisation des données RH.

Au total, j'ai identifié six unités de sens distinctes, réparties en trois catégories qui correspondent aux trois séquences du processus de construction des données RH (*QCR*). Ils révèlent comment le réseau d'agents économiques utilise les compromis issus de controverses pour progresser séquentiellement dans la conception de la fonction épistémique des données RH et affirmer leur singularité en tant que biens économiques.

Afin de retracer les trajectoires de ces six unités de sens, j'ai codé toutes les citations pertinentes dans *Microsoft Excel* à partir de mes données, incluant les entretiens, les notes de terrain, les échanges formels et informels, ainsi que la documentation. Sur la base de cette analyse, j'ai extrait les dialogues, mes

observations et les fragments d'histoires pour chaque unité de sens afin de construire une narration composite pour chacune d'elles (Langley, 1999). Cela a constitué la base pour analyser la trajectoire de ces unités de sens au fil du temps, ainsi que les modes d'existence des différents agents économiques impliqués dans le projet et leurs interactions.

Ce processus itératif de comparaison des données collectées et le raffinement du processus de construction des données RH s'est poursuivi jusqu'à la formulation des résultats. La Figure 9 présente le processus de construction des données RH, les unités de sens associées à son réseau et leur transition vers les résultats empiriques. Par la suite, le Tableau 10 détaille le codage en intégrant certaines des citations correspondantes.

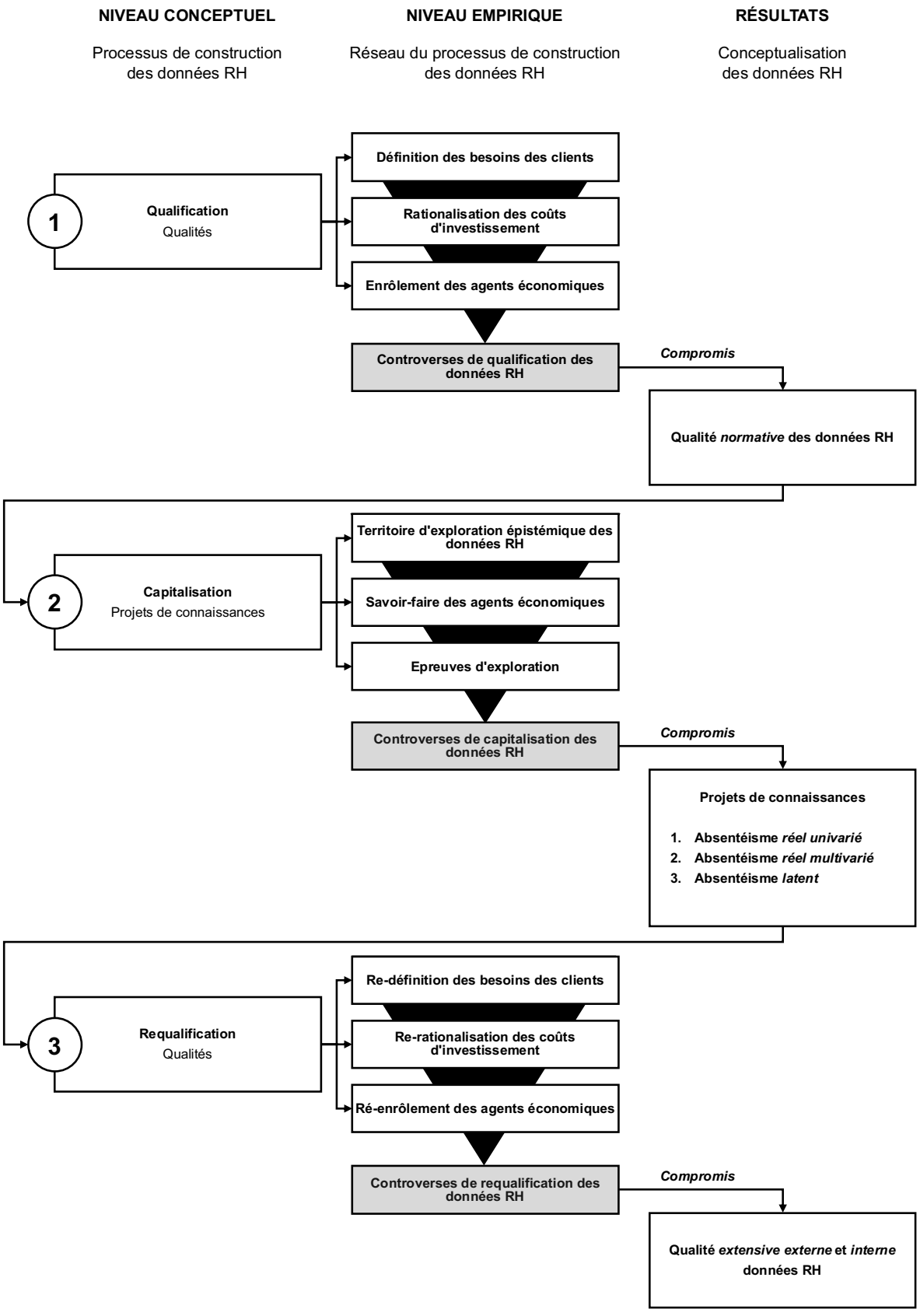


Figure 9 : Représentation conceptuelle et empirique du processus de construction des données RH

Tableau 10 : Représentation conceptuelle et empirique détaillée du processus de construction des données RH

Niveau conceptuel	Citations correspondantes	Niveau empirique	Résultats
QUALIFICATION	*« Ne me dites pas que vous ne pouvez pas proposer mieux 😉. »*	Définir les besoins des clients	Qualité *normative* des données RH
	« […] *Je le dis sans arrogance, j'espère. Mais je n'ai pas encore vu d'offre, dans ce que j'ai regardé, qui permette vraiment de partir de la data et d'arriver à l'élever jusqu'à l'action opérationnelle »*		
	« Le sujet de l'absentéisme est très bien identifié par les entreprises, les RH et toutes les sociétés de conseil qui traitent les sujets RH. […] *Il est peu probable de sortir des choses nouvelles sur le sujet car tout a été un peu dit… »*		
	« […] *assurer une lecture à plusieurs niveaux et partager une vision commune* [de l'absentéisme] *»*		
	« […] *Ce qui était toute la complexité. Et souvent avec des données qui n'étaient pas conformes : des gens qui arrivent dans l'entreprise après en être partis, des gens qui avaient moins de 15 ans, des salaires négatifs, plein de choses comme ça. »*	Rationaliser les coûts d'investissement	
	« […] *C'est la notion de standard. Il n'y avait pas besoin de l'adapter à chaque fois. Ça permettait de gagner des coûts aussi pour nous et pour le client. »*		
	« Quelle est la démarche la plus pragmatique et qui mobilise le moins de ressources ? »	Enrôler des agents économiques	
	« C'est du win-win. […] *tu vas compléter mon portefeuille d'offres. »*		

Niveau conceptuel	Citations correspondantes	Niveau empirique	Résultats
	« Pour aller sur le marché, on voulait des gens qui connaissaient le marché… »		
CAPITALISATION	*« C'est une intuition. Moi, je sais faire le calcul de l'analyse de l'absentéisme* […] *»*	Territoire d'exploration épistémique des données RH	Premier projet de connaissances : l'absentéisme *réel univarié*
	« Ça, lui [l'entreprise] permet de s'éclairer, de se dire ok, j'ai un absentéisme qui est quand même un peu trop important… »		
	« Est-ce que tu prends le salaire divisé par 12 ou tu prends le salaire annuel ? Tu prends le salaire avec prime ? […] *Tu vois pleins de choses… »*	Savoir-faire des agents économiques	
	« […] *Il se trouve que plus on va faire des moyennes sur un grand groupe, plus on va commencer à avoir des résultats qui sont fiables… »*		
	« Il manquait certaines valeurs, certaines variables socio-démo[graphiques] *qu'on aurait aimées avoir… »*	Epreuves d'exploration	
	« Une absence qui commençait le 21 décembre de l'année et qui durait six mois […] *C'était un peu absurde de considérer que c'était une absence de trois jours sur l'année où l'absentéisme est arrivé… »*		

Niveau conceptuel	Citations correspondantes	Niveau empirique	Résultats
	« Le but, c'est de donner les clés aux managers opérationnels pour se dire : j'ai vu l'absentéisme dimension par dimension. Est-ce qu'il y a des combinaisons de facteurs qui sur-concentrent l'absentéisme ? »	Territoire d'exploration épistémique des données RH	Deuxième projet de connaissances : l'absentéisme *réel multivarié*
	« […] *côté QIA c'est un algo[rithme] qu'on a maintenant l'habitude de manipuler »*	Savoir-faire des agents économiques	
	« […] *Mais en réalité, au-delà de quatre critères, ça n'a aucun sens. »*	Epreuves d'exploration	
	« Ce n'est pas trop dur d'interpréter les sous-groupes, parce qu'il y a une trop grande complexité au niveau des règles [conventions de mesure] *? »*		
	« Le but est de dire : quel est l'absentéisme auquel je t'attends compte tenu de tes caractéristiques socio-démographiques ?»	Territoire d'exploration épistémique des données RH	Troisième projet de connaissances : l'absentéisme *latent*
	« […] *Si l'absentéisme réel est largement supérieur à l'absentéisme attendu, ça signifie que l'entité ou le métier est anormal et que cet absentéisme s'explique par d'autres facteurs que les caractéristiques socio-démographiques… »*		
	« […] *l'absentéisme ne dépend pas [par exemple] de l'âge de façon linéaire. Peut-être que l'écart entre 30 et 25 ans n'est pas le même que l'écart entre 50 et 45 ans* […]*. Il peut y avoir un effet de plafond et donc limitant, parce que les effets* [présumés des données RH] *sont linéaires… »*	Savoir-faire des agents économiques	

Niveau conceptuel	Citations correspondantes	Niveau empirique	Résultats
	« […] *Je trouvais qu'il y avait en même temps un sens* [métier] *et un challenge technique qui étaient intéressants. »*		
	« Ça pose un challenge technique parce qu'il faut réussir à le faire »		
	« C'est ce qui manquait. Il fallait que ça prenne [en compte] plus de facteurs explicatifs pour que ce[la] *soit utile. »*	Epreuves d'exploration	
	« […] *Je me suis cassé les dents [à la difficulté] de [combiner] la fréquence et la sévérité avec effet mixte et de pouvoir les mélanger à la fin.* […] *dès que j'ai récupéré les données, j'ai essayé mais je n'ai pas réussi* […]. *»*		
	« […] *Par exemple, attribuer une crédibilité de 5000 % [de la modélisation de l'absentéisme [latent] à un établissement [entité] n'a pas de sens... »*		
REQUALIFICATION	*« Si tu déroules un processus courant, je pense que tu commences justement par te comparer à l'extérieur : par rapport aux autres vous avez plus d'absentéisme. Et puis après, tu descends de plus en plus finement… »*	Re-définir les besoins des clients	Qualité *extensive externe* des données RH
	« L'idée est d'arriver avec un truc clé en main, avec un modèle qui a été calibré sur tout le monde et sur un panel varié […] *la plus-value métier est assez claire… »*		

Niveau conceptuel	Citations correspondantes	Niveau empirique	Résultats
	« […] *une sorte de service freemium, c'est-à-dire je fais le benchmark standard gratuitement. Et pour les entreprises qui veulent s'améliorer, elles peuvent accéder à une version sur mesure* [premium] *qui permet d'assigner des analyses. »*	Re-rationaliser les coûts d'investissement	
	« On faisait le bench[mark] gratuit parce qu'on n'avait en vrai jamais utilisé de données DSN pour nous. Et du coup, c'était l'occasion quand même de saisir de vraies données […] *»*		
	« […] *le truc dans la démarche commerciale, c'est vraiment de commencer par le carnet d'adresses. Donc, c'est le mien, c'est celui de QIA… »*	Ré-enrôler des agents économiques	
	« […] *Travailler avec la CNPR, c'est quand même un facteur de sérieux sur le marché* […] *Ils disposent de 1,5 million d'entreprises qui envoient leur DSN chaque mois. »*		
	« […] *par l'association IISS-QIA, le développement de benchmark offrira à chaque entreprise un positionnement stratégique face à son secteur et lui permettra d'enclencher des analyses spécifiques si elle le souhaite. »*		
	« […] *On vous propose exactement ce dont vous avez besoin et pas une machine de guerre pour tuer une mouche… »*	Re-définir les besoins des clients	Qualité *extensive interne* des données RH
	« […] *Le sur-mesure va aller [par exemple] prendre des pointages dans des outils de production et toutes les affectations… »*		

Niveau conceptuel	Citations correspondantes	Niveau empirique	Résultats
	« En fait, DSN égal partout pareil, égal je peux faire un produit. À partir du moment où chaque SIRH n'a pas les mêmes informations, pas structurées de la même façon, tu ne peux pas faire un produit plug-and-play… »	Re-rationaliser les coûts d'investissement	
	« Ça restera le cœur. Déjà 80 % des informations qui ont de la valeur, elles sont dans la DSN. Ensuite, 90% de la valeur est extraite de la DSN parce que la DSN est de bonne qualité… »		
	« […] *un travail empirique de lobbying visant à exercer une influence sur les influenceurs potentiels pour créer une sorte de bruit de fond. »*	Ré-enrôler des agents économiques	
	« Je me suis dit : je pense qu'il faut d'abord cibler les boîtes qui ont un vrai problème d'absentéisme et ayant un fils interne qui galère dans les hôpitaux, il n'arrête pas… »		
	« Je pense que les interlocuteurs qu'on avait étaient très métiers et pas data […] *s'il n'y a aucune personne data, ils ne peuvent pas juger comme il faut notre offre. »*		

7. Conclusion

L'approche méthodologique choisie pour explorer le processus de construction des données RH est fondée sur une approche qualitative et séquentielle. Elle s'inscrit dans un continuum entre recherche-action et ethnographie. Elle est rendue possible par une immersion de trois ans et sept mois au sein d'un cabinet de conseil en *data science*, dans le cadre d'une convention *CIFRE*. Le processus de construction des données RH est analysé à travers la conception de *DSN Analytics*, un outil d'IA destiné à l'analyse de l'absentéisme.

La validité de cette approche est assurée par la transparence du travail de recherche, soulignant que, malgré son caractère évolutif, elle reste rigoureuse et crédible. De plus, sa flexibilité inhérente s'avère indispensable pour naviguer à travers la « *science en train de se faire* » (Latour, 2005a, p. 29).

La démarche de recherche est structurée en trois grandes phases :

1. La phase exploratoire ;
2. La phase de collecte ;
3. La phase d'analyse.

De la phase d'analyse émergent six unités de sens regroupées en trois catégories et correspondant aux trois séquences du processus de construction des données RH (*QCR*). Ces catégories soulignent des controverses dont les compromis contribuent à développer progressivement la fonction épistémique des données RH, facilitant ainsi leur transformation en biens économiques.

Enfin, l'exploration du processus de construction des données RH conduit également à questionner comment la dimension « RH » des données est préservée ou modifiée. Plus spécifiquement, cette question se concentre sur les représentations des agents économiques à l'égard de la fonction RH, qualifiée de « traditionnelle », qui appartient au contexte « d'origine » de ces données.

Partie II. PROCESSUS DE CONSTRUCTION DES DONNES RH

Chapitre 3. Séquence de Qualification des données RH

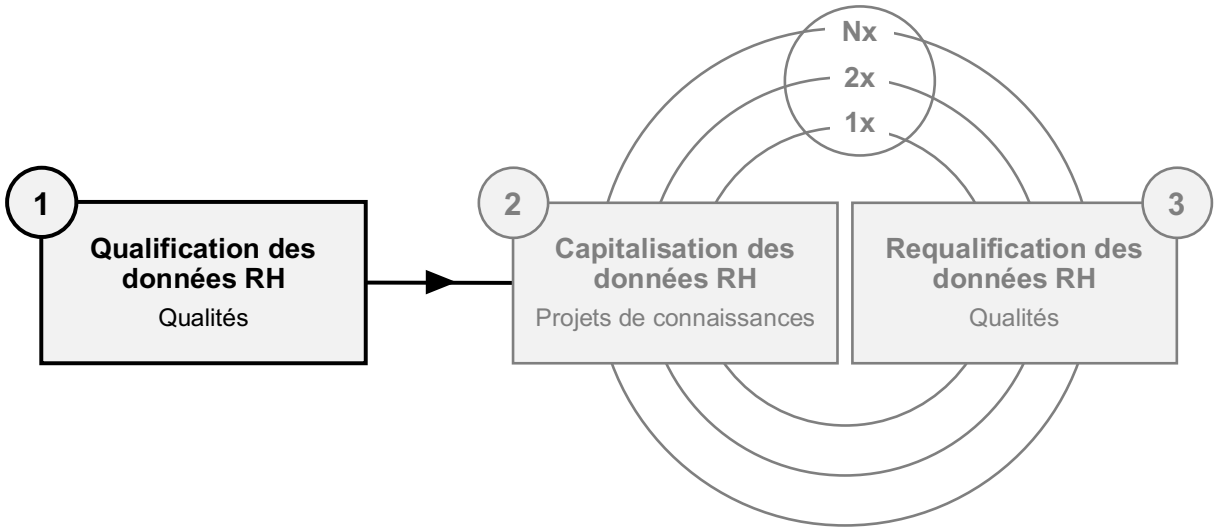


Figure 10 : Séquence de Qualification des données RH

1. Introduction

Ce troisième chapitre se concentre sur la première séquence du processus de construction des données RH : la Qualification. Celle-ci concerne l'évaluation des qualités des données RH, en déterminant les critères initiaux qui définissent leur singularité potentielle.

La qualification met en lumière une première qualité : la qualité *normative* des données RH.

Cette séquence est abordée à travers trois actes clés :

1. La définition des besoins des clients, présentée dans la première section.
2. La rationalisation des coûts d'investissement, examinée dans la deuxième section.
3. L'enrôlement des agents économiques, détaillé dans la troisième section.

L'analyse des controverses, au cours desquelles ces agents économiques négocient la qualité *normative* des données RH, est ensuite approfondie.

2. Qualité *normative* des données RH

2.1. Acte I : définition des besoins des clients

La qualification d'un bien économique nécessite la prise en compte des idées préconçues sur les caractéristiques du marché cible. En imaginant les clients potentiels avec leurs préférences et leurs motivations, les *data scientists* anticipent également leurs besoins actuellement non satisfaits et les origines de ce manque.

Dans ce premier acte de qualification, les *data scientists* définissent les besoins des clients en se basant sur deux mythes qui serviront de fondement à la réforme de la qualification traditionnellement établie des données RH.

2.1.1. Mythe I : le manque d'expertise en données RH

Le 19 janvier 2021, Xavier (directeur de projet) partage l'offre commerciale d'un concurrent, *Fast MS*. Cette offre, qui inclut un outil exploitant également les données *DSN*, est qualifié par ce dernier de « standard ».

Fast MS propose une analyse détaillée de l'absentéisme aux grandes entreprises et à leurs courtiers (avec un minimum de 500 salariés). La plateforme offre un accès à plus de 400 indicateurs sociaux ainsi qu'à différents rapports RH. Leur outil se compose de trois écrans principaux, comme présenté dans le Tableau 11.

Fiches	**Données RH**
Identité entreprise	Répartition des effectifs par sexe
	Age moyen des salariés
	Ancienneté moyenne par statut
	Rémunération moyenne des salariés
	Évolution des effectifs
	Synthèse des effectifs
Identité absentéisme	Taux d'absentéisme par cause d'absence
	Taux d'absentéisme par entité
	Effectif, nombre d'absence, nombre de salariés absents
	Nombre de jours d'absence mensuelle par type d'absence
	Nombre de journées d'absence par motif
	Coût direct des journées d'absence par motif
Rapport complet sur l'absentéisme (ce dernier exportable en PDF)	

Tableau 11: Éléments tirés de l'offre commerciale « Analyse de l'absentéisme » de *Fast MS* (2020)

Dans ce même courriel, Xavier (directeur de projet) rapporte un commentaire légèrement provocateur de Benoît (partenaire en actuariat) qui lui avait initialement partagé l'offre de *Fast MS* : « *Ne me dites pas que vous ne pouvez pas proposer mieux* 😉. » (Courriel envoyé le 19/01/2021).

Le premier acte de qualification des données RH, découlant du courriel de Xavier (directeur de projet), consiste donc à définir les besoins des clients en utilisant l'offre standard de *Fast MS* comme point de référence dans le marché RH. Cela implique l'alignement d'une offre prospective adaptée aux besoins imaginés des clients potentiels et une demande présupposée de ces derniers pour l'offre de *Fast MS*.

Ce premier acte adopte ainsi une approche dualiste : d'une part, il s'efforce de différencier les données RH en les positionnant de manière unique sur le marché RH afin de combler les lacunes non couvertes par *Fast MS* ; d'autre part, il cherche à les aligner avec celles de son concurrent pour répondre aux attentes des clients potentiellement intéressés par ce dernier. C'est à travers cette dualité de différence-similitude, de singularité-comparaison, que Xavier (directeur de projet) et son équipe visent à réformer le marché RH existant :

- Justine (directrice de projets stratégiques) : « *Effectivement, très facile de refaire ça avec l'expérience qu'on a eue* […] *et en développant une bonne connaissance des données de la DSN…* » (Courriel envoyé le 20/01/2021).
- Xavier (directeur de projet) : « *Catherine, essaie d'avoir autant que possible une vision des données présentes dans la DSN d'ici-là* 😉. » (Courriel envoyé le 19/01/2021).

Quels besoins sont identifiés par Xavier (directeur de projet) dans cette séquence ?

« […] *Je le dis sans arrogance, j'espère. Mais je n'ai pas encore vu d'offre, dans ce que j'ai regardé, qui permette vraiment de partir de la data et d'arriver à l'élever jusqu'à l'action opérationnelle »* (Xavier, directeur de projet, entretien no1, 31/10/2023).

L'évaluation que réalise Xavier (directeur de projet) du marché RH actuel révèle une lacune dans les offres existantes, mettant en évidence l'absence d'un lien direct entre les données RH et les actions opérationnelles. Cette observation suggère que l'offre qu'il proposera avec son équipe ne reposera pas uniquement sur la gestion de l'absentéisme, mais plutôt sur un savoir-faire avancé dans le traitement des données RH. L'absence de ce savoir-faire crée un manque, et donc un besoin, sur le marché. Face à cette lacune, Xavier (directeur de projet) vise à construire des données qui dépassent les simples connaissances *descriptives*, telles que celles proposées par *Fast MS*, pour véritablement transformer le marché RH.

Cette évaluation est par ailleurs corroborée par Pierre, l'un des fondateurs du cabinet de conseil, qui insiste sur l'importance de mettre en avant ce savoir-faire spécifique comme un élément différenciateur de l'offre :

« *Le sujet de l'absentéisme est très bien identifié par les entreprises, les RH et toutes les sociétés de conseil qui traitent les sujets RH.* […] *Il est peu probable de sortir des choses nouvelles sur le sujet car tout a été un peu dit… Qu'est-ce que nous devons mettre en avant et qui est très différenciant de tout ce qu'on peut trouver sur le marché:* […] *l'expertise de QIA en valorisation de données complexes ;* […] *Notre savoir-faire dans le développement d'applications digitales métier, déployables facilement et très opérationnelles pour les métiers ;* […] *Notre connaissance sectorielle et nos 13 ans d'expérience en Machine Learning dans les secteurs les plus exigeants et les plus contraignants du BIG DATA* » (Pierre, fondateur, courriel envoyé le 30/11/2021).

L'adéquation entre l'offre et la demande, envisagée par Xavier (directeur de projet) et Pierre (fondateur) s'opère donc au sein d'un marché RH à la fois perçu comme déficient et saturé. D'après Pierre (fondateur) cette adéquation doit impérativement s'articuler autour d'une qualité qui met en valeur un savoir-faire distinctif, afin de susciter l'intérêt des clients potentiels.

2.1.2. Mythe II : le potentiel de généralisation des données RH

« […] *Pour aller plus loin, il faut réfléchir aux variables qu'on aura la capacité de recréer avec les données disponibles, en prenant l'exemple de celles qu'on avait créées pour GSM* […] *Celles qui me semblent accessibles a priori : toutes les variables d'absences ; subro*[gation] *; toutes les variables de carrière (classification, qualification, échelon, position, ancienneté, emploi / métier, type contrat de travail, convention collective, salaire, etc.…) ; les variables descriptives du travail (temps de travail, lieu de travail / société de rattachement, pénibilité du travail, fatigue) ; les variables démographiques (âge, sexe, handicap) ; éventuellement les données de formation (au minimum on doit pouvoir les approcher par le nb* [nombre] *de jours passés en formation, qui doit figurer comme motif d'absence)…* » (Justine, directrice de projet stratégiques, courriel envoyé le 20/01/2021).

Le recensement fait par Justine (directrice de projet stratégiques) des données RH d'intérêt, découle d'un premier projet sur l'absentéisme au sein de *GSM*, une entreprise spécialisée dans le nettoyage industriel. Délégué par Benoît (partenaire en

actuariat), ce projet avait pour objectif d'évaluer le risque d'absentéisme et d'identifier les facteurs déterminants en exploitant les données disponibles dans les SI de l'entreprise. Si certaines de ces données étaient issues du SIRH (cf. : données socio-démographiques), d'autres provenaient de sources différentes (cf. : management).

Dans ce projet, le travail des *data scientists* consistait à modéliser plus de vingt types de données différentes concernant les salariés et leurs absences, dans le but de détecter des sous-groupes présentant un risque élevé d'absentéisme. La segmentation a été effectuée selon trois catégories principales de risques présentée dans le Tableau 12.

Catégories de risques	Définitions
Abus	Cas d'absentéisme opportuns, comme des arrêts de travail alignés avec les périodes de maintien de salaire ou planifiés autour des congés payés.
Actions locales	Situations spécifiques liées à une entité ou un métier, telles que le manque de formation, la faible mobilité professionnelle ou une ancienneté limitée.
Situations complexes de personnes mal préparées	Défis organisationnels, tels que les projets d'envergure, les conditions de travail difficiles, les transitions managériales, ou encore les faibles marges de manœuvre, qui ont un impact particulièrement significatif sur les salariés confrontés à des handicaps, à un manque de formation, à une gestion inexpérimentée ou à une faible ancienneté.

Tableau 12 : Catégories de risques issues d'un premier projet sur l'absentéisme (2019)

Les conclusions tirées de ce projet sont de nature double. Premièrement, Xavier (directeur de projet) conçoit l'absentéisme non pas comme un phénomène spécifique et circonscrit à une entreprise, mais plutôt comme une « réalité » généralisée à l'ensemble des secteurs jugés pénibles. Par conséquent, les données RH exploitées durant cette étude et présentes dans la *DSN* sont considérées comme ayant un potentiel de généralisation lié à ce besoin sectoriel :

« [...] *la propreté dans le secteur agroalimentaire ou hospitalier sont des secteurs très pénibles* [...] *c'est aussi là où on a beaucoup d'absentéisme.* […] *Quand on*

analyse l'absentéisme [de *GSM*]*, ce n'est pas un sujet de propreté* [nettoyage industriel]*, c'est un sujet de pénibilité.* » (Xavier, directeur de projet, entretien no1, 31/10/2023).

Deuxièmement, les décisions opérationnelles relatives à la gestion de l'absentéisme proviennent principalement des cadres managériaux plutôt que la fonction RH qui assume un rôle de support. Les managers et directeurs opérationnels, étant directement impliqués dans le quotidien des salariés, détiennent une compréhension plus approfondie des motifs d'absences :

« […] *l'explication du pourquoi les gens s'absentent, elle est chez les managers et les directeurs d'entité et les RH ne font qu'accompagner les leviers* [d'action] *qui sont aussi, pour la plupart, du côté des managers*. » (Xavier, directeur de projet, entretien no1, 31/10/2023).

La problématique des clients potentiels cibles et de leurs usages émerge alors : Pierre (fondateur) définit la fonction RH comme une cible, tandis que Xavier (directeur de projet) privilégie davantage les managers et les directeurs opérationnels. Cette divergence de perspectives est clairement explicitée dans l'une des diapositives de la présentation de l'offre commerciale. Celle-ci détaille les usages de trois domaines fonctionnels :

1. Les directions RH ;
2. Les directions financières ;
3. Les directions opérationnelles.

Elle stipule que les données RH sont destinées à être exploitées par l'ensemble de ces directions dans l'objectif d'« *assurer une lecture à plusieurs niveaux et partager une vision commune* [de l'absentéisme] » (extrait issu de la présentation de l'offre, 13/01/2022).

La réticence de Xavier (directeur de projet) à positionner la fonction RH comme l'unique cible dans l'analyse de l'absentéisme, activité traditionnellement dévolue à cette fonction, est principalement motivée par deux considérations spécifiques. Ces dernières reflètent également celles des autres dirigeants de *QIA*.

La première considération est que les dirigeants estiment que la fonction RH perçoit de manière négative le terme « absentéisme ». Selon eux, cette perception négative dissuade la fonction RH de s'engager dans ce type de projet :

- Michel (directeur général, *QIA*) : « *Ça* [les conclusions de l'analyse de l'absentéisme] *va leur revenir dessus...* »
- Pierre (fondateur) : « *Oui… Il y a un intérêt possible pour le DRH mais qui peut craindre un retour de frappe. C'est un enjeu caché parce que ce sont d'énormes chantiers qui devront être mis en place...* » (Notes prises lors d'un échange informel, 06/07/21).

La deuxième considération est le constat par Xavier (directeur de projet) et les autres dirigeants de *QIA* d'un déficit significatif de compétences analytiques au sein de la fonction RH. Cette carence, aggravée par une potentielle aversion pour les technologies et les méthodes d'analyse avancées, constitue un ensemble de facteurs qui limitent la capacité de la fonction RH à conduire ce genre de projet :

« […] *Ne pas oublier que nous nous adressons surtout à des interlocuteurs non analytiques, non digitales native, et très peu portés sur les sujets Tec*[hnologiques], algo[rithimiques], *IA et tout ce qui tournent autour, je pense même que ces personnes ont plutôt un naturel réticent à ces solutions tec*[hnologiques]… » (Pierre, fondateur, courriel envoyé le 30/11/2021).

Cette perspective est corroborée par les interrogations de Xavier (directeur de projet) concernant l'aptitude de la fonction RH à exploiter et interpréter les données RH :

« *Les DRH n'arriveront pas à se saisir des chiffres, s'il n'y a pas de conseil sur la lecture des chiffres* […] *C'est une question de compétences analytiques. Le côté stratégique, c'est qu'il embarque les autres directions...* » (Xavier, directeur de projet, notes prises lors d'un échange téléphonique, 23/08/2021).

Le contexte de qualification des données RH met ainsi en lumière les perceptions de nombreux décideurs chez *QIA*, qui attestent de l'incapacité de la fonction RH à être le client cible de ce genre d'offre, la considérant davantage comme une courroie de transmission. Cette vision révèle un défaut d'alignement avec les exigences

analytiques et stratégiques nécessaires pour mener à bien la réforme de qualification proposée par le cabinet.

De plus, le choix de cibler un large éventail de directions parmi les clients potentiels suscite des questionnements quant à son impact sur la qualification des données RH. L'intérêt de Xavier (directeur de projet) à englober un public aussi vaste peut être interprété comme une phase exploratoire visant à sonder le marché RH. Cette démarche peut également refléter une prise de conscience de la nécessité d'approfondir sa compréhension des besoins spécifiques des clients potentiels afin d'assurer un éventuel succès commercial, tout en évitant un investissement financier substantiel.

Dans ce premier acte de qualification, Xavier (directeur de projet) et son équipe définissent les besoins des clients en se basant sur deux mythes qui visent à réformer la qualification traditionnellement établie en matière d'absentéisme. Le premier mythe identifie des lacunes dans le savoir-faire en *data science* sur le marché RH actuel, en prenant comme référence l'offre standard de *Fast MS*. Le second mythe, fondée sur leur expérience lors d'un projet antérieur sur l'absentéisme, révèle deux conclusions principales. Premièrement, les données RH présentes dans la *DSN* sont perçues comme ayant un potentiel de généralisation en raison d'un besoin émanant des secteurs jugés pénibles. Deuxièmement, les décisions opérationnelles concernant la gestion de l'absentéisme proviennent principalement des cadres managériaux, tandis que la fonction RH assume un rôle de soutien, montrant ainsi la transversalité du phénomène à l'échelle de l'entreprise.

Face à ce constat, les *data scientists* structurent ce premier acte de qualification de manière à répondre aux exigences de trois directions : (1) RH, (2) financières et (3) opérationnelles. Cette stratégie, visant à étendre leur portée pour évaluer le marché RH, permet également de minimiser les investissements financiers substantiels.

2.2. Acte II : rationalisation des coûts d'investissement

Comme pour tout projet, il est impératif de démontrer la viabilité économique en justifiant l'allocation optimale des ressources et minimiser les investissements nécessaires. Deux éléments clés permettent aux *data scientists* d'assurer ce deuxième acte de qualification des données RH :

1. La réduction des contraintes usuelles de qualification des données RH ;
2. L'exploitation de quatre outils préexistants qui facilite leur travail de qualification.

2.2.1. Réduction des contraintes usuelles de qualification des données RH

Comme exposé précédemment, l'analyse de l'absentéisme nécessite l'exploitation de multiples bases de données provenant des SI des entreprises. Concernant le SIRH, cette analyse implique généralement l'extraction et la consolidation de données RH provenant de trois sources principales :

1. Une base de données démographiques ;
2. Une base de données arrêts de travail ;
3. Une base de données invalidité.

Ces données RH sont ensuite traitées conformément aux conventions de mesure ou aux « règles métier » définies par les *data scientists* afin d'assurer leur uniformité et leur cohérence. Cependant, cette standardisation se heurte à divers obstacles :

« *Ils* [la direction SI] *devaient construire ce format-là à partir de ce qu'ils avaient et c'était parfois incomplet. Ce qui était toute la complexité. Et souvent avec des données qui n'étaient pas conformes : des gens qui arrivent dans l'entreprise après en être partis, des gens qui avaient moins de 15 ans, des salaires négatifs, plein de choses comme ça.* » (Tristan, chef de projet, entretien no1, 15/04/2021).

Face aux défis usuels que rencontrent les *data scientists* lors de la collecte de données RH, la *DSN*, caractérisée par la qualité *normative* de ses données puisque régulée par des conventions juridiques, émerge comme une alternative prometteuse. Actualisée mensuellement par les gestionnaires de la paie, sa disponibilité généralisée au sein des entreprises françaises la rend particulièrement intéressante. Xavier (directeur de projet) et Tristan (chef de projet) évaluent ainsi favorablement son potentiel à optimiser le temps de travail nécessaire pour le traitement et le retraitement des données RH d'intérêt, en réduisant les incohérences (comme leur incomplétude et leur non-conformité) pour l'analyse de l'absentéisme :

« *C'est la rapidité... Et puis surtout le fait de ne pas avoir à recommencer un traitement une fois qu'il avait été fait... C'est la notion de standard. Il n'y avait pas*

besoin de l'adapter à chaque fois. Ça permettait de gagner des coûts aussi pour nous et pour le client. » (Xavier, directeur de projet, entretien no2, 21/11/2023).

L'acte de rationaliser les coûts d'investissement est ainsi guidé par la recherche d'un équilibre entre les bénéfices attendus et les coûts associés à la qualification des données RH. En tirant parti des données *DSN* préalablement normalisées, Xavier (directeur de projet) et son équipe réduisent leurs coûts en économisant du temps lors du (re)traitement des données RH, leur permettant d'éviter de multiples itérations. Cette stratégie renforce la singularité des données RH sur le marché en les rendant plus accessibles financièrement.

2.2.2. Exploitation des outils préexistants pour faciliter la qualification des données RH

L'engagement de Xavier (directeur de projet) envers les données *DSN*, est également motivé par l'opportunité de minimiser les coûts d'investissement grâce à l'utilisation d'outils préexistants. Cette opportunité se matérialise par le biais de quatre intermédiaires principaux :

1. Une base de données RH fictive à des fins d'entraînement pour les modèles analytiques ;
2. Un script de structuration des données RH ;
3. Une documentation technique des données RH ;
4. La maîtrise d'un outil de visualisation des données RH.

2.2.2.1. Une base de données RH fictive d'entraînement pour les modèles analytiques

La base de données RH fictive est principalement choisie en raison de sa disponibilité immédiate, facilitant ainsi considérablement l'acte de rationaliser les coûts d'investissement. Du fait de sa disponibilité, ce premier intermédiaire permet d'économiser le temps des *data scientists* et de minimiser les efforts nécessaires pour solliciter l'accès à des bases de données *DSN* ou à d'autres nouvelles sources de données pertinentes dans le cadre de l'analyse de l'absentéisme.

Cette base de données couvre la période de janvier 2017 à juin 2020 et regroupe les données RH de 64 856 salariés. Pour l'année 2020, les données *DSN* sont limitées

aux six premiers mois et elles incluent des informations spécifiques à la pandémie de COVID-19 à partir du mois mars.

Guidé par l'intuition de Tristan (chef de projet), l'objectif principal des *data scientists* est de rassembler une vaste gamme de données *DSN* concernant la catégorie de risques : actions locales (cf. Tableau 12) en se référant à deux critères principaux :

1. Les entités auxquelles les salariés sont rattachés ;
2. Les métiers des salariés.

La base de données RH fictive comprend donc 963 entités géographiques réparties sur l'ensemble du territoire français. Il est toutefois important de noter que les *data scientists* disposent seulement de neuf catégories de métiers distinctes pour classer l'ensemble de ces salariés. Représentées dans le Tableau 13, les données RH suivantes sont choisies comme étant d'intérêt pour l'analyse de l'absentéisme.

Données RH	**Types**
Âge	Numérique
Salaire	Numérique
Facteur_ETP (Équivalent Temps Plein)	Numérique
Ancienneté	Numérique
Nombre d'enfant	Numérique
Métier	Catégorielle
Établissement (lieu de travail)	Catégorielle
Type de contrat de travail	Catégorielle
Statut de travail	Catégorielle
Nombre d'arrêts de travail	Numérique
Durée totale des arrêts de travail	Numérique
Jours de présence	Numérique

Tableau 13 : Données RH sélectionnées par les *data scientists* dans la base de données *DSN* fictive d'entraînement

Malgré la disponibilité de cette base de données RH fictive, les *data scientists* sont confrontés à une limitation importante : l'absence de données brutes, imposée par le

devoir de conformité au *RGPD*[18]. Cette contrainte restreint la portée et la diversité des projets de connaissances pouvant être menés durant la phase de capitalisation, retardant ainsi leur avancement jusqu'à ce que l'accès aux données *DSN* complètes soit possible après l'engagement d'un premier client (post-étude) :

« *Elles limitaient* [les données RH fictives] *nos actions à ce moment-là, parce qu'on avait que les données fictives qui ne suffisaient pas. Mais de toute façon, les limites de ces données ne nous intéressaient pas tant que ça. On était projeté sur : on va avoir des clients et là on pourra faire différentes choses.* [C'est plutôt] *qu'est-ce qu'on peut faire en attendant ? Est-ce qu'on peut affiner nos modèles* [analytiques]*, travailler, faire des choses ? Mais l'idée, dans la construction du produit, c'est d'avoir accès à l'entièreté de la DSN.* » (Luc, *data scientist* senior, entretien no1, 02/11/2023).

Néanmoins, l'accès à une base de données RH préexistante permet tout de même de réduire considérablement les délais associés à la compréhension et à la préparation initiale des données :

« […] c'est *une phase qui occupe environ 80 % du temps de travail d'un data scientist…* » (Anatole, stagiaire *data scientist*, notes issues du journal de bord, 15/06/2023).

Cette préparation préalable facilite ainsi la transition vers la demande d'un premier client, favorisant une réactivité accrue à ce stade du processus. Malgré la limitation initiale des données *DSN* aux données RH socio-démographiques et relatives aux arrêts de travail, leur disponibilité a tout de même permis une avancée plus rapide vers la séquence de capitalisation, qui constitue le cœur du savoir-faire des *data scientists*.

2.2.2.2. Un script de structuration des données RH

La capacité des *data scientists* à garantir la rationalisation des coûts d'investissement dans la qualification des données RH est rendue possible grâce à leur collaboration avec Benoît (partenaire en actuariat). En effet, ce dernier a

[18] Le *RGPD*, ou Règlement Général sur la Protection des Données, constitue le cadre législatif de l'Union européenne qui régule le traitement et le transfert des données personnelles des individus au sein de l'UE. Adopté en 2016, ce règlement est entré en vigueur le 25 mai 2018. Pour plus d'informations, voir Conseil de l'Union européenne (2018). *Règlement général sur la protection des données (RGPD)*. https://www.consilium.europa.eu/fr/policies/data-protection/data-protection-regulation/ (consulté le 05/11/2024).

développé un script *Python* spécifiquement dédié à la structuration des données *DSN*, lesquelles sont inexploitables dans leur format d'origine. Ce deuxième intermédiaire permet ainsi de convertir les données *DSN* (XBRL) en un format tabulaire (CSV), format largement adopté dans le travail des données.

Bien que le script présente de nombreuses imperfections, principalement dues à de mauvaises pratiques de programmation, sa pérennisation est essentiellement justifiée par sa préexistence :

« *Normalement, il* [Benoît] *code sur VBA*[19]*, mais pour l'occasion, il s'est dit : tiens, je vais utiliser Python parce qu'avec les bases* [de données *DSN*]*, cela aurait été trop complexe. Du coup, il a écrit un code inutilement complexe.* […] *Nous, derrière, on a récupéré ce code et avons construit très rapidement des scripts pour tester des concepts* [sur l'absentéisme]*. Mais bon…c'est un peu le propre du conseil, c'est toujours un peu en crash test…* » (Anatole, stagiaire *data scientist*, entretien no2, 27/06/2023).

Ainsi, la décision de le maintenir en usage découle du choix de Xavier (directeur de projet). D'après lui, une refonte complète, malgré un potentiel à générer un intermédiaire à la fois plus efficace et plus précis, ne justifie pas l'investissement en temps et en ressources. Par conséquent, plutôt que d'entreprendre une refonte intégrale, il est choisi de maintenir le script existant, qui remplit ses fonctions principales, tout en effectuant des corrections ciblées là où des problèmes se manifestent.

L'existence préalable de ce deuxième intermédiaire est perçue comme conférant un avantage substantiel, en permettant aux *data scientists* de gagner du temps ; un temps qui serait normalement consacré à la compréhension de la structure des données *DSN*. En effet, l'arborescence complexe des blocs de données et leurs différentes cardinalités représentent un défi significatif en matière de programmation, notamment en raison de l'opacité du traitement requis par le format *XBRL*. L'utilisation du script existant permet ainsi de maximiser la valeur des données *DSN* en

[19] *Visual Basic for Applications* (VBA) est un langage de programmation intégré à *Microsoft Excel*.

concentrant le temps et les ressources sur la capitalisation de leurs qualités, séquence qui nécessite plus spécifiquement le savoir-faire des *data scientists*.

L'optimisation du temps de travail consacré à la structuration des données *DSN* a toutefois pour conséquence de maintenir l'opacité concernant le fonctionnement du script :

« […] *la partie qui pose problème et qui est un peu bloquante,* [celle] *sur laquelle Benoît* (partenaire en actuariat) *est le seul à avoir réellement de la connaissance, c'est la lecture des* [données] *DSN. Cette fonction de lecture des DSN* […] *est un peu une black box,* [elle] *n'est pas transparente...* » (Anatole, stagiaire *data scientist*, entretien no2, 27/06/2023).

Cette opacité induite renforce non seulement la dépendance vis-à-vis de Benoît (partenaire en actuariat) pour l'utilisation du script, mais elle met également en exergue la nécessité de se fier à sa capacité de comprendre les interactions entre les données *DSN*.

2.2.2.3. Une documentation technique des données RH

La qualification des données *DSN* est rendue possible grâce à des conventions juridiques françaises. Par conséquent, cette observation souligne que la pertinence des données *DSN* est intrinsèquement liée au contexte français, les ancrant ainsi dans une dimension culturelle spécifique.

La norme *NEODeS* (Norme d'Échanges Optimisée des Données Sociales) fournit un cadre structuré pour le traitement et le partage des données *DSN*. Accessible publiquement sur Internet, cette norme propose une documentation complète sur les spécifications de la *DSN*, incluant des informations sur les modifications apportées aux données, comme l'ajout ou la suppression de codes, ainsi que les ajustements des blocs d'informations. La norme *NEODeS* est continuellement mise à jour pour suivre les changements réglementaires, témoignant ainsi des évolutions dans les pratiques fiscales et administratives des entreprises françaises. Son actualisation représente ainsi un avantage significatif pour les *data scientists* qui nécessitent un accès à des données de qualité et uniformément structurées :

« *Benoît* (partenaire en actuariat) *a commencé à prendre la description de la DSN* [norme NEODeS] *et à construire un script Python qui lisait la DSN et qui reconstruisait*

toutes les variables dont il avait besoin. […] *à ce moment-là, on s'est dit* […]*, c'est énorme parce qu'on peut l'utiliser* [script] *pour sortir les données DSN de n'importe quelle entreprise, on peut en faire plein de choses puisqu'on a les informations sur l'absentéisme et sur les salariés. On peut même les enrichir...* » (Tristan, chef de projet, entretien no1, 15/12/21).

Grâce au développement du script basé sur cette documentation technique, Benoît (partenaire en actuariat) a acquis une importante compréhension des relations entre les blocs des données *DSN* et leurs rubriques. Cette compréhension le distingue comme un expert unique, possédant une maîtrise des subtilités de la norme *NEODeS*, ce qui en fait une figure clé dans la séquence de qualification des données RH. En effet, les données *DSN* ne se caractérisent pas par une stabilité immuable, mais s'inscrivent plutôt dans une dynamique d'évolution continue. Chaque année, elles subissent des transformations, telles que l'ajout, la modification ou la suppression de codes et de rubriques, s'adaptant constamment aux exigences émergentes et au changement des conventions juridiques françaises.

2.2.2.4. La maîtrise d'un outil de visualisation des données RH

La sélection de l'outil *Power BI* de *Microsoft*, spécialisé dans la visualisation des données, repose sur les compétences des *data scientists*, principalement celles de Tristan (chef de projet), dans la maîtrise de cet outil. Cette décision assure la rationalisation des coûts d'investissement, ce dernier étant déjà utilisé comme principal intermédiaire dans la visualisation des données pour les autres projets de *QIA*. Cette expertise interne facilite donc une mise en œuvre efficace, visant à élaborer une qualification des données RH rapide et abordable dans le but d'attirer les clients potentiels :

« […] *Parce que Tristan est le master de Power BI et qu'il l'a fait sous Power BI... Ça permettait de faire le développement très rapide d'une interface et d'avoir quelque chose de propre et de montrable à tous les clients avec un effort limité. On savait en plus qu'on avait des opportunités pour intégrer ces* [onglets] *Power BI dans des pages Web, donc on pouvait potentiellement aller plus loin au niveau de l'industrialisation derrière.* » (Tristan, chef de projet, entretien no1, 15/12/2021).

Cette stratégie permet ainsi de réduire le temps nécessaire à la formation des équipes et de minimiser les risques associés à l'introduction de nouveaux outils technologiques.

Cette approche pragmatique, axée sur l'optimisation des ressources internes, est principalement motivée par la phase de croissance continue au sein du cabinet. Comme l'indique Emilie (directrice associée) :

« *QIA se met en déséquilibre pour trouver une finalité. Nous, on n'est pas dans une phase de business stable. On est dans une phase de croissance. Il faut tester plein de choses…* » (entretien no1, 31/05/2021).

Ce déséquilibre, tout en facilitant la qualification des données RH pour l'analyse de l'absentéisme, expose également *QIA* à un risque financier, nécessitant une gestion prudente. Dans cette situation, Xavier (directeur de projet) doit ainsi jongler entre innovation et contraintes financières. Par conséquent, avant de lancer de nouveaux investissements, il est impératif qu'il valide et promeuve la qualification des données RH auprès des clients potentiels afin d'assurer et justifier auprès de la direction de *QIA* l'intérêt financier de ses initiatives :

« […] *il faut d'abord vendre le truc avant d'investir, c'est clair…* » (Pierre, fondateur, notes prises lors d'une réunion, 15/04/2022).

En conclusion, en ce qui concerne la rationalisation des coûts d'investissement, deux éléments clés permettent aux *data scientists* d'assurer ce deuxième acte de qualification des données RH. Premièrement, la réduction des contraintes usuelles de qualification. En effet, la *DSN*, régulée par des conventions juridiques, se présente comme une alternative prometteuse face aux défis traditionnels du (re)traitement des données RH. Deuxièmement, l'engagement des *data scientists* envers les données *DSN* est également motivé par l'opportunité de minimiser les coûts d'investissement. Cela est rendu possible grâce à l'utilisation de quatre intermédiaires principaux :

1. Une base de données RH fictive à des fins d'entraînement pour les modèles analytiques ;
2. Un script de structuration des données RH ;
3. Une documentation technique ;
4. La maîtrise d'un outil de visualisation.

Cependant, cet acte de rationalisation des coûts repose fortement sur la relation de proximité avec Benoît (partenaire en actuariat), créant ainsi une relation de dépendance des *data scientists* à son égard dans la séquence de qualification des données RH.

2.3. Acte III : enrôlement des agents économiques

Pour que les données RH s'inscrivent dans un processus d'économisation, elles doivent être intégrées dans un réseau d'agents économiques qui qualifient leurs singularités et substituabilités en tant que biens. Par conséquent, l'enrôlement d'agents économiques est essentiel puisque c'est à travers ce réseau que se façonne la co-construction de l'offre et de la demande.

Ainsi, pour répondre aux exigences de réussite de la construction des données RH tout en préservant l'efficacité des opérations rentables existantes chez *QIA*, Xavier (directeur de projet) met en œuvre, dès la séquence de qualification, trois stratégies d'enrôlement :

1. La stratégie d'intégration académique ;
2. La stratégie de mobilisation « en temps masqué » ;
3. La stratégie d'apport d'affaires.

Ces stratégies, basées sur l'optimisation sous contrainte, visent à minimiser les interférences avec les projets en cours au sein du cabinet afin de réduire au maximum le risque financier.

2.3.1. La stratégie d'intégration académique

La première stratégie d'enrôlement d'agents économiques repose sur l'intégration systématique d'étudiants en apprentissage : stagiaires en master spécialisés en *data science* ou au doctorat en GRH.

L'intégration de stagiaires en master constitue un élément central du modèle d'affaires de *QIA*. Ces stagiaires, issus de prestigieuses écoles, participent à des projets pratiques au cours de leur stage de fin d'études au sein du cabinet.

Dans le contexte du projet *DSN Analytics*, plusieurs *data scientists* se succèdent pour contribuer à la construction des données RH. De février à juillet 2021, Olivier, un

data scientist junior issu de la filiale santé, est affecté temporairement au projet, palliant ainsi une pénurie de ressources. En juillet 2021, Anatole prend sa relève en tant que stagiaire et poursuit le travail jusqu'en juillet 2022, suivi par Marco, qui assume le rôle jusqu'en novembre 2022. Les responsabilités de ces stagiaires incluent la préparation de la base de données *DSN* fictive, l'enrichissement du script développé par Benoît (partenaire en actuariat) ainsi que le développement de projets de connaissances qui visent à capitaliser sur les qualités des données *DSN*.

Ce système de rotation des stagiaires est caractéristique de l'organisation chez *QIA* : une fois leur stage terminé, les stagiaires embauchés sont habituellement orientés vers des projets plus lucratifs.

Mon recrutement chez *QIA*, dans le cadre d'une thèse *CIFRE* en GRH, a également marqué une étape importante pour le projet. Reconnue sous le titre de *HR business analyst*, ce rôle m'a rapidement valu la réputation d'être l'« experte RH » au sein de l'équipe. Cette reconnaissance n'est pas basée sur une expérience pratique antérieure, mais plutôt sur mon statut de doctorante spécialisée en GRH. Mon implication dans le développement d'une offre commerciale RH élargie a également joué un rôle dans le renforcement de cette perception. Doté d'une fondation théorique sur les interactions entre les activités RH et les données RH, mon objectif principal est d'approfondir ma compréhension des données *DSN*. Cela nécessite notamment une familiarisation approfondie avec la documentation technique, dans le but de diversifier l'expertise et de réduire la dépendance exclusive à l'égard de Benoît (partenaire en actuariat) comme seul expert de la *DSN*.

Étant donné que cette première stratégie d'enrôlement cible des agents économiques à l'expérience restreinte, il s'est avéré impératif de recruter d'autres agents dotés d'une expertise plus importante pour assurer la réussite de la construction des données RH.

2.3.2. La stratégie de mobilisation « en temps masqué »

La deuxième stratégie d'enrôlement consiste à mobiliser les salariés de *QIA* pour qu'ils consacrent du temps au projet pendant les périodes où ils ne sont pas affectés à d'autres tâches :

« […] *vu que c'est un projet interne, ils le font sur le temps durant lequel ils ne sont pas staffés en projets clients. Donc, de base, ce sont des types de projets qui prennent du temps à aboutir parce que tu le fais sur du temps en inter-contrat. Donc, il suffit que tu aies un rush côté client, enfin un « vrai » autre projet. Du coup, tu ne le fais pas... Tu ne peux pas, en conseil, bloquer autant de personnes. Donc, tu fais forcément avec ce qu'il y a.* » (Julie, cheffe de projet, entretien no1, 17/10/2023).

Cette stratégie vise ainsi à optimiser l'allocation des ressources sans perturber négativement les autres projets du cabinet de conseil. Plusieurs membres de l'équipe ont contribué à différentes étapes du projet : Tristan assume le rôle de chef de projet jusqu'à son départ du cabinet en mai 2022, puis Julie prend la relève. Justine, directrice des projets stratégiques, contribue également de manière ponctuelle en raison de son rôle en tant que chef de projet sur le premier projet de *QIA* sur l'absentéisme. Enfin, Luc, *data scientist* senior, recruté chez *QIA* en octobre 2021, participe également au projet.

L'enjeu de la temporalité revêt une importance capitale pour les *data scientists*, qui doivent qualifier rapidement les données RH afin de se confronter au marché sans un investissement excessif en ressources. Les étudiants constituent la principale force de production, tandis que les salariés, mobilisés en « temps masqué », encadrent la construction des données lorsqu'ils disposent de temps libre. Cet enjeu est étroitement lié aux impératifs de coût, de délais et de qualité, comme le souligne Emilie, directrice associée de *QIA* :

« *Quelle est la démarche la plus pragmatique et qui mobilise le moins de ressources ?* » (Entretien no1, 31/05/2021).

Ainsi, il est nécessaire de « faire plus avec moins ». Cela implique que les *data scientists* doivent également enrôler des agents économiques externes à *QIA* ayant une connaissance approfondie du marché RH, sans que cette stratégie d'enrôlement n'entraîne de coûts supplémentaires.

2.3.3. La stratégie d'apport d'affaires

Cette troisième et dernière stratégie se trouve au cœur de la séquence de qualification des données RH. En effet, l'enrôlement de Valérie de *Diag'santé* (partenaire en gestion des conditions de travail) et de Benoît d'*Assur'act* (partenaire

en actuariat) s'avère être le catalyseur du processus d'économisation des données RH par *QIA* :

« C'est du win-win. […] *tu vas compléter mon portefeuille d'offres. »* (Xavier, directeur de projet, entretien no2, 21/11/2023).

L'apport d'affaires, tel qu'illustré par la citation de Xavier (directeur de projet), repose sur une synergie au sein de laquelle chaque agent économique apporte une valeur ajoutée à l'autre. Cette stratégie d'enrôlement vise à créer un « cercle vertueux » où les agents se bénéficient mutuellement. C'est le principe du « win-win », évoqué par Xavier (directeur de projet), où chaque agent bénéficie non seulement de l'élargissement de son offre, mais aussi du renforcement de la fidélité du client par la création de nouvelles valeurs :

- Benoît (partenaire en actuariat) : « *Nous débriefions à chaud avec Xavier* (directeur de projet) *juste après le call, et nous pensions en effet* […qu'] *il serait utile de brainstormer avec vous sur des idées que nous pourrions avoir de sorte à monter une solution commune « plus clé en main » sur le plan diagnostic/Chiffre* [de l'absentéisme]*. Cela pourrait être intéressant… »* (Courriel envoyé le 27/01/2021).
- Valérie (partenaire en gestion des conditions de travail) : *« Oui tout à fait, je trouve l'approche tout à fait pertinente. »* (Courriel envoyé le 27/01/2021).

La stratégie d'apport d'affaires, orchestrée par les trois agents économiques : Xavier (directeur de projet), Benoît (partenaire en actuariat) et Valérie (partenaire en gestion des conditions de travail), repose donc principalement sur la consolidation d'un réseau d'affaires aux compétences conjointes. Elle implique de combler les lacunes dans les offres des deux agents : *Diag'santé* et *Assur'act,* déjà établis sur le marché RH, en évitant tout caractère concurrentiel. En s'appuyant sur leur soutien, Xavier (directeur de projet) peut légitimer le potentiel d'une réforme de la qualification des données RH, renforçant ainsi sa position face aux éventuels sceptiques :

« Pour aller sur le marché, on voulait des gens qui connaissaient [déjà, le marché]*. Diag'santé, c'est une boîte qui fait de la QVT* [Qualité de Vie au Travail]*, donc ils ont des clients qui ont des besoins et on est complémentaires parce qu'ils ne font pas de la data. Benoît* (partenaire en actuariat)*, on est complémentaires entre actuariat et données RH opérationnelles. Ce sont des gens qui pouvaient amener du business. C'est une des raisons d'aller sur l'offre, c'est l'écosystème. C'est-à-dire que j'ai une*

capacité à aller sur le marché. J'avais commencé par ces deux-là parce qu'il y a un trou des deux côtés. On peut faire un truc conjoint : Diag'santé, sur les actions concrètes, Assur'act sur la partie coût et nous sur la partie opérationnelle. Ça fait un truc assez complet. » (Xavier, directeur de projet, entretien no1, 31/10/2023).

Ainsi, cette troisième et dernière stratégie d'enrôlement sous-tend la conviction de Xavier (directeur de projet) qu'une coalition d'expertises hétérogènes est essentielle non seulement pour singulariser le positionnement des données RH, mais aussi pour répondre aux exigences d'un marché à la fois déficient et saturé.

Pour ce troisième et dernier acte de qualification centré sur l'enrôlement d'agents économiques, trois stratégies principales sont mises en œuvre :

1. La stratégie d'intégration académique : implique systématiquement des étudiants en apprentissage.
2. La stratégie de mobilisation « en temps masqué » : consiste à solliciter les salariés de *QIA* pour qu'ils consacrent du temps au projet durant les périodes où ils ne sont pas affectés à d'autres tâches.
3. La stratégie d'apport d'affaires : repose sur une synergie où chaque agent économique apporte une valeur ajoutée à l'autre et bénéficie mutuellement.

Ces stratégies, fondées sur une optimisation sous contraintes, visent à minimiser les interférences avec les projets en cours au sein du cabinet de conseil. Elles requièrent par conséquent de « faire plus avec moins », ce qui implique que Xavier (directeur de projet) et son équipe doivent développer leur réseau sans entraîner de coûts supplémentaires.

3. Controverses de qualification des données RH

L'analyse des espaces de négociation permet de mieux comprendre les mécanismes par lesquels les agents économiques, impliqués dans la qualification des données RH, négocient leur qualité *normative*. Ces espaces, désignés sous le terme de « controverses », constituent des cadres privilégiés où se forgent les compromis nécessaires à la transition de la qualification à la séquence de capitalisation.

Dans ce cadre, le réseau de qualification des données RH se compose de six modes d'existence distincts : (1) Commercial, (2) Développement, (3) Marché, (4)

Réglementaire, (5) Stratégique et (6) Technique. Chacun de ces modes, adoptant une perspective distincte, joue un rôle déterminant dans les négociations entourant la qualification des données RH. Ces modes d'existence sont décrits dans le Tableau 14.

Tableau 14 : Rôles des différents modes d'existence dans le réseau de qualification des données RH

Modes d'existence	Rôles
Commercial	Assurer le développement commercial des données RH et renforcer les relations d'affaires.
Développement	Garantir la coordination et l'harmonisation entre les développements technique et commercial.
Marché	Donner accès à des connaissances liées au marché RH.
Réglementaire	Veiller à la conformité des données RH avec les exigences réglementaires.
Stratégique	Orienter les investissements afin d'assurer leur alignement avec la stratégie globale du cabinet.
Technique	Assurer le développement technique des données RH.

Dans le contexte de cette séquence de qualification des données RH, trois controverses principales se manifestent :

1. L'ambiguïté dans la définition du client cible ;
2. L'incomplétude de la base de données RH fictive d'entraînement pour les modèles analytiques ;
3. La dépendance critique à l'égard d'une expertise externe.

Le Tableau 15 présente ces controverses, en mettant en évidence les perspectives propres à chaque mode d'existence concerné, ainsi que les compromis envisagés pour leur résolution.

Tableau 15 : Controverses pour la séquence de qualification des données RH

Sujets des controverses		**Ambiguïté dans la définition du client cible**	**Incomplétude de la base de données RH fictive d'entraînement pour les modèles analytiques**	**Dépendance critique à l'égard d'une expertise externe**
Perspectives spécifiques aux modes d'existence	**Commercial**	Fonction RH en support	Complétude simulée	Dépendance rentable
	Développement			
	Marché	Fonction RH prédominante		Dépendance profitable
	Réglementaire		Incomplétude imposée	
	Stratégique	Fonction RH vulnérable		
	Technique		Incomplétude accommodée	Dépendance limitante
Compromis négociés issu des controverses		**Approche multi-ciblage**	**(In)complétude fonctionnelle**	**Dépendance contrôlée**

3.1. Controverse I : l'ambiguïté dans la définition du client cible

Afin de mieux cerner les besoins des clients, les données RH sont analysées à travers un cadre de similitudes et de différences. D'un côté, la similitude se manifeste dans l'analyse de l'absentéisme, un domaine largement investi par de nombreux acteurs - dont *Fast MS* - ce qui a conduit à une saturation du marché. De l'autre côté, la différence réside dans le savoir-faire des *data scientists*, dont l'expertise est largement absente dans ce secteur. Pour que cette différenciation devienne une opportunité stratégique, il s'avère nécessaire de cibler des clients capables de reconnaître et de valoriser la contribution des données RH.

Trois modes d'existence incarnent des perspectives spécifiques par rapport à cette controverse :

1. Le mode d'existence Commercial : définit la fonction RH comme un support, jouant un rôle de relais sans implication directe dans le processus décisionnel. Dans une optique commerciale, la cible prioritaire reste les cadres managériaux, identifiés comme principaux décideurs opérationnels en matière de gestion de l'absentéisme.
2. Le mode d'existence Marché : cible principalement la fonction RH, historiquement responsable de la gestion de l'absentéisme. Dans ce segment du marché, cette activité est perçue comme un prolongement légitime et naturel des responsabilités inhérentes à la fonction RH.
3. Le mode d'existence Stratégique : considère la fonction RH comme une cible à la fois essentielle et vulnérable, en raison des risques de répercussions négatives liés à une gestion inadéquate de l'absentéisme. Par ailleurs, ce mode estime que la fonction RH manque de compétences analytiques suffisamment solides, ce qui accentue sa fragilité dans ce contexte.

Bien que les perspectives des différents modes d'existence divergent, le compromis adopté repose sur une approche commerciale multi-cibles. Cette approche vise à cibler simultanément les directions RH, financière et opérationnelle, afin de maximiser les opportunités de capter rapidement l'intérêt d'un client potentiel, qu'il relève de la fonction RH ou non.

3.2. Controverse II : l'incomplétude de la base de données RH fictive d'entraînement pour les modèles analytiques

L'absence d'accès aux données RH brutes, en raison des exigences du *RGPD*, limite la profondeur des analyses et freine les avancées des *data scientists*. En effet, les données complètes issues de la *DSN* ne deviendront accessibles qu'après l'acquisition d'un premier client.

Pour répondre à la contrainte du mode d'existence Réglementaire, qui impose le respect strict du *RGPD* et l'effacement des données brutes, deux autres modes d'existence apportent des perspectives différentes à cette controverse :

1. Le mode d'existence Commercial : estime qu'il est essentiel de ne pas différer les analyses en attendant l'accès à une base de données RH complète, afin de garantir les avancées commerciales par la présentation de résultats tangibles aux clients potentiels. La base est donc simulée comme complète pour satisfaire cette exigence et renforcer la crédibilité de l'offre commerciale.
2. Le mode d'existence Technique : reconnaît que, bien que les données soient limitées et restreignent la portée des analyses, leur disponibilité immédiate permet de gagner du temps dans le développement technique.

Face à l'incomplétude des données RH disponibles, le compromis adopté est une (in)complétude fonctionnelle, s'appuyant sur le mode d'existence Technique, tout en intégrant les impératifs du mode Commercial. Ce dernier mise sur les données partielles pour optimiser les modèles et préparer les analyses, permettant ainsi de démontrer rapidement ses capacités analytiques aux clients. Cette stratégie respecte les exigences du *RGPD* et soutient un rythme d'apprentissage continu pour les *data scientists* avec ces nouvelles données.

3.3. Controverse III : la dépendance critique à l'égard d'une expertise externe

L'initiative *DSN Analytics*, résultant de la collaboration initiale entre Benoît (partenaire en actuariat) et Xavier (directeur de projet), engendre une dépendance

critique à l'égard de l'expertise de Benoît. Ce dernier occupe une position centrale grâce à son script Python, conçu pour convertir les données *DSN* (XBRL) au format CSV. Sa maîtrise exclusive des relations entre les données RH, fondée sur sa connaissance approfondie de la documentation technique *NEODeS*, le rend indispensable au projet. Cette situation souligne la vulnérabilité du processus de construction, inhérente à la dépendance vis-à-vis d'une expertise externe.

Dans ce contexte, le mode d'existence Marché, représenté par Benoît (partenaire en actuariat), tire profit de la dépendance générée par son expertise et de son rôle central dans l'orientation de la qualification des données RH. En parallèle, deux autres modes d'existence offrent des perspectives divergentes face à cette controverse :

1. Le mode d'existence Commercial : considère la dépendance à l'expertise de Benoît (partenaire en actuariat) comme étant rentable. En effet, bien que la refonte complète du script puisse théoriquement accroître l'efficacité et diminuer la dépendance à son expertise, le mode Commercial estime que l'investissement nécessaire, tant en temps qu'en ressources, ne serait pas économiquement justifiable.
2. Le mode d'existence Technique : perçoit la dépendance à l'expertise de Benoît (partenaire en actuariat) comme limitante. Ce mode éprouve un inconfort avec le script développé par Benoît, critiqué pour son manque de structure et son désordre. Ces déficiences entravent la maintenance et l'adaptation du code, rendant le travail de qualification des données RH plus ardu.

Confrontés à une dépendance critique vis-à-vis de l'expertise de Benoît (partenaire en actuariat), les modes d'existence adoptent un compromis de dépendance contrôlée. Le script Python, bien qu'imparfait, est maintenu en l'état puisqu'il assure les fonctions essentielles requises par le projet. Le mode d'existence Technique se concentre sur des interventions ponctuelles pour rectifier les dysfonctionnements, tandis que le mode Commercial intensifie ses efforts pour consolider la collaboration avec Benoît. Ce compromis garantit la continuité du projet et permet d'exploiter judicieusement l'expertise de Benoît, tout en atténuant les risques économiques associés à une éventuelle refonte complète du script, en attendant l'arrivée d'un premier client.

En résumé, la séquence de qualification des données RH fait émerger trois controverses principales sur :

1. L'ambiguïté dans la définition du client cible ;
2. L'incomplétude de la base de données RH fictive d'entraînement pour les modèles analytiques ;
3. La dépendance critique à l'égard d'une expertise externe.

Ces controverses sont intrinsèquement liées à gestion des ressources limitées tout en assurant un déploiement rapide sur le marché (*go-to-market*). Elles soulignent les tensions entre la recherche d'une croissance rapide et les contraintes imposées par les ressources disponibles. La qualification des données RH, à la fois exploratoire et axé sur l'efficience, vise à maximiser les opportunités tout en évitant des investissements trop précoces.

4. Conséquences pour la fonction RH

L'analyse de cette séquence initiale de qualification des données RH révèle l'absence de la fonction RH. En effet, les seules figures considérées « représentatives » de cette fonction sont Xavier (directeur de projet) et moi-même (*HR business analyst*). Cette absence interpelle et invite à une réflexion sur les causes sous-jacentes, lesquelles se structurent en trois catégories principales :

1. La gestion opérationnelle de l'absentéisme par les managers et les directeurs ;
2. La connotation négative de l'absentéisme, qui suscite des craintes de représailles ;
3. Un manque de capacités analytiques, aggravé par une aversion potentielle pour les technologies.

Ces trois causes révèlent un défaut d'alignement de la fonction RH avec les exigences analytiques et stratégiques nécessaires pour mener à bien la réforme de qualification des données RH proposée par le cabinet.

Bien que la fonction RH ne soit pas explicitement impliquée dans la séquence de qualification, son expertise est tout de même reconnue, se manifestant à travers deux domaines distincts : théorique et pratique.

1. Théorique par le recrutement d'une doctorante en GRH : qui vise principalement à édifier une base conceptuelle sur les interactions entre le phénomène de l'absentéisme et les données RH.
2. Pratique par l'expérience du directeur de projet en tant que praticien RH : qui cible l'utilisation des données RH pour élaborer des plans d'action opérationnels dans la gestion de l'absentéisme :

« [...] *je suis un praticien des RH. Je manage depuis longtemps, depuis 2008.* [...] *je formais des gens, je recrutais. J'étais dans les processus d'évaluation et dans les comités où ça se décidait.* [...] *Et je vois bien que la fonction RH, il y a une articulation avec l'opérationnel qui est très importante. Je suis un praticien. Je ne connais pas bien les théories RH* […] *C'est ma conviction* […], *je pense que si tu retiens les gens par les salaires, ils partent pour 5% d'augmentation le lendemain. C'est quelque chose d'assez dangereux. Il vaut mieux, pour moi, les retenir par l'humain, par les conditions de travail, par le fait d'arriver à créer une cohésion de groupe.* » (Xavier, directeur de projet, entretien no1, 31/10/2023).

En somme, les représentations des différents agents économiques révèlent une fonction RH *fragmentée*. En effet, le « transfert » de son expertise théorique et pratique vers des agents extérieurs entraîne une marginalisation compensatoire de la dimension « RH » des données.

5. Conclusion

Ce chapitre porte sur la première séquence du processus de construction des données RH : la Qualification. Celle-ci aborde l'évaluation des qualités des données RH, en déterminant les critères initiaux qui définissent leur singularité potentielle.

Trois actes clés sont abordés dans cette séquence :

1. Définir les besoins des clients ;
2. Rationaliser les coûts d'investissement ;
3. Enrôlement des agents économiques.

La qualification met en lumière une première qualité des données RH : leur qualité *normative*.

Le choix de cette qualité s'appuie sur deux mythes : (1) un manque d'expertise en données RH sur le marché actuel et (2) leur potentiel de généralisation. Ces mythes orientent le choix de trois cibles principales parmi les clients potentiels :

1. Les directions RH ;
2. Les directions financières ;
3. Les directions opérationnelles.

Les données RH, issues de la *DSN*, sont privilégiées pour leur standardisation et leur faible risque économique grâce à l'exploitation de quatre outils préexistants. Elles facilitent également l'enrôlement d'agents, structuré autour de trois stratégies principales :

1. Stratégie d'intégration académique ;
2. Stratégie de mobilisation « en temps masqué » ;
3. Stratégie d'apport d'affaires.

Ces stratégies visent à minimiser les interférences avec les projets en cours en demandant aux *data scientists* de « faire plus avec moins » en les incitant à élargir leur réseau sans encourir de coûts supplémentaires.

Trois controverses émergent de cette première séquence, chacune faisant l'objet de négociations. Les compromis atteints sont les suivants :

1. Une approche de multi-ciblage ;
2. Une (in)complétude fonctionnelle ;
3. Une dépendance contrôlée.

Ces compromis incarnent une première stratégie de marchandisation des données RH, centrée sur une optimisation efficiente. Elle vise à maximiser l'efficacité tout en limitant strictement les investissements tant que la demande du marché demeure incertaine. Elle constitue le socle qui permet le passage à la deuxième séquence du processus de construction des données RH : la Capitalisation.

Chapitre 4. Séquence de Capitalisation des données RH

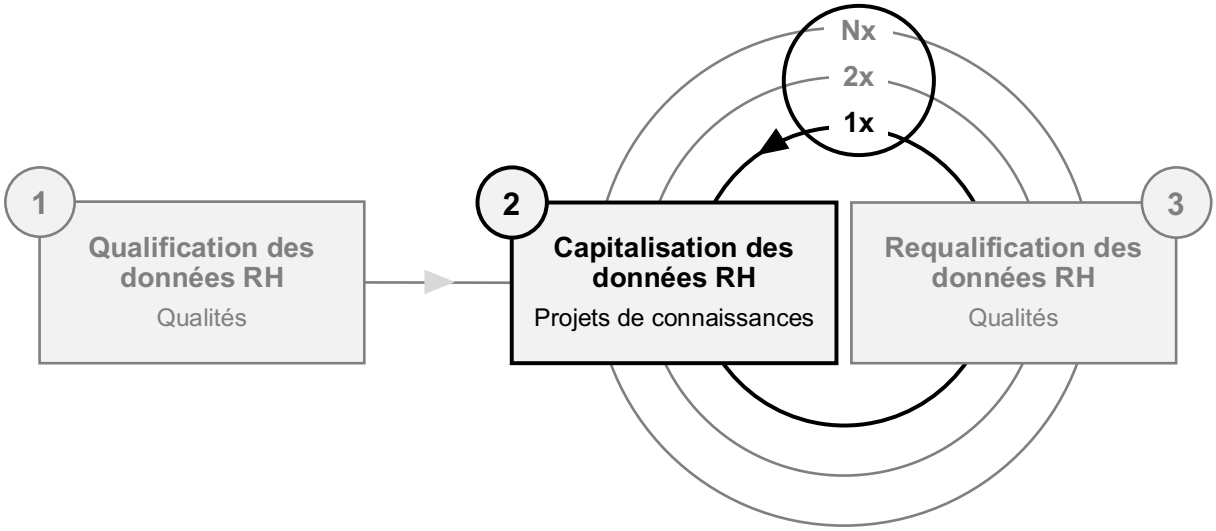


Figure 11 : Séquence de Capitalisation des données RH

1. Introduction

Le troisième chapitre s'est concentré sur la qualification des données RH, mettant en lumière une première qualité : leur qualité *normative*. Ce choix de qualité est guidé par une première stratégie de marchandisation axée sur l'optimisation efficiente des ressources humaines, techniques et financières.

Ce quatrième chapitre se penche sur la deuxième séquence du processus de construction des données RH : la Capitalisation. Celle-ci examine les projets qui s'appuient sur la qualité *normative* de ces données pour produire de nouvelles connaissances. L'objectif de cette séquence est de concevoir la fonction épistémique des données RH, permettant leur singularisation en tant que biens économiques.

Ce chapitre se décline en trois projets de connaissances distincts, chacun caractérisé par un type spécifique de connaissances sur l'absentéisme :

1. L'absentéisme *réel univarié*, qui fournit des connaissances *descriptives*, est présenté dans la première section ;
2. L'absentéisme *réel multivarié*, qui offre des connaissances *explicatives* est ensuite exploré dans la deuxième section ;
3. L'absentéisme *latent*, qui apporte connaissances *prédictives* est enfin détaillé dans la troisième section.

Les controverses, au cours desquelles ces agents économiques négocient le développement de ces trois projets de connaissances, font ensuite l'objet d'une analyse approfondie.

2. Projet de connaissances I : l'absentéisme *réel univarié*

Pour rendre compte de l'absentéisme réel univarié en tant que premier projet de connaissances, il convient d'investiguer trois dimensions :

1. Le territoire d'exploration épistémique des données RH ;
2. Le savoir-faire des agents économiques ;
3. Les épreuves d'exploration qui en découlent.

2.1. Territoire d'exploration épistémique des données RH

Le premier projet de connaissances, développé par les *data scientists*, a pour objectif de produire une connaissance *descriptive* de l'absentéisme. Il repose sur le calcul du taux d'absentéisme formalisé à travers la formule mathématique présentée ci-dessous :

$$Taux\ d'absentéisme = \frac{Nombre\ de\ jours\ d'arrêts}{Nombre\ de\ jours\ travaillés}\ x\ 100$$

Le taux d'absentéisme se calcule en divisant le volume total d'absences par le nombre estimé de jours de travail, sur une base annuelle pour une population de salariés donnée.

À première vue, le calcul paraît simple : il consiste en une division suivie d'une multiplication par 100. Toutefois, la définition choisie par les *data scientists* pour le numérateur et le dénominateur introduit diverses subtilités. Étant donné l'abondance des méthodes disponibles et l'absence de conventions juridiques dans le calcul de l'absentéisme, comment choisir ? La réponse de Tristan (chef de projet) tient en un principe : celui de la simplicité.

1. Le numérateur : correspond à la durée totale des absences, déterminée par la somme des périodes d'absence enregistrées au cours d'une année spécifique. Le calcul se fait en mesurant, pour chaque absence, l'intervalle de temps entre la date de début et la date de fin. Cette méthode ne prend toutefois pas en compte les jours de repos théoriques qui peuvent coïncider avec les périodes d'absence des salariés.
2. Le dénominateur : représente le nombre total de jours ouvrables, déduit des 365 jours de l'année, fournissant ainsi un dénominateur constant qui uniformise la période d'observation sur l'ensemble d'une année calendaire. Cette approche est largement adoptée par les statisticiens analysant l'absentéisme et est conforme

aux prescriptions médicales habituelles, où un arrêt de travail peut couvrir une semaine (soit sept jours) du lundi matin au dimanche soir[20].

Cette convention de mesure simplifie le calcul du taux d'absentéisme en éliminant le besoin d'ajuster les données RH pour les années bissextiles ou les variations régionales et sectorielles des jours ouvrables. Utiliser les jours calendaires pour le calcul ne rend pas l'estimation plus ou moins précise que l'utilisation des jours ouvrés ; cette approche est plus facile à gérer techniquement.

L'absentéisme est donc quantifié par cette formule mathématique. Comme nous le savons, les conventions choisies pour mesurer les phénomènes peuvent modifier significativement l'échelle de ce qui est observé en (re)configurant les propriétés des données. L'influence de cette convention sur les résultats est donc significative. Par exemple, la somme des durées d'absence est calculée en accumulant les interruptions de travail, sans tenir compte des interruptions régulières comme les week-ends ou les jours fériés. De plus, la décision de quantifier les absences en jours entiers plutôt qu'en fractions (par exemple, en centièmes de jour) affecte également la précision des mesures. Ainsi, si un salarié travaille une demi-journée, cette contribution est omise dans le calcul d'une journée complète de présence, ce qui conduit à une sous-évaluation des présences réelles et, par conséquent, à une surestimation du taux d'absentéisme.

Sur la base de cette définition calculée de l'absentéisme, Tristan (chef de projet) va principalement orienter le premier projet de connaissances vers deux types d'exploration épistémique des données RH :

1. Une exploration interne, spécifique à l'entreprise ;
2. Une exploration externe, destinée à sa comparaison nationale et sectorielle.

2.1.1. Exploration épistémique interne des données RH

L'exploration épistémique interne qui vise l'absentéisme au sein de l'entreprise, se concentre sur l'analyse des données socio-démographiques des salariés absents au cours d'une année donnée. Les *data scientists* examinent comment l'absentéisme

[20] Selon Bernon et Pertinant (2023), cette approche est couramment utilisée par les statisticiens pour analyser l'absentéisme.

varie en fonction de chaque donnée RH prise individuellement (d'où l'emploi du terme « *univarié* » pour décrire ce projet). Ces dernières sont représentées dans le Tableau 16 ci-dessous.

Tableau 16: Données RH sélectionnées pour l'absentéisme *réel univarié*

Données RH	Segments
Âge	18 ans et moins
	Entre 18-25 ans
	Entre 26-35 ans
	Entre 36-45 ans
	Entre 46-55 ans
	55 ans et plus
Salaire	Moins de 10K €
	Entre 10K et 30K €
	Entre 30K et 60K €
	Entre 60K et 120K €
	120K € et plus
Ancienneté	Moins de 1 ans
	Entre 1-3 ans
	Entre 4-10 ans
	Entre 11-20 ans
	20 ans et plus
Sexe	Homme
	Femme
Statut	AM (agent de maîtrise)
	CAD (cadre)
	EMP (employé)
Contrat	CDD (contrat à durée déterminée)
	CDI (contrat à durée indéterminée)
	Stage
	Autre
Parents	Avec enfant
	Sans enfant
Situation familiale	Célibataire
	Marié

Le choix des données socio-démographiques dans l'analyse de l'absentéisme, repose sur une approche univariée, où chaque donnée est étudiée indépendamment

pour comprendre son influence sur l'absentéisme. Cette méthode permet de détecter des tendances spécifiques au sein de chaque grand groupe socio-démographique. Tristan (chef de projet) utilise l'analogie des actifs financiers individuels dans un portefeuille pour illustrer l'importance de l'utilisation des données socio-démographiques pour traiter les questions d'absentéisme :

« *C'est une intuition. Moi, je sais faire le calcul de l'analyse de l'absentéisme* […] *Et donc, je me suis projeté sur un RH qui dit à son manager et mais regarde toi t'as beaucoup d'absentéisme, c'est un peu chaud. Premier réflexe du manager c'est : ben ouais mais moi j'ai des gens comme si comme ça. De même que dans la banque, quand tu lui dis : ouais ton équipement crédit conso[mmation], il est trop faible, il dit : ouais mais moi mon portefeuille, il y a des gens pauvres, etc.* » (Tristan, chef de projet, entretien no1, 15/12/2021).

Le choix des données socio-démographiques, telles que l'âge, le salaire, l'ancienneté, le sexe et le statut, se réalise dans un cadre imposé. Ce choix est fondé sur l'utilisation de données déjà sélectionnées au sein d'une base de données RH fictive. Par conséquent, ce processus décisionnel, au lieu de découler d'une concertation délibérée, est guidé par des paramètres prédéfinis :

« Il manquait certaines valeurs, certaines variables socio-démo[graphiques] *qu'on aurait aimées avoir.* […] *on se posait des questions et on savait que la donnée était dispo*[nibles] *dans la DSN mais on ne pouvait pas la récupérer parce que les données avaient été effacées conformément au RGPD. »* (Luc, *data scientist* senior, entretien no1, 02/11/2023).

Bien que le cadre soit contraint, certaines des données RH présentes dans la base, telles que le nombre d'enfants et la situation maritale, sont néanmoins débattue en raison de leur nature sensible. Jean (partenaire RH) souligne notamment l'importance de ces données, en mettant plus spécifiquement l'accent sur les données relatives à la maternité. Il évoque notamment ses préoccupations antérieures liées au recrutement d'un grand nombre de femmes, évoquant un risque de pénurie de salariés. Selon lui, l'exploitation de ce type de données est impérative car la finalité de l'utilisation des données RH transcende le cadre individuel pour appréhender les comportements collectifs :

- Jean (partenaire RH) : « *Même si on utilise les données individuelles ça reste dans l'objectif de comprendre le comportement collectif, c'est une démarche collective. Il faut trouver des garde-fous contractuels*… »
- Emilie (directrice associée) : « *Si la donnée est disponible, les leviers* [d'action] *doivent cependant rester à la maille des entités. Le signe avant-coureur de l'absentéisme, ce n'est pas à l'échelle des données, c'est à l'échelle des leviers. Il faut exploiter le signal* [des données RH]*, la frontière se trouvant dans l'exploitation et dans l'enseignement qu'on en tire ; c'est toujours une question de finalité. L'outil doit uniquement générer des connaissances et c'est l'entreprise qui décide ensuite comment les exploiter…* » (Notes prises lors d'une réunion, 21/07/21).

Emilie (directrice associée) met ainsi en évidence que, bien que les données RH soient disponibles et exploitables, il est nécessaire de distinguer les données utiles de celles qui ne le sont pas pour l'analyse de l'absentéisme. Elle insiste sur l'importance d'identifier et de convertir le signal des données en connaissances, en veillant à ce que la finalité de leur utilisation guide constamment leur exploitation.

2.1.2. Exploration épistémique externe des données RH

L'exploration épistémique externe revêt une importance significative pour Xavier (directeur de projet), car elle offre à l'entreprise l'opportunité de se positionner au sein de son environnement et de se comparer aux autres entreprises du même secteur.

Pour ce faire, Tristan (chef de projet) se tourne vers des données RH externes, et plus spécifiquement celles fournies par *Alior*, un cabinet de conseil international. Ce cabinet publie annuellement un baromètre qui rend compte des taux d'absentéisme parmi quatre macro-secteurs :

1. Le commerce ;
2. L'industrie ;
3. La santé ;
4. Les services.

Ces données RH sectorielles servent de points de repère, aidant les entreprises à déterminer si leur taux d'absentéisme est excessif ou conforme à celui de leur secteur :

« *Ça, lui* [l'entreprise] *permet de s'éclairer, de se dire ok, j'ai un absentéisme qui est quand même un peu trop important…* » (Tristan, chef de projet, notes prises lors d'une réunion, 15/04/2022).

Pour opérationnaliser la demande de Xavier (directeur de projet), Tristan (chef de projet) et son équipe choisissent de faire une comparaison évolutive entre le taux moyen d'absentéisme de l'entreprise et les moyennes nationale et sectorielle sur une période de quatre ans. Ils structurent ces analyses en trois volets :

1. Le taux moyen d'absentéisme de l'entreprise comparé à la moyenne nationale en 2020.
2. L'évolution du taux d'absentéisme de l'entreprise par rapport à la moyenne nationale de 2017 à 2020.
3. L'évolution du taux d'absentéisme de l'entreprise par rapport à la moyenne par secteur (commerce, industrie, santé, et services) de 2017 à 2020.

Cette approche comparative et diachronique offre ainsi une perspective sur la position de l'entreprise par rapport à l'évolution des taux d'absentéisme au niveau national et sectoriel. Bien que l'utilisation d'un cadre de comparaison externe semble indispensable pour décrire l'absentéisme, des réserves subsistent chez *QIA*, quant à la pertinence des données RH publiées gracieusement par *Alior* :

- Pierre (fondateur) : « *Pour un DRH […] ça n'a pas de sens de comparer son taux d'absentéisme au taux national, il n'en a strictement rien à foutre, sincèrement, ce qu'il veut, c'est par rapport au taux de son secteur et des entreprises similaires. Tu peux être au-dessus, tu peux être en dessous du taux national et être bien supérieur au taux qui concerne ton secteur. Elles sont toujours comme ça les entreprises, elles veulent toujours se comparer avec leur voisin concurrent : Je suis comment ? Pour moi, la valeur qu'il faut mettre, ce n'est pas tellement le taux national, c'est le taux du secteur.* »
- Tristan (chef de projet) : « *On a mis les taux par secteur qui sont les taux disponibles sur Internet. Alior nous fournit quatre secteurs.* »
- Pierre (fondateur) : « *Il n'y a que quatre secteurs ?* »
- Tristan (chef de projet) : « *Les données publiques gratuites que tu vas trouver, c'est ça.* »

- Xavier (directeur de projet) : « *Pierre, ça va inclure des éléments comme des arrêts liés aux congés maternité alors que ce n'est pas le niveau d'absentéisme sur lequel une entreprise veut agir. Le sujet, c'est l'accès ; L'accès à ces données…* » (Notes prises lors d'une réunion, 15/04/2022).

Les données diffusées par *Alior* imposent des restrictions aux *data scientists*, en particulier leur capacité à isoler certaines catégories d'absences, telles que les congés parentaux. Xavier (directeur de projet) souligne l'importance de l'accès à des données RH complètes pour développer des conventions de mesure personnalisées. Ceci permettrait de classer de telles absences en tant qu'« incompressibles », signifiant des absences sur lesquelles les interventions managériales ne peuvent exercer aucune influence.

2.2. Savoir-faire des agents économiques

Dans le cadre du calcul de l'absentéisme *réel univarié*, le savoir-faire des *data scientists* réside dans leur capacité à extraire des informations à partir des données *DSN*, une tâche rendue complexe par la structure même de ces dernières. Pour ce faire, les *data scientists* doivent identifier avec exactitude les blocs et les rubriques au sein de la *DSN* qui renferment les données RH d'intérêt et les associer à leurs données principales afin qu'elles soient liées une fois extraite. Un cas illustratif de ce savoir-faire est le traitement des données *DSN* relatives à l'ancienneté.

Exemple de l'interprétation de l'ancienneté à partir des données *DSN*

L'interprétation de l'ancienneté nécessite l'exploration des diverses rubriques du bloc **S21.G00.86**. La rubrique S21.G00.86.**001** est indispensable pour préciser le type d'ancienneté considéré, qu'il s'agisse de l'ancienneté depuis le début du contrat de travail, de l'entrée du salarié dans l'entreprise ou de son intégration dans le secteur d'activité. Ensuite, il est essentiel de consulter la rubrique S21.G00.86.**002**, laquelle précise l'unité de mesure employée pour le calcul de l'ancienneté, telle que les jours plutôt qu'une autre unité temporelle. Enfin, il convient d'examiner la rubrique codifiée S21.G00.86.**003**, qui dévoile la valeur exacte de l'ancienneté.

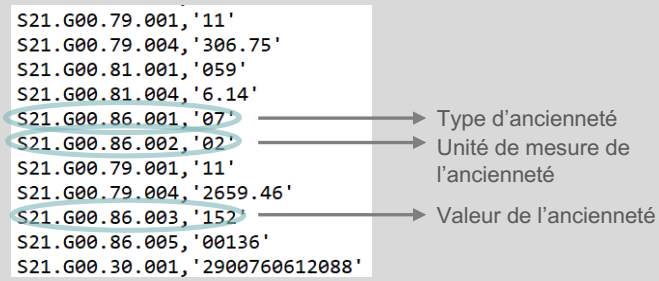


Figure 12: Interprétation de l'ancienneté (issue d'une présentation, 07/02/2023)

Pour garantir la fiabilité des données RH, une compréhension approfondie de la structure des données *DSN* est essentielle, car les interactions entre leurs blocs et rubriques peuvent aisément mener à des erreurs de calcul. La nécessité d'établir des conventions de mesure, essentielles dans le travail des données, s'avère donc indispensable pour assurer une interprétation de ces dernières. Tristan (chef de projet) illustre ce point avec l'exemple des données salariales, où des choix de conventions de mesure sont requis pour naviguer à travers la complexité des données RH :

« *Est-ce que tu prends le salaire divisé par 12 ou tu prends le salaire annuel ? Tu prends le salaire avec prime ?* […] *Tu vois pleins de choses…* » (Tristan, chef de projet, entretien no1, 15/12/2021).

Les conventions de mesure, adoptées par les *data scientists*, pour le projet d'absentéisme *réel univarié* reposent sur un principe de simplicité. Par exemple, le « salaire », initialement exprimé sur une base mensuelle dans les données *DSN*, est

converti en une valeur annuelle en le multipliant par douze. Cependant, comme l'illustre Benoît (partenaire en actuariat) dans un courriel intitulé « *Traitement salaire : quelles règles chez nous ?* », le traitement des données salariales peut présenter des défis, notamment en présence d'anomalies :

- Tristan (chef de projet) : « *Je discutais ce matin avec Benoît (partenaire en actuariat) qui se posent des questions sur la manière de faire évoluer ses traitements de salaire. Je voulais donc savoir quelle règle nous utilisions… !* » (Courriel envoyé le 07/02/2022).
- Anatole (stagiaire *data scientist*) : « *Pour le calcul du salaire annuel, on va utiliser le salaire de base, donc le bloc 51 et le type 010. On va ensuite simplement multiplier par 12 ce salaire mensuel pour l'annualiser.* » (Courriel envoyé le 07/02/2022).
- Benoît (partenaire en actuariat) : « *Merci de l'info. C'est aussi la rubrique que j'utilise par principe … quand elle est renseignée, ce qui n'est pas mon cas par exemple avec X. D'où la nécessité de se rabattre sur les autres, et les traitements de nettoyage que je dois bâtir ! Un peu sport…* » (Courriel envoyé le 07/02/2022).

Les conventions de mesure adoptées pour les données RH peuvent ainsi être mises à l'épreuve par diverses anomalies, telles que des données mal renseignées.

Parmi les conventions de mesure adoptées par les *data scientists*, l'agrégation des données RH mensuelles en valeurs annuelles revêt une importance particulière. Cette approche s'applique non seulement aux salaires, mais également à l'ensemble des données socio-démographiques utilisées telles que l'âge, l'ancienneté ou bien les arrêts de travail. Ce choix, effectué par Tristan (chef de projet), d'agréger les données RH puis de calculer des moyennes sur de plus grands groupes de salariés, vise à éviter la production de résultats trop erronés :

« […] *Il se trouve que plus on va faire des moyennes sur un grand groupe, plus on va commencer à avoir des résultats qui sont fiables, c'est-à-dire un effet en moyenne vrai à l'échelle d'un groupe alors qu'à l'échelle individuelle, celui-ci reste faux… »*

L'absentéisme étant un phénomène multifactoriel, le choix de l'agrégation comme convention de mesure permet aux *data scientists* de produire des résultats plus fiables. Cela révèle des tendances « en moyenne vraies » à l'échelle du groupe, même si les calculs individuels restent relativement peu représentatifs du phénomène étudié.

Cependant, l'agrégation des données RH peut entraîner des représentations simplifiées. Cela peut notamment se traduire par une incapacité à refléter les variations et les changements survenus au cours de l'année ou d'une année à l'autre, dissimulant possiblement les véritables dynamiques de l'absentéisme :

« *Une absence qui commençait le 21 décembre de l'année et qui durait six mois* […] *C'était un peu absurde de considérer que c'était une absence de trois jours sur l'année où l'absentéisme est arrivé. Pour une absence qui était à cheval sur deux ans, on en mettait une partie en 2021, une partie en 2022. Ça, c'est assez critiquable je pense. Ce n'est pas comme ça que tu essaies de modéliser un phénomène. L'absence, elle arrive l'année où elle est arrivée, 2022 n'est pas du tout responsable de l'absence qu'elle récupère de 2021…* » (Luc, *data scientist* senior, entretien no1, 02/11/2023).

En plus de l'agrégation, les conventions de mesure choisies par Tristan (chef de projet) incluent également la concaténation des données RH. La Figure 13, ci-dessous, illustre la manière dont les arrêts de travail sont concaténés.

Elle présente deux arrêts de travail, chacun délimité par un intervalle spécifique sur un calendrier qui s'étend de mi-janvier à fin février. Ces deux arrêts sont concaténés en une seule représentation continue, s'étendant du début du premier arrêt à la fin du second. La concaténation permet d'analyser les données de manière continue, facilitant ainsi l'évaluation de l'impact total des arrêts de travail sur la période observée. Toutefois, cette convention de mesure peut paraître trop simpliste, masquant les nuances entre les périodes d'arrêt de travail et pouvant potentiellement fausser l'analyse des tendances et des causes sous-jacentes de l'absentéisme. Par exemple, lorsque deux types d'arrêt de travail s'enchaînent, seul le premier type est retenu par les *data scientists*, ce qui occulte potentiellement les différences entre les diverses périodes d'arrêt de travail et complexifier la représentativité du phénomène.

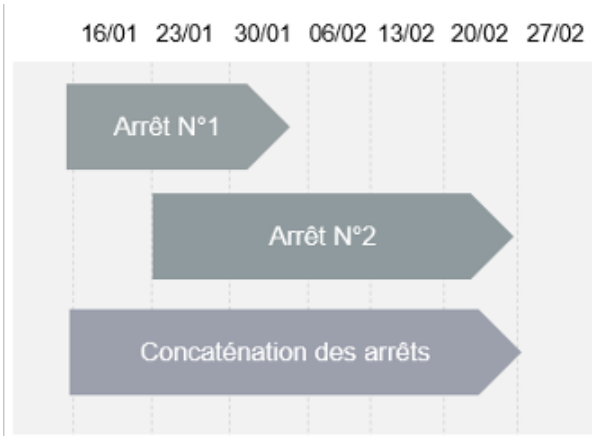


Figure 13 : Concaténation des arrêts de travail (issue d'une présentation, 07/02/2023)

Ce premier projet de connaissances révèle ainsi que les *data scientists* s'efforcent de maintenir un équilibre délicat : simplifier les conventions pour faciliter l'interprétation des résultats tout en assurant que ces résultats reflètent fidèlement les données RH.

2.3. Epreuves d'exploration

Dans le cadre du développement de l'absentéisme *réel univarié*, les *data scientists* se trouvent confrontés à deux grandes catégories d'épreuves. Une synthèse des épreuves est illustrée par le Tableau 17.

1. L'accès aux données RH brutes : le manque d'accès aux données socio-démographiques restreint les analyses, empêchant l'évaluation de l'influence « réelle » des données telles que l'âge, le salaire, l'ancienneté, le sexe ou le statut des salariés sur l'absentéisme. Cette épreuve résulte d'une controverse de qualification ayant conduit à un compromis pragmatique : poursuivre l'optimisation des modèles et la préparation des analyses malgré l'incomplétude des données. À cette base déjà partielle s'ajoutent les restrictions imposées par *Alior* sur les données sectorielles, ce qui limite les comparaisons externes et l'identification de certaines absences spécifiques, comme les congés parentaux, affectant également la précision du projet de connaissances.
2. La représentativité des conventions de mesure : en l'absence de directives officielles sur le calcul de l'absentéisme, les conventions de mesure reposent sur

la simplicité, ce qui peut conduire à des représentations simplifiées, voire trompeuses, des données RH. Par exemple, quantifier les absences en jours entiers plutôt qu'en fractions peut sous-évaluer les présences réelles et gonfler artificiellement les taux d'absentéisme. De même, l'agrégation mensuelle en valeurs annuelles, bien qu'utile pour une vue d'ensemble, peut masquer les dynamiques internes. Enfin, concaténer les périodes d'arrêt de travail en une seule séquence continue peut également dissimuler les nuances entre différentes périodes d'absence, surtout lorsque différents types d'arrêts se succèdent, compliquant ainsi l'interprétation des données RH.

Ces épreuves révèlent les limites du déterminisme technique appliqué aux données RH, car elles simplifient les dynamiques sociales sous-jacentes à l'absentéisme. Cette simplification découle notamment des compromis établis lors de la séquence de qualification. Ainsi, le choix des données socio-démographiques et des conventions de mesure n'est pas neutre, mais résulte de décisions méthodologiques influencées par des arbitrages.

Tableau 17 : Synthèse des épreuves d'exploration issues du premier projet de connaissances

Epreuves d'exploration	Descriptions des épreuves
L'accès aux données brutes	**Restrictions des données socio-démographiques :** les données issues de la base de données RH fictive sont confinées et restreintes en portée.
	Contraintes des données RH externes : les données publiques fournies par Alior ne facilitent pas l'exclusion sélective de certaines données lors de l'analyse.
La représentativité des conventions de mesure	**Défi de précision avec l'agrégation :** l'agrégation annuelle peut masquer des dynamiques internes importantes, altérant la précision des analyses.
	Défi d'interprétation avec la concaténation : La concaténation en séquences continues risque de masquer les distinctions entre périodes/types d'absence, réduisant l'exactitude des interprétations des données RH.

Cette première capitalisation des données RH se concentre sur l'absentéisme *réel univarié*. Développée par les *data scientists*, elle vise à produire une connaissance *descriptive* de l'absentéisme. Pour analyser ce projet de connaissances, trois dimensions sont explorées :

1. Le territoire d'exploration épistémique des données RH ;
2. Le savoir-faire des agents économiques impliqués ;
3. Les épreuves d'exploration qui en découlent.

Les *data scientists* orientent d'abord ce projet vers deux types d'exploration :

1. Une exploration interne, spécifique à l'entreprise ;
2. Une exploration externe, destinée à la comparaison nationale et sectorielle.

En résumé, les deux types d'épreuves rencontrées par les *data scientists* révèlent que les analyses des données RH sont limitées. D'une part, l'accès restreint aux données brutes freine les analyses. D'autre part, les conventions de mesure simplifiées réduisent la capacité à saisir la complexité des absences. En conséquence, la précision des résultats et la représentativité « réelle » de l'absentéisme sont compromises. Les *data scientists* doivent donc trouver un équilibre entre la simplification des conventions et la représentativité des données RH.

3. Projet de connaissances II : l'absentéisme réel multivarié

À la suite de l'absentéisme *réel univarié*, un deuxième projet est développé par les *data scientists* : l'absentéisme *réel multivarié.* Pour rendre compte de ce dernier, il est nécessaire d'investiguer les trois mêmes dimensions :

1. Le territoire d'exploration épistémique des données RH ;
2. Le savoir-faire des agents économiques impliqués ;
3. Les épreuves d'exploration qui en découlent.

3.1. Territoire d'exploration épistémique des données RH

Le deuxième projet de connaissances a pour objectif de produire une connaissance *explicative* de l'absentéisme. Ce projet, axé sur l'absentéisme *réel multivarié*, repose sur le développement d'un modèle analytique servant d'intermédiaire pour analyser

les relations entre les différentes données RH. Ce modèle permet d'identifier les spécificités de différents groupes de salariés, distinguant ceux présentant un fort absentéisme de ceux présentant un faible absentéisme :

« *Le but, c'est de donner les clés aux managers opérationnels pour se dire : j'ai vu l'absentéisme dimension par dimension. Est-ce qu'il y a des combinaisons de facteurs* [données] *qui sur* [ou sous] *-concentrent l'absentéisme ?* » (Tristan, chef de projet, notes prises lors d'une réunion, 15/04/2022).

Ce deuxième projet de connaissances se concrétise donc par la segmentation en sous-groupes de salariés, en combinant les différentes données socio-démographiques présentées dans le cadre du premier projet de connaissances. Ces dernières sont : l'âge, le salaire, l'ancienneté, le sexe, le statut, le type de contrat, le nombre d'enfants et la situation familiale. Leur combinaison permet ainsi une distinction plus précise entre les sous-groupes de salariés à fort et faible absentéisme. L'objectif est de doter les managers d'une « loupe », leur permettant de caractériser ces différents groupes et d'identifier les données RH qui sur-concentrent l'absentéisme.

Les *data scientists* choisissent de décliner ce deuxième projet de connaissances en trois représentations graphiques :

1. Un tableau des sous-groupes de salariés : regroupe les salariés selon des données socio-démographiques communes, en indiquant le nombre de salariés et le taux d'absentéisme moyen pour chaque sous-groupe.
2. Une catégorisation des sous-groupes par taux d'absentéisme : catégorisés en fonction de leur taux d'absentéisme, le « point de référence zéro » étant le taux d'absentéisme moyen de l'entreprise.
3. Une échelle de couleur : allant du vert au rouge pour représenter la concentration relative par rapport au taux d'absentéisme moyen. Les teintes de rouge indiquent une sur-représentation (taux d'absentéisme supérieur à la moyenne de l'entreprise), tandis que les teintes de vert indiquent une sous-représentation (taux d'absentéisme inférieur à la moyenne de l'entreprise).

L'absentéisme *réel multivarié*, en tant que deuxième projet de connaissances, se concrétise par la segmentation en sous-groupes de salariés. Cette approche permet ainsi une évaluation comparative et plus nuancée des taux d'absentéisme par rapport

à la moyenne de l'entreprise, mettant en évidence les spécificités de chaque sous-groupe.

3.2. Savoir-faire des agents économiques

Toujours basé sur un principe de simplicité, Tristan (chef de projet) décide, dans le cadre du développement de l'absentéisme *réel multivarié*, de réutiliser un modèle analytique déjà employé lors de ses précédents projets chez *QIA* : le modèle « *target* ».

Ce modèle, initialement élaboré dans le cadre de projets financiers, a été développé pour alimenter un nouveau dispositif de vente proactive destiné aux particuliers. Il découle de l'hybridation de deux modèles :

1. Un modèle de ciblage des sous-populations ;
2. Un modèle de scoring.

Il fonctionne comme un ensemble d'arbres de décision aléatoires, sélectionnant à chaque phase d'entraînement un sous-ensemble aléatoire de données, assurant ainsi une analyse robuste et variée des données RH. Les arbres sont ensuite examinés pour identifier les nœuds, qui représentent des points de décision spécifiques dans la structure de l'arbre. Les chemins conduisant à ces nœuds sont définis pour formuler des conventions de mesure. Par conséquent, un ensemble de conventions est généré et les plus pertinentes sont automatiquement sélectionnées :

« […] *la détection de sous-groupes nous a permis d'identifier des populations de salariés avec un très fort ou faible absentéisme par rapport à la moyenne. Les variables qui sont revenus majoritairement pour réaliser les séparations sont le sexe, l'âge, l'ancienneté, le type de contrat ainsi que la présence d'enfants à charge...* » (Marco, stagiaire *data scientist*, notes prises lors d'une réunion, 07/02/2023).

La sélection des conventions de mesure repose sur deux aspects principaux.

1. La réduction des chevauchements de groupes de salariés en analysant les intersections entre les différentes conventions établies. Cette méthode permet de clarifier et de spécifier les distinctions entre les groupes en éliminant les groupes redondants par l'identification des salariés en commun.

2. Le « *lift* » qui mesure l'importance de la relation entre les sous-groupes définis par la convention et l'absentéisme moyen de l'entreprise. Cette approche permet ainsi de récupérer les groupes les plus pertinents, en mettant en évidence les sur- et sous-concentrations d'absentéisme, en se focalisant sur les combinaisons de données RH présentant un absentéisme différent de la moyenne.

« […] *côté QIA c'est un algo*[rithme] *qu'on a maintenant l'habitude de manipuler* » (Anatole, stagiaire *data scientist*, notes prises lors d'une réunion, 15/04/2022).

Cependant, l'intégration de plus de quatre données RH dans le modèle *target* accroît sa complexité et réduit sa capacité de généralisation :

- Mathieu (*data scientist*, *QIA*) : *« Ce n'est pas trop dur d'interpréter les sous-groupes, parce qu'il y a une trop grande complexité au niveau des règles* [conventions de mesure] *?* »
- Tristan (chef de projet) : « *Si, c'est le bordel et on n'y comprend rien quand il y a plus de cinq critères* [données]*, ça c'est sûr. En fait, le but d'avoir un peu plus de critères ; c'est pour dépasser les critères usuels qui nous met sur le salaire ou l'âge. Mais en réalité, au-delà de quatre critères, ça n'a aucun sens.* » (Notes prise lors d'une réunion, 15/04/2022).

En somme, l'intégration de plus de quatre types de données RH peut non seulement diminuer la précision du modèle, mais également accroître sa complexité, rendant son interprétation plus difficile. Cette limitation soulève des questions quant à la représentativité des données RH.

3.3. Epreuves d'exploration

Dans le cadre du développement de l'absentéisme *réel multivarié*, les *data scientists* se trouvent confrontés à deux grandes catégories d'épreuves. Une synthèse des épreuves est illustrée par le Tableau 18.

1. La représentativité du modèle analytique : initialement conçu pour des projets financiers et orienté vers la vente proactive, le modèle *target* a été transposé à l'analyse de l'absentéisme, soulevant des questions quant à sa pertinence dans ce nouveau contexte. L'utilisation de plus de quatre données (sexe, âge, ancienneté, type de contrat ou présence d'enfants à charge) peut limiter sa capacité de

généralisation, affectant ainsi la précision des prédictions sur de nouvelles données.

2. La représentativité des méthodes de segmentation des sous-groupes : l'utilisation des quatre données principales dans les méthodes de segmentation soulève également des enjeux de simplification. Bien que ces données permettent de réduire les chevauchements entre groupes et de clarifier certaines distinctions, elles risquent de masquer des dynamiques plus complexes, notamment lorsque des salariés appartiennent à plusieurs sous-groupes. Cela peut affecter la granularité et la précision de l'analyse. Par ailleurs, l'utilisation du *lift* pour identifier des sous-groupes présentant des écarts significatifs par rapport à l'absentéisme moyen ne garantit pas une explication causale. Cette approche risque donc de produire des interprétations réductrices, compromettant ainsi la représentativité et la pertinence des sous-groupes identifiés.

Ces épreuves révèlent ainsi les limites inhérentes à une approche déterministe des données RH. Le choix du modèle analytique et des méthodes de segmentation traduit ainsi des décisions méthodologiques conditionnées par des compromis, plutôt que purement objectives.

Tableau 18 : Synthèse des épreuves d'exploration issues du deuxième projet de connaissances

Epreuves d'exploration	Descriptions des épreuves
La représentativité du modèle analytique	**Problème de généralisation** : le modèle *target*, conçu pour des projets financiers, est limité par l'utilisation de plus de quatre données, ce qui entrave sa généralisation en GRH.
La représentativité des méthodes de segmentation des sous-groupes	**Simplification excessive des chevauchements** : la réduction des chevauchements simplifie l'analyse, mais peut masquer des dynamiques importantes.
	Biais du *lift* : le *lift* identifie des sous-groupes significatifs sans garantir une causalité claire.

Cette deuxième capitalisation des données RH se concentre sur l'absentéisme *réel multivarié*. Celle-ci vise à produire une connaissance *explicative* de l'absentéisme. Pour analyser ce projet de connaissances, trois dimensions sont explorées :

1. Le territoire d'exploration épistémique des données RH ;
2. Le savoir-faire des agents économiques impliqués ;
3. Les épreuves d'exploration qui en découlent.

Tout d'abord, l'absentéisme *réel multivarié* permet une évaluation comparative et détaillée des taux d'absentéisme par rapport à la moyenne de l'entreprise, mettant en lumière les sous-groupes de salariés à fort et faible absentéisme. Pour assurer le développement et garantir la crédibilité de ce deuxième projet, Tristan (chef de projet) décident de réutiliser un modèle élaboré lors de précédents projets menés par *QIA*, à savoir le modèle *target*.

En somme, les deux catégories d'épreuves montrent que l'utilisation d'un modèle et de méthodes de segmentation limités par un petit nombre de données (quatre données principales) peut nuire à la complexité et à la granularité requises pour une analyse fine des données RH, rendant difficile une interprétation précise de l'absentéisme.

4. Projet de connaissances III : l'absentéisme *latent*

A la suite de l'absentéisme *réel multivarié*, un troisième et dernier projet de connaissances est développé par les *data scientists* : l'absentéisme *latent*. Pour analyser ce projet, il convient d'explorer les trois mêmes dimensions :

1. Le territoire d'exploration épistémique des données RH ;
2. Le savoir-faire des agents économiques impliqués ;
3. Les épreuves d'exploration qui en découlent.

4.1. Territoire d'exploration épistémique des données RH

Ce troisième et dernier projet vise à produire une connaissance *prédictive* de l'absentéisme afin de cibler les causes sous-jacentes du phénomène :

« Le but est de dire : quel est l'absentéisme auquel je t'attends compte tenu de tes caractéristiques socio-démographiques ? Ça va permettre de nuancer la lecture [de l'absentéisme]*. »* (Tristan, chef de projet, notes prises lors d'une réunion, 26/11/2021).

Afin d'assurer la pertinence de ce troisième projet de connaissances, les *data scientists* optent pour une capitalisation des données RH selon deux catégories distinctes :

1. Les entités géographiques ;
2. Les métiers.

Sous l'initiative de Tristan (chef de projet), le taux d'absentéisme *latent* est défini comme l'écart entre le taux absentéisme *attendu* simulé et le taux d'absentéisme *réel* :

« On va modéliser, pour un métier ou pour une entité donnée, l'absentéisme attendu en fonction des caractéristiques socio-démographiques de la population concernée. Et après [...] *on va venir comparer cet absentéisme attendu avec l'absentéisme réellement observé. Si l'absentéisme réel est largement supérieur à l'absentéisme attendu, ça signifie que l'entité ou le métier est anormal et que cet absentéisme s'explique par d'autres facteurs que les caractéristiques socio-démographiques… »* (Anatole, stagiaire *data scientist*, notes prises lors d'une réunion, 15/04/2022).

La représentation schématique de l'absentéisme *latent* est illustrée à la Figure 14[21].

[21] Il convient de préciser que cette représentation schématique a pour but de faciliter la compréhension du lecteur. Elle ne reflète pas la visualisation réelle développée par les *data scientists*, laquelle est confidentielle. C'est pourquoi elle est accompagnée d'une description détaillée de la visualisation originale.

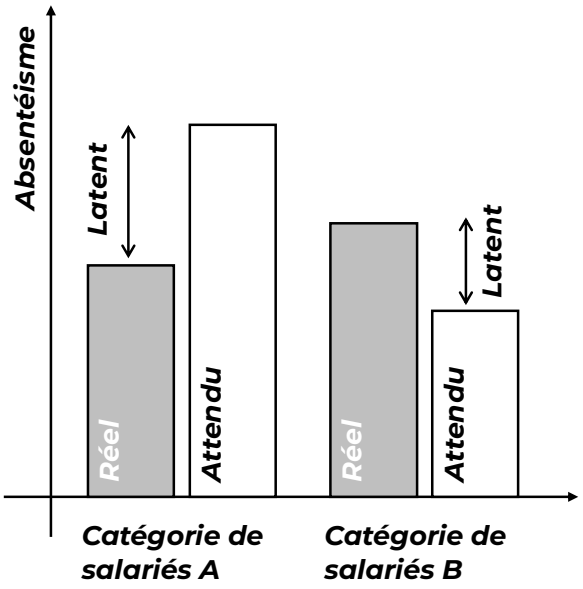


Figure 14 : L'absentéisme *latent*

Pour illustrer visuellement l'absentéisme *latent*, ils choisissent de le décliner en quatre représentations graphiques distinctes :

1. Un graphique à bulles : classe les entités géographiques et les métiers en fonction de l'importance de l'absentéisme *latent*, c'est-à-dire l'écart entre leur taux d'absentéisme *attendu* (sur l'axe des ordonnées) et leur taux d'absentéisme *réel* (sur l'axe des abscisses).
2. Une proportion des salariés par groupe : est représentée par la taille des bulles pour chaque groupe (entités géographiques et métiers).
3. Une échelle de couleur : allant du vert au rouge pour les bulles, indiquant l'ampleur de l'absentéisme *latent*.
4. Une droite d'équivalence : ou une droite « invisible », définie par la convention de mesure Y (taux d'absentéisme *attendu*) = X (taux d'absentéisme *réel*), et utilisée comme référence pour distinguer les entités géographiques et les métiers ayant des « bonnes pratiques » (situées en dessous de la droite) de celles ayant des « mauvaises pratiques » (situées au-dessus de la droite) en termes de gestion de l'absentéisme.

L'absentéisme *latent* repose ainsi sur l'ampleur de l'écart entre l'absentéisme *attendu* et l'absentéisme *réel* des entités géographiques et des métiers. Un écart

négatif important indique que le taux d'absentéisme *réel* est significativement plus élevé que celui *attendu*, mettant en évidence des anomalies ou des situations préoccupantes propres à une entité géographique ou à un métier.

Avec ce troisième projet de connaissances, les *data scientists* cherchent ainsi à reproduire la complexité de l'absentéisme lequel présente un caractère difficilement quantifiable en raison de sa nature intrinsèquement elusive :

« […] *la modélisation d'effets RH est très peu précise. Les métriques qu'on a sur les modèles d'absentéisme ne sont pas très bonnes et c'est normal parce que l'effet est essentiellement dérivé par des choses qui ne sont pas dans les données, qui sont très individuelles. Est-ce que tu as plutôt un physique fragile donc tu es souvent malade ? Est-ce que tu as une propension à l'absentéisme parce que tu es désengagé de ton entreprise ? Ce sont toutes ces choses-là qui vont évidemment expliquer 80 % de si tu es absent ou pas. Et en plus, il y a un effet complètement aléatoire. Tu as été absent en juillet ? C'est imprévisible… »* (Tristan, chef de projet, notes prises lors d'une réunion, 15/04/2022).

En outre, à l'instar des autres projets de connaissances, la modélisation des écarts de l'absentéisme *latent* dépend étroitement des données issues de la base de données RH. Cette base détermine explicitement quelles données peuvent être quantifiées, introduisant ainsi une forte variabilité selon les bases de données utilisées. Concernant les projets sur l'absentéisme, ces derniers s'appuient sur deux types de bases :

1. Une base fictive, utilisée pour l'entraînement des modèles, qui est confrontée à des défis d'exhaustivité, un problème déjà évoqué lors du premier projet de connaissances.
2. La base de données *DSN*, qui ne sera accessible qu'avec l'engagement d'un premier client, présente également des défis d'exhaustivité, comme l'identifie Justine (directrice des projets stratégiques) :

« […] *les variables qu'on ne pourra pas approcher avec la DSN - à moins de faire une étude sur mesure - sont globalement : les variables relatives au management (relation managériale et variables descriptives du manager)* […] *;* [et] *les variables relatives à la nature elle-même du travail (ex. sur GSM les employés travaillaient sur des « chantiers » de différentes nature (nettoyage, maintenance, entretien, etc.…)*

rattachés à des clients, eux-mêmes travaillant dans un certain secteur d'activité, qui était extrêmement déterminant dans les causes d'absentéisme). » (Justine, courriel envoyé le 20/01/2021).

Étant donné l'importance de l'exhaustivité des bases de données pour analyser un phénomène aussi complexe et multifactoriel que l'absentéisme, il semble contradictoire que les *data scientists* se limitent à une unique base de données *normative*, telle que la *DSN*. Ce choix peut masquer certaines dimensions critiques du phénomène, notamment parce que des données RH pertinentes, comme celles liées au management ou au contexte du travail, ne sont pas incluses dans la base utilisée.

4.2. Savoir-faire des agents économiques

Dans l'objectif de développer l'absentéisme *latent*, les *data scientists* vont utiliser de manière chronologique, quatre modèles analytiques différents :

1. Le modèle linéaire ;
2. Le modèle additif généralisé ;
3. Le modèle fréquence/sévérité ;
4. Le modèle mixte.

4.2.1. Le modèle linéaire pour une représentation « simplifiée » de l'absentéisme *latent*

Entre février et juillet 2021, Olivier (*data scientist* junior) élabore un premier modèle analytique « linéaire » sous la supervision de Tristan (chef de projet). Ce choix a pour objectif de rendre le modèle accessible facilitant ainsi l'interprétation de ses résultats. Il permet d'analyser l'impact de chaque donnée RH en quantifiant son influence (positive ou négative) et d'évaluer sa contribution relative au taux d'absentéisme *latent*.

Étant donné l'appartenance d'Olivier à une autre filiale de *QIA*, Tristan (chef de projet) opte pour ce modèle afin de garantir une progression rapide jusqu'à la finalisation du projet de connaissances. L'objectif demeure, avant toute chose, d'acquérir un premier client. Il est impératif d'avoir quelque chose à présenter ! Ce choix vise ainsi à assurer que la participation d'Olivier (*data scientist* junior) ne soit pas brusquement interrompue en milieu de développement. De plus, en raison de son

statut de junior, la préférence est également donnée à un modèle simplifié pour préserver sa clarté et sa compréhension.

Cependant, cette quête de simplicité s'avère être une illusion trompeuse. L'illusion réside dans l'idée qu'un modèle analytique « simple » peut représenter effectivement l'absentéisme. En effet, la quête de simplicité peut masquer les variations non linéaires présentes dans les données RH, compliquant ainsi leur analyse et leur interprétation :

« […] *l'absentéisme ne dépend pas* [par exemple] *de l'âge de façon linéaire. Peut-être que l'écart entre 30 et 25 ans n'est pas le même que l'écart entre 50 et 45 ans* […]*. Il peut y avoir un effet de plafond et donc limitant, parce que les effets* [présumés des données RH] *sont linéaires… »* (Luc, *data scientist* senior, entretien no1, 11/02/2023).

Il est alors évident que cet intermédiaire ne renforce pas la crédibilité de l'absentéisme *latent*. En effet, le modèle linéaire est limité par sa capacité à modéliser uniquement des relations strictement linéaires, ce qui le rend peu flexible pour capturer les dynamiques complexes présentes dans les données RH. Les relations entre ces dernières ne suivent pas une tendance linéaire claire : elles peuvent être courbes, irrégulières ou fluctuantes, ne correspondant pas au cadre rigide de ce premier modèle. En conséquence, Tristan (chef de projet) conclue que ce dernier est trop rigide, imposant des contraintes qui ne capturent pas de manière adéquate les interactions sous-jacentes entre les données RH ; notamment par la prise en compte d'un nombre limité de ces dernières :

« La régression linéaire ne permettait pas de faire des interprétations par métier ni par entités. C'est ce qui manquait. Il fallait que ça prenne [en compte] *plus de facteurs explicatifs pour que ce*[la] *soit utile. Si tu prends* [en compte] *juste les quatre premières variables, le salaire, le nombre d'enfants et l'âge… C'est limité. Tu ne peux pas interpréter* [l'absentéisme] *parce que, tu vas toujours te retrouver avec les jobs physiques et pas bien payé… »* (Luc, *data scientist* senior, entretien no1, 11/02/2023).

Au regard du manque de représentativité du modèle linéaire dans l'analyse de l'absentéisme, les *data scientists* se voient donc contraints de développer un modèle plus sophistiqué. Ce nouvel intermédiaire vise ainsi à modéliser plus fidèlement les relations entre ces dernières.

4.2.2. Le modèle additif généralisé pour une représentation « souple » de l'absentéisme *latent*

De juillet à octobre 2021, Tristan (chef de projet) et Anatole (stagiaire *data scientist*) - lors de son recrutement en tant que stagiaire chez *QIA* - entreprennent l'exploration de modèles plus avancées pour atténuer les limitations imposées par le modèle linéaire. Leur recherche les a tout d'abord orientés vers les modèles additifs généralisés (MAG) :

« Si tu voulais interpréter les effets du métier [et de l'entité géographique]*, il n'y avait pas le choix que d'aller vers ce type de modèle. »* (Luc, *data scientist* senior, entretien no1, 02/11/2021).

Le choix des *data scientists* se base principalement sur la performance des modèles MAG en comparaison aux modèles linéaires traditionnels. Les MAG offrent une flexibilité dans la modélisation des relations plus complexes et non linéaires, ce qui permet de mieux appréhender les variations entre les données RH, même si leur interprétation peut s'avérer plus complexe.

Le développement de cet intermédiaire semble ainsi permettre aux *data scientists* d'assurer la crédibilité de l'absentéisme *latent*. Malgré leur intérêt envers ce modèle, celui-ci présente tout de même d'importants défis pour Anatole (stagiaire *data scientist*) en raison de son niveau d'expérience : il est intrinsèquement plus complexe, nécessite une compréhension approfondie de son fonctionnement, et les implémentations disponibles en langage Python étaient soit inexistantes, soit jugées insuffisamment configurables. Cet épisode de turbulence est cependant rapidement résolu par l'arrivée de Luc en tant que *data scientist* senior en octobre 2021 :

« J'aimais bien le projet RH : comprendre les facteurs qui drivent l'absentéisme, l'approche QVT et tout ça me semblait être un bon produit […]*. Je trouvais qu'il y avait en même temps un sens* [métier] *et un challenge technique qui étaient intéressants. »* (Luc, *data scientist* senior, entretien no1, 02/11/2021).

Le recrutement de Luc, en tant que *data scientist* senior, a considérablement influencé le choix des modèles pour ce troisième projet de connaissances. Non seulement il développe le modèle MAG, mais il introduit également un nouveau modèle : celui de fréquence/sévérité. Ce dernier permet d'aborder les relations entre

les données RH de manière plus pragmatique, en capturant leurs dynamiques sous-jacentes selon deux dimensions distinctes.

4.2.3. Le modèle fréquence/sévérité pour une représentation « pragmatique » de l'absentéisme *latent*

« Ça pose un challenge technique parce qu'il faut réussir à le faire » (Luc, *data scientist* senior, entretien no1, 02/11/2023).

Pour que le développement de l'absentéisme *latent* reflète fidèlement la réalité du phénomène, Luc (*data scientist* senior) avec l'aide d'Anatole (stagiaire *data scientist*), formule une nouvelle hypothèse technique : combiner la souplesse du modèle GAM avec la précision du modèle « fréquence/sévérité » :

« [Il y a] *deux avantages à réaliser ce type de modélisation :* [le premier,] *améliorer nos performances et* [le second,] *avoir une interprétation plus fine parce qu'on peut interpréter différemment la fréquence et la sévérité* […]*. La seule condition pour pouvoir faire ce type de modélisation, est* [que ces deux dimensions] *ne soient pas parfaitement corrélées.* [Cette] *condition est respectée dans nos données. »* (Anatole, stagiaire *data scientist*, notes prises lors d'une réunion, 15/04/2022).

Cette hypothèse technique trouve ses racines dans l'expérience professionnelle antérieure de Luc (*data scientist* senior), où celui-ci s'adonnait à la modélisation des sinistres. Largement répandu dans le secteur de l'assurance, le modèle fréquence/sévérité est utilisé pour évaluer deux dimensions distinctes inhérentes à un même événement :

1. La fréquence : mesurant la récurrence d'un événement ;
2. La sévérité : quantifiant le coût associé à cet événement.

L'approche pragmatique des *data scientists*, renforcée par l'expérience antérieure de Luc (*data scientist* senior) avec ce modèle, les amène à envisager la transposition de celui-ci à l'absentéisme, pouvant également être caractérisé selon deux dimensions :

1. La fréquence : le nombre d'arrêt de travail attendu sur l'année ;
2. La sévérité : la durée attendue de cet arrêt de travail.

Ce dernier permet ainsi de modéliser la durée totale des arrêts de travail d'un salarié sur l'année :

« [Par exemple,] *on observe que les jeunes parents ont une fréquence d'absence plus élevée que les gens proches de la retraite, mais qu'au final l'absentéisme moyen est plus faible pour les jeunes parents parce qu'ils ont plein de petites absences courtes tandis que les gens proches de la retraite ont moins d'absences mais elles peuvent durer trois mois… »* (Luc, *data scientist* senior, entretien no1, 02/11/2023).

Bien que la combinaison des deux modèles soit pertinente, elle ne saisit toutefois qu'une fraction de la complexité du phénomène de l'absentéisme. En se concentrant exclusivement sur les effets « fixes », c'est-à-dire les relations constantes entre les données RH et les entités géographiques et les métiers, elle néglige les variations « aléatoires » d'un salarié à l'autre. Les effets fixes capturent des aspects prévisibles et systématiques (tels que l'impact général de l'âge) sur l'absentéisme. Ainsi, la seule combinaison des modèles GAM et fréquence/sévérité restreint leur capacité à offrir une représentation plus exhaustive du phénomène.

Face à ce nouvel impératif, Luc (*data scientist* senior) suggère également l'incorporation d'un troisième modèle, appelé modèle « mixte », afin d'intégrer les effets « aléatoires » dans le développement de l'absentéisme *latent*. Le choix de ce modèle vise à améliorer la gestion de la variabilité des données RH au sein d'un même groupe de salariés, une variabilité qui ne peut être expliquée uniquement par les données RH dans la *DSN*.

4.2.4. Le modèle mixte pour une représentation « intégrative » de l'absentéisme *latent*

« [L'absentéisme] *correspond à un phénomène physique avec des effets aléatoires.* […] [Il existe] *des métiers hyper fatigants où tout le monde est absent trois fois plus que les autres et* [d'autres au sein desquels] *une absence longue* [peut] *complètement driver* [influencer] *ton effet* [dans les données RH] […]*. Il faut prendre ça en compte quand tu fais ta modélisation… »* (Luc, *data scientist* senior, entretien no1, 02/11/2021).

Pour assurer la crédibilité de l'absentéisme *latent*, les *data scientists* doivent considérer l'impact de diverses influences, notamment celles des données RH

absentes de la *DSN*. En intégrant les effets aléatoires, la combinaison des modèles gagne en flexibilité, permettant ainsi de tenir compte des variations individuelles qui échappent aux données RH socio-démographiques issues de la *DSN*. Cette approche permet de capturer l'incertitude ainsi que les particularités et caractéristiques spécifiques des différents groupes (entités géographiques et métiers), rendant les modèles plus conformes à la « réalité ». Ce choix de modèle revêt une importance particulière dans la compréhension de l'absentéisme, étant donné que celui-ci intègre de nombreux microphénomènes interdépendants.

Une collaboration intensive d'environ trois semaines entre Anatole (stagiaire *data scientist*) et Luc (*data scientist* senior) est lancée pour associer les trois intermédiaires : les modèles MAG, fréquence/sévérité et mixte, dans le développement de l'absentéisme *latent*. Toutefois, en raison de la complexité des données RH et du manque de temps alloué à ce travail - accentuées par la priorisation de projets plus lucratifs - l'engagement des *data scientists* est réduit dès novembre 2021. Jusqu'à mai 2022, une interphase technique s'installe, en attendant l'arrivée de Marco en tant que stagiaire *data scientist* :

« […] *Je me suis cassé les dents* [à la difficulté] *de* [combiner] *la fréquence et la sévérité avec effet mixte et de pouvoir les mélanger à la fin.* […] *dès que j'ai récupéré les données, j'ai essayé mais je n'ai pas réussi* […]. *Donc c'est de là* [que] *viennent les objectifs donnés à Marco* (stagiaire *data scientist*) *… »* (Luc, *data scientist* senior, entretien no1, 02/11/2021).

4.2.5. Le travail de combinaison des modèles pour une représentation « holistique » de l'absentéisme *latent*

Afin d'assurer la crédibilité de l'absentéisme *latent*, les *data scientists* doivent minimiser l'erreur associée à leur modélisation (cf. combinaison des trois modèles) pour les deux catégories choisies : (1) les entités géographiques et (2) les métiers.

Cette erreur représente la différence entre les prévisions de la modélisation et les données *réelles* observées.

En ce qui concerne la catégorie des métiers, ceux-ci disposent d'un nombre suffisant de salariés par métier (soit neuf métiers pour 64 859 salariés). Cela permet d'obtenir des résultats crédibles pour chaque métier grâce à un échantillon adéquat.

Toutefois, la répartition est moins homogène au niveau des entités géographiques, où le nombre de salariés fluctue significativement (963 entités géographiques pour 64 859 salariés). Dans ce cas, la modélisation pour une entité comptant seulement quelques salariés s'avère peu judicieuse en raison de la grande variabilité inhérente à un si petit échantillon. Cette variabilité excessive peut entraîner des interprétations erronées ou des conclusions biaisées sur l'absentéisme latent au sein de l'entité géographique. En effet, dans les petits groupes de salariés, quelques absences peuvent fortement influencer le taux d'absentéisme global :

« Une entité qui a quatre salariés, dont un qui est absent toute l'année, aura à minima, un absentéisme de 25 % sans pour autant que l'entité soit responsable de cette absence. » (Marco, stagiaire *data scientist*, notes issues du journal de bord, 15/09/2022).

En raison de l'hétérogénéité de la taille des entités géographiques qui complique la modélisation, les comparaisons entre entités peuvent s'avérées trompeuses. Les *data scientists* doivent donc tenir compte de la variabilité des groupes de salariés au niveau des entités géographiques et des métiers. Leur approche consiste alors à accorder moins de poids aux groupes de petite taille, car ils sont plus susceptibles de présenter des variations aléatoires. Pour ce faire, ils optent pour une partition de l'erreur de la modélisation en deux composantes :

1. L'absentéisme attribué aux groupes selon les entités géographiques ou les métiers ;
2. Le bruit dans les données qui renvoie aux incertitudes ou variations aléatoires inexpliquées.

« Une entité [ou un métier] *avec seulement quatre salariés aura moins de crédibilité par rapport à une autre plus grande avec un nombre plus important de salariés. Cela signifie que* la portion *de l'absentéisme ou de l'absence d'absentéisme attribuée à l'entité* [ou le métier] *plus petite sera réduite en raison de sa crédibilité moindre, tandis que l'entité* [ou le métier] *plus grand aura une crédibilité plus élevée et influencera davantage le modèle. »* (Marco, stagiaire *data scientist*, notes issues du journal de bord, 15/09/2022).

Lors du travail de combinaison des modèles, tandis que le modèle de fréquence démontre une crédibilité convaincante pour le développement de l'absentéisme *latent*,

le modèle de sévérité, en revanche, révèle des niveaux de crédibilité anormaux, voire absurdes :

« En examinant les valeurs, il apparaît que la modélisation fonctionne correctement dans la majorité des cas. Toutefois, pour les valeurs extrêmes, l'interprétation de la crédibilité devient problématique. Par exemple, attribuer une crédibilité de 5000 % à la modélisation d'une entité est dénuée de sens. Par conséquent, cette partie du modèle, impliquant les modèles de fréquence/sévérité avec des effets aléatoires, est jugée défaillante. Cela nous conduit à la décision de ne plus utiliser ces modèles pour la modélisation des données… » (Marco, stagiaire *data scientist*, notes prises lors d'une réunion, 07/02/2023).

Déconcerté par l'échec de ce modèle pour l'analyse de l'absentéisme, Luc (*data scientist* senior) n'arrive toujours pas à en identifier précisément les causes :

« J'ai l'intuition que c'est plus mathématique que métier. Je ne pense pas que ce soit l'histoire des absences. […] *ça devrait fonctionner. Mais il y a un truc qui ne marche pas. Je n'ai pas de certitude, mais à mon avis c'est mathématique… »* (Luc, *data scientist* senior, notes prises lors d'une réunion, 07/02/2023).

A la suite de cet échec, les *data scientists* optent finalement pour la combinaison des modèles GAM et mixte avec un modèle plus compréhensible, centré sur la durée des arrêts de travail, pour développer l'absentéisme *latent* (post-étude) :

« […] *cela reste une petite satisfaction, car le modèle fonctionne tout de même avec l'effet aléatoire, permettant encore l'analyse croisée de l'absentéisme par métier et entité…* » (Marco, stagiaire *data scientist*, notes prises lors d'une réunion, 07/02/2023).

Ainsi, le troisième projet de connaissances, centré sur le développement de l'absentéisme *latent*, démontre que différents modèles analytiques peuvent être utilisés seuls ou combinés pour représenter les relations entre les données RH. Toutefois, leur utilisation ne garantit pas nécessairement leur fiabilité. De nombreux défis subsistent, tels que la dynamique entre les données quantifiables et non quantifiables, qui introduit une dimension fortement aléatoire, ainsi que la capacité à obtenir un nombre suffisant de salariés par groupe pour assurer la crédibilité des résultats.

Le Tableau 19 synthétise les divers modèles utilisés par les *data scientists* en mettant en lumière leurs avantages et leurs limites.

Tableau 19 : Synthèse des modèles utilisés par les *data scientists* pour l'absentéisme *latent*

Périodes	Agents	Modèles	Avantages	Inconvénients
Février à juillet 2021	Olivier (et Tristan - chef de projet - en soutien)	Modèle linéaire	- Facilité de compréhension et d'interprétation ; - Applicable à des contextes variés grâce à sa simplicité.	- Limitation aux relations strictement linéaires ; - Manque de flexibilité pour modéliser les dynamiques de l'absentéisme.
Novembre 2021	Luc (*data scientist* senior) (et Anatole - stagiaire *data scientist* - en soutien)	Modèle additif généralisé (MAG)	- Capacité à modéliser des relations non linéaires entre les données ; - Apporte une plus grande souplesse d'analyse.	- Complexité accrue dans l'interprétation des résultats.
Novembre 2021	Luc - *data scientist* senior (et Anatole - stagiaire *data scientist* - en soutien)	Modèle fréquence/sévérité	- Analyse approfondie du phénomène de l'absentéisme par la séparation des composantes de fréquence et de sévérité ; - Permet une modélisation plus représentative de l'absentéisme.	- Sensibilité aux variations dans les données pouvant induire une instabilité dans les résultats.
Juin 2022 – fin de l'étude	Marco (stagiaire *data scientist*) (et Luc - *data scientist* senior - en soutien)			
Novembre 2021	Luc - *data scientist* senior (et Anatole - stagiaire *data scientist* - en soutien)	Modèle mixte	- Intégration de la variabilité intra groupes par l'application d'effets fixes et aléatoires ; - Pertinent pour les données RH (hiérarchiques ou longitudinales).	- L'ajout de complexité au modèle MAG, notamment par les effets aléatoires, ne garantit pas pleinement sa crédibilité, illustré par le modèle de sévérité.
Juillet 2022 – fin de l'étude	Marco (stagiaire *data scientist*) (et Luc - *data scientist* senior - en soutien)			

4.3. Epreuves d'exploration

Dans le cadre du développement de ce troisième et dernier projet de connaissances, les *data scientists* se trouvent confrontés à deux grandes catégories d'épreuves. Une synthèse des épreuves est illustrée par le Tableau 20.

1. La représentativité des modèles analytiques : divers modèles analytiques peuvent être utilisés pour explorer les relations entre les données RH. Cependant, en raison de la nature multifactorielle de l'absentéisme, de nombreux défis persistent, tels que la variabilité et l'hétérogénéité des données, compliquant la fiabilité des résultats. Aucun modèle n'est parfait et chacun présente ses avantages et ses inconvénients. Le choix des modèles est souvent influencé par des contraintes spécifiques, telles que le temps disponible, comme dans le cas du modèle linéaire utilisé par Olivier (*data scientist* junior), ou l'expérience professionnelle, pour le modèle fréquence/sévérité développé par Luc (*data scientist* senior).
2. Les ressources allouées au développement du projet de connaissances : le manque de temps alloué au projet de connaissances, en raison de la priorité accordée aux projets clients, constitue une barrière significative à son développement. Cette situation découle directement du compromis établi lors de la séquence initiale de qualification, visant à maximiser l'efficience en limitant les investissements tant que la demande n'était pas clairement justifiée. Bien que ce compromis ait permis de répondre aux exigences immédiates, il génère des difficultés persistantes pour les *data scientists*, qui en subissent encore les effets. Le recours majoritaire aux stagiaires comme ressource principale pour ce projet met en évidence un manque de personnel qualifié, restreignant ainsi la possibilité d'une réflexion approfondie sur la modélisation optimale pour l'analyse de l'absentéisme *latent*. Cette épreuve résulte ainsi d'une controverse de qualification partiellement résolue, dont les effets persistent et continuent de peser sur le développement du projet.

En somme, les choix opérés par les *data scientists* dans le développement de ce projet de connaissances révèlent la subjectivité inhérente à leur démarche, influencée par les compromis établis lors de la séquence de qualification. Contraints de jongler en permanence avec le triptyque coût, délais et qualité, ils se trouvent face à un dilemme constant entre la précision et la simplicité des modèles. Cette gestion des

priorités met en évidence les limites d'une approche déterministe, en montrant que les modèles ne sont pas des représentations parfaites de l'absentéisme, mais des constructions approximatives façonnées par des choix méthodologiques et contextuels. Comme l'énonce le célèbre adage de Box (1976) : « *tous les modèles sont faux, mais certains sont utiles* ».

Tableau 20 : Synthèse des épreuves d'exploration issues du troisième projet de connaissances

Epreuves d'exploration	Descriptions des épreuves
La représentativité des modèles analytiques	**Manque de flexibilité (modèle linéaire)** : limité aux relations linéaires pour modéliser la complexité des données RH.
	Complexité d'interprétation (modèle GAM) : modélise des relations non linéaires, mais rend l'interprétation plus difficile.
	Sensibilité aux variations des données (modèle fréquence/sévérité) : permet de distinguer fréquence et sévérité de l'absentéisme, mais les résultats sont instables en fonction des variations des données.
	Clarté des résultats compromise (modèle mixte) : intègre des effets fixes et aléatoires pour une analyse plus fine, mais la complexité accrue peut nuire à la clarté des résultats.
Les ressources liées au développement du projet de connaissances	**Peu de temps alloué au projet** : le projet est relégué en second plan, car la priorité est donnée aux projets clients, ce qui freine son développement.
	Stagiaires peu expérimentés comme principal moteur de production : limite la capacité à approfondir la modélisation et l'analyse des données RH.

5. Controverses de capitalisation des données RH

À l'instar de la séquence de qualification, les controverses liées à la capitalisation des données RH mettent en évidence les compromis nécessaires à la transition vers la requalification.

Dans ce cadre, le réseau de capitalisation se compose de six modes d'existence distincts : (1) Commercial, (2) Développement, (3) Marché, (4) Réglementaire, (5) Stratégique et (6) Technique. Chacun de ces modes joue un rôle spécifique dans les négociations entourant la capitalisation des données RH. Ces modes d'existence sont décrits dans le Tableau 21.

Tableau 21 : Rôles des différents modes d'existence dans le réseau de capitalisation des données RH

Modes d'existence	Rôles
Commercial	Définir les attentes commerciales afin d'orienter la capitalisation.
Développement	Garantir l'adéquation entre les attentes commerciales et les projets de connaissances.
Marché	Apporter des connaissances liées au marché RH pour guider la capitalisation.
Réglementaire	Veiller à la conformité des données RH avec les exigences réglementaires.
Stratégique	Formuler des suggestions sur la cohérence de la capitalisation et leur alignement avec la stratégie globale du cabinet.
Technique	Superviser le développement technique de la capitalisation.

Au sein de la séquence de capitalisation des données RH, deux principales controverses émergent :

1. L'absence de convention de mesure pour l'absentéisme ;
2. Des modèles fondés sur l'expérience des ressources disponibles.

Le Tableau 22 présente ces controverses, en mettant en évidence les perspectives propres à chaque mode d'existence concerné, ainsi que les compromis envisagés pour leur résolution.

Tableau 22 : Controverses pour la séquence de capitalisation des données RH

Sujets des controverses		**Absence de convention de mesure pour l'absentéisme**	**Modèles fondés sur l'expérience des ressources disponibles**
Perspectives spécifiques aux modes d'existence	**Commercial**	Conventions opérationnelles idéales	
	Développement	Conventions pragmatiques	Modèles interprétables et fonctionnels
	Marché	Conventions personnalisées	
	Réglementaire	Absence de conventions	
	Stratégique	Conventions sectorielles adaptées	
	Technique		Modèles adaptables sous contraintes
Compromis négociés issu des controverses		**Conventions pragmatiques évolutives**	**Modèles à complexité croissante**

5.1. Controverse I : l'absence de convention de mesure pour l'absentéisme

Les conventions adoptées pour mesurer l'absentéisme peuvent significativement influencer les résultats en modifiant ou en reconfigurant les propriétés des données RH. L'absence de conventions officielles au sein du mode d'existence Réglementaire aggrave cette problématique, entraînant des variations dans le calcul de l'absentéisme selon les méthodes employées et soulevant des questions de comparabilité des données RH.

Outre le mode d'existence Réglementaire, quatre autres modes offrent des perspectives spécifiques concernant cette controverse :

1. Le mode d'existence Commercial : privilégie une sélection stratégique des données en excluant les types d'absences considérées comme hors du champ d'action des interventions de gestion, telles que les congés parentaux. Son objectif est de maximiser l'utilité opérationnelle des données RH, en centrant l'analyse sur celles qui permettent des actions directes et ciblées, optimisant ainsi l'efficacité des initiatives mises en œuvre. Cette approche repose sur des conventions opérationnelles idéales.
2. Le mode d'existence Développement : s'appuie principalement sur les données RH disponibles, qu'elles soient complètes ou non. Il favorise la rapidité et l'efficacité des calculs, en adaptant les conventions de mesure à la nature fragmentaire des données, qu'elles proviennent de bases internes ou de sources gratuites en ligne. Cette approche s'appuie sur des conventions pragmatiques.
3. Le mode d'existence Marché : adopte une méthodologie exhaustive, intégrant un large éventail de données RH, y compris celles que le mode d'existence Commercial souhaite exclure, telles que les absences parentales. Cette approche se base sur des conventions personnalisées.
4. Le mode d'existence Stratégique : met l'accent sur l'importance de comparaisons sectorielles, jugées plus spécifiques et pertinentes que celles proposées par le mode Marché, qui s'appuie sur des données publiées gratuitement en ligne. Pour le mode Stratégique, les comparaisons sectorielles constituent la référence principale pour interpréter les données d'absentéisme. Cette approche est axée sur des conventions sectorielles adaptées.

Les comparaisons entre les données RH n'étant pertinentes que dans des contextes uniformes, un écart se creuse entre les ambitions des modes d'existence Commercial et Stratégique et les contraintes du mode Développement. Ce dernier doit naviguer entre ces ambitions et les possibilités offertes par les données RH disponibles. Limité par la nature incomplète et disparate des données, le mode Développement peine à appliquer le principe du « toutes choses égales par ailleurs », en particulier lorsqu'il intègre des données issues du mode d'existence Marché. Face à ces contraintes, un compromis fondé sur des conventions pragmatiques évolutives est privilégié, en mettant l'accent sur les analyses réalisables et sur l'acquisition progressive de données pour enrichir les analyses futures.

5.2. Controverse II : des modèles fondés sur l'expérience des ressources disponibles

La complexité de l'absentéisme repose sur des dynamiques imprévisibles, où les relations entre les données RH échappent aux tendances préétablies et peuvent adopter des formes non linéaires, irrégulières ou évolutives. L'analyse de ce phénomène multifactoriel, composé de nombreux microphénomènes interdépendants, nécessite donc une approche analytique flexible et adaptable.

Deux modes d'existence principaux reflètent des perspectives spécifiques face à cette controverse :

1. Le mode Développement : met l'accent sur l'interprétabilité et l'utilité opérationnelle des modèles analytiques, avec pour objectif de les rendre facilement exploitables. Pour garantir cette accessibilité et assurer une exécution rapide, ce mode privilégie l'utilisation de modèles déjà éprouvés dans d'autres contextes. Cette approche permet de générer des résultats rapidement interprétables, facilitant ainsi la collaboration avec les *data scientists* moins expérimentés du mode Technique (comme les juniors ou stagiaires). En outre, elle ne se limite pas aux objectifs d'apprentissage interne, mais permet également de fournir des preuves concrètes de la fonctionnalité aux clients potentiels, bien que la pertinence de ces modèles dans le contexte spécifique de l'absentéisme puisse parfois être limitée.
2. Le mode Technique : vise à adapter et améliorer les modèles analytiques en tenant compte des contraintes techniques et contextuelles propres au projet. Il intègre

diverses dimensions afin d'ajuster les modèles aux spécificités des données RH tout en maximisant l'utilisation des ressources disponibles. La sélection des modèles est influencée par l'expérience des *data scientists* et la complexité requise par le projet. Cependant, le développement de ces modèles - nécessitant des ressources importantes - conduit à une hiérarchisation des projets, ce qui entraîne le report de certains développements et limite la création de modèles plus exhaustifs. Cette situation génère ainsi une frustration au sein du mode Technique.

Compte tenu du faible investissement dans la capitalisation, le mode d'existence Développement privilégie une approche pragmatique, centrée sur des modèles à la fois interprétables et fonctionnels. Le mode d'existence Technique, quant à lui, doit concilier les contraintes de délais et de qualité, naviguant entre précision et simplicité. Chaque modèle implique des compromis spécifiques, dictés par les ressources disponibles. Ce processus souligne la subjectivité des choix opérés, car chaque modèle est façonné par les exigences du contexte. La priorisation repose sur un compromis : des modèles simples satisfont les besoins immédiats, tandis que des modèles plus complexes sont introduits progressivement, bien que la représentation « parfaite » de l'absentéisme reste hors de portée.

En résumé, la séquence de capitalisation des données RH fait apparaître deux controverses principales sur :

1. L'absence de convention de mesure pour l'absentéisme ;
2. Des modèles fondés sur l'expérience des ressources disponibles.

Ces deux controverses révèlent ainsi un fil conducteur : la tension entre ambition et faisabilité, où les arbitrages entre attentes élevées et contraintes pratiques sont indispensables pour maximiser l'efficience des projets de connaissances dans un contexte de ressources limitées.

6. Conséquences pour la fonction RH

La fragmentation de la fonction RH lors de la séquence de qualification est résolue dans cette séquence par l'enrôlement de Jean (partenaire RH). Jean, ancien DRH d'une banque et connaissance de Michel (directeur général de *QIA*), devait initialement devenir DRH du cabinet. Bien que cela n'ait pas abouti, sa présence dans le projet incarne, pour Xavier (directeur de projet), la science holistique de la GRH.

En plus de son expertise, Jean apporte également un réseau professionnel, pouvant enrichir la portée des projets de connaissances. Sa présence reste toutefois aléatoire :

« [Jean possède] *une vraie capacité d'analyse métier* [mais] *Il ne fait pas l'effort* […] *de rentrer dans le détail et de façon un peu régulière dans les analyses data qu'on fait* […] *s'il était plus présent, il pourrait apporter plus, c'est sûr* […] [Quand] *il nous donne quelques billes elles sont précieuses. Mais c'est plus aléatoire…* » (Xavier, directeur de projet, entretien no1, 31/10/2023).

Pour certains, cette présence, perçue comme sporadique et insaisissable, remet en question la capacité de Jean (partenaire RH) à faire valoir la perspective RH dans le développement des projets de connaissances :

« […] *c'est un fantôme pour moi* […] *Jean faisait un peu notre bêta- testeur, enfin, notre crash- test* […]. *Mais au global, on a présenté ce qu'on voulait…* » (Julie, cheffe de projet no2, entretien no1, 17/10/2023).

Bien que les raisons spécifiques de l'absence de Jean (partenaire RH) restent inconnues, un constat général se dégage : la construction des données RH n'est pas jugée comme prioritaire, soit par manque d'intérêt - et donc de temps -, soit par manque de compétences.

En résumé, les représentations des différents agents économiques mettent en évidence une fonction RH qui apparaît à la fois *fragmentée* et *fantomatique*. Toutefois, contrairement à la qualification, qui entraîne une marginalisation compensatoire de cette fonction, la séquence de capitalisation révèle une marginalisation volontaire de la dimension « RH » des données.

7. Conclusion

Ce chapitre se concentre sur la deuxième séquence du processus de construction des données RH : la Capitalisation. Celle-ci explore les projets qui s'appuient sur la qualité *normative* de ces données pour produire de nouvelles connaissances. L'objectif de cette séquence est de concevoir la fonction épistémique des données RH, permettant leur singularisation en tant que biens économiques.

Trois projets distincts de connaissances sur l'absentéisme sont développés, chacun caractérisé par un type spécifique de connaissances :

1. L'absentéisme *réel univarié* : apporte des connaissances *descriptives* ;
2. L'absentéisme *réel multivarié* : fournit des connaissances *explicatives* ;
3. L'absentéisme *latent* : génère des connaissances *prédictives*.

L'analyse de ces projets s'articule autour de trois dimensions clés :

1. Le territoire d'exploration épistémique des données RH ;
2. Le savoir-faire des agents économiques impliqués ;
3. Les épreuves d'exploration qui en découlent.

La progression séquentielle des projets de connaissances est conçue pour développer une fonction épistémique offrant une granularité analytique variable et complémentaire. Cette capacité à ajuster la granularité à différentes échelles – macroscopique, mésoscopique et microscopique - permet aux *data scientists* d'effectuer des analyses à divers niveaux de détails, alternant d'une perspective plus générale à une vue plus spécifique du phénomène. Cette approche incrémentale vise à enrichir les connaissances de l'absentéisme à travers une exploration adaptative des données RH.

Bien que la qualité *normative* de ces données facilite le développement d'un « zoom multi-échelle » à travers ces trois projets, elle confronte également les *data scientists* à de multiples épreuves d'exploration.

Deux controverses émergent de cette deuxième séquence, chacune faisant l'objet de négociations. Les compromis atteints sont les suivants :

1. Des conventions pragmatiques évolutives ;
2. Des modèles à complexité croissante.

Ces compromis incarnent une deuxième stratégie de marchandisation, centrée sur un ajustement progressif. Elle vise à concilier l'ambition de développer des projets de connaissances qui se distinguent sur le marché, tout en faisant face aux contraintes de faisabilité imposées par des ressources humaines, techniques et financières limitées. Cette stratégie forme le socle qui facilite le passage à la troisième séquence du processus de construction des données RH : la Requalification.

Chapitre 5. Séquence de Requalification des données RH

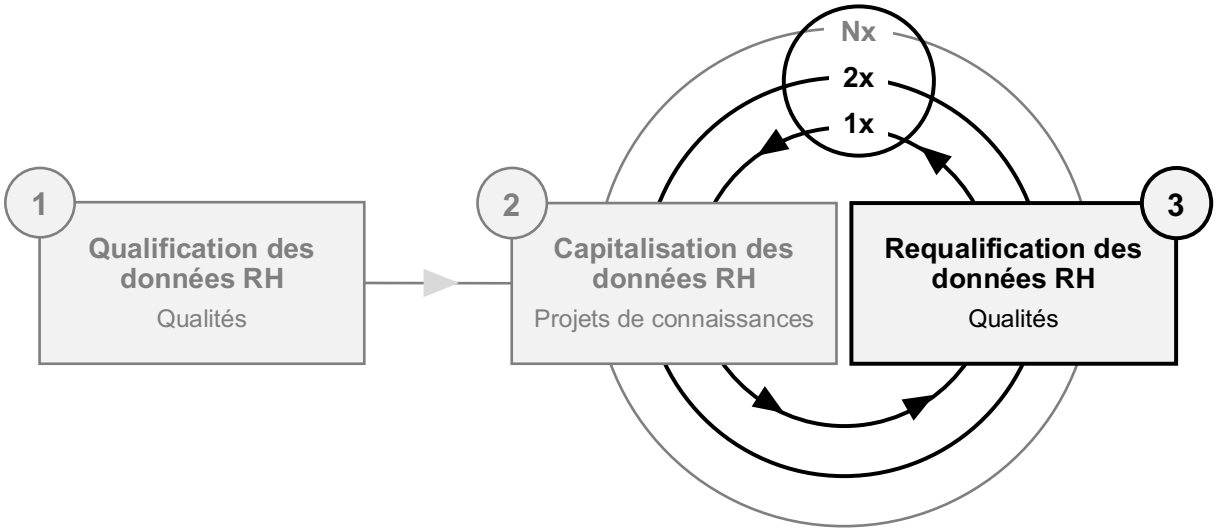


Figure 15 : Séquence de Requalification des données RH

1. Introduction

« *On a développé un produit* [*DSN Analytics], mais on ne l'avait jamais testé chez un client. On ne savait pas si ça allait marcher... Et surtout, on n'avait aucun retour du marché : est-ce qu'ils en ont vraiment besoin ? Est-ce qu'ils vont l'utiliser ? Parce que si c'est un truc récurrent, est-ce que ça va vraiment marcher ou pas ?* […] *Ce n'est pas évident que ça marche du premier coup. Donc le but ultime, c'était vraiment d'avoir le premier client.* » (Julie, chef de projet no2, entretien no1, 17/10/2023).

Le quatrième chapitre s'est penché sur les projets de connaissances résultant de la qualité *normative* des données RH. Cette qualité a permis le développement de trois types de connaissances distincts :

1. Les connaissances *descriptives* par le projet sur l'absentéisme *réel univarié* ;
2. Les connaissances *explicatives* par le projet sur l'absentéisme *réel multivarié* ;
3. Les connaissances *prédictives* par le projet sur l'absentéisme *latent*.

Cette approche incrémentale vise à enrichir la compréhension de l'absentéisme par une exploration adaptative et multi-échelle des données RH. Ce choix repose sur une stratégie de marchandisation qui privilégie l'ajustement progressif des projets de connaissances, en réponse aux contraintes de faisabilité dictées par des ressources limitées.

Ce cinquième chapitre se concentre sur la troisième séquence de construction des données RH : la Requalification. Elle consiste à requalifier la qualité *normative* des données RH à travers les projets issus de la capitalisation.

Cette séquence met en lumière la qualité *extensive* des données RH. Cette qualité se manifeste sous deux formes : *externe* et *interne*.

L'analyse des controverses, au cours desquelles les agents économiques négocient cette qualité *extensive*, est ensuite approfondie.

2. Qualité *extensive externe* des données RH

La requalification *extensive externe* des données RH entraîne un changement de perspective, passant d'un référentiel de comparaison interne des données RH au sein de l'entreprise à un référentiel de comparaison externe sectoriel. Ce déplacement de

la qualification des données RH nécessite non seulement l'accès à un large volume de données RH *normatives*, mais aussi le développement de nouvelles conventions de mesure sectorielles.

Pour la requalification des données RH, trois actes sont nécessaires :

1. Re-définir les besoins des clients ;
2. Re-rationaliser les coûts d'investissement ;
3. Ré-enrôler des agents économiques.

2.1. Acte I : re-définition des besoins des clients

Ce premier acte de requalification des données RH se concentre sur la précision des besoins des clients, en particulier leur nécessité de se situer par rapport à leur marché sectoriel. Pour Xavier (directeur de projet) et son équipe, cet impératif de comparaison externe peut conférer une valeur substantielle aux données RH, tout en érigeant une barrière à l'entrée pour leurs concurrents potentiels :

« *Si tu déroules un processus courant, je pense que tu commences justement par te comparer à l'extérieur : par rapport aux autres vous avez plus d'absentéisme. Et puis après, tu descends de plus en plus finement…* » (Luc, *data scientist* senior, notes prises lors d'une réunion, 15/12/2021).

Le premier projet de connaissances sur l'absentéisme *réel univarié* est d'ores et déjà confronté à une contrainte majeure : son exploration externe des données RH se base exclusivement sur des données publiques annuelles, données fournies par *Alior*, un cabinet de conseil international. Cette limitation entrave sa pertinence, car les conventions de mesure des taux d'absentéisme diffèrent entre celles souhaitées par les *data scientists* dans le projet et celles d'*Alior*. Notamment, l'inclusion des congés parentaux complique la comparaison des données RH entre elles puisqu'il s'agit de données que les *data scientists* veulent isoler dans l'analyse de l'absentéisme.

Pour ces derniers, un référentiel de comparaison sectorielle revêt un rôle majeur dans la compréhension de l'absentéisme conjoncturel en fournissant des repères externes aux entreprises. L'absentéisme conjoncturel, caractérisé par sa nature temporaire, découle généralement de facteurs externes tels que les conditions économiques, socio-politiques ou météorologiques. Ce référentiel permet à

l'entreprise de cibler son absentéisme en le contextualisant par rapport aux tendances observées dans son secteur grâce à des conventions de mesure établies :

« L'idée est d'arriver avec un truc clé en main, avec un modèle qui a été calibré sur tout le monde et sur un panel varié […] *la plus-value métier est assez claire, tu vas voir une entreprise et tu fais : voilà, vous êtes en dessous de la norme* [ou] *vous êtes au-dessus de la norme et pas uniquement en fonction du taux* [d'absentéisme] *général, mais en fonction de qui sont vos salariés*. » (Luc, *data scientist* senior, notes prises lors d'une réunion, 15/04/2022).

Ainsi, l'établissement de nouvelles conventions de mesure, visant à créer un nouveau référentiel de comparaison sectorielle externe, nécessite l'accès à un large volume de données *normative*s. En effet, pour que cette requalification soit effective, les données doivent être comparables, et la qualité *normative* impose cette exigence :

« […] *le benchmark, en fait, nécessite d'avoir beaucoup de données et qu'elles soient toutes traitées de la même façon. Qui dit benchmark dit forcément outil standard. »* (Xavier, directeur de projet, entretien no2, 21/11/2023).

2.2. Acte II : re-rationalisation des coûts d'investissement

Pour accroître les opportunités de positionner les données RH en tant que biens sur un marché RH saturé mais déficient, Xavier (directeur de projet) s'engage dans l'élaboration d'une stratégie en deux volets, en développant deux modèles économiques :

1. Un modèle *freemium* ;
2. Un modèle *premium*.

« […] *une sorte de service freemium, c'est-à-dire je fais le benchmark standard gratuitement. Et pour les entreprises qui veulent s'améliorer, elles peuvent accéder à une version sur mesure* [premium] *qui permet d'assigner des analyses. »* (Xavier, directeur de projet, entretien no2, 21/11/2023).

Le Tableau 23 détaille le contenu de chacun des modèles économiques proposés par Xavier (directeur de projet) et son équipe.

Tableau 23 : Modèles économiques de requalification *extensive externe* des données RH : *freemium* et *premium* (issus d'une présentation, 05/04/2022)

<table>
<tr><th>Modèles économiques</th><th>Modèle freemium</th><th>Modèle premium</th></tr>
<tr><td>Description</td><td>Mise à disposition de benchmarks pour toutes les entreprises.</td><td>Analyse détaillée pour une entreprise qui souhaite comprendre ses spécificités.</td></tr>
<tr><td>Formes de connaissances</td><td>Descriptives</td><td>Descriptives, explicatives et prédictives</td></tr>
<tr><td>Financement</td><td>Par QIA et une « tierce partie »</td><td>Par une entreprise</td></tr>
<tr><td rowspan="3">Projets de connaissances des données RH</td><td>Accès au projet sur l'absentéisme réel univarié.</td><td>Accès aux trois projets de connaissances sur l'absentéisme réel univarié, multivarié et latent.</td></tr>
<tr><td>Additionnel : taux d'absentéisme par métier</td><td rowspan="2">Additionnel : comparaison des données RH avec des éléments de benchmark sectoriel</td></tr>
<tr><td>Additionnel : filtres possibles des données RH selon les caractéristiques des entreprises (secteur, taille, etc.)</td></tr>
</table>

Ces deux modèles économiques intègrent les trois projets de connaissances préalablement développés dans la séquence de capitalisation. Le modèle *freemium* s'inscrit comme une extension du premier projet de connaissances, à savoir l'absentéisme *réel univarié*, en proposant un service de base permettant une comparaison générale et le suivi des tendances de l'absentéisme. En revanche, le modèle *premium* offre une analyse personnalisée et approfondie, incluant l'ensemble des trois projets de connaissances : (1) l'absentéisme *réel univarié, (2)* l'absentéisme *réel multivarié* et (3) *latent*.

Dans le cadre de la mise en œuvre du modèle *freemium*, l'échange de données RH contre des analyses gratuites constitue la pierre angulaire de la stratégie de Xavier (directeur de projet). Cette approche vise à ériger une barrière à l'entrée pour les

concurrents potentiels et engager les futurs clients en leur offrant des services sans coût initial.

Deux avantages découlent de ce modèle :

1. Une facilité d'accès aux services : constitue un avantage pour Xavier (directeur de projet) lui permettant de toucher un large éventail de clients potentiels. Les services initiaux gratuits réduisent les obstacles à l'engagement en offrant aux clients la possibilité d'évaluer la valeur des données avant de décider d'investir dans des technologies plus coûteuses. En offrant un accès aux connaissances *descriptives* de l'absentéisme, Xavier (directeur de projet) et son équipe pourraient attirer des clients qui n'auraient pas envisagé ces services autrement. Cette approche leur permettrait d'élargir leur marché cible et d'accroître leur portée.
2. Un incitatif pour la promotion de ventes additionnelles : désigne la stratégie consistant à proposer initialement un service de base, puis encourager les clients à opter pour des versions plus avancées. Dans ce contexte, après avoir expérimenté et constaté la valeur du service de base, les clients sont plus susceptibles de percevoir la valeur ajoutée des services payants, qui sont plus approfondis et personnalisés. Ainsi, l'offre *freemium* sert de levier pour démontrer la valeur des connaissances produites, créant une opportunité pour Xavier (directeur de projet) de vendre le modèle *premium*, augmentant à la fois ses revenus et la fidélisation des clients.

Le modèle *freemium* remplit ici une double fonction : d'une part, il permet aux entreprises de bénéficier d'un service initial gratuit, facilitant ainsi leur appréciation de la valeur des connaissances produites sans encourir de coûts. D'autre part, ce modèle accorde aux *data scientists* un accès élargi aux données RH, leur offrant ainsi une plus grande latitude dans le procesus de construction de données RH.

Cette latitude reflète la volonté sous-jacente des *data scientists* d'acquérir une autonomie dans l'établissement des conventions de mesure. Cette autonomie leur permettrait de traiter efficacement une large gamme de cas spécifiques relatifs aux données RH, augmentant ainsi la robustesse de leurs projets de connaissances :

« On faisait le bench[mark] *gratuit parce qu'on n'avait en vrai jamais utilisé de données DSN. Et du coup, c'était l'occasion quand même de saisir de vraies données* […] *On voulait avoir le plus de DSN possible pour traiter le plus de cas particuliers*

possible et pouvoir avoir un outil qui soit robuste. » (Julie, cheffe de projet, entretien no1, 17/10/2023).

Le modèle *freemium* sert ainsi de fondation pour établir la crédibilité des *data scientists* sur le marché RH, une crédibilité renforcée par la robustesse des données qu'ils construisent. Cependant, cette même fondation révèle une tension inhérente : bien que les *data scientists* recherchent l'autonomie dans la construction des données RH, cette autonomie leur confère également un contrôle sur ce qui compte et ce qui ne compte pas en termes de connaissances sur l'absentéisme. Cette quête de contrôle se manifeste particulièrement dans leur aspiration à accéder à un large volume de données RH, un atout essentiel dans leur ambition de dominer le marché :

« Être capable de calculer nous-mêmes le taux d'absentéisme sur tous les secteurs, sur tous les métiers, sur les différentes tailles d'entreprises... » (Tristan, chef de projet, notes prises lors d'une réunion, 15/12/2021).

Étant donné que le modèle freemium implique une structure de cofinancement avec une tierce partie agissant comme intermédiaire entre les entreprises et les *data scientists*, il est essentiel de déterminer comment cette entité intermédiaire gère et détient les données RH collectées auprès des entreprises participantes. Cependant, une question demeure : comment la gestion et l'accès à ces données sont-ils effectivement mis en œuvre ?

2.3. Acte III : ré-enrôlement des agents économiques

Dans le cadre de la requalification *extensive externe*, aux trois stratégies d'enrôlement initiales (l'intégration académique, la mobilisation « en temps masqué » et l'apport d'affaires) s'ajoute la stratégie de cofinancement, qui se concentre sur l'enrôlement d'une tierce partie. Cette entité est essentielle pour garantir aux *data scientists* l'accès aux données RH.

La réussite de cette requalification dépend fortement de l'exploitation des réseaux d'alliances de proximité pour attirer de nouveaux agents économiques. Le recours aux contacts professionnels est un point de départ privilégié pour Xavier (directeur de projet) :

« […] *le truc dans la démarche commerciale, c'est vraiment de commencer par le carnet d'adresses. Donc, c'est le mien, c'est celui de QIA…* » (Xavier, directeur de projet, entretien no2, 21/11/2023).

Cette pratique permet non seulement une réduction des délais de négociation et des risques associés à l'enrôlement de nouveaux agents, mais peut également ouvrir la voie à de nouvelles opportunités sans coût d'investissement supplémentaire. Ainsi, l'enrôlement de Jean (partenaire RH), lors de la séquence de capitalisation, a non seulement permis l'accès à un nouveau réseau mais a aussi catalysé une collaboration fructueuse :

« *Jean* […] *a amené à Sandrine. Et* […] *la personne avec qui ça a marché le mieux, c'est Sandrine en apport d'affaires.* » (Xavier, directeur de projet, entretien no1, 31/10/2023).

Sandrine est une experte en droit social international, avec une spécialisation en gestion des droits humains et restructurations, ainsi qu'en gestion des risques liés aux données nominatives. Forte d'une expérience approfondie en mobilités internationales, elle opère également dans le dialogue social, notamment dans la négociation d'accords transnationaux. Bien qu'elle ne possède pas de compétences spécifiques dans le travail des données, elle manifeste un intérêt marqué pour approfondir ses connaissances dans ce domaine :

« […] *j'ai rencontré Jean* (partenaire RH) et […] *parce que lui, c'est un pur RH, il m'a dit mais il je faut que je te fasse connaître à l'équipe RH, parce qu'il y a une partie réglementaire sur l'absentéisme, il n'y a pas que la partie RH* […]. *J'ai été vraiment incroyablement séduite par le produit parce que moi ça m'ouvrait des champs qui était complètement nouveau pour moi, ça me donnait une seconde jeunesse si je puis dire. J'ai tout de suite vu finalement l'intérêt de matcher mes compétences juridiques avec ma pratique, puisque moi j'ai vécu en entreprise, autant au sein de directions juridiques que directions RH* […] *Mais bon mes clients principaux c'étaient quand même les RH…* » (Sandrine, partenaire en droit social, entretien no1, 21/11/2023).

Grâce à un réseau d'alliances renforcé par une présence RH plus affirmée, Xavier (directeur de projet), formule deux hypothèses concernant le choix des agents économiques à enrôler dans cette requalification extensive *externe* :

1. Un organisme gestionnaire des données RH ;
2. Un groupe de réflexion spécialisé dans la protection sociale.

2.3.1. Hypothèse I : ré-enrôler un organisme gestionnaire des données RH

Les relations professionnelles préexistantes de Justine (directrice de projets stratégiques) avec la *CNPR* facilitent l'intégration de Xavier (directeur de projet) et son équipe dans leur « groupe de travail data ». Ce groupe de travail (GT) - dans lequel les *data scientists* animent un atelier d'émergence dans l'utilisation des données *DSN* - sert de tremplin pour susciter l'intérêt autour de la requalification *extensive externe* des données RH :

Charles (*CNPR*) : « *Bonjour Xavier, j'ai sans doute une opportunité à t'offrir pour un pitch autour de l'offre DSN de QIA auprès de responsables data et d'experts métier DSN.* » (Courriel envoyé le 14/10/2021).

Xavier (directeur de projet) : « […] *Il a enfin répondu à mes relances* 😀 *Ils doivent être affûtés sur la DSN. On se fera une réunion de préparation. J'imagine qu'il faut mettre en avant des capacités en audit / contrôle des données DSN en plus des analyses fonctionnelles...* » (Courriel envoyé à l'équipe interne de *QIA* le 14/10/2021).

Pour enrôler le *CNPR* dans leur réseau d'alliances, les *data scientists* doivent ainsi démontrer leur savoir-faire dans l'audit et le contrôle des données *DSN* afin de renforcer leur crédibilité auprès de ce nouvel agent économique. L'accent de cette initiative ne réside donc pas sur l'analyse de l'absentéisme en lui-même, mais plutôt sur la démonstration des compétences des *data scientists* dans le traitement des données RH. Cela a par conséquent pour effet de reléguer le phénomène de l'absentéisme au second plan :

« […] *Travailler avec la CNPR, c'est quand même un facteur de sérieux sur le marché* […] *Ils disposent de 1,5 million d'entreprises qui envoient leur DSN chaque mois. C'est-à-dire ils recueillent énormément de DSN. Aujourd'hui, c'est vu comme un outil de simple reporting* [données DSN]*, y compris au CNPR, on ne se rend pas compte de la richesse opérationnelle… »* (Xavier, directeur de projet, notes prises lors d'une réunion, 15/04/2022).

Le besoin d'accès à un large volume de données RH pour la requalification *extensive* conduit ainsi les *data scientists* à formuler « *deux opportunités de valorisation des données DSN »* (extrait issu de la présentation faite au *CNPR*, 18/03/2022). Ces opportunités, proposées au *CNPR,* visent non seulement à valoriser leurs services auprès de l'organisme, mais également à établir un partenariat par le biais de son enrôlement, notamment grâce à la stratégie de cofinancement. Celles-ci sont présentées dans le Tableau 24.

Malgré cette proposition : « […] *ça n'a pas été possible…* » (Xavier, directeur de projet, entretien no2, 21/11/2023).

Une explication potentielle de cet arrêt réside dans le caractère exploratoire et non prioritaire du GT data. En effet, ce groupe s'est formé sur la base d'un constat interne à la *CNPR* : « *la nécessité de transversalité dans les usages et l'exploitation des données des différentes directions* » (issu du compte-rendu de l'atelier d'émergence du GT data, 31/03/2022). Ainsi, la logique flexible et multidisciplinaire de la *data science*, ainsi que la nécessité d'un travail exploratoire pour comprendre ce nouveau domaine, ont probablement contribué à sa faible priorité.

Tableau 24 : Opportunités de valorisation des données *DSN* pour le *CNPR* (issues d'une présentation, 18/03/2022)

Opportunités	Critères	Détails
Mise à disposition d'un baromètre annuel de l'absentéisme (Cf. modèle *freemium*)	Bénéfices	- Valoriser le positionnement du *CNPR* en tant qu'acteur apportant des services à valeur ajoutée aux entreprises. - Granularité de l'information inégalée : par sous-secteur d'activité, par métier, etc. - Possibilité d'analyser le turnover de l'entreprise, notamment par métier et par entité, avec un croisement possible avec les données d'absentéisme.
	Modalités possibles	- Rapport statique non spécifique à une entreprise. - Réalisation à financer par le *CNPR.*

Opportunités	Critères	Détails
Mise à disposition de la plateforme *DSN Analytics* à vos entreprises (Cf. modèle *premium*)	Bénéfices	- Analyse détaillée et comparée de l'absentéisme d'une entreprise en particulier au regard de son secteur, de sa géographie et de sa population de salariés de façon à isoler les phénomènes propres à l'entreprise. - Comparaison des métriques de l'entreprise à des métriques de groupes d'entreprises comparables.
	Modalités possibles	- Diffusion par *CNPR* de la plateforme *DSN Analytics* en marque blanche contre rémunération. - Mise à disposition des données *DSN* par le *CNPR* pour diffusion par *QIA* auprès des entreprises suivant des conditions préférentielles à définir.

2.3.2. Hypothèse II : ré-enrôler un groupe de réflexion spécialisé dans la protection sociale

L'enrôlement de Sandrine (partenaire en droit social) dans le réseau d'alliances, offre notamment aux *data scientists* un accès précieux à son carnet d'adresses. Plus spécifiquement, sa collaboration antérieure avec l'*IISS* lui permet de jouer un rôle d'intermédiaire dans les discussions portant sur un éventuel partenariat. Ce partenariat, également axé sur l'apport d'affaires, a pour objectif d'accélérer la requalification *extensive externe* des données RH. Pour ce faire, il vise à établir des liens avec des responsables du secteur de la santé, en vue de trouver une tierce partie appropriée :

« *L'ISS si tu veux, a un ancrage énorme dans le monde de la santé, donc ce que j'ai vu, c'est un carnet d'adresse formidable, tu vois.* […] *je me disais pourquoi QIA ne se ferait pas connaître ? Moi, je pensais que QIA avait un problématique de visibilité. En fait, ce que j'avais pressenti, c'est que l'outil était formidable, mais il souffrait d'une image*. » (Sandrine, partenaire en droit du travail, entretien no1, 21/11/2023).

La proposition d'apport d'affaires soumise à l'*IISS* est la suivante : « *par l'association IISS-QIA, le développement de benchmark offrira à chaque entreprise un*

positionnement stratégique face à son secteur et lui permettra d'enclencher des analyses spécifiques si elle le souhaite. » (Extrait issu de la présentation faite à l'*IISS*, 05/04/2022). Ses modalités sont présentées dans le Tableau 25 ci-dessous.

Tableau 25 : Proposition d'apport d'affaires soumise à l'*IISS* (issue d'une présentation, 05/04/2022)

Objectifs	Bénéfices
Collaboration à des publications	- Notre benchmark favorisera le développement de travaux de R&D appliquée au phénomène de l'absentéisme. - *QIA* allie chercheurs, consultants, partenaires privés et publics à la production de connaissances actionnables. - La collaboration pourra prendre la forme de publications, livres blancs ou autres.
Accès aux résultats	- Dans l'optique d'une pérennisation du benchmark, il sera proposé sous la forme d'un modèle *freemium*. - Certaines fonctionnalités seront en libre accès, d'autres seront proposées sous la forme d'un abonnement *premium* à un tarif préférentiel. - En fonction des besoins, il sera possible de réaliser des analyses à différents degrés de granularité pour une compréhension plus approfondie du phénomène de l'absentéisme.

Alors que le modèle *freemium* est initialement bâti sur les intérêts technico-économiques des *data scientists*, l'enrôlement de ce nouvel agent économique introduit de nouveaux intérêts qui déplacent la requalification *extensive externe* des données RH. Ces nouveaux intérêts politico-médiatiques visent à renforcer la visibilité publique de *QIA*. Cette perspective est notamment mise en lumière lors d'une conversation avec Fabrice (*IISS*) lors d'une réunion (23/03/2022) :

- Fabrice (*IISS*) : « *Il semblerait que notre participation soit plus pertinente dans le cadre du modèle freemium, qui permettrait une réflexion d'envergure nationale. C'est particulièrement vrai dans le secteur hospitalier, où l'absentéisme est non*

seulement un enjeu RH mais aussi un sujet à résonance médiatique et politique. Il faut jouer la corde sensible et l'enrober dans du velours… »

- Fabrice (*IISS*) : « *Quelle est la réelle valeur ajoutée de ce baromètre ? Il faut maintenir une perspective nationale sans entrer dans des comparaisons qui pourraient alarmer les directeurs d'hôpitaux. Lorsqu'ils sont nommés par le conseil des ministres, ils préfèrent cachés leurs problèmes : ils ne veulent pas que l'on parle d'eux*. »
- Fabrice (*IISS*) : « *L'IISS peut vous soutenir en vous aidant à envelopper votre présentation dans une format plus « politique ». Il faut différencier le benchmark* [modèle *freemium*] *de l'outil* [modèle *premium*]*. Le baromètre* [de l'absentéisme] *doit être politique.* »
- Xavier (directeur de projet) : « *Peu importe l'acteur* [organisme destinataire de la *DSN* ou acteur sectoriel] *celui qui est intéressé est le premier qui l'emporte.* » (Notes prise lors d'une réunion, 23/03/2023).

La stratégie d'apports d'affaires est fondée sur le principe de réciprocité, où l'offre d'un service est conditionnée par l'obtention d'une opportunité en retour. Bien que ce nouveau partenariat introduise des intérêts politico-médiatiques pour les données RH, l'intérêt sous-jacent à la requalification *extensive externe* de ces dernières reste fondamentalement économique :

« *Alors après, pourquoi ça n'a pas abouti ? Parce qu'évidemment, l'ISS, il faut être membre et membre, ça veut dire qu'il faut donner une obole qui n'est pas qu'une obole, d'ailleurs, je crois que c'était 10 000 € qui leur demandait, et la moitié pour une demi-année, et bien il fallait les sortir.* […] *C'était une question financière avant tout pour un rapport sur investissement immédiat pas perceptible...* » (Sandrine, partenaire en droit social, entretien no1, 21/10/2023).

La requalification *extensive externe* des données RH marque une transition du référentiel de comparaison, évoluant d'un cadre interne propre à l'entreprise vers un cadre externe sectoriel. Cette transition s'organise autour de trois actes :

1. La re-définition des besoins des clients ;
2. La re-rationalisation des coûts d'investissement ;
3. Le ré-enrôlement des agents économiques ;

Pour les *data scientists*, il est impératif d'établir un référentiel sectoriel qui permette aux clients potentiels de situer leur gestion de l'absentéisme par rapport à leurs concurrents. Cette nécessité dicte l'établissement de nouvelles conventions de mesure qui dépassent le périmiètre initial de l'absentéisme r*éel univarié*, pour mieux comprendre et surtout distinguer les fluctuations structurelles et conjoncturelles de l'absentéisme.

Pour faciliter cette transition, les *data scientists* développent une stratégie bifocale, englobant le développement de deux modèles économiques :

1. Un modèle *freemium* ;
2. Un modèle *premium*.

Contrairement au modèle *premium* qui intègre les acquis des trois projets antérieurs de capitalisation, le modèle *freemium* est spécifiquement conçu pour cette séquence dans l'objectif d'engager les clients en offrant initialement des services gratuits en échange d'un accès aux données.

En complément, l'introduction d'une stratégie de cofinancement pour l'enrôlement d'une entité neutre vient consolider ce premier type de requalification. Cette démarche vise l'enrôlement d'organismes spécialisés, tels qu'un gestionnaire des données RH ou un groupe de réflexion sur la protection sociale, pour assurer un accès continu aux données RH nécessaires.

En définitive, la motivation profonde derrière la requalification *extensive externe* des données RH reste de nature économique. Les efforts déployés visent à instituer de nouvelles conventions de mesure qui exige l'accès à un volume substantiel de données *normatives*.

3. Qualité *extensive interne* des données RH

La requalification *extensive interne* des données RH induit également un changement de perspective, passant de la combinaison d'un même type de données RH (*normatives* dans le cas de la requalification *extensive externe*) à une combinaison avec d'autres types de données. Ce déplacement dans la requalification des données RH exige, quant à lui, un accès à une diversité de sources de données.

Ce déplacement s'organise autour des trois mêmes actes de requalification :

1. Re-définir les besoins des clients ;
2. Re-rationaliser les coûts d'investissement ;
3. Ré-enrôlement des agents économiques.

3.1. Acte I : re-définition des besoins des clients

Bien que les données *DSN* demeurent le socle de la requalification *extensive interne*, Xavier (directeur de projet) et son équipe envisagent également l'intégration de données externes à la *DSN* afin de répondre plus précisément aux besoins des clients potentiels :

« […] *On vous propose exactement ce dont vous avez besoin et pas une machine de guerre pour tuer une mouche… »* (Julie, cheffe de projet, entretien no1, 17/10/2023).

Une attention particulière est portée sur l'investigation de sources de données complémentaires susceptibles d'enrichir les données RH *normatives* et ainsi dépasser le cadre rigide des données issues de la *DSN* :

« […] *Le sur-mesure va aller* [par exemple] *prendre des pointages dans des outils de production et toutes les affectations. La différence entre les deux* [données de pointage et données *DSN*]*, c'est que quand on utilise uniquement la DSN, on va avoir* […]*, des motifs d'absence qui vont être agrégés dans la DSN. Alors que des fois, on veut pouvoir sélectionner plus finement les motifs d'absence qu'on veut prendre ou pas : les congés maternité, les [autres types de] congés, les grèves, etc. Donc ça, ça peut fausser. Et ensuite, on est sur un taux d'absentéisme qui est forcément un petit peu théorique, puisqu'on ne va pas savoir pour chaque personne si elle travaille le samedi, le dimanche… La vérité des jours travaillés, en fait. Donc pour ça, il faut prendre des plannings de prod*[uction]. » (Xavier, directeur de projet, entretien no2, 21/11/2023).

Tout comme l'intégration des données de production dans la requalification *extensive interne*, d'autres données supplémentaires émergent, résultant de discussions avec un large éventail d'agents économiques. Ces échanges ont lieu soit par l'intermédiaire d'alliés, notamment Sandrine (partenaire en droit social), soit dans le cadre de tentatives d'enrôlement de nouveaux agents.

Alors que l'intégration de ces nouvelles données demeure toujours à l'état de chimère, elle sert toutefois de prétexte à Xavier (directeur de projet) et Julie (cheffe de projet) pour engager plus facilement le dialogue avec les clients potentiels. En effet, cette stratégie vise à adopter une approche moins prescriptive quant à la définition des besoins. Les données supplémentaires sont présentées dans le Tableau 26.

Tableau 26 : Données complémentaires optionnelles proposées dans la requalification *extensive interne* des données RH (issue d'une présentation, 24/06/2022)

Données complémentaires optionnelles	**Descriptions**
Suivi des actions	Analyse de l'impact des actions mises en place sur le taux d'absentéisme.
Facteurs explicatifs	Analyse approfondie des variables internes expliquant le taux d'absentéisme.
Taux de rotation	Corrélations observées entre l'absentéisme et le taux de rotation des salariés.
Localisation et transport	Impact du temps de transport et du télétravail sur le taux d'absentéisme.
Aide à la planification	Ajustement de la planification en fonction de l'absentéisme prévisionnel.
Reportings personnalisés	PDF téléchargeables mensuellement pour suivre l'évolution des KPIs.

La requalification *extensive interne* est ainsi développée dans le but d'enrichir le : « *storytelling avec de la crème autour* [pour] *se différencier et donner de la valeur* [aux données RH] … » (Pierre, fondateur, courriel envoyé le 30/11/ 2021).

Cependant, la persistance de Xavier (directeur de projet) et son équipe à se focaliser essentiellement sur les données socio-démographiques issues de la *DSN* réduit de manière significative leur possibilités d'interagir avec certains clients, notamment ceux issus du secteur public. La limitation de leur historique des données

DSN à deux ans[22] exige un renforcement des investissements dans le développement de sources de données alternatives.

L'exemple d'un appel d'offres, destiné à une collectivité de près de 2000 agents, illustre les défis auxquels les *data scientists* sont confrontés en raison de leur dépendance exclusive aux données *DSN*. Cette dépendance engendre une rigidité dans l'approche adoptée pour analyser l'absentéisme :

« C'est très tourné *DSN* alors que l'investissement d'analyse se fera hors *DSN* mais avec requête SIRH… ce n'est pas très clair ce que l'on fait… » (Courriel envoyé par Benoît, partenaire en actuariat, le 17/02/2022).

Cette approche affecte négativement la stratégie de ciblage des secteurs à forte pénibilité. Cela réduit le champ d'action des *data scientists*, notamment dans le secteur de la santé (cf. les hôpitaux), considéré dès la séquence de qualification comme un point d'entrée stratégique. Face à ce constat, un changement de stratégie s'observe chez Xavier (directeur de projet), non pas en termes de source de données principale - qui demeure la *DSN* - mais de cibles choisies :

« […] Nous venons seulement de comprendre que la DSN ne fonctionne pas avec les hôpitaux publics. Pour l'instant, nous allons seulement orienter nos efforts vers les hôpitaux privés. » (Notes issues du journal de bord, 17/05/2022).

Cette persistance, bien qu'avantageuse à court terme pour évaluer le coût d'entrée sur le marché RH peut, à plus long terme, potentiellement entraver le développement de nouvelles opportunités commerciales.

3.2. Acte II : re-rationalisation des coûts d'investissement

Les intérêts technico-économiques des *data scientists* en faveur des données *DSN* repose sur leur qualité *normative*. En effet, dans le cadre de la requalification *extensive*

[22] Depuis 2017, la *DSN* est obligatoire pour les entreprises du secteur privé et cette obligation s'étend désormais aux entreprises du secteur public depuis janvier 2022. Toutefois, l'exploitation des données entre 2020 et 2021 est compliquée par les perturbations dues à la pandémie de COVID-19, rendant actuellement impossible l'obtention de résultats fiables pour le secteur public.

interne des données RH, la diversification des sources de données entrave leur ambition de concevoir un outil « *plug-and-play* » :

« *En fait, DSN égal partout pareil, égal je peux faire un produit. À partir du moment où chaque SIRH n'a pas les mêmes informations, pas structurées de la même façon, tu ne peux pas faire un produit plug-and-play parce que tu as forcément une phase de : je restructure mes données.* » (Julie, cheffe de projet, entretien no1, 17/10/2023).

Chaque adaptation devient ainsi une opération spécifique, exigeant une planification, une gestion et une allocation des ressources. Chaque client, avec son propre SIRH ou ses structures de données distinctes, demande un investissement additionnel pour l'adaptation et l'intégration de ses données. Par conséquent, la variabilité des systèmes et des données introduit des complexités supplémentaires et entraîne des coûts additionnels pour les clients, rendant la requalification *extensive interne* des données RH moins accessible financièrement et donc potentiellement moins attractive sur le marché. Ainsi pour ne pas exposer *QIA* à un risque financier, Xavier (directeur de projet) et son équipe s'attachent à préserver les données *DSN* comme élément central de la requalification des données RH, en restreignant leurs ajustements aux dimensions périphériques :

« Ça restera le cœur. Déjà 80 % des informations qui ont de la valeur, elles sont dans la DSN. Ensuite, 90% de la valeur est extraite de la DSN parce que la DSN est de bonne qualité. Tu sais, tu ne pourras pas savoir ce que tu vas récupérer par ailleurs, mais ce sera souvent assez foireux. Et ensuite, l'effort pour récupérer et mettre en qualité les autres données sera très important. Donc en fait la valeur marginale qui est créée là-dessus sera forcément plus faible. Après tout, ce raisonnement évidemment n'est valable que quand on a la DSN… » (Tristan, chef de projet, entretien no1, 15/12/2021).

Lorsque Tristan (chef de projet) évalue l'effort nécessaire par rapport à la valeur ajoutée des autres sources de données, il estime que les investissements supplémentaires ne sont pas justifiés par les gains marginaux obtenus. Ainsi, cette approche pragmatique, caractérisée par une prudence face aux risques financiers, influence de manière significative la construction des données RH en tant que biens économiques.

3.3. Acte III : ré-enrôlement des agents économiques

Le réseau d'alliances est considérablement renforcé par l'enrôlement de Jean (partenaire RH) et Sandrine (partenaire en droit social) qui agissent en tant qu'intermédiaires en utilisant leur carnet d'adresses. Ces partenaires jouent un rôle crucial dans le projet en permettant aux *data scientists* d'accéder au marché RH. Jean (partenaire RH) conçoit cette démarche de mise en réseaux comme : « *un travail empirique de lobbying visant à exercer une influence sur les influenceurs potentiels pour créer une sorte de bruit de fond.* » (Notes prises lors d'une réunion, 20/12/2021).

Ce « lobby de fond » est également souligné par Sandrine (partenaire en droit social), qui travaille avec ténacité à l'enrôlement de nouveaux agents économiques :

« […] *Je comprends qu'une boîte qui est petite, elle a besoin d'un retour sur investissement rapidement. C'est pour ça que je me suis mise en quatre pour trouver tout un tas de personnes qui puissent rapidement rapporter des contrats tout en sachant que le produit n'était pas mûr. Il n'était pas fini. On continuait à l'améliorer* [et] *on n'avait pas non plus pléthore d'expériences sur le sujet* [de l'absentéisme] *...* » (Sandrine, partenaire en droit social, entretien no1, 21/11/2023).

C'est donc majoritairement par l'intermédiaire de Sandrine (partenaire en droit social) que les *data scientists* accèdent à différents rendez-vous, dont beaucoup proviennent du secteur de la santé (e.g. entreprises privées) :

« Je me suis dit : je pense qu'il faut d'abord cibler les boîtes qui ont un vrai problème d'absentéisme et ayant un fils interne qui galère dans les hôpitaux, il n'arrête pas. Il est même brancardier à ses heures tellement il y a une pénurie due à l'absentéisme dans les hôpitaux, c'est criant. Pareil pour les Ehpads parce que malheureusement, vu mon âge, je suis confrontée directement au grand âge… » (Sandrine, partenaire RH, entretien no1, 21/11/2023).

L'intérêt pour ces secteurs découlent d'une crise manifeste, perçue comme incitant leurs entreprises à des investissements substantiels dans l'objectif de résoudre la problématique de l'absentéisme. En effet, sans une problématique clairement définie, la pertinence d'une requalification *extensive interne* des données RH risque de diminuer :

« […] *secteur en crise égal potentiellement beaucoup d'argent à mettre dans de ce domaine d'activité […] Il faut, de toute façon s'attaquer à des secteurs où il y a des problèmes d'absentéisme. Si tu vas dans un secteur où il n'y a pas de problème d'absentéisme, ils te vont dire : qu'est-ce que tu veux que j'en fasse ?* » (Julie, cheffe de projet, entretien no1, 17/10/2023).

Malgré un accueil favorable mais réservé lors des différentes présentations, la progression subséquente des rendez-vous reste cependant marquée par une indifférence générale de la part des clients potentiels :

« Je pense que les interlocuteurs qu'on avait étaient très métiers et pas data... » (Julie, cheffe de projet, entretien no1, 17/10/2023).

Ainsi, Julie (cheffe de projet) attribue l'indifférence des interlocuteurs à une orientation prédominante vers un savoir-faire RH plutôt qu'un savoir-faire technique. Cette disparité entre les deux domaines de compétences entrave la pleine reconnaissance de la valeur des données RH, telle que considérée par les *data scientists*, pour l'analyse de l'absentéisme :

« *J'ai remis un coup de manivelle* [pour les Ephads] *en septembre dernier parce que je suivais ce qu'ils faisaient et qu'ils ont négocié un accord sur l'absentéisme. C'est te dire que c'est vraiment un sujet chez eux. Mais ils gèrent ça en famille et ils pensent qu'ils vont y arriver…* […] *il faut vendre du rêve*. […] *C'est de la séduction qu'il faut faire. Alors je ne dis pas qu'il faut être des bonimenteurs, mais c'est presque ça. Aujourd'hui je pense que ce produit doit être packagé parce que si on en reste sur une approche technique, ils en perçoivent pas l'intérêt, ça les attire pas plus que ça* […] [il faut leur montrer] *que ça va* [leur] *amener quelque chose de révolutionnaire. Et ça, on sait pas le faire.* » (Sandrine, partenaire en droit social, entretien no1, 21/11/2023).

Ainsi, la difficulté à lier les deux domaines de compétences peut potentiellement s'expliquer par la manière dont les données RH sont présentées. En effet, les interlocuteurs RH ont du mal à saisir les qualités techniques des données RH si celles-ci ne sont pas « emballées » de manière attrayante pour montrer leur potentialité « révolutionnaire ». Les données RH doivent donc être présentées de manière à créer un effet « wow » :

« [Mais] *qu'est-ce qu'ils appellent un effet « wow » ? Si un effet « wow », c'est : j'en ai absolument besoin maintenant et c'est urgent.* […] *il faut que tu comprennes que la data va t'aider à faire des choses. Mais c'est surtout à toi de te saisir du sujet. Ce n'est qu'une aide en fait…* » (Xavier, directeur de projet, entretien no2, 21/11/2023).

Il y a donc une tension qui s'installe entre deux approches : d'une part, créer un effet « wow » à travers les données RH pour convaincre les clients potentiels de l'importance de l'absentéisme ; et d'autre part, reconnaître que cet effet « wow » devrait plutôt émaner de la prise de conscience de l'importance de ce sujet par les clients eux-mêmes.

La requalification *extensive interne* des données RH induit une évolution de perspective, marquant la transition de l'utilisation exclusive de données RH *normative*s, propre à la requalification *extensive externe*, vers l'intégration de sources variées. Cette transformation nécessite un accès élargi à diverses sources de données et s'articule autour de trois actes :

1. Re-définir les besoins des clients ;
2. Re-rationaliser les coûts d'investissement ;
3. Ré-enrôlement des agents économiques.

Quoique que les données issues de la *DSN* demeurent centrales, les *data scientists* explorent aussi l'intégration d'autres données pour mieux cerner les attentes des clients potentiels. Cette intégration, encore en phase exploratoire, permet d'adopter une démarche moins directive et plus consultative dans la définition des besoins des clients. Cependant, la focalisation prédominante sur les données RH *normative* restreint leurs interactions avec certains clients, notamment ceux du secteur public.

Les *data scientists*, en évaluant l'effort requis par rapport à la valeur ajoutée des sources alternatives, concluent que les investissements supplémentaires ne se justifient pas par les bénéfices marginaux, reflétant ainsi une approche pragmatique et une aversion au risque économique.

Ainsi, pour renforcer leur réseau d'alliances, ils mobilisent donc les réseaux professionnels de deux partenaires, Sandrine (partenaire en droit du travail) et Jean

(partenaire RH), qui facilitent l'accès au marché. Jean qualifie cette mise en réseau comme un « effort empirique de lobbying ».

Bien que les premières interactions soient favorablement accueillies, les échanges ultérieurs souffrent d'une indifférence notable. Les *data scientists* attribuent cette indifférence à la prédominance d'interlocuteurs issus des RH « traditionnelles », manquant de compréhension technique. Ce déficit de compréhension entrave le processus de requalification *extensive* des données RH, limitant ainsi la capacité à exploiter pleinement leur potentiel pour l'analyse de l'absentéisme.

4. Controverses de requalification des données RH

Les controverses liées à la requalification des données RH soulignent les compromis nécessaires à leur transformation en biens économiques. Si cette transformation échoue après une première application du processus *QCR*, des cycles répétés entre requalification et capitalisation seront réalisés jusqu'à ce que les données RH soient finalement reconnues, ou non, en tant que biens économiques.

Dans le cadre de cette troisième séquence, le réseau de requalification est structuré autour de cinq modes d'existence : (1) Commercial, (2) Développement, (3) Marché, (4) Réglementaire et (5) Stratégique. Chacun de ces modes joue un rôle spécifique dans les négociations liées à la requalification des données RH. Ces modes d'existence sont décrits dans le Tableau 27.

Tableau 27 : Rôles des différents modes d'existence dans le réseau de requalification des données RH

Modes d'existence	Rôles
Commercial	Assurer le développement commercial des données RH et renforcer les relations d'affaires.
Développement	Garantir la coordination et l'harmonisation entre les développements technique et commercial.
Marché	Donner accès à des connaissances liées au marché RH.
Réglementaire	Veiller à la conformité des données RH avec les exigences réglementaires.

Modes d'existence	Rôles
Stratégique	Orienter les investissements afin d'assurer leur alignement avec la stratégie globale du cabinet.

Au sein de la séquence de requalification des données RH, quatre principales controverses émergent :

1. L'usage légal des données *DSN* limitant d'autres applications ;
2. Le désalignement des intérêts entre partenaires ;
3. La diversification des données comme obstacle à un outil « *plug-and-play* » ;
4. Le décalage entre perceptions internes et attentes du marché.

Le Tableau 28 présente ces controverses, en mettant en évidence les perspectives propres à chaque mode d'existence concerné, ainsi que les compromis envisagés pour leur résolution.

Tableau 28 : Controverses pour la séquence de requalification des données RH

Sujet des controverses		**Usage légal des données *DSN* limitant d'autres applications**	**Désalignement des intérêts entre partenaires**	**Diversification des données comme obstacle à un outil « *plug-and-play* »**	**Décalage entre perceptions internes et attentes du marché**
Perspectives spécifiques aux modes d'existence	**Commercial**	Usage opportuniste et diversifié	Intérêts technico-économiques à court terme	Diversification instrumentale	Décalage entre urgence et importance de l'absentéisme
	Développement			Diversification risquée	Décalage entre compréhension métiers et data
	Marché	Usage traditionnel et exploratoire	Intérêts politico-médiatiques à long terme		Décalage entre innovation et attrait des données RH
	Réglementaire	Usage légal restrictif		Diversification progressive	
	Stratégique		Intérêts financiers et de rentabilité		
Compromis négociés issu des controverses		**Non établi**	**Non établi**	**Diversification exclusive**	**Décalage limité et ciblé**

4.1. Controverse I : l'usage légal des données *DSN* limitant d'autres applications

La *DSN* est conçue pour améliorer la gestion des données de paie par les organismes de protection sociale. Ces organismes, en tant que principaux détenteurs des données *DSN*, bénéficient directement de la dématérialisation qui simplifie le traitement des déclarations. Cependant, bien qu'ils disposent d'un volume considérable de données *DSN*, leur rôle se limite principalement à gérer et traiter ces données dans un cadre légal et déclaratif.

Le mode Réglementaire restreint l'accès aux données *DSN* aux seuls organismes officiellement désignés pour leur gestion, imposant un cadre strict qui garantit leur utilisation dans les limites légales et déclaratives prévues. Face à cette restriction, deux autres modes d'existence émergent, chacun proposant une perspective distincte sur cette controverse :

1. Le mode d'existence Commercial : valorise l'utilisation des données *DSN* en mobilisant des compétences spécialisées en audit et contrôle qui permettent non seulement de sécuriser l'accès à ces données, mais également de légitimer leur usage opportuniste au-delà de leur cadre légal initial. Cette approche se concrétise par l'élaboration de deux modèles économiques - *freemium* et *premium* - destinés à établir un partenariat avec un organisme détenteur. Ce mode cherche ainsi à démontrer le potentiel économique encore inexploité de ces données et à ouvrir de nouvelles opportunités commerciales.
2. Le mode Marché : s'appuyant sur un usage traditionnel des données *DSN*, cherche à explorer de nouvelles possibilités par un usage exploratoire, intégrant les modes Commercial et Développement au sein d'un groupe de travail dédié. Cette démarche exige toutefois un investissement en temps et repose sur une coordination transversale, mobilisant des données issues de diverses directions au sein de l'organisme détenteur. La complexité de cette coordination, combinée au caractère complémentaire de l'application, relègue l'initiative à une priorité secondaire par rapport à l'usage traditionnel des données *DSN*.

Cette controverse met en lumière un décalage entre les restrictions réglementaires imposées aux données *DSN* et les aspirations à élargir leur usage vers des

applications plus diversifiées. Ce désalignement révèle les difficultés d'adaptation des données *DSN* à des finalités non prévues initialement.

Aucun compromis n'émerge de cette première controverse, car l'usage exploratoire au-delà du cadre légal n'est pas considéré comme prioritaire. Cette faible priorité s'explique par les délais requis pour sa mise en œuvre et par le rôle secondaire qu'il occupe dans les politiques de l'organisme responsable des données *DSN*.

4.2. Controverse II : le désalignement des intérêts entre partenaires

La nécessité d'accroître la visibilité et de construire un réseau solide est essentielle dans les partenariats commerciaux, souvent basés sur une stratégie d'apport d'affaires reposant sur la réciprocité : fournir un service en échange d'une contrepartie. Cependant, l'alignement des intérêts entre les partenaires n'est pas toujours garanti, créant ainsi des défis quant à la compatibilité des objectifs de chaque partie.

Cette dynamique mobilise trois modes d'existence distincts, chacun adoptant une perspective spécifique sur cette controverse :

1. Le mode d'existence Commercial : privilégie une gestion prudente des coûts pour minimiser les risques financiers. Il est motivé par des intérêts technico-économiques, visant à acquérir rapidement des clients tout en réduisant au maximum les dépenses. Il cherche également à maximiser sa visibilité à moindre coût, d'où l'adoption de la stratégie d'apport d'affaires visant à limiter les investissements.
2. Le mode d'existence Marché : conditionne son appui à l'adhésion du mode Commercial en tant que membre de son réseau. Il adopte un modèle d'affaires orienté vers des intérêts politico-médiatiques et s'efforce de maintenir cette ligne directrice. En évitant toute implication directe dans des activités commerciales, il cherche à préserver une position d'intermédiaire « neutre », en se focalisant principalement sur le soutien au développement d'un baromètre national.
3. Le mode d'existence Stratégique : adopte une posture d'attente en raison de ses intérêts financiers et de rentabilité, suspendant les investissements jusqu'à la marchandisation effective des données RH. Cette approche vise à optimiser le

retour sur investissement en limitant les engagements financiers tant que la viabilité commerciale des données n'est pas confirmée.

Ces divergences illustrent des stratégies contrastées dans la gestion des partenariats, notamment en matière de coûts et d'alignement des intérêts. La controverse persiste, faute de compromis, privilégiant les intérêts de rentabilité différée et de prudence budgétaire en raison de l'absence de retour sur investissement immédiat.

4.3. Controverse III : la diversification des données comme obstacle à un outil « *plug-and-play* »

La diversification des données RH constitue un frein important aux ambitions des *data scientists*, qui privilégient l'utilisation des données *DSN* en raison de leur standardisation. Cette standardisation est perçue comme un facteur clé dans le développement d'outils « *plug-and-play* » capables de s'adapter à divers contextes sans nécessiter d'ajustements majeurs. Cependant, dans le cadre de la requalification *extensive* des données RH, la diversité des sources de données freine cette aspiration. En effet, les SIRH présentent des variations dans la structuration et le contenu des données, ce qui rend indispensable une phase préalable de restructuration.

Face à cette controverse, trois modes d'existence expriment des perspectives distinctes :

1. Le mode d'existence Commercial : valorise les données *DSN* pour leur capacité à limiter les risques financiers, grâce à une standardisation qui élimine le besoin de retraitement. La diversification, en revanche, est perçue comme dévalorisante en raison des coûts élevés qu'elle implique, bien que d'autres types de données - telles que les données de production - soient reconnues pour leur précision potentiellement supérieure. Elle prend ainsi une dimension instrumentale, servant de prétexte pour initier un dialogue avec les clients potentiels et explorer leurs attentes tout en mettant en avant les potentialités des données *DSN*.
2. Le mode d'existence Développement : considère que les données issues de la *DSN* fournissent l'essentiel des informations pertinentes pour analyser l'absentéisme. Par conséquent, les autres sources de données, souvent jugées

peu fiables, sont reléguées au second plan. La diversification est ainsi vue comme risquée, car les données *DSN* répondent à l'essentiel des besoins informationnels.

3. Le mode d'existence Réglementaire : a introduit progressivement la *DSN* dans le secteur privé en 2017, puis dans le secteur public en 2020, révélant ainsi une différence de maturité entre les deux secteurs (cf. historique de données). Cette diversification progressive complique la standardisation des données *DSN*, rendant difficile l'adoption d'un cadre uniforme pour l'ensemble des secteurs.

Le compromis de cette controverse repose sur une diversification exclusive, impliquant l'exclusion temporaire du secteur public. Cette approche privilégie la maîtrise des coûts et la réduction des complexités liées à la diversité des données, en concentrant les efforts sur le secteur privé. L'objectif est d'atteindre la marchandisation des données RH en consolidant d'abord l'outil et en assurant un premier engagement client dans le secteur privé, avant d'envisager une extension vers le secteur public.

4.4. Controverse IV : le décalage entre perceptions internes et attentes du marché

Bien que les *data scientists* soient convaincus que leur proposition répond aux besoins et attentes du marché RH, les échanges ultérieurs ont révélé une indifférence notable de la part des clients potentiels. Cette situation met en lumière un décalage possible entre la perception interne de l'offre et les attentes réelles du marché, soulevant ainsi des interrogations quant au positionnement et à la pertinence des données RH sur ce marché.

Trois principaux modes d'existence illustrent des perspectives spécifiques à l'égard de cette controverse :

1. Le mode d'existence Commercial : considère que la gestion de l'absentéisme incombe avant tout aux clients, les données RH étant perçues comme un simple outil d'appui. Dans cette optique, il perçoit un décalage entre l'urgence et l'importance de l'absentéisme. En effet, malgré son importance, l'appropriation des données par les clients ne se produira véritablement qu'en situation d'urgence, lorsque ces derniers seront confrontés à une pression accrue pour gérer efficacement leur absentéisme.

2. Le mode d'existence Développement : soutient que les représentants des clients potentiels sont principalement des opérationnels dépourvus d'expertise en données, entraînant un décalage entre la compréhension des métiers et celle des *data scientists*. Cette situation limite leur capacité à reconnaître et à apprécier pleinement la valeur ajoutée de l'offre associée aux données RH, qui demeure ainsi sous-évaluée.
3. Le mode d'existence Marché : exprime une indifférence à l'égard des données RH, ne percevant ni innovation majeure ni effet marquant dans l'offre actuelle. Il remet également en question la proposition du mode Commercial, qu'il juge incapable de susciter un véritable intérêt. Ce décalage entre innovation et attrait des données RH souligne que, selon ce mode, une simple approche technique des données ne suffit pas à capter l'attention du marché ; il est nécessaire de les rendre plus attractives. Ainsi, ce mode plaide pour une approche plus « révolutionnaire » de l'analyse de l'absentéisme, un défi que les modes Commercial et Développement ont du mal à relever.

Le compromis qui découle de cette controverse repose sur une stratégie visant à poursuivre les démarches de prospection tout en surveillant les évolutions des besoins liés à la gestion de l'absentéisme. Cette approche proactive inclut des relances régulières, créant ainsi un décalage limité et ciblé par rapport aux urgences potentielles. L'objectif est de se positionner stratégiquement pour intervenir au moment le plus opportun, lorsque les conditions d'urgence ou de changement rendent l'offre plus pertinente et attractive.

En résumé, la séquence de requalification des données RH fait émerger quatre controverses principales sur :

1. L'usage légal des données *DSN* limitant d'autres applications ;
2. Le désalignement des intérêts entre partenaires ;
3. La diversification des données comme obstacle à un outil « *plug-and-play* » ;
4. Le décalage entre perceptions internes et attentes du marché.

Ces controverses sont toutes liées par la nécessité de trouver un compromis entre les ambitions de valorisation des données RH et les contraintes pratiques, qu'elles soient réglementaires, financières ou techniques. Elles illustrent ainsi les défis d'allier innovation et faisabilité dans un environnement contraignant.

5. Conséquences pour la fonction RH

Les deux premières séquences du processus de construction des données mettent en lumière une représentation de la fonction RH *fragmentée* et *fantomatique*. Dans cette troisième séquence - et pour combler ces lacunes -, Sandrine, spécialiste en droit social et connaissance de Jean (partenaire RH), rejoint le réseau, cherchant à se donner « *une deuxième jeunesse* » (entretien no1, 21/11/2023).

Bien que Sandrine (partenaire en droit social) possède une expertise substantielle en GRH, elle éprouve cependant des difficultés à établir une connexion entre son savoir-faire et les projets de connaissances développés par les *data scientists* :

« [...] *au début, elle s'y connaissait beaucoup en RH, mais je pense qu'elle ne comprenait pas beaucoup ce qu'on faisait* [...] *c'était flou. Je pense qu'elle ne se rendait pas compte de ce qu'était un absentéisme attendu* [deuxième projet de connaissances]*. Au début, je pense qu'elle n'avait pas du tout un pied dans la data...* » (Julie, cheffe de projet, entretien no1, 17/10/2023).

Toutefois, l'engagement croissant et régulier de Sandrine (partenaire en droit social) avec l'équipe conduit à un développement progressif de sa compréhension :

« […] *avec Sandrine, ça a bien marché. Elle rentre dans le détail un peu avec nous et de façon un peu régulière dans les analyses data qu'on fait.* » (Xavier, chef de projet, entretien no2, 21/11/2023).

Cette évolution démontre sa capacité à acquérir graduellement les compétences nécessaires pour une interaction active avec les données RH, passant d'une expertise centrée principalement sur la GRH à des compétences plus hybrides. Toutefois :

« […] *si on a* [que] *des équivalents de Sandrine (partenaire en droit social), et qu'il n'y a aucune* [autre] *personne data, ils ne peuvent pas juger comme il faut notre offre.* » (Julie, cheffe de projet, 17/10/2023).

Cette observation montre que, malgré les progrès de Sandrine (partenaire en droit social), la fonction RH au sein du projet reste *embryonnaire*, limitant la valorisation de la GRH dans la construction des données.

Avec cette troisième séquence, une nouvelle dynamique émerge : alors que la qualification et la capitalisation révèlent une marginalisation de la fonction RH, la

requalification, par l'investissement actif de cette dernière, amorce une réintégration progressive bien que fragile de la dimension « RH » des données.

6. Conclusion

Ce chapitre se penche sur la troisième séquence de construction des données RH : la Requalification. Elle vise à requalifier la qualité *normative* des données RH en s'appuyant sur les projets de capitalisation.

Trois actes clés sont abordés dans cette séquence :

4. Re-définir les besoins des clients ;
5. Re-rationaliser les coûts d'investissement ;
6. Ré-enrôlement des agents économiques.

La requalification met en lumière la qualité *extensive* des données RH, laquelle se manifeste sous deux formes : (1) *externe* et (2) *interne*

La requalification *extensive externe* représente un changement de perspective, passant d'un référentiel de comparaison interne à un cadre externe sectoriel. Elle nécessite un accès étendu à des données RH *normative*s et le développement de nouvelles conventions de mesure spécifiques à chaque secteur.

Parallèlement, la requalification *extensive interne* exige de combiner les données RH *normative*s avec d'autres types de données. Elle requiert l'accès à une diversité de sources, favorisant ainsi une compréhension plus fine des besoins des clients grâce à l'intégration de critères variés.

La qualité *extensive* des données RH est évaluée en fonction des gains marginaux attendus. À partir de cette évaluation, les *data scientists* concluent que les bénéfices potentiels de sa forme interne ne justifient pas les coûts des données non issues de la *DSN*. Par conséquent, ils optent pour une stratégie bifocale centrée sur les données *normatives* avec deux modèles : (1) *freemium* et (2) *premium*.

Les interactions entre les *data scientists* et le marché RH se caractérisent cependant par une indifférence notable. Ces derniers attribuent notamment cette réaction à la prépondérance d'interlocuteurs RH « traditionnels », qui affichent un manque de compréhension technique.

Quatre controverses émergent de cette troisième séquence, chacune faisant l'objet de négociations. Les compromis atteints sont les suivants :

1. Non établi ;
2. Non établi ;
3. La diversification exclusive ;
4. Le décalage limité et ciblé.

Ces compromis, ou plutôt cette absence de compromis, mettent en lumière une stratégie de marchandisation des données RH basée sur une sélection contrainte. Axée sur les données *DSN*, cette stratégie transforme leur fonction initiale : d'instruments épistémiques pour l'analyse de l'absentéisme, elles deviennent progressivement l'élément central de l'offre commerciale, reléguant ainsi l'absentéisme à un rôle secondaire. En outre, cette stratégie réduit non seulement les interactions avec les clients potentiels, mais limite également l'influence des *data scientists* à un réseau restreint. En raison de leur incapacité à transformer les données RH en biens économiques, un nouveau cycle entre Capitalisation et Requalification s'engage.

Partie III. DISCUSSION ET CONCLUSION

Chapitre 6. Discussion

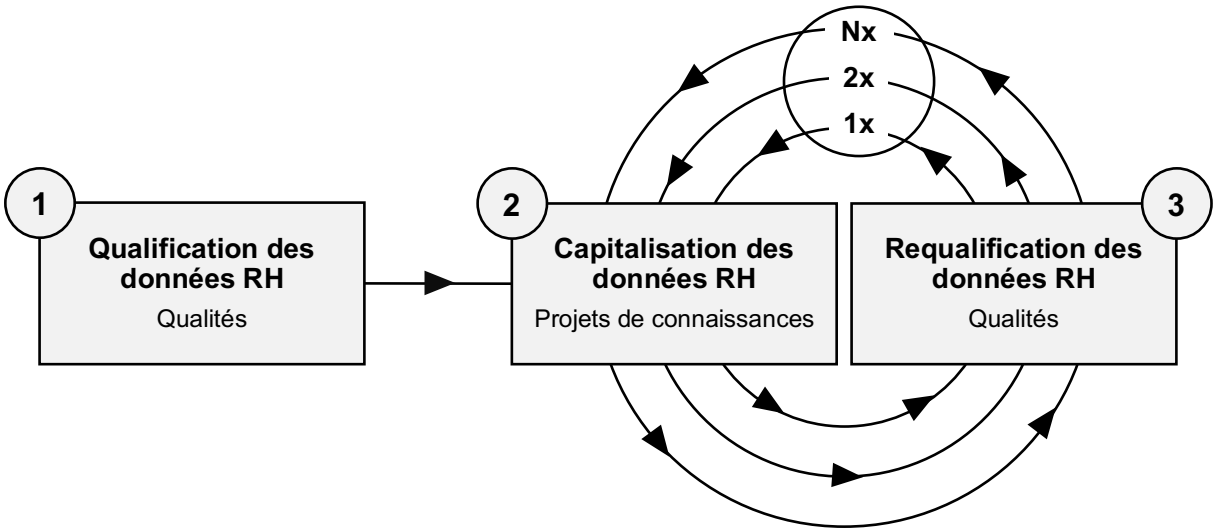


Figure 16 : Processus de construction des données RH

1. Introduction

Dans un contexte caractérisé par une accélération exponentielle de la spirale *data-driven*, les données RH sont de plus en plus reconnues comme des instruments du capital. De ce fait, il est essentiel de reconsidérer la construction de ces données au sein de réseaux dédiés à leur valorisation et leur intégration dans des circuits commerciaux déjà établis.

Le processus de construction des données RH, qui est théorisé dans cette thèse, se structure autour de trois séquences clés - Qualification, Capitalisation et Requalification (*QCR*). Chacune de ces séquences correspond à un chapitre de résultats distinct :

1. La Qualification : qui révèle la qualité *normative* des données RH.
2. La Capitalisation : qui met en évidence trois projets de connaissances distincts, chacun caractérisé par un type spécifique de connaissances sur l'absentéisme :
    1. L'absentéisme *réel univarié* : connaissances *descriptives* ;
    2. L'absentéisme *réel multivarié* : connaissances *explicatives* ;
    3. L'absentéisme *latent* : connaissances *prédictives*.
3. La Requalification : qui met en lumière la nouvelle qualité *extensive* des données RH, se manifestant sous deux formes : *externe* et *interne*.

Au sein de chacune de ces trois séquences, l'analyse des espaces de négociation, désignés comme des « controverses », permet d'identifier trois stratégies de marchandisation impliquées dans la construction des données RH :

1. L'optimisation efficiente pour la qualification ;
2. L'ajustement progressif pour la capitalisation ;
3. La sélection contrainte pour la requalification.

Ce sixième chapitre est consacré à la discussion théorique et empirique des résultats obtenus dans le cadre de cette thèse, ainsi qu'à la présentation des différentes contributions, limites et perspectives de recherche.

2. Discussion des travaux théoriques

Un vaste corpus de recherches en SI s'intéressent aux données (Alaimo et al., 2020; Alaimo & Kallinikos, 2022; Monteiro & Parmiggiani, 2019; Pachidi et al., 2021;

Parmiggiani et al., 2022). Selon ces auteurs, les données ne constituent pas des entités autonomes. Elles sont profondément imbriquées dans un réseau d'éléments contextuels, humains et techniques qui interviennent à chaque étape de leur cycle de vie, depuis leur construction jusqu'à leur exploitation. La partialité qui les caractérise résulte de la complexité inhérente à leur écosystème, qu'il soit matériel ou symbolique (Kitchin & Lauriault, 2018).

Ces études, contestant l'idée de données neutres ou purement techniques et révélant leur intégration dans des dynamiques plus vastes, forment le socle de cette discussion. Elle se déploie autour de trois axes principaux, qui définissent également les principales contributions de cette recherche :

1. La conceptualisation des données RH en tant qu'objets d'étude en GRH ;
2. La conceptualisation des dispositifs socio-numériques des données RH à travers la spirale *data-driven* ;
3. La théorisation du processus de construction des données RH articulée en trois séquences (*QCR*).

2.1. La conceptualisation des données RH en tant qu'objets d'étude en GRH

Le manque de reconnaissance des données en tant qu'objets d'étude en GRH, notamment dans l'e-GRH et l'instrumentation de GRH (Coron, 2022; Garcia-Arroyo & Osca, 2019; Margherita, 2022; Zhang et al., 2021), constitue un frein majeur à notre compréhension des dynamiques de transformation de ce domaine. En effet, cette lacune perpétue une vision réductrice des données, perçues comme le reflet neutre et objectif de la « réalité » des activités et phénomènes RH (Coron, 2019g; Desrosières, 2013b). Or, une telle vision occulte la complexité de l'écosystème socio-numérique qui sous-tend la construction de ces données, en particulier dans un contexte marquée par l'expansion rapide du marché numérique RH (Kowu, 2024).

Cette expansion montre que les principaux leviers des technologies RH ne sont pas uniquement techniques, mais profondément sociaux et économiques. Les données RH sont construites non seulement pour appuyer le développement technologique mais aussi pour stimuler une croissance soutenue des profits générés par ces technologies (Ruppert et al., 2017). Par conséquent, les données RH évoluent

en actifs et en marchandises activement achetés et vendus, influençant directement les dynamiques de transformation de la GRH (Alaimo & Kallinikos, 2022; Sadowski, 2019).

Dans ce cadre, il devient impératif de questionner les données perçues comme ayant une valeur ou un potentiel de valorisation futur pour les agents économiques opérant sur le marché (Alaimo et al., 2020). Cette interrogation justifie la nécessité de conceptualiser les données RH en tant que véritables objets d'étude pour mieux comprendre et anticiper leur rôle dans la transformation de la GRH.

Répondant à cet impératif, mon étude contribue, d'une part, à la définition des caractéristiques techniques des données RH, telles qu'identifiées par Kitchin (2022d) : leurs formes et structures, leurs sources, leurs producteurs et les types de données. Illustrées par des exemples concrets, ces caractéristiques révèlent la nature intrinsèquement malléable des données RH (Kallinikos et al., 2013). Cette malléabilité se manifeste dans ma recherche par l'ambition des *data scientists* d'élargir l'utilisation des données *DSN*, initialement conçues pour la gestion administrative, vers des applications diversifiées telles que l'analyse de l'absentéisme. Cette extension illustre le potentiel économique encore largement inexploité des données RH, orientant ainsi leur construction vers des usages plus variés.

En reconnaissant les données RH comme des biens économiques, cette étude contribue, d'autre part, à démontrer que les données ne sont pas simplement « données » ; elles sont des produits construits, imprégnés de valeurs, d'influences et de rationalités spécifiques (Ruppert et al., 2017).

Dans ce cadre, les *data scientists*, en négociant les qualités *normative* et *extensive* de ces données, cherchent à répondre à des intérêts technico-économiques bien précis. L'adoption de standards nationaux participe ainsi à maximiser la rentabilité des données RH tout en réduisant les incohérences souvent rencontrées dans les projets en *data science*, telles que l'incomplétude et la non-conformité. Cette approche vise également à élargir l'utilisation de ces données standards au-delà des frontières internes de l'entreprise pour atteindre une portée sectorielle. Au regard de cette recherche, les données RH ne représentent donc pas une manifestation transparente de l'absentéisme ; elles sont construites de manière subjective afin d'influencer activement les dynamiques du marché RH dédiée l'analyse de ce phénomène.

En résumé, mon premier apport réside dans la conceptualisation des données RH en tant qu'objets théoriques. En examinant leurs caractéristiques techniques et, plus largement, leurs qualités en tant que biens économiques, cette étude met en lumière comment les connaissances, intérêts et rationalités des agents orientent le processus de construction des données RH.

2.2. La conceptualisation des dispositifs socio-numériques des données RH à travers la spirale *data-driven*

Le mythe de la donnée « naturelle » (Power, 2023, p. vii) véhicule, à tort, l'idée selon laquelle les données existent comme des entités brutes, alors qu'elles sont le produit d'une démarche humaine et organisée (Alaimo & Kallinikos, 2024; Kitchin, 2022d). Les données sont étroitement liées aux idées, technologies, acteurs humains, systèmes et contextes qui interviennent à chaque étape de leur cycle de vie (Kitchin & Lauriault, 2018). Remettre en question le déterminisme technique prévalant en GRH et démontrer la construction sociale des données nécessitent une exploration des dispositifs socio-numériques dans lesquels elles s'inscrivent (Coron, 2019d; Desrosières, 2013a; Diaz-Bone & Didier, 2016). En effet, les données agissent comme des relais de ces dispositifs, qui encadrent et guident leur construction (Kitchin & Lauriault, 2018).

Pour approfondir cette perspective, j'ai analysé les corpus de littérature susceptibles de définir l'écosystème sous-jacent à ce que l'on nomme « données RH ». En rapprochant les domaines des SI et de GRH, j'ai structuré ma revue de littérature en quatre actes distincts, chaque acte augmentant progressivement en complexité pour illustrer l'interaction entre ces deux champs :

1. Définition des données RH ;
2. Transformation des données RH ;
3. Valorisation des données RH ;
4. Construction des données RH.

À travers ces quatre actes, mon étude vise à établir un rapport conceptuel entre les SI et la GRH. La littérature en GRH, en particulier celle portant sur l'e-GRH et l'instrumentation de GRH, présente des limites importantes. Bien que la sélection des

données RH soit abordée dans l'analyse des activités et phénomènes RH, elle reste largement sous-explorée (Coron, 2022). Cette littérature se focalise principalement sur les effets des technologies numériques, souvent considérées comme des « boîtes noires », et tend à négliger l'exploration du processus de construction des données RH. Or, ce processus est essentiel pour appréhender pleinement l'impact de ces données à travers l'usage des technologies numériques (Tambe et al., 2019).

Le rapprochement entre les SI et la GRH a ainsi permis le développement d'un cadre structuré et exhaustif que j'ai conceptualisé sous la forme de la spirale *data-driven*. Cette spirale, régie par une logique capitaliste d'accumulation et de circulation des données RH (Ruppert et al., 2017), illustre le rôle prépondérant et omniprésent des données, qui sont désormais reconnues comme un élément structurant de notre société contemporaine.

La portée de cette spirale est d'autant plus significative qu'elle permet de réunir des phénomènes de transformation des données RH, souvent traités de manière isolée, à l'échelle de l'organisation (numérisation, digitalisation, datafication) et à celle de la GRH (*big data* RH, analytique RH et métrique RH, ainsi que l'épistémologie positiviste et formelle). Elle met en lumière un renforcement mutuel entre ces phénomènes, qui se soutiennent et s'alimentent réciproquement. Ces transformations plurielles, intensifiées par l'essor et la sophistication croissante des technologies numériques, favorisent également l'émergence de nouveaux acteurs (*data scientists*) dont les logiques professionnelles guident le développement des différentes fonctions (instrumentale et épistémique) que les données sont appelées à performer.

La conceptualisation de la spirale *data-driven* offre donc un cadre visant à combler les lacunes de la littérature en GRH, où les données sont perçues comme des objets athéoriques. Ce manque de fondements théoriques pourrait, en outre, contribuer à une marginalisation de la fonction RH dans la construction de ses propres données, un phénomène que cette étude met en évidence[23]. En effet, l'incapacité de la fonction RH à s'imposer pleinement dans ce processus démontre que les données RH ne sont pas simplement une représentation transparente des activités et phénomènes RH. Au

[23] Il convient de rappeler que ces observations sont complémentaires aux résultats principaux et que l'hypothèse d'une possible relation entre ces deux éléments n'est pas suffisamment approfondie pour être affirmée avec certitude.

contraire, elles sont profondément modelées par les dynamiques des dispositifs socio-numériques.

En synthèse, le rapprochement entre les domaines des SI et de GRH constitue le socle de mon second apport : la conceptualisation de la spirale *data-driven*. À travers cette spirale en quatre actes, mon étude contribue ainsi une compréhension plus approfondie et nuancée des dispositifs socio-numériques qui régissent la construction des données RH.

2.3. La théorisation du processus de construction des données RH articulée en trois séquences (*QCR*)

Les données RH sont utilisées pour capturer, représenter, connaître et agir sur les activités et phénomènes RH (Alaimo & Kallinikos, 2022; Østerlie & Monteiro, 2020). Elles sont intégrées dans des dispositifs socio-numériques, qui incluent des systèmes, technologies, outils, professions, processus et pratiques (Alaimo & Kallinikos, 2022, 2024; Cadin et al., 2012). Ces données ne sont toutefois pas intrinsèquement utiles ; leur valeur découle plutôt de la capacité de ces dispositifs à les construire de manière à répondre à la demande du marché RH (Greasley & Thomas, 2020).

De nombreuses études critiques montrent que les données dépassent leur simple valeur d'origine pour devenir des actifs et des marchandises échangés et vendus (Alaimo & Kallinikos, 2022; Sadowski, 2019). Situées à l'interface entre l'offre et la demande du marché numérique RH, les données sont construites selon des logiques de valorisation qui agissent au carrefour de leurs dimensions sociales, techniques et économiques (Ruppert et al., 2017). Dès lors, il devient impératif d'explorer comment s'opère la construction de ces données au sein de réseaux dédiés à leur valorisation et à leur intégration dans des circuits commerciaux déjà établis (Callon & Muniesa, 2005; Muniesa et al., 2007).

En réponse à cet impératif, cette recherche propose d'établir un cadre conceptuel pour théoriser le processus de construction des données RH en tant que biens économiques. Ce cadre combine la production de connaissances (Latour, 2005a, 2007) et l'économie des qualités (Callon et al., 2000, 2002) créant ainsi une synergie conceptuelle structurée en trois séquences : la Qualification, la Capitalisation et la Requalification.

Ce processus, désigné sous le nom de processus *QCR*, se caractérise par un déplacement itératif et expansif des connaissances visant à construire les données RH de manière à les aligner avec la demande du marché. Ce déplacement se manifeste par un va-et-vient entre l'intérieur (offre) et l'extérieur (demande) du réseau de construction des données RH. Il influence directement les connaissances que les *data scientists* possèdent sur ces données et le cadre dans lequel elles peuvent s'inscrire pour répondre à la demande du marché RH.

Ainsi, le déplacement et, par conséquent, l'enrichissement des connaissances, au cours du processus de construction des données RH, repose sur la mobilisation de ressources extérieures (humaines et non humaines) au réseau initial. Comme le souligne Latour (2005a, p. 373) : « *la notion de spécialiste isolé est une contradiction dans les termes. De deux choses l'une : ou bien vous êtes maintenu dans l'isolement et vous perdez très vite la qualité de spécialiste, ou vous demeurez spécialiste mais cela implique que vous n'êtes pas isolé*. ».

Dès lors, afin que les connaissances des *data scientists* soient reconnues - qu'elles soient descriptives, explicatives, ou prédictives dans cette étude - et que la singularisation des données RH devienne effective, celles-ci doivent être enrichies par de nouvelles ressources au sein d'un réseau élargi. Ce déplacement, qui relie l'intérieur et l'extérieur du réseau de construction des données RH, favorise l'alignement des intérêts et permet une adéquation entre l'offre (intérieur) et la demande (extérieur), conduisant ainsi à la transformation des données RH en biens économiques.

Au regard de mon étude, ce déplacement est toutefois porté par des controverses. Ces dernières naissent de l'équilibre fragile entre, d'une part, l'ambition de développement et l'exploitation des opportunités de marché et, d'autre part, la nécessité de gérer prudemment les risques et les investissements.

Pour résoudre ces controverses, les *data scientists* adoptent trois types de stratégies de marchandisation des données RH : (1) d'optimisation efficiente pour la qualification, (2) d'ajustement progressif pour la capitalisation et (3) de sélection contrainte pour la requalification. Ces stratégies, qui reflètent principalement des intérêts internes au réseau (qualités *normative* et *extensive* des données) sont issues d'une approche pragmatique centrée sur la maîtrise des coûts. En effet, les compromis

réalisés sont orientés par le mode d'existence Stratégique qui conditionne les investissements à une marchandisation préalable, puis encadrés par le mode d'existence Commercial qui justifie les choix de construction des données RH selon leur rentabilité potentielle.

Cette configuration cloisonnée restreint les interactions avec l'extérieur (à l'exemple du secteur public, qui se trouve exclu) et entrave le partage entre l'intérieur et l'extérieur du réseau. En conséquence, la singularité des données RH, essentielle pour en faire un point de passage obligé, est affaiblie, fragilisant ainsi leur valorisation économique : les partenaires et clients potentiels se désengagent, leurs intérêts se désalignent et l'« effet wow » ne se manifeste pas.

En définitive, les choix et arbitrages liés à la construction des données RH s'apparentent à ceux engagés dans la marchandisation de tout type de données. Concevoir leur processus de construction suppose de les inscrire dans un réseau qui établit un lien entre un intérieur (offre) et un extérieur (demande). Ce lien, structuré en trois séquences clés (*QCR*), se renforce par un partage fondé sur des qualités « débattues », « contestées » et « négociées » des données, contribuant ainsi à leur singularité.

En mobilisant les principes de l'ANT, cette théorisation peut enrichir d'autres cadres théoriques, tels que la sociologie de la quantification, en apportant une compréhension approfondie du processus de construction des données RH (Coron, 2019d; Desrosières, 2013a; Diaz-Bone & Didier, 2016). Comme le soulignent Desrosières et Kott (2005, p. 2) à propos des données quantifiées : « […] *il existe une série de conventions préalables, de négociations, de compromis, de traductions, d'inscriptions, de codages et de calculs conduisant à la mise en nombre.* […] *L'usage du verbe quantifier attire l'attention sur la dimension, socialement et cognitivement créatrice, de cette activité. Celle-ci ne fournit pas seulement un reflet du monde (point de vue méthodologique usuel), mais elle le transforme, en le reconfigurant autrement.* ». Desrosières et Kott mobilisent des concepts clés empruntés à l'ANT, tels que les compromis, la traduction, entre autres. Cette perspective, en accord avec les résultats de cette étude, s'oppose au mythe de la donnée « naturelle » (Power, 2023, p. vii), qui suppose une objectivité intrinsèque. Elle démontre que les données RH sont le fruit d'un processus socialement construit, reposant sur des conventions et des négociations spécifiques.

La complémentarité entre l'ANT et la sociologie de la quantification permet de proposer une conceptualisation plus concrète de la quantification « *en train de se faire* » (Latour, 2005a, p. 29). Cette approche met en lumière l'imbrication des pratiques calculatoires et non calculatoires au sein d'un réseau d'éléments contextuels, humains et techniques, structuré autour des controverses qui jalonnent le processus de mise en nombre. Ces controverses illustrent comment les tensions influencent ce processus en révélant les compromis qui affectent les pratiques de quantification des acteurs concernés.

En résumé, mon troisième apport réside dans la théorisation du processus *QCR*, qui se décompose en trois séquences clés : la Qualification, la Capitalisation et la Requalification. En associant la production de connaissances (Latour, 2005a, 2007) à l'économie des qualités (Callon et al., 2000, 2002), ce cadre permet de mieux appréhender les dynamiques sous-jacentes à la construction des données RH en tant que biens économiques.

En conclusion, cette thèse présente trois contributions théoriques :

1. La conceptualisation des données RH en tant qu'objets d'étude en GRH ;
2. La conceptualisation des dispositifs socio-numériques des données RH à travers la spirale *data-driven* ;
3. La théorisation du processus de construction des données RH articulée en trois séquences.

Ainsi, la convergence de ces contributions met en visibilité les nombreuses dynamiques en jeu dans l'étude des données RH et souligne l'importance d'explorer les dispositifs socio-numériques qui sous-tendent leur construction.

3. Discussion des travaux empiriques

Mon immersion de trois ans et sept mois au sein du cabinet de conseil *QIA* m'a permis de me rapprocher de l'équipe de *data scientists* et de suivre de près le processus de construction des données RH. Cette expérience pratique sert de fondation à ma discussion, laquelle s'articule autour de trois axes principaux représentant également les contributions principales de cette recherche :

1. Pour la gestion de projet : en identifiant les facteurs limitants et les facteurs de succès du processus de construction des données RH ;
2. Pour les *data scientists* : en renforçant la sensibilisation et la formation en GRH afin de garantir une contextualisation adéquate des données RH ;
3. Pour la fonction RH : en garantissant une sensibilisation et une formation en *data science* pour assurer un contrôle adéquat sur la construction des données RH.

3.1. Pour la gestion de projet : identifier les facteurs limitants et les facteurs de succès du processus de construction des données RH

Situées à la jonction de l'offre et de la demande sur le marché numérique, les données RH résultent de dynamiques de valorisation qui opèrent à l'intersection de leurs dimensions sociales, techniques et économiques. Dans ce cadre, la gestion de projet joue un rôle clé dans l'orchestration de ces dynamiques, garantissant le partage entre l'intérieur (offre) et l'extérieur (demande) du réseau de construction de ces données.

En tant que *HR business analyst* impliquée dans ce réseau, ma position centrale au sein de la gestion de projet m'a permis d'alterner entre une participation active et une observation (in)directe. Cette double approche a également facilité le suivi rapproché des facteurs limitants et des facteurs de succès essentiels au processus de construction des données RH.

Les facteurs limitants découlent principalement de la logique qui gouverne le processus de construction des données RH, laquelle est soumise au mode d'existence Stratégique et encadrée par le mode d'existence Commercial. Cette logique, centrée sur la maîtrise des coûts, entrave l'exploration épistémique des données, créant un cercle vicieux où les contraintes entourant les connaissances existantes limitent également l'apprentissage des *data scientists* impliqués. Ces facteurs se répartissent en deux catégories principales :

1. Facteur contextuel : les données RH, extraites d'une base de données fictive conçue pour l'entraînement de modèles analytiques, se trouvent déconnectées de leur contexte d'utilisation réel. Cette origine engendre des interrogations concernant leur validité et leur applicabilité dans des situations pratiques.

2. Facteurs organisationnels :
   2.1. Composition de l'équipe : l'équipe, composée principalement d'agents économiques issus de l'intégration académique (cf. stagaires *data scientists*) et la mobilisation « en temps masqué », est soumise à des fluctuations de composition dues à leur repositionnement vers des projets jugés plus rentables. Ces changements perturbent la continuité et la cohérence du processus de construction des données RH.
   2.2. Formation de l'équipe : les membres de l'équipe manquent de formation spécialisée en GRH, ce qui limite leur capacité à mettre en œuvre efficacement les connaissances accumulées sur les données RH et à les transformer en biens économiques.

Ainsi, pour compenser les facteurs limitants du processus de construction des données RH, il est essentiel : d'enrôler des médiateurs data RH.

Ce facteur de succès est nécessaire pour maximiser l'efficacité de la transformation des données RH en biens économiques. Les stratégies d'enrôlement doivent se concentrer sur l'intégration de médiateurs qualifiés. Il ne s'agit pas seulement d'enrôler des agents issus de la fonction RH, mais de recruter des médiateurs capables de naviguer entre la *data science* et la GRH[24]. Ces médiateurs jouent un rôle déterminant en se positionnant à l'interface des données et des connaissances en GRH, assurant ainsi une intégration harmonieuse de ces deux domaines dans le processus de construction des données.

En synthèse, mon premier apport empirique se concentre sur la gestion de projet et met en évidence les facteurs limitants ainsi que le facteur de succès qui influencent le processus de construction des données RH.

[24] Ce constat est renforcé par l'observation des choix d'enrôlement d'agents économiques issus de la fonction RH (cf. Jean et Sandrine) qui met en évidence leurs difficultés à opérer à l'interface entre les domaines de la *data science* et de la GRH.

3.2. Pour les *data scientists* : renforcer la sensibilisation et la formation en GRH afin de garantir une contextualisation adéquate des données RH

Les données RH ne reflètent pas simplement la « réalité » des activités ou des phénomènes RH étudiés. Elles sont intrinsèquement façonnées par les dispositifs socio-numériques qui président et encadrent leur construction.

Travailler avec les données RH nécessite une compréhension approfondie de la complexité et de la sensibilité de ces dispositifs, particulièrement en raison de leur lien direct avec la gestion des salariés. Bien que le travail sur ces données présente de nombreux avantages, il soulève également des dilemmes éthiques significatifs pour la GRH. Ces dilemmes peuvent notamment comprendre :

- Le profilage des salariés : les données RH peuvent conduire à un profilage basé sur des caractéristiques telles que l'âge, la nationalité, le genre ou le statut socio-économique, ce qui soulève des préoccupations éthiques.
- La distorsion de la réalité en GRH : en l'absence de contextualisation adéquate, les données RH peuvent fournir une vision biaisée ou incomplète des activités ou phénomènes RH, pouvant influencer négativement les décisions de gestion.

Ainsi, il est crucial de sensibiliser et former les *data scientists* à l'importance de la contextualisation des données RH. Ces actions doivent inclure :

- La compréhension des dispositifs socio-numériques en GRH : les *data scientists* doivent être sensibilisés et formés pour comprendre l'influence des dispositifs socio-numériques sur la construction des données RH et savoir comment les cartographier.
- L'intégration de la fonction RH : les *data scientists* doivent être sensibilisés et formés à collaborer avec des fonctions non spécialistes, tels que la fonction RH, pour s'assurer que les données sont construites de manière à refléter fidèlement leurs réalités et leurs enjeux spécifiques, notamment dans le choix des conventions de mesure.

- La prise en compte des dimensions éthiques en GRH : les *data scientists* doivent être sensibilisés et formés aux implications éthiques de leur travail sur les données RH et adopter des pratiques de collaboration et de recontextualisation des données qui minimisent les risques de profilage et de biais.

En synthèse, mon deuxième apport empirique repose sur la sensibilisation et la formation des *data scientists* à l'importance de la contextualisation des données RH, qui sont indispensables pour garantir une construction de données à la fois éthique et pertinente. Cette approche permet une meilleure intégration des réalités et des enjeux spécifiques de la GRH dans le travail des *data scientists*, tout en prévenant les risques éthiques associés.

3.3. Pour la fonction RH : garantir une sensibilisation et une formation en *data science* pour assurer un contrôle sur la construction des données RH

La fonction RH souffre d'un déficit de compétences analytiques et d'une perte de légitimité quant à sa capacité à s'impliquer dans des projets en *data science*. Ces lacunes se manifestent dans cette étude, qui met en évidence une fonction RH :

1. *Fragmentée* (Qualification) ;
2. *Fantomatique* (Capitalisation) ;
3. *Embryonnaire* (Requalification).

Ce constat indique une incapacité de la fonction RH à s'affirmer pleinement dans le processus de construction des données RH. On peut croire à une logique « *data-driven* » inversée, voire « *data-lagging* » qui illustre une réalité où la fonction RH reste en retard et progresse lentement, incapable de s'adapter aux exigences d'un environnement de plus en plus axé sur les données. Ce déficit de compétences peut conduire la fonction RH à déléguer entièrement la responsabilité de la construction des données RH à des acteurs externes, comme les *data scientists*. Cependant, cette délégation présente plusieurs risques, notamment :

- La perte de contrôle sur la construction des données RH : la fonction RH pourrait perdre son autorité et sa maîtrise sur la construction de ses propres données, ce

qui est particulièrement problématique dans un contexte où les décisions sont de plus en plus basées sur les données (cf. la spirale *data-driven*).

- La fragmentation des connaissances en GRH : en confiant la construction des données à des acteurs externes, il y a un risque que les connaissances et les compétences nécessaires pour comprendre et utiliser ces données restent fragmentées, empêchant ainsi une intégration pertinente en GRH.

Ainsi, la fonction RH doit être sensibilisée et formée pour garantir un contrôle éthique et pertinent dans la construction des données. Ces actions doivent inclure :

- L'éducation à l'éthique des données : la fonction RH doit être sensibilisée et formée pour comprendre les implications éthiques de la construction des données RH. Cette formation aidera à minimiser les risques de profilage et de biais, en assurant que les données sont construites de manière responsable.
- L'encouragement à la collaboration avec les *data scientists* : la fonction RH doit être sensibilisée et formée pour collaborer étroitement avec les *data scientists*. Cette collaboration permettra de garantir que les besoins et les enjeux spécifiques en GRH soient bien intégrés dans le processus de construction des données.
- Le développement de compétences analytiques : la fonction RH doit être sensibilisée et formée pour renforcer ses compétences analytiques. Cela permettra une meilleure compréhension et une meilleure utilisation des données, tout en garantissant une participation active et informée dans les projets de *data science*.

En synthèse, mon troisième apport empirique réside dans l'importance de sensibiliser et de former la fonction RH pour garantir que la construction des données RH soit basée sur une expertise appropriée et adaptée aux contextes spécifiques. Ceci est crucial pour minimiser les risques associés à une délégation complète de cette responsabilité à des acteurs externes.

En conclusion, cette thèse permet de dégager trois contributions empiriques :

1. Pour la gestion de projet : en identifiant les facteurs limitants et les facteurs de succès du processus de construction des données RH ;
2. Pour les *data scientists* : en renforçant la sensibilisation et la formation en GRH afin de garantir une contextualisation adéquate des données RH ;

3. Pour la fonction RH : en garantissant une sensibilisation et une formation en *data science* pour assurer un contrôle adéquat sur la construction des données RH.

La combinaison de ces trois apports souligne la nécessité d'une approche intégrée pour assurer une construction pertinente et éthique des données RH. En effet, la valeur des données dépend directement de l'intégrité des dispositifs socio-numériques qui les sous-tendent.

4. Limites et perspectives de recherche

Ce travail de thèse comporte plusieurs limites importantes qui doivent être prises en compte pour contextualiser les résultats et les discussions présentés. Ces limites se répartissent en deux grandes catégories : théorique et empirique.

4.1. Limite théorique : dilution du pouvoir entre les agents économiques du réseau

La dynamique entre ceux qui cartographient (les *data scientists*) et ceux qui sont cartographiés (la fonction RH) illustre une lutte de pouvoir dans la symétrie des connaissances. Cette lutte ne se déroule pas sur des terrains cognitifs ou culturels isolés, mais dans un espace où les connaissances doivent être accumulées, stabilisées et intégrées au sein de réseaux fragiles et en constante évolution (Latour, 2005a). Il n'existe donc pas de séparation nette entre les connaissances « universelles » - et décontextualisées - des *data scientists* et les connaissances locales de la fonction RH. Au contraire, tous les acteurs se déplacent à l'intérieur de réseaux étroits et fragiles qu'ils doivent stabiliser et faire converger en accumulant des connaissances par le biais d'alliances stratégiques.

L'approche de l'ANT adoptée dans cette thèse est fréquemment critiquée pour sa tendance à diluer le concept de pouvoir en l'attribuant de manière équivalente à tous les acteurs, qu'ils soient humains ou non-humains. Cette approche tend ainsi à minimiser les dynamiques de pouvoir et les inégalités sociales existantes entre les agents économiques étudiés. En mettant davantage l'accent sur la question du pouvoir des données, notamment en réutilisant le concept d'assemblage tel que proposés par Carter (2018) et Kitchin & Lauriault (2018), cette perspective théorique pourrait permettre de mieux interroger l'absence de la fonction RH dans le projet de

construction des données RH. Elle permettrait ainsi d'analyser plus finement les implications de cette absence sur la dynamique globale du projet et sur la nature des données produites.

4.2. Limite empirique : influence d'une implication active

La présente étude, fondée sur l'analyse d'un cas unique, comporte des limites en termes de portée et de transférabilité directe des conclusions à d'autres contextes organisationnels. Néanmoins, la limite la plus significative réside dans ma posture en tant que chercheuse intégrée au terrain.

Pendant trois ans et sept mois, j'ai été intégrée à l'équipe en tant que *HR business analyst*, participant activement à divers projets (quatre au total) et apportant des contributions à la fois théoriques et pratiques. Cette proximité avec le terrain m'a permis de documenter en détail le processus de construction des données RH, mais elle a également influencé la perception qu'avaient les autres membres de l'équipe de mon rôle. Mon statut de référente RH était reconnu, non seulement par mon équipe immédiate, mais aussi par un ensemble plus large d'acteurs, comme en témoigne un courriel reçu le 22/12/2021 : « […] *Pour rappel, il s'agit aussi pour toi d'être capable de porter l'offre RH plus large, de pouvoir parler non seulement de DSN Analytics mais aussi des leviers RH, du dimensionnement, des formations, etc.* » Cette reconnaissance de mon rôle s'inscrivait dans une fonction stratégique au sein de *QIA*, visant à faciliter la pénétration du marché RH.

Cette position de référente RH, intégrée au projet selon une stratégie d'intégration académique, revêt également une dimension symbolique. Mon affiliation à l'ESCP Business School en tant que doctorante était perçue comme un gage de légitimité et de crédibilité, renforçant ainsi la validité du processus de construction des données RH et contribuant à leur singularisation. Comme en témoignent des extraits de la plaquette de l'offre RH : *« Interventions régulières de nos experts dans l'enseignement* […] *au sein de l'ESCP Business School* » ou encore « *Association aux programmes de recherche* ». Ce positionnement institutionnel permettait d'ancrer la reconnaissance et la différenciation des données RH sur le marché.

Bien que mon implication ait été active et visible à certains moments, jai délibérément choisi de m'effacer des résultats présentés dans cette thèse, qui exposent les arbitrages relatifs à la construction des données RH. Cette décision était motivée par le désir de maintenir une posture analytique distanciée, tout en reconnaissant que ma participation, que ce soit au niveau de la qualification (participation directe) ou de la capitalisation et de la requalification (participations indirectes) des données RH, aurait pu introduire des biais ou influencer les résultats. L'influence réciproque entre l'équipe et moi a certainement façonné les dynamiques observées, rendant complexe la dissociation entre mon rôle de chercheuse et celui de membre actif de l'équipe.

4.3. Perspectives de recherche additionnelles

Afin de dépasser les limites identifiées et d'enrichir les connaissances à l'intersection des domaines des SI et de GRH, deux perspectives de recherche additionnelles peuvent être envisagées. Ces perspectives se divisent également en deux grandes catégories : théorique et empirique.

4.3.1. Perspective théorique : la qualification des « bonnes » données RH

Le processus de construction met en lumière que le choix des qualités « débattues », « contestées » et « négociées » des données RH n'est moralement pas neutre. En conséquence, une perspective de recherche théorique intéressante serait d'approfondir la question de la valeur morale qui émerge des « bonnes » données RH dans le cadre de l'étude d'une activité ou d'un phénomène RH. Elle pourrait être réalisée par le biais d'une revue de littérature systématique visant à définir ce qu'est une « bonne » donnée. Une telle revue permettrait non seulement de clarifier ce concept, mais également de porter un regard critique sur les notions émergentes telles que le « soin » des données (Loukissas, 2019; Zakharova & Jarke, 2024). Ces nouvelles notions sont sujettes à de nombreuses interprétations, ce qui peut rendre difficile leur compréhension et leur application. Une analyse approfondie pourrait aider à mieux définir et comprendre leur rôle dans la qualification des « bonnes » données RH.

4.3.2. Perspective empirique : la fétichisation des données RH par leur accumulation

La littérature sur les données RH en tant qu'instruments illustre la diversité des débats sur l'accumulation des données ainsi que la volonté, à la fois interne et externe à l'organisation, de développer une GRH guidée par et pour ses données. Alignée avec cette littérature (Sadowski, 2019), la requalification *extensive externe* et *interne* des données RH issue de mes résultats révèle une tendance à la valorisation excessive des données par leur accumulation. Cette accumulation sans discernement conduit à un phénomène que l'on pourrait qualifier de « blanchiment éthique », où les considérations morales sur les données sensibles, telles que les données RH, sont négligées au profit du simple amassage de données.

Pour aborder cette problématique, une recherche empirique pourrait explorer les effets de cette accumulation sur les qualités des données. Cette exploration pourrait se concentrer sur la manière dont les données sont collectées, stockées et utilisées, ainsi que sur les implications de ces pratiques pour les salariés concernés voire pour la société dans son ensemble. Une meilleure qualification des données, tenant compte de leur valeur morale, pourrait ainsi être promue comme une alternative à la simple accumulation quantitative.

5. Conclusion

Ce sixième chapitre est consacré à la discussion théorique et empirique des résultats obtenus dans le cadre de cette thèse, ainsi qu'à la présentation des différentes contributions, des limites et des perspectives de recherche.

De la discussion théorique développée dans cette thèse émergent trois contributions principales :

1. La conceptualisation des données RH en tant qu'objets d'étude en GRH ;
2. La conceptualisation des dispositifs socio-numériques des données RH à travers la spirale *data-driven* ;
3. La théorisation du processus de construction des données RH articulée en trois séquences (*QCR*).

L'articulation de ces contributions met en évidence les dynamiques complexes et interdépendantes inhérentes à l'étude des données RH. Elle souligne l'importance d'une approche intégrée qui prend en compte les données ainsi que les dispositifs socio-numériques qui participent à leur construction.

En outre, cette thèse a également permis de dégager trois contributions empiriques à différents niveaux :

1. Pour la gestion de projet : en identifiant les facteurs limitants et les facteurs de succès du processus de construction des données RH ;
2. Pour les *data scientists* : en renforçant la sensibilisation et la formation en GRH afin de garantir une contextualisation adéquate des données RH ;
3. Pour la fonction RH : en garantissant une sensibilisation et une formation en *data science* pour assurer un contrôle adéquat sur la construction des données RH.

Ces contributions montrent que la valeur des données dépasse leur seule dimension technique et repose avant tout sur l'intégrité et la cohérence des dispositifs socio-numériques qui les encadrent. Cela met en évidence l'importance d'une approche collaborative et interdisciplinaire, mobilisant conjointement les expertises en *data science* et en GRH, afin d'optimiser leur construction et d'assurer leur pertinence contextuelle.

Cette thèse présente également des limites théoriques et empiriques. Sur le plan théorique, elle questionne la dilution du pouvoir entre les agents économiques du réseau, une dimension que l'ANT n'aborde pas directement et qui est souvent critiquée pour sa tendance à attribuer le pouvoir de manière équivalente à tous les acteurs, qu'ils soient humains ou non-humains.

Empiriquement, mon implication active sur le terrain a influencé les dynamiques observées, rendant complexe la distinction entre mon rôle de chercheuse et celui de membre actif de l'équipe. Cette position, reconnue comme stratégique au sein de *QIA*, visait à faciliter la pénétration du marché RH, illustrant ainsi l'enchevêtrement entre ma posture et les logiques de valorisation des données RH.

Enfin, les perspectives de recherche proposent des pistes de réflexion théoriques et empiriques complémentaires. D'un point de vue théorique, il s'agit d'examiner la valeur morale associée à la qualification des « bonnes » données RH dans l'analyse

d’activités ou de phénomènes RH. Empiriquement, mes résultats révèlent une tendance à la fétichisation des données RH par leur accumulation, souvent opérée sans discernement. Ce phénomène peut conduire à un « blanchiment éthique », où les considérations morales liées aux données sensibles, comme les données RH, sont éclipsées par une logique d’accumulation. Ces deux perspectives convergent, offrant des opportunités pour approfondir la compréhension et l’usage des données RH dans un contexte de marchandisation.

Conclusion

Les données étaient autrefois perçues comme des entités « naturelles », destinées uniquement à la production de connaissances. Cette vision a cependant évolué avec la révolution numérique, qui a profondément transformé notre rapport à ces dernières.

Des volumes considérables de données sont désormais générés quotidiennement, constituant à la fois le moteur et le produit de cette révolution. En parallèle, elle permet l'émergence de nouveaux acteurs - tels que les plateformes et les cabinets de conseil spécialisés -, dont les modèles économiques reposent essentiellement sur les données.

Les données dépassent aujourd'hui leur valeur initiale pour être perçues comme des actifs et des marchandises, échangés et vendus de façon intensive. En occupant un rôle central dans la dynamique économique - notamment en facilitant la production et la circulation du capital - elles induisent également une transformation des modalités de production des connaissances. Ce phénomène a des répercussions sur l'ensemble des domaines, la GRH ne faisant pas exception.

Dans le prolongement de ces études, cette thèse visait à transcender le déterminisme technique souvent associé à la GRH, en illustrant que les données RH ne sont pas simplement des entités « données ». Loin d'être brutes, elles émergent de pratiques - calculatoires ou non -, à la fois discrétionnaires et situées.

Compte tenu des dynamiques impulsées par la révolution numérique et du rôle de plus en plus prépondérant des données dans la GRH, cette thèse s'est focalisée sur une question centrale :

« ***Comment les données RH sont-elles construites dans un contexte de marchandisation ?*** »

Pour répondre à cette question, j'ai d'abord effectué une revue de littérature centrée sur la conceptualisation des données RH. Cette analyse avait pour but de définir et d'examiner l'écosystème entourant les données RH, ainsi que les contextes et infrastructures dans lesquels elles évoluent.

En rapprochant les corpus de littérature en SI et en GRH j'ai pu conceptualiser la spirale *data-driven* structurée en quatre actes :

1. Définition des données RH ;
2. Transformation des données RH ;
3. Valorisation des données RH ;
4. Construction des données RH.

Afin de saisir la complexité des données RH et de leurs dispositifs socio-numériques, j'ai ensuite développé un cadre théorique qui combine la production de connaissances (Latour, 2005a, 2007) et l'économie des qualités (Callon et al., 2000, 2002). Cette synergie conceptuelle - désignée sous le terme de processus *QCR* - décrit la manière dont les données RH sont construites pour acquérir le statut de biens économiques. Elle s'articule autour de trois séquences clés :

1. La Qualification : concerne l'évaluation des qualités des données RH. Il s'agit de déterminer les critères initiaux qui définissent leur singularité potentielle.
2. La Capitalisation : examine les projets qui s'appuient sur la qualification pour développer des connaissances. L'objectif est de concevoir la fonction épistémique des données RH, permettant leur singularisation en tant que biens économiques.
3. La Requalification : consiste à requalifier les données RH à partir des projets de capitalisation.

Pour mettre en œuvre et valider ce cadre théorique, l'étude s'est appuyée sur une méthodologie qualitative et séquentielle, s'inscrivant dans un continuum entre recherche-action et ethnographie. Cette approche a été rendue possible grâce à une immersion de trois ans et sept mois au sein d'un cabinet de conseil en *data science*, dans le cadre d'une convention *CIFRE*. Le processus de construction des données RH est analysé à travers la conception de *DSN Analytics*, un outil d'IA destiné à l'analyse de l'absentéisme.

L'application du processus de construction des données RH a ensuite été formalisée autour des trois séquences clés - conformément à la théorisation proposée - chacune constituant un chapitre de résultats distinct :

1. La Qualification : met en lumière la qualité *normative* des données RH.

Cette qualité permet un traitement fiable et économique en raison de sa standardisation à l'échelle nationale. Elle se trouve toutefois au cœur de trois

controverses, dont les compromis incarnent une stratégie de marchandisation axée sur une optimisation efficiente.

2. La Capitalisation : met en évidence trois projets de connaissances distincts, chacun caractérisé par un type spécifique de connaissances sur l'absentéisme :
    1. L'absentéisme *réel univarié* : connaissances *descriptives* ;
    2. L'absentéisme *réel multivarié* : connaissances *explicatives* ;
    3. L'absentéisme *latent* : connaissances *prédictives*.

Ces projets se situent au centre de deux controverses, dont les compromis reflètent une stratégie de marchandisation axée sur un ajustement progressif.

3. La Requalification : qui révèle la qualité *extensive* des données RH, se manifestant sous deux formes : *externe* et *interne*.

La requalification *extensive externe* des données RH nécessite un accès élargi aux données *normatives* ainsi que l'élaboration de conventions de mesure spécifiques. La requalification *interne*, pour sa part, exige un accès à diverses sources de données.

La qualité *extensive* des données RH se trouve également au cœur de quatre controverses, dont les compromis incarnent une stratégie de marchandisation fondée sur une sélection contrainte.

Les résultats obtenus mettent en évidence que les données RH, par leurs qualités *normative* et *extensive*, s'intègrent dans une dynamique capitaliste d'accumulation et de circulation. Initialement extraites de leur contexte d'origine - la gestion administrative -, ces données acquièrent une valeur marchande qui dépasse leur usage initial. Ce phénomène de décontextualisation conduit les données RH à devenir progressivement l'élément central de l'offre commerciale, reléguant ainsi l'absentéisme à un rôle secondaire. Cette transformation illustre non seulement leur statut de biens économiques mais aussi la spirale *data-driven*, telle que définie dans la littérature existante.

Sur la base de ce constat et, plus généralement, de l'ambition de cette thèse de déconstruire le mythe des données « naturelles », trois contributions théoriques ont pu être dégagées :

1. La conceptualisation des données RH en tant qu'objets d'étude en GRH ;

2. La conceptualisation des dispositifs socio-numériques des données RH à travers la spirale *data-driven* ;
3. La théorisation du processus de construction des données RH articulée en trois séquences (*QCR*).

L'articulation de ces contributions met en évidence les dynamiques complexes et interdépendantes inhérentes à l'étude des données RH. Elle souligne ainsi l'importance d'une approche intégrée qui prend en compte les données et les dispositifs socio-numériques participant à leur construction.

En outre, cette étude a également permis de mettre en évidence trois contributions empiriques, offrant des perspectives variées sur la construction des données RH :

1. Pour la gestion de projet : en identifiant les facteurs limitants et les facteurs de succès du processus de construction des données RH ;
2. Pour les *data scientists* : en renforçant la sensibilisation et la formation en GRH afin de garantir une contextualisation adéquate des données RH ;
3. Pour la fonction RH : en garantissant une sensibilisation et une formation en *data science* pour assurer un contrôle adéquat sur la construction des données RH.

Afin de bien contextualiser l'ensemble de ces contributions, j'ai identifié les limites liées à cette recherche, réparties en deux grandes catégories : la première est de nature théorique et découle de la dilution du pouvoir entre les agents économiques au sein du réseau de construction des données RH. La seconde est de nature empirique et concerne l'influence de mon implication active dans l'étude.

Enfin, deux perspectives de recherche ont été envisagées, réparties en deux catégories principales : la première, théorique, interroge la qualification des « bonnes » données RH, tandis que la seconde, empirique, examine la fétichisation des données RH à travers leur accumulation.

En conclusion, le processus *QCR* peut être envisagé comme applicable à l'ensemble des données. En effet, le phénomène de décontextualisation remet en question la signification du terme « RH », censé qualifier ces données. Dans cette étude, le terme « RH » - associé à la fonction RH en raison du contexte - apparaît *fragmenté* lors de la qualification, *fantomatique* au moment de la capitalisation, et *embryonnaire* durant la requalification. Ces trois stades mettent en évidence les

difficultés à stabiliser une définition univoque des qualités « RH » des données. Ils révèlent également les défis auxquels la fonction RH se heurte pour légitimer son rôle dans la construction de ses propres données. Cela amène à se demander si une essence « RH » des données existe réellement, ou si elle n'est qu'une construction contingente et contextuelle façonnée par les agents économiques et les dynamiques du réseau.

Ainsi, l'étude des données RH invite à s'interroger sur le rôle de la fonction RH dans leur construction, révélant la fragilité de son intégration. Pour renforcer l'intégrité et la pertinence contextuelle des données RH, une recontextualisation fondée sur une approche collaborative et interdisciplinaire s'avère nécessaire. Dans ce cadre, la fonction RH joue un rôle central pour garantir que cette construction reflète ses réalités et réponde aux enjeux éthiques, techniques et sociaux auxquels elle fait face.

Bibliographie

Aaltonen, A., & Tempini, N. (2014). Everything counts in large amounts: A critical realist case study on data-based production. *Journal of Information Technology, 29*(1), 97–110. https://doi.org/10.1057/jit.2013.29

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy, 47*, 1–6. https://doi.org/10.1016/j.infoecopol.2019.05.001

Akrich, M., Callon, M., & Latour, B. (2006). *Sociologie de la traduction : Textes fondateurs*. Presses des Mines.

Alaimo, C., & Kallinikos, J. (2017). Computing the everyday: Social media as data platforms. *The Information Society, 33*, 175–191. https://doi.org/10.1080/01972243.2017.1318327

Alaimo, C., & Kallinikos, J. (2022). Organizations decentered: Data objects, technology and knowledge. *Organization Science, 33*(1), 19–37. https://doi.org/10.1287/orsc.2021.1552

Alaimo, C., & Kallinikos, J. (2024). Introduction. Dans *Data Rules: Reinventing the Market Economy* (pp. 1–18). MIT Press.

Alaimo, C., Kallinikos, J., & Aaltonen, A. (2020). Data and value. Dans S. Nambisan, K. Lyytinen, & Y. Yoo (dir.), *Handbook of Digital Innovation* (pp. 162–178). Edward Elgar Publishing.

Alavi, M., & Leidner, D. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 1*(10), 107–136. https://doi.org/10.2307/3250961

Alvesson, M., & Gabriel, Y. (2013). Beyond formulaic research: In praise of greater diversity in organizational research and publications. *Academy of Management Learning & Education, 12*(2), 245–263. http://www.jstor.org/stable/43696557

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. *Human Resource Management Journal, 26*(1), 1–11. https://doi.org/10.1111/1748-8583.12090

Atasoy, Ö., & Morewedge, C. (2017). Digital goods are valued less than physical goods. *Journal of Consumer Research, 44*(6), 1343–1357. https://doi.org/10.1093/jcr/ucx102

Baker, T., & Nelson, R. E. (2005). Creating something from nothing: Resource construction through entrepreneurial bricolage. *Administrative Science Quarterly, 50*(3), 329–366. https://doi.org/10.2189/asqu.2005.50.3.329

Belizón, M. J., & Kieran, S. (2022). Human resources analytics: A legitimacy process. *Human Resource Management Journal, 32*(3), 603–630. https://doi.org/10.1111/1748-8583.12417

Bell, E., & Willmott, H. (2020). Ethics, politics and embodied imagination in crafting scientific knowledge. *Human Relations, 73*(10), 1366–1387. https://doi.org/10.1177/0018726719876687

Bernon, J., & Pertinant, G. (2023). Les indicateurs fondamentaux. Dans *Le sens de l'absence : Prévenir durablement l'absentéisme* (pp. 143–153). ObjectifQVT Éditions.

Billon, J. (2024, 29 février). Les réseaux sociaux en France : Nombre d'utilisateurs, temps passé, usages… *Blog du Modérateur.* https://www.blogdumoderateur.com/classement-reseaux-sociaux-france-2024/.

Bolin, G. (2022). The value dynamics of data capitalism: Cultural production and consumption in a datafied world. Dans A. Hepp, J. Jarke, & L. Kramp (dir.), *New Perspectives in Critical Data Studies: The Ambivalences of Data Power* (pp. 167–186). Springer International Publishing.

Borgman, C. L. (2017a). Provocations. Dans *Big Data, Little Data, No Data: Scholarship in the Networked World* (pp. 1–16). MIT Press.

Borgman, C. L. (2017b). What are data. Dans *Big Data, Little Data, No Data: Scholarship in the Networked World* (pp. 17–29). MIT Press.

Bowker, G. C., & Star, S. L. (1999). Why classifications matter. Dans *Sorting Things Out: Classification and Its Consequences* (pp. 319–326). MIT Press.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bradbury-Huang, H. (2010). What is good action research? Why the resurgent interest? *Action Research, 8*(1), 93–109. https://doi.org/10.1177/1476750310362435

Broek, E., Sergeeva, A., & Huysman, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly, 45*(3), 1557–1580. https://doi.org/10.25300/MISQ/2021/16559

Brown, J. S., & Duguid, P. (2000). Limits to information. Dans *The Social Life of Information* (pp. 11–34). Harvard Business Review Press.

Cabantous, L., & Gond, J.-P. (2012). Du mode d'existence des théories dans les organisations : La fabrique de la décision comme praxis performative. *Revue française de gestion, 225*(6), 61–81. https://doi.org/10.3166/RFG.225.61-81

Cadin, L., Guérin, F., & Pigeyre, F. (2012). Le capitalisme : Cadre(s) de la gestion des ressources humaines. Dans *Pratiques et éléments de théorie GRH - Gestion des ressources humaines* (4e éd., pp. 27–78). Dunod.

Caldwell, R. (2003). The changing roles of personnel managers: Old ambiguities, new uncertainties. *Journal of Management Studies, 40*(4), 983–1004. https://doi.org/10.1111/1467-6486.00367

Callon, M. (1986). Éléments pour une sociologie de la traduction : La domestication des coquilles Saint-Jacques et des marins-pêcheurs dans la baie de Saint-Brieuc. *L'Année sociologique, 36*, 169–208.

Callon, M., & Ferrary, M. (2006). Les réseaux sociaux à l'aune de la théorie de l'acteur-réseau. *Sociologies pratiques, 13*(2), 37–44. https://doi.org/10.3917/sopr.013.0037

Callon, M., Méadel, C., & Rabeharisoa, V. (2000). L'économie des qualités. *Politix, 13*(52), 211–239. https://doi.org/10.3406/polix.2000.1126

Callon, M., Méadel, C., & Rabeharisoa, V. (2002). The economy of qualities. *Economy and Society, 31*(2), 194–217. https://doi.org/10.1080/03085140220123126

Callon, M., & Muniesa, F. (2005). Peripheral vision: Economic markets as calculative collective devices. *Organization Studies, 26*(8), 1229–1250. https://doi.org/10.1177/0170840605056393

Cappelli, P. (2017). There's no such thing as big data in HR. *Harvard Business Review.* https://hbr.org/2017/06/theres-no-such-thing-as-big-data-in-hr

Cardon, D. (2019). Introduction. Dans *Culture numérique* (pp. 5–13). Presses de Sciences Po.

Carter, D. (2018). Reimagining the big data assemblage. *Big Data & Society, 5*(2). https://doi.org/10.1177/2053951718818194

Chalutz Ben-Gal, H. (2019). An ROI-based review of HR analytics: Practical implementation tools. *Personnel Review, 48*(6), 1429–1448. https://doi.org/10.1108/PR-11-2017-0362

Chandler, A. D. (1990). Introduction: Scale and scope. Dans *Scale and Scope: The Dynamics of Industrial Capitalism* (pp. 1–46). Harvard Business Review Press.

Coron, C. (2019a). Analytique et big data en ressources humaines - Une étude au prisme de la notion de justification. *Revue Française de Gestion, 45*(280), 55–72. https://doi.org/10.3166/rfg.2019.00319

Coron, C. (2019b). Big data et pratiques de GRH. *Management & Data Science, 3*(1). https://halshs.archives-ouvertes.fr/halshs-01961214

Coron, C. (2019c). De la mise en statistique du travail aux algorithmes en ressources humaines : Les différents usages de la quantification. Dans *Quantification en ressources humaines. Usages et analyses* (Vol. 2, pp. 15–55). ISTE Éditions.

Coron, C. (2019d). Introduction. Dans *Quantification en ressources humaines. Usages et analyses* (Vol. 2, pp. 3–14). ISTE Éditions.

Coron, C. (2019e). Le « Big data RH » : Vers une nouvelle convention de quantification ? *Annales des Mines - Gérer et comprendre, 137*(3), 27–38. https://doi.org/10.3917/geco1.137.0027

Coron, C. (2019f). Les enjeux éthiques de la quantification. Dans *Quantification en ressources humaines. Usages et analyses* (Vol. 2, pp. 157–186). ISTE Éditions.

Coron, C. (2019g). Quantification et prise de décision. Dans *Quantification en ressources humaines. Usages et analyses* (Vol. 2, pp. 57–95). ISTE Éditions.

Coron, C. (2022). Quantifying human resource management : A literature review. *Personnel Review, 51*(4), 1386–1409. https://doi.org/10.1108/PR-05-2020-0322

Cousineau, L., Ollier-Malaterre, A., & Parent-Rocheleau, X. (2023). Employee surveillance technologies: Prevalence, classification, and invasiveness. *Surveillance & Society, 21*(4), 447–468. https://doi.org/10.24908/ss.v21i4.15763

Czarniawska, B. (2006). Book review: Bruno Latour: Reassembling the social: An introduction to actor-network theory. *Organization Studies, 27*(10), 1553–1557. https://doi.org/10.1177/0170840606071164

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical data studies: A dialog on data and space. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951716648346

Dambrin, C., & Grall, B. (2021). Bruno Latour, penseur des modes d'existence. Dans P. Gilbert & D. Mourey (dir.), *Philosophie et outils de gestion. Entre dévoilement des impensés et nouvelles potentialités de théorisation* (pp. 249–271). Éditions EMS.

Davenport, T. H. (2014). How strategists use "big data" to support internal business decisions, discovery and production. *Strategy & Leadership, 42*(4), 45–50. https://doi.org/10.1108/SL-05-2014-0034

Davenport, T. H., Harris, J., & Shapiro, J. (2010). Competing on talent analytics. *Harvard Business Review.* https://hbr.org/2010/10/competing-on-talent-analytics

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review.* https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Delbridge, R., & Fiss, P. C. (2013). Editors' comments: Styles of theorizing and the social organization of knowledge. *Academy of Management Review, 38*(3), 325–331. https://doi.org/10.5465/amr.2013.0085

Desrosières, A. (2013a). Classer et mesurer : Les deux faces de l'argument statistique. Dans *Pour une sociologie historique de la quantification* (pp. 120–141). Presses des Mines.

Desrosières, A. (2013b). Pour une politique des outils du savoir : Le cas de la statistique. In *Pour une sociologie historique de la quantification : L'Argument statistique I* (pp. 57–76). Presses des Mines.

Diaz-Bone, R., & Didier, E. (2016). The sociology of quantification - perspectives on an emerging field in the social sciences. *Historical Social Research, 41*(2), 7–26. https://doi.org/10.12759/hsr.41.2016.2.7-26

Dumez, H. (2011). L'Actor-Network-Theory (ANT) comme technologie de la description. *Le Libellio d'Aegis, 7*(4), 27–38. http://crg.polytechnique.fr/v2/aegis.html#libellio

Eden, C., & Huxham, C. (1996). Action research for management research. *British Journal of Management, 7*(1), 75–86. https://doi.org/10.1111/j.1467-8551.1996.tb00107.x

Eisenhardt, K. M. (1989). Building theories from case study research. *The Academy of Management Review, 14*(4), 532–550. https://doi.org/10.2307/258557

Eisenhardt, K. M. (2021). What is the Eisenhardt method, really? *Strategic Organization, 19*(1), 147–160. https://doi.org/10.1177/1476127020982866

Eisenhardt, K. M., Graebner, M. E., & Sonenshein, S. (2016). Grand challenges and inductive methods: Rigor without rigor mortis. *Academy of Management Journal, 59*(4), 1113–1123. https://doi.org/10.5465/amj.2016.4004

Eisenstein, E. L. (1980). The unacknowledged revolution. Dans *The Printing Press as an Agent of Change* (pp. 3–42). Cambridge University Press.

Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology, 24*, 313–343. https://www.jstor.org/stable/223484

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization, 28*(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Fitz-enz, J. (2010). *The new HR analytics: Predicting the economic value of your company's human capital investments.* American Management Association.

Floridi, L. (2010). *Information: A Very Short Introduction.* Oxford University Press.

Flyverbom, M. (2019). Digital and datafied spaces. Dans *The Digital Prism: Transparency and Managed Visibilities in a Datafied World* (pp. 25–38). Cambridge University Press.

Fourcade, M., & Healy, K. (2016). Seeing like a market. *Socio-Economic Review, 15*(1), 9–19. https://doi.org/10.1093/ser/mww033

Garcia-Arroyo, J., & Osca, A. (2019). Big data contributions to human resource management: A systematic review. The International Journal of Human Resource Management, 32(20), 4337–4362. https://doi.org/10.1080/09585192.2019.1674357

Gilbert, P. (2021). L'instrumentation de GRH. Penser l'instrument pour penser la gestion. Dans *Les grands courants en gestion des ressources humaines* (pp. 288–304). EMS Éditions.

Gillespie, T. (2014). The relevance of algorithms. Dans T. Gillespie, P. J. Boczkowski, & K. A. Foot (dir.), *Media Technologies : Essays on Communication, Materiality, and Society* (pp. 167–193). The MIT Press.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods, 16*(1), 15–31. https://doi.org/10.1177/1094428112452151

Gitelman, L., & Jackson, V. (2013). Introduction. Dans L. Gitelman (dir.), *« Raw Data » Is an Oxymoron* (pp. 1–14). The MIT Press.

Greasley, K., & Thomas, P. (2020). HR analytics: The onto-epistemology and politics of metricised HRM. *Human Resource Management Journal, 30*(4), 494–507. https://doi.org/10.1111/1748-8583.12283

Greenwood, D. J., & Levin, M. (2007). *Introduction to Action Research: Social Research for Social Change* (2e éd.). SAGE Publications.

Harley, B. (2015). The one best way? 'Scientific' research on HRM and the threat to critical scholarship. *Human Resource Management Journal, 25*(4), 399–407. https://doi.org/10.1111/1748-8583.12082

Hermans, M., & Ulrich, M. D. (2021). How symbolic human resource function actions affect the implementation of high-performance work practices: The mediating effect of influence on strategic decision-making. *Human Resource Management Journal, 31*(4), 1063–1081. https://doi.org/10.1111/1748-8583.12361

Herr, K., & Anderson, G. (2015). *The Action Research Dissertation: A Guide for Students and Faculty* (2e éd.). SAGE Publications.

Hinds, P., & Kiesler, S. (2002). *Distributed Work.* MIT Press.

Hinings, B., Gegenhuber, T., & Greenwood, R. (2018). Digital innovation and transformation: An institutional perspective. *Information and Organization, 28*(1), 52–61. https://doi.org/10.1016/j.infoandorg.2018.02.004

Kallinikos, J. (2009). On the computational rendition of reality: Artefacts and human agency. *Organization, 16*(2), 183–202. https://doi.org/10.1177/1350508408100474

Kallinikos, J., Aaltonen, A., & Marton, A. (2013). The ambivalent ontology of digital artifacts. *MIS Quarterly, 37*(2), 357–370. https://www.jstor.org/stable/43825913

Kapoor, B., & Kabra, Y. (2016). Current and future trends in human resources analytics adoption. *Journal of Cases on Information Technology, 16*(1), 50–59. https://doi.org/10.4018/jcit.2014010105

Kellogg, K. C., Valentine, M. A., & Christin, A. (2019). Algorithms at work: The new contested terrain of control. *Academy of Management Annals, 14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1–12. https://doi.org/10.1177/2053951714528481

Kitchin, R. (2022a). Business. Dans *The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures* (2e éd., pp. 143–158). SAGE Publications.

Kitchin, R. (2022b). Data analytics. Dans *The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures* (2e éd., pp. 97–111). SAGE Publications.

Kitchin, R. (2022c). Data ethics and data governance. Dans *The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures* (2e éd., pp. 267–283). SAGE Publications.

Kitchin, R. (2022d). Introducing data. Dans *The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures* (2e éd., pp. 1–20). SAGE Publications.

Kitchin, R. (2022e). Small data and data infrastructure. Dans *The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures* (2e éd., pp. 45–59). SAGE Publications.

Kitchin, R. (2022f). The epistemology of academic research. Dans *The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures* (2e éd., pp. 113–126). SAGE Publications.

Kitchin, R., & Lauriault, T. (2018). Toward critical data studies: Charting and unpacking data assemblages and their work. Dans J. Eckert, A. Shears, & J. Thatcher (dir.), *Thinking Big Data in Geography: New Regimes, New Research* (pp. 3–20). University of Nebraska Press.

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal, 80*(4), 463–475. https://doi.org/10.1007/s10708-014-9601-7

Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951716631130

Kowu, K. (2024). *Ressources humaines & digital : données de marché de 2022 à 2027.* Markess by Exaegis. https://www.markess.com/rh/ressources-humaines-digital-donnees-de-marche-de-2022-a-2027/

LaForgia, M., Rosenberg, M., & Dance, G. J. X. (2019, 13 mars). Facebook's data deals are under criminal investigation. *The New York Times.* https://www.nytimes.com/2019/03/13/technology/facebook-data-deals-investigation.html

Laney, D. (2001, 6 février). *3-D data management: Controlling data volume, velocity and variety.* Application Delivery Strategies, META Group, 949. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Langley, A. (1999). Strategies for theorizing from process data. *The Academy of Management Review, 24*(4), 691–710. https://doi.org/10.2307/259349

Langley, A., & Abdallah, C. (2011). Templates and turns in qualitative studies of strategy and management. Dans D. D. Bergh & D. J. Ketchen (dir.), *Building Methodological Bridges* (Vol. 6, pp. 201–235). Emerald.

Latour, B. (1992). *Aramis ou l'amour des techniques.* La Découverte.

Latour, B. (2005a). *La science en action : Introduction à la sociologie des sciences* (traduit par M. Biezunski). La Découverte. (Ouvrage original publié en 1987).

Latour, B. (2005b). On the difficulty of being an ANT: An interlude in the form of a dialog. Dans *Reassembling the Social: An Introduction to Actor-Network-Theory* (pp. 141–156). Oxford University Press.

Latour, B. (2007). A textbook case revisited. Knowledge as mode of existence. Dans E. J. Hackett, M. Lynch, J. Wajcman, & O. Amsterdamska (dir.), *The Handbook of Science and Technology Studies* (3e éd., pp. 83–112). The MIT Press.

Latour, B. (2012). *Enquête sur les modes d'existence : Une anthropologie des Modernes.* La Découverte.

Lawler III, E. E., Levenson, A., & Boudreau, J. W. (2004). HR metrics and analytics: Use and impact. *Human Resource Planning, 27*(4), 27–35.

Legner, C., Eymann, T., Hess, T., Matt, C., Böhmann, T., Drews, P., Mädche, A., Urbach, N., & Ahlemann, F. (2017). Digitalization: Opportunity and challenge for the business and information systems engineering community. *Business & Information Systems Engineering, 59*(4), 301–308. https://doi.org/10.1007/s12599-017-0484-2

Leonardi, P. M., & Treem, J. W. (2020). Behavioral visibility: A new paradigm for organization studies in the age of digitization, digitalization, and datafication. *Organization Studies, 41*(12), 1601–1625. https://doi.org/10.1177/0170840620970728

Leonelli, S. (2016). Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160122. https://doi.org/10.1098/rsta.2016.0122

Leung, E., Cito, M. C., Paolacci, G., & Puntoni, S. (2022). Preference for material products in identity-based consumption. *Journal of Consumer Psychology, 32*(4), 672–679. https://doi.org/10.1002/jcpy.1272

Lévi-Strauss, C. (1962). La science du concret. Dans *La pensée sauvage* (pp. 11–49). Plon.

Lewis, P. (2007). Using ANT ideas in the managing of systemic action research. *Systems Research and Behavioral Science, 24*(6), 589–598. https://doi.org/10.1002/sres.832

Lindén, L. (2021). Moving evidence: Patients' groups, biomedical research, and affects. *Science, Technology, & Human Values, 46*(4), 815–838. https://doi.org/10.1177/0162243920948126

Lismont, J., Vanthienen, J., Baesens, B., & Lemahieu, W. (2017). Defining analytics maturity indicators: A survey approach. *International Journal of Information Management, 37*(3), 114–124. https://doi.org/10.1016/j.ijinfomgt.2016.12.003

Loukissas, Y. A. (2019). Introduction. Dans *All Data Are Local* (pp. 1–12). The MIT Press.

Lupton, D. (2020). Thinking with care about personal data profiling: A more-than-human approach. *International Journal of Communication, 14*. https://ijoc.org/index.php/ijoc/article/view/13540

Lusch, R., & Nambisan, S. (2015). Service innovation: A service-dominant logic perspective. *MIS Quarterly, 39*(1), 155–175. https://doi.org/10.25300/MISQ/2015/39.1.07

Lüscher, L. S., & Lewis, M. W. (2008). Organizational change and managerial sensemaking: Working through paradox. *Academy of Management Journal, 51*(2), 221–240. https://doi.org/10.5465/amj.2008.31767217

Lycett, M. (2013). 'Datafication': Making sense of (big) data in a complex world. *European Journal of Information Systems, 22*(4), 381–386. https://doi.org/10.1057/ejis.2013.10

Madsen, D. Ø., & Slåtten, K. (2019). An examination of the current status and popularity of HR analytics. *International Journal of Strategic Management, 19*(2), 17–38. https://doi.org/10.18374/IJSM-19-2.2

Manroop, L., Malik, A., & Milner, M. (2024). The ethical implications of big data in human resource management. *Human Resource Management Review, 34*(2), 101012. https://doi.org/10.1016/j.hrmr.2024.101012

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011, juin). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute. https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20di

gital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_full_report.pdf

Margherita, A. (2022). Human resources analytics: A systematization of research topics and directions for future research. *Human Resource Management Review, 32*(2), 100795. https://doi.org/10.1016/j.hrmr.2020.100795

Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR analytics. *The International Journal of Human Resource Management, 28*(1), 3–26. https://doi.org/10.1080/09585192.2016.1244699

Marler, J. H., Cronemberger, F., & Tao, C. (2017). HR analytics: Here to stay or short-lived management fashion? Dans T. Bondarouk, H. J. M. Ruël, & E. Parry (dir.), *Electronic HRM in the Smart Era* (pp. 59–85). Emerald Publishing Limited.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. John Murray Publishers.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review.* https://hbr.org/2012/10/big-data-the-management-revolution

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951716650211

Miller, H. J. (2010). The data avalanche is here, shouldn't we be digging? *Journal of Regional Science, 50*(1), 181–201. https://doi.org/10.1111/j.1467-9787.2009.00641.x

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

Monteiro, E., & Parmiggiani, E. (2019). Synthetic knowing: The politics of the internet of things. *MIS Quarterly, 43*(1), 167–184. https://doi.org/10.25300/MISQ/2019/13799

Morozov, E. (2015, 23 juin). Digital technologies and the future of data capitalism. *Social Europe.* https://www.socialeurope.eu/digital-technologies-and-the-future-of-data-capitalism

Muniesa, F., Millo, Y., & Callon, M. (2007). An Introduction to Market Devices. The Sociological Review, 55(2_suppl), 1-12. https://doi.org/10.1111/j.1467-954X.2007.00727.x

Nafus, D., & Sherman, J. (2014). This one does not go up to 11: The Quantified Self movement as an alternative big data practice. *International Journal of Communication, 8*, 1784-1794. https://ijoc.org/index.php/ijoc/article/view/2170/1157

National Institute of Standards and Technology. (2019, 19 décembre). NIST study evaluates effects of race, age, sex on face recognition software. https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software

Nolin, J. M. (2019). Data as oil, infrastructure or asset? Three metaphors of data as economic value. *Journal of Information, Communication and Ethics in Society, 18*(1), 28–43. https://doi.org/10.1108/JICES-04-2019-0044

Normann, R. (2001a). Chained to the value chain? Dans *Reframing Business: When the Map Changes the Landscape* (pp. 49–60). John Wiley & Sons.

Normann, R. (2001b). The offering as a tool to organize coproduction. Dans *Reframing Business: When the Map Changes the Landscape* (pp. 113–129). John Wiley & Sons.

OCDE (2022). *Skills for the Digital Transition: Assessing Recent Trends Using Big Data*. https://doi.org/10.1787/38c36777-en

Olivier De Sardan, J.-P. (2018). Le « je » méthodologique. Implication et explicitation dans l'enquête de terrain. Dans *La rigueur du qualitatif : Les contraintes empiriques de l'interprétation socio-anthropologique* (pp. 165–207). Academia Bruylant.

Orlikowski, W. J., & Scott, S. V. (2015). Exploring material-discursive practices. *Journal of Management Studies, 52*(5), 697–705. https://doi.org/10.1111/joms.12114

Østerlie, T., & Monteiro, E. (2020). Digital sand: The becoming of digital representations. *Information and Organization, 30*(1), 100275. https://doi.org/10.1016/j.infoandorg.2019.100275

Pachidi, S., Berends, H., Faraj, S., & Huysman, M. (2021). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science, 32*(1), 18–41. https://doi.org/10.1287/orsc.2020.1377

Parker, G., Van Alstyne, M. W., & Jiang, X. (2016). Platform ecosystems: How developers invert the firm. *Boston University Questrom School of Business Research Paper*, 2861574. http://dx.doi.org/10.2139/ssrn.2861574

Parmiggiani, E., Østerlie, T., & Almklov, P. (2022). In the backrooms of data science. *Journal of the Association for Information Systems, 23*(1), 139–164. https://doi.org/10.17705/1jais.00718

Passi, S., & Jackson, S. (2017). Data vision: Learning to see through algorithmic abstraction. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2436–2447. https://doi.org/10.1145/2998181.2998331

Passi, S., & Jackson, S. J. (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction, 2*(136), 1–28. https://doi.org/10.1145/3274405

Passi, S., & Sengers, P. (2020). Making data science systems work. *Big Data & Society, 7*(2). https://doi.org/10.1177/2053951720939605

Pfeffer, J., & Sutton, R. I. (2006). Evidence-based management. *Harvard Business Review.* https://hbr.org/2006/01/evidence-based-management

Piovesan, F. (2022). Reflections on combining action research and actor-network theory. *Action Research, 20*(4), 363–379. https://doi.org/10.1177/1476750320919167

Porter, T. M. (1995). Introduction: Cultures of objectivity. Dans *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (pp. 3–8). Princeton University Press.

Power, M. (2023). Foreword. Dans C. Alaimo & J. Kallinikos (dir.), *Data rules: Reinventing the market economy* (pp. vii–ix)

Pratt, M. G., Sonenshein, S., & Feldman, M. S. (2022). Moving beyond templates: A bricolage approach to conducting trustworthy qualitative research. *Organizational Research Methods, 25*(2), 211–238. https://doi.org/10.1177/1094428120927466

Ranck, J. (2012). *The wearable computing market: A global analysis.* Gigaom Pro. https://www.shoutwiki.com/w/images/wearable/e/e4/Wearable-computing-the-next-big-thing-in-tech.pdf

Rasmussen, T., & Ulrich, D. (2015). Learning from practice: How HR analytics avoids being a management fad. *Organizational Dynamics, 44*(3), 236–242. https://doi.org/10.1016/j.orgdyn.2015.05.008

Resseguier, A., & Ufert, F. (2024). AI research ethics is in its infancy: The EU's AI Act can make it a grown-up. *Research Ethics, 20*(2), 143–155. https://doi.org/10.1177/17470161231220946

Ribes, D., & Jackson, S. J. (2013). Data bite man: The work of sustaining a long-term study. Dans L. Gitelman (dir.), *« Raw Data » Is an Oxymoron* (pp. 147–166). The MIT Press.

Ruppert, E., Isin, E., & Bigo, D. (2017). Data politics. *Big Data & Society, 4*(2). https://doi.org/10.1177/2053951717717749

Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. *Big Data & Society, 6*(1). https://doi.org/10.1177/2053951718820549

Saifer, A., & Dacin, M. T. (2022). Data and organization studies: Aesthetics, emotions, discourse and our everyday encounters with data. *Organization Studies, 43*(4), 623–636. https://doi.org/10.1177/01708406211006250

Saija, L. (2014). Writing about engaged scholarship: Misunderstandings and the meaning of "quality" in action research publications. *Planning Theory & Practice, 15*(2), 187–201. https://doi.org/10.1080/14649357.2014.904922

Saltz, J. S., & Grady, N. W. (2017). The ambiguity of data science team roles and the need for a data science workforce framework. *IEEE Big Data Conference Proceedings*, 2355–2361. https://doi.org/10.1109/BigData.2017.8258190

Sandberg, J., & Tsoukas, H. (2011). Grasping the logic of practice: Theorizing through practical rationality. *Academy of Management Review, 36*(2), 338–360. https://doi.org/10.5465/amr.2009.0183

Smaldone, F., Ippolito, A., Lagger, J., & Pellicano, M. (2022). Employability skills: Profiling data scientists in the digital labour market. *European Management Journal, 40*(5), 671–684. https://doi.org/10.1016/j.emj.2022.05.005

Strohmeier, S. (2020). Big HR data. Dans T. Bondarouk & S. Fisher (dir.), *Encyclopedia of Electronic HRM* (pp. 259–264). De Gruyter.

Sturdy, A., Clark, T., Fincham, R., & Handley, K. (2009). Between innovation and legitimation—Boundaries and knowledge flow in management consultancy. *Organization, 16*(5), 627–653. https://doi.org/10.1177/1350508409338435

Suchman, L. A., & Trigg, R. H. (1993). Artificial intelligence as craftwork. Dans J. Lave & S. Chaiklin (dir.), *Understanding Practice: Perspectives on Activity and Context* (pp. 144–178). Cambridge University Press.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *The Academy of Management Review, 20*(3), 571–610. https://doi.org/10.2307/258788

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*. https://doi.org/10.1177/0008125619867910

Tilson, D., Lyytinen, K., & Sørensen, C. (2010). Research commentary—Digital infrastructures: The missing IS research agenda. *Information Systems Research, 21*(4), 748–759. https://doi.org/10.1287/isre.1100.0318

Tootell, B., Blackler, M., Toulson, P., & Dewe, P. (2009). Metrics: HRM's holy grail? A New Zealand case study. *Human Resource Management Journal, 19*(4), 375–392. https://doi.org/10.1111/j.1748-8583.2009.00108.x

Treem, J. W., & Leonardi, P. M. (2012). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Communication Yearbook, 36*, 143–189. https://doi.org/10.2139/ssrn.2129853

Tsoukas, H., & Vladimirou, E. (2001). What is organizational knowledge? *Journal of Management Studies, 38*(7), 973–993. https://doi.org/10.1111/1467-6486.00268

Van den Heuvel, S., & Bondarouk, T. (2017). The rise (and fall?) of HR analytics: A study into the future application, value, structure, and system support. *Journal of Organizational Effectiveness: People and Performance, 4*(2), 157–178. https://doi.org/10.1108/JOEPP-03-2017-0022

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society, 12*(2), 197–208. https://doi.org/10.24908/ss.v12i2.4776

van Dijck, J. (2018). Platform mechanisms. Dans J. van Dijck, T. Poell, & M. de Waal (dir.), *The Platform Society* (pp. 31–48). Oxford University Press.

Weinberger, D. (2010). The problem with the data-information-knowledge-wisdom hierarchy. *Harvard Business Review.* https://hbr.org/2010/02/data-is-to-info-as-info-is-not

Woodcock, J. (2021). Workers inquiry and the experience of work: Using ethnographic accounts of the gig economy. Dans J. Aroles, F.-X. de Vaujany, & K. Dale (dir.), *The Cambridge Handbook of the Gig Economy* (pp. 136–156). Cambridge University Press.

Wright, C. (2008). Reinventing human resource management: Business partners, internal consultants and the limits to professionalization. *Human Relations, 61*(8), 1063–1086. https://doi.org/10.1177/0018726708094860

Yoo, Y., Henfridsson, O., & Lyytinen, K. (2010). Research commentary: The new organizing logic of digital innovation: An agenda for information systems research. *Information Systems Research, 21*(4), 724–735.

Zakharova, I., & Jarke, J. (2024). Care-ful data studies: Or, what do we see, when we look at datafied societies through the lens of care? *Information, Communication & Society, 27*(4), 651–664. https://doi.org/10.1080/1369118X.2024.2316758

Zhang, Y., Xu, S., Zhang, L., & Yang, M. (2021). Big data and human resource management research: An integrative review and new directions for future research. *Journal of Business Research, 133*, 34–50. https://doi.org/10.1016/j.jbusres.2021.04.019

Zittrain, J. (2008). Cybersecurity and the generative dilemma. In *The Future of the Internet and How to Stop It* (pp. 36–61). Yale University Press. https://dash.harvard.edu/handle/1/4455262

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology, 13*(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399

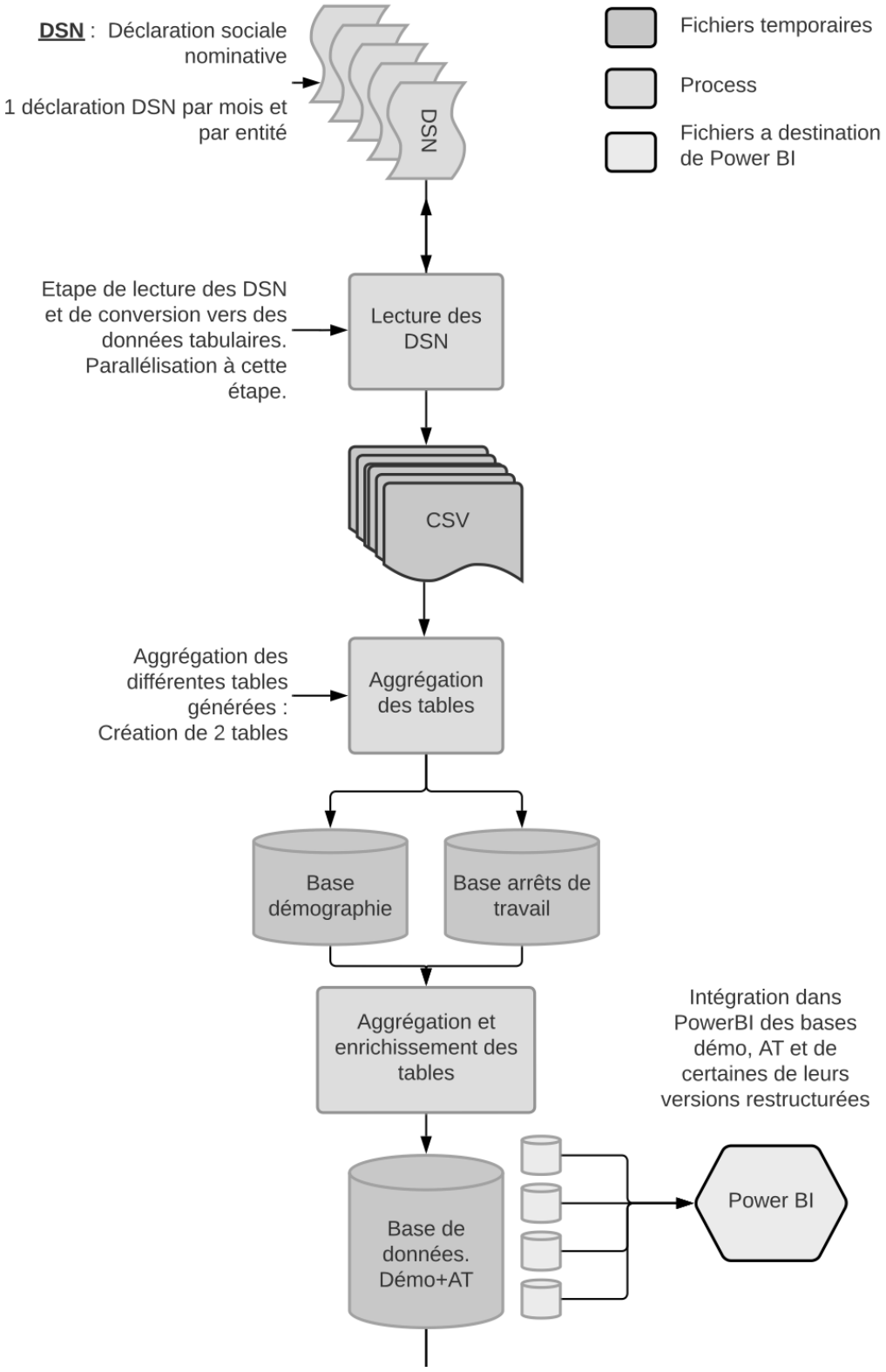
Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology, 30*(1), 75–89. https://doi.org/10.1057/jit.2015.5

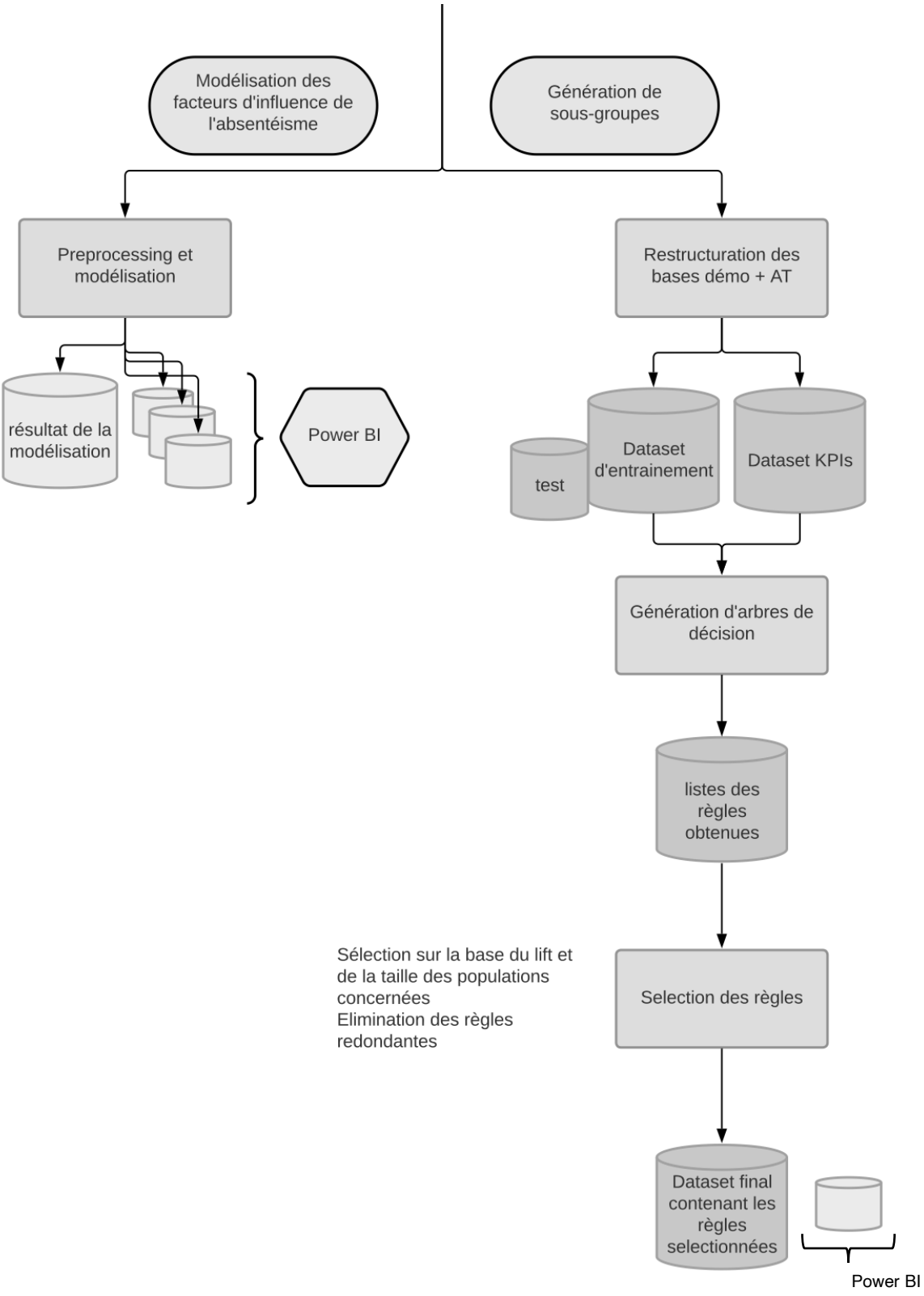
Zuboff, S. (2019). Surveillance capitalism and the challenge of collective action. *New Labor Forum, 28*(1), 10–29. https://doi.org/10.1177/1095796018819461

Zwitter, A. (2014). Big data ethics. *Big Data & Society, 1*(2), 2053951714559253. https://doi.org/10.1177/2053951714559253

Annexes

Annexe I. Processus de construction technique des données RH




Modélisation des facteurs d'influence de l'absentéisme
Génération de sous-groupes
Preprocessing et modélisation
Restructuration des bases démo + AT
résultat de la modélisation
Power BI
test
Dataset d'entrainement
Dataset KPIs
Génération d'arbres de décision
listes des règles obtenues
Sélection sur la base du lift et de la taille des populations concernées
Elimination des règles redondantes
Selection des règles
Dataset final contenant les règles selectionnées
Power BI

Annexe II. Représentations des réseaux d'alliances dans le processus de construction des données RH

Les Figure 17, Figure 18 et Figure 19 illustrent les réseaux d'alliances pour les trois séquences du processus de construction des données RH : la qualification, la capitalisation et la requalification. Ces réseaux s'organisent selon trois flux principaux :

1. (Re)définir les besoins des clients ;
2. (Re)rationaliser les coûts d'investissement ;
3. (Ré)enrôler des agents économiques.

En outre, ces réseaux mettent en évidence la diversité des agents et leurs modes d'existence respectifs dans le projet, tout en soulignant l'importance des intermédiaires qui facilitent les interactions au sein de ces réseaux.

Le mode d'existence Scientifique dans ces trois réseaux représente l'ensemble des agents ayant participé à la production et à la diffusion de connaissances spécialisées à l'intersection de la GRH et des données. Ces agents sont abordés dans les sections traitant des conséquences sur la fonction RH (voir chapitres de résultats).

1. Représentation du réseau d'alliances dans la qualification des données RH

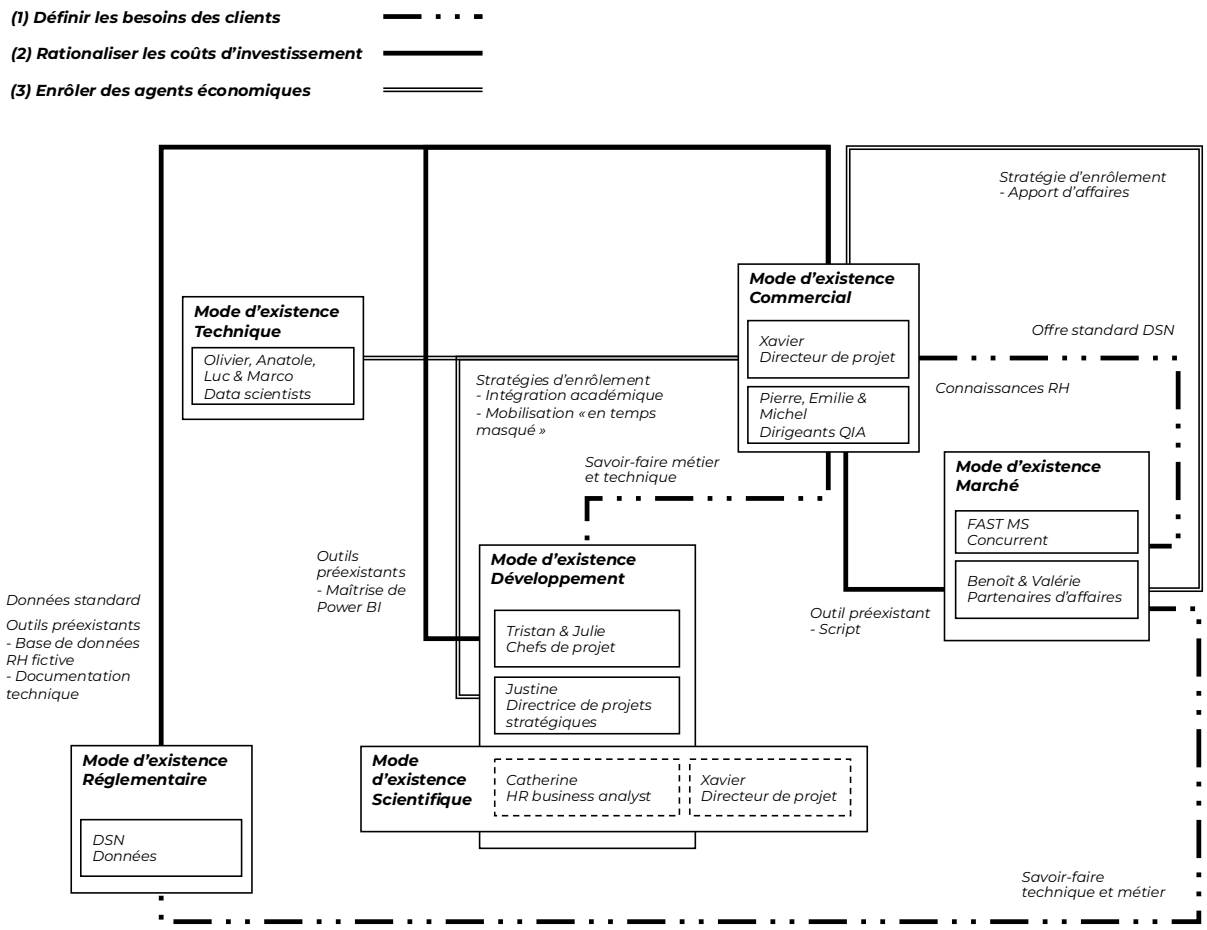


Figure 17 : Réseau d'alliances pour la séquence de Qualification

2. Représentation du réseau d'alliances dans la capitalisation des données RH

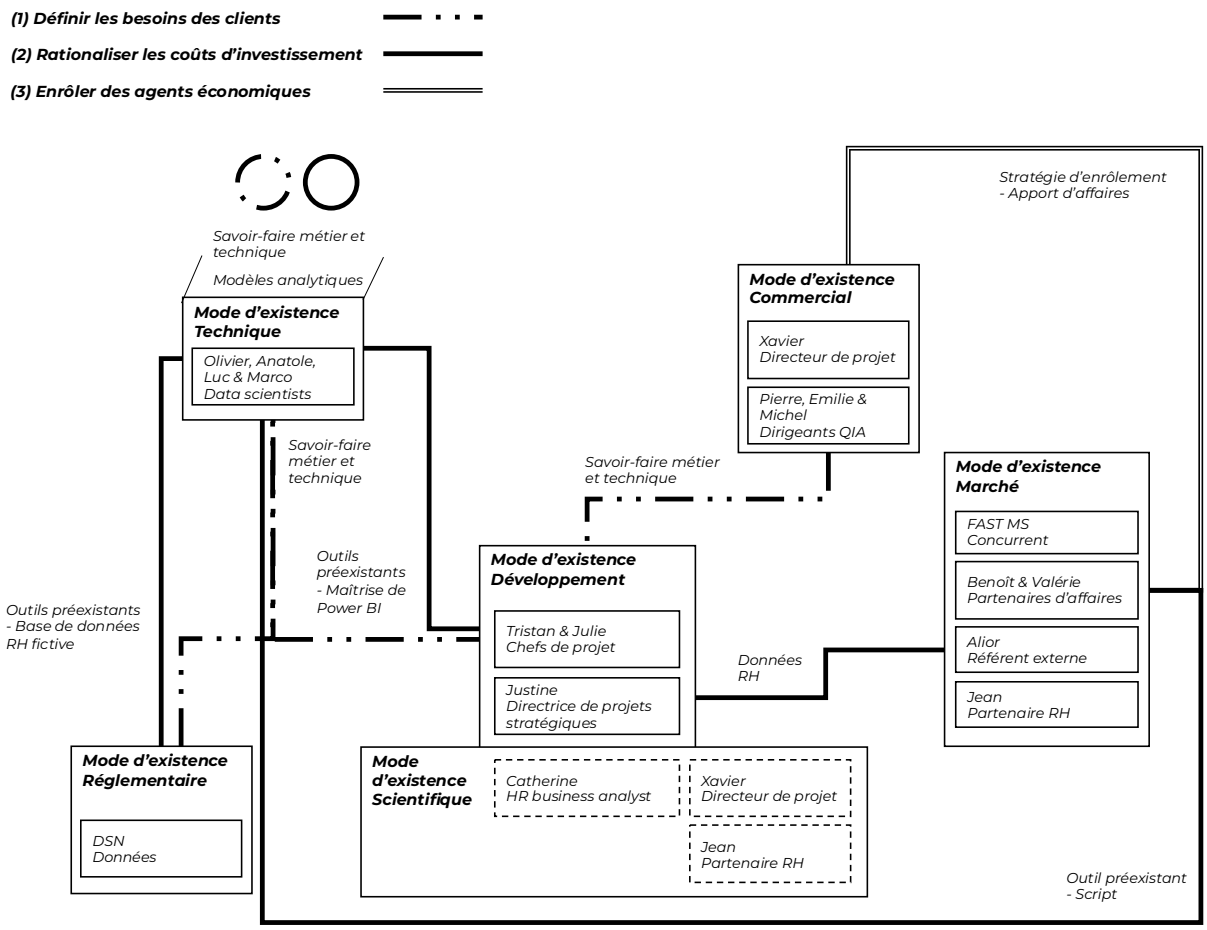


Figure 18 : Réseau d'alliances pour la séquence de Capitalisation

3. Représentation du réseau d'alliances dans la requalification des données RH

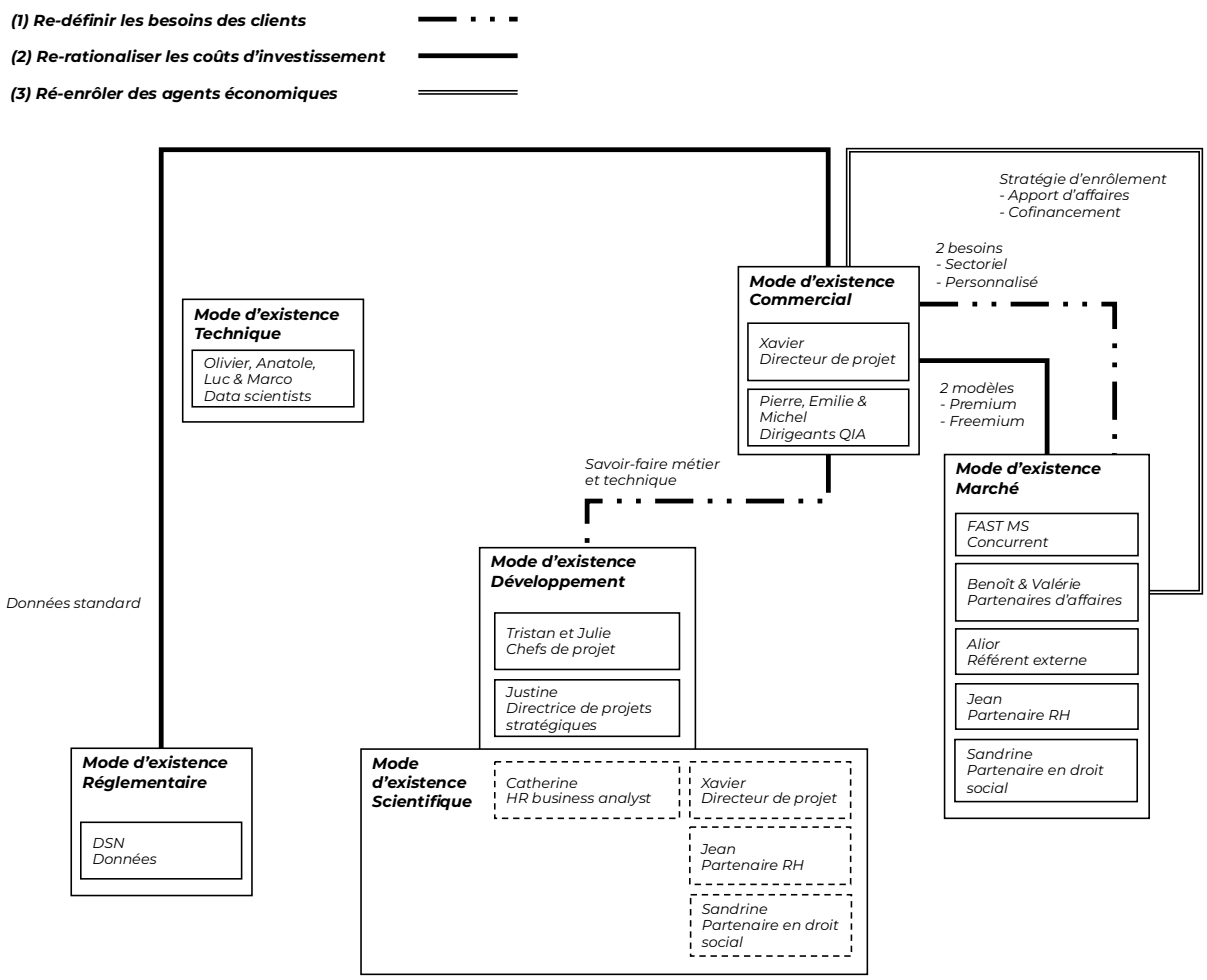


Figure 19 : Réseau d'alliances pour la séquence de Requalification

Annexe III. Extrait de la description de la phase de modélisation co-construite avec les *data scientists*

Système de couleurs par version (V)

V1

V2

V3

V4

V5

En vue d'initier la phase de modélisation des données *DSN*, une préparation des données s'imposait, condition sine qua non à l'entraînement des modèles. Cette étape émanait de la requête formulée par Tristan, qui requérait des ressources humaines pour s'y investir. Ce besoin a notamment trouvé son origine dans le manque de compétences techniques de la doctorante en GRH. Dans ce contexte, Tristan sollicita Bianca, gestionnaire de la filiale santé, pour déterminer si des ressources humaines extérieures à *QIA* étaient disponibles, dans le but d'apporter un appui au projet *DSN Analytics*. C'est ainsi qu'Olivier, *data scientist* junior, entra en jeu. Chez *QIA* depuis mai 2020 dans le cadre d'un stage de fin d'études, Olivier venait alors de finaliser un projet et disposait de temps pour contribuer au projet RH. Compte tenu de son manque d'expérience, Bianca a vu l'occasion de lui permettre de progresser en l'impliquant dans la formation technique d'un membre moins expérimenté de l'équipe. Par conséquent, ses objectifs étaient de guider la doctorante dans son apprentissage lors de la préparation des données et de prendre en charge une première version concrète d'un PoC « *Preuve de Concept* ».

Janvier 2021 à mai 2021

Préparation des données RH : la base de données d'apprentissage pour l'entraînement des modèles

La préparation des données est une étape fondamentale mais chronophage, absorbant généralement la majeure partie du temps d'un *data scientist*, estimée à environ 80 % de sa charge. De plus, cette étape est étroitement corrélée au type de modélisation envisagée.

Cette phase de travail fut ainsi initiée au cours du mois de février 2021. Grâce à la structuration préalable des données *DSN* réalisée par le script de Denis, celles-ci

étaient déjà agrégées et organisées dans des tables CSV. Ainsi, Olivier n'avait pas besoin de se familiariser avec les données brutes issues de la *DSN*, que ce soit en consultant la documentation ou en comprenant le script de lecture. Sa tâche se focalisait sur la préparation des données ainsi que sur l'entraînement et le développement subséquent du modèle. Les tables au format CSV englobaient des informations différentes consolidées à la maille de l'entité. Dans le but de fournir des éléments d'apprentissage au modèle de *Machine Learning*, des données *DSN* dérivées furent générées en fonction de critères préalablement définis, en vue de construire les premières variables pertinentes :

- Le nombre d'employés par entité ;
- Le nombre d'employés en CDD et en CDI ;
- Le nombre d'Employés Temps Plein (ETP) total par entité (comprenant les facteurs ETP) ;
- Le nombre de jours travaillés par entité ;
- L'ancienneté moyenne des employés ;
- Le nombre d'employés avec plus de 20 ans d'ancienneté ;
- Le nombre d'employés ayant moins de 5 ans d'ancienneté ;
- L'âge moyen ;
- Le nombre d'employés âgés de plus de 50 ans :
- Le nombre d'employés âgés de moins 30 ans ;
- Le Code de la Convention Collective Nationale (CCN) de l'entité (en vérifiant si différentes conventions collectives sont fréquemment utilisées) ;
- Le nombre d'employés pour chaque métier par entité ;
- Etc.

Les variables étaient générées de manière successive par Olivier et la doctorante au sein d'un premier « *Jupyter notebook* ». Un notebook Jupyter est un document numérique interactif qui combine du code informatique, du texte explicatif et des éléments visuels. Il est couramment utilisé pour explorer des données, développer du code et partager des résultats. L'objectif premier résidait dans l'accumulation d'une vaste gamme d'informations relatives à l'entité de rattachement des employés. Les variables ont principalement été identifiées par Tristan et la doctorante, en s'appuyant également sur les discussions avec Justine, responsable du premier projet sur l'absentéisme de *QIA*. Les décisions ont notamment été prises en considération du contenu du « *databook* » associé à ce projet. Le databook est un tableur, utilisé dans les projets, qui répertorie toutes les données collectées en spécifiant leur niveau de qualité et leur pertinence.

Durant la phase de préparation des données *DSN*, il a été essentiel de traiter les occurrences de valeurs manquantes. Par exemple, il a été observé que des entrées associées à des employés présentaient des absences d'informations dans diverses colonnes. Étant donné qu'un modèle ne peut apprendre de données incomplètes, ces entrées ont été supprimées. De plus, un processus de marquage des « *outliers* », a été appliqué. Celui-ci consistait à repérer des données qui se démarquent nettement

du reste de l'ensemble. Ces données atypiques peuvent résulter d'erreurs, de valeurs extrêmes ou de variations inhabituelles. Le but était de les repérer en vue de déterminer si elles nécessitaient une correction, un examen plus approfondi, ou si elles devaient être exclues de l'analyse, en se basant sur des critères prédéfinis. Dans notre cas, cette identification a été réalisée conformément aux critères suivants :

- Les employés dont les dates étaient discordantes entre elles ;
- Les employés avec des salaires négatifs, ainsi que toute autre valeur numérique négative qui ne devrait pas l'être (ex : âge ou ancienneté).

L'évaluation de la performance du modèle a été réalisée à l'aide d'indicateurs de performance classiques pour les régressions linéaires, tels que l'erreur moyenne, le coefficient de détermination (R^2) et l'erreur quadratique moyenne (RMSE). Ces métriques conventionnelles ont servi à déterminer les performances du modèle (la comparaison des prédictions obtenus à la réalité observée). Cette évaluation peut refléter la pertinence des données employées dans la modélisation. La régression linéaire n'est pas principalement choisie en raison de son caractère linéaire. Sa linéarité présente des limitations en ce sens qu'elle n'est pas nécessairement applicable de manière linéaire. Par exemple, l'absentéisme ne dépend pas exclusivement de l'âge de manière linéaire, car les variations entre les âges de 30 et 25 ans peuvent différer de celles entre 50 et 45 ans. Ainsi, il devient évident que la relation n'est pas strictement linéaire. La régression linéaire peut atteindre un plafond dans ces situations, car elle suppose des effets linéaires et ne permet pas de modéliser adéquatement les interactions complexes. En conséquence, en prenant en compte les limitations inhérentes à l'utilisation de la régression linéaire pour modéliser des phénomènes qui ne présentent pas nécessairement une corrélation linéaire avec la variable à expliquer, il est important de noter que la régression linéaire impose un cadre rigide qui exige que les relations soient linéaires. En revanche, d'autres modèles statistiques offrent la flexibilité nécessaire pour prendre en compte des interactions, des effets non linéaires et des relations multivariées, ce qui les rend souvent plus appropriés d'un point de vue méthodologique. Toutefois, lors de la sélection d'un modèle, il est essentiel de trouver un équilibre entre trois critères cruciaux : la simplicité, l'interprétabilité et la performance. Cette approche combinait donc simplicité et interprétabilité. Tristan et Olivier ont opté pour cette approche en raison de qualités susmentionnées mais également pour la facilité d'application pour Olivier, qui était encore relativement junior et peu familier avec un large éventail de modèles.

Annexe IV. Grille d'entretien sur la fonction RH

1. **Questions introductives**
   - 1.1. Nom, prénom :
   - 1.2. Âge :
   - 1.3. Ancienneté dans l'entreprise :
   - 1.4. Fonction (grade et missions) :
   - 1.5. Ancienneté dans cette fonction :
   - 1.6. Pourriez-vous me décrire votre parcours dans la société ?
2. **Facteurs environnementaux de la fonction RH / Eléments de contexte**
   - 2.1. Quels sont, selon vous, les grands facteurs qui influencent le plus votre fonction aujourd'hui ?
3. **Belonging (Identité) : qu'est-ce qui caractérise la GRH ?**
   - 3.1. Comment définissez-vous les données RH ? Depuis quand les qualifiez-vous ainsi ?
   - 3.2. Quelles sont les données RH que vous identifiez dans votre organisation ?
   - 3.3. Qui utilise ces données ? À quelles fins ?
   - 3.4. Quels sont vos besoins en termes d'exploitation des données RH ?
4. **Learning (connaissance) : A quel moment la GRH doit-elle acquérir de nouvelles connaissances ?**
   - 4.1. Comment définissez-vous votre position aujourd'hui ? Conformiste ou innovatrice ? Pourquoi ?
   - 4.2. Que signifie, pour vous, l'Intelligence Artificielle ? Pourriez-vous me préciser votre propre définition de l'IA ?
   - 4.3. Possédez-vous des outils d'IA actuellement ? Dans quelles activités ?
   - 4.4. En quoi les outils d'IA peuvent-ils, selon vous, être pertinents dans l'exploitation de vos données RH ? En quoi vous intéressent-ils ? Pourquoi ?
   - 4.5. Selon vous, en quoi les outils d'IA dans les RH peuvent-ils être une opportunité et/ou un risque pour votre fonction ?
   - 4.6. Des compétences techniques/analytiques sont-elles nécessaires ?
   - 4.7. Quels sont/ou devraient être les facteurs clés de succès des outils d'IA dans les RH ? Pourquoi ? (Compréhension conjointe entre concepteurs/utilisateur ?)

4.8. L'exploitation des données RH peut-elle être la source d'un changement pour la fonction RH ? Dans quelle mesure ?

5. Organizing (conception des postes) : comment la GRH devrait-elle organiser ses processus ?

5.1. Le traitement et l'analyse des données RH se font-ils dans votre département ou les sous-traitez-vous à une autre direction ou à une entreprise externe ?

5.2. Considérez-vous que cette exploitation soit actuellement au maximum de son potentiel ? Pourquoi ?

5.3. Les données RH contribuent-elles aux usages des autres directions ? Lesquelles ?

5.4. En tant que producteur de données RH, vous sentez-vous propriétaire de ces données ? Pourquoi ?

5.5. Selon vous, les autres directions sont-elles légitimes pour exploiter les données RH ? Sentez-vous le besoin d'être impliqué dans le processus d'exploitation ? Pour l'interprétation des données par exemple et pourquoi ?

6. Performing (Objectifs) : quels sont les objectifs principaux de la GRH ?

6.1. Quelles données sont, selon vous, stratégiques pour la performance de l'entreprise ?

6.2. Quelles données sont, selon vous, stratégiques pour la performance des RH ?

6.3. Quel est votre ressenti sur l'importance d'exploiter les données RH ? Quels sont les impacts sur la fonction RH ?

6.4. Croyez-vous que l'exploitation des données RH par la fonction RH représente un défi aujourd'hui ?

7. Questions additionnelles (possibles selon les cas

7.1. Possibilité de me mettre en contact avec d'autres DRH ?

7.2. Possibilité de réaliser une observation courte pour rendre compte de l'utilisation des outils dans vos pratiques quotidiennes ?

EXPLORATION DU PROCESSUS DE CONSTRUCTION DES DONNEES RH – QUALIFICATION, CAPITALISATION ET REQUALIFICATION

Cette thèse de doctorat examine le rôle des données dans la conception d'outils d'intelligence artificielle (IA) appliqués à la gestion des ressources humaines (GRH). Elle se concentre plus spécifiquement sur le développement d'un outil destiné à l'analyse de l'absentéisme. Elle explore comment, au cœur de ce processus, les *data scientists* déploient un ensemble de pratiques calculatoires et non calculatoires dans la construction des données RH. Bien que la littérature académique abonde en discussions sur les approches positivistes et « basées sur les preuves » dans ce domaine, les modalités par lesquelles ces pratiques opèrent de l'intérieur restent sous-explorées.

Cette recherche repose sur une méthodologie qualitative et séquentielle, s'inscrivant dans un continuum entre recherche-action et ethnographie, enrichie par une immersion de trois ans et sept mois au sein d'un cabinet de conseil spécialisé en *data science*. Elle apporte une contribution aux domaines de la gestion électronique des ressources humaines (e-GRH) et de l'instrumentation de GRH, en mettant en lumière la dimension socio-numérique des données RH. En remettant en question le déterminisme technique prévalant dans ce domaine, cette étude propose une nouvelle conceptualisation du processus de construction des données RH en trois séquences : Qualification, Capitalisation et Requalification. Ce processus conduit à reconnaître les données comme des biens économiques utilisés en tant qu'instruments épistémiques. Ces instruments, dont les qualités sont « débattues », « contestées » et « négociées », interrogent ce que signifie être une « bonne » donnée dans le cadre de l'analyse de l'absentéisme.

Mots clés : Données RH, Gestion des Ressources Humaines, Systèmes d'Information, Théorie de l'Acteur-Réseau

---

EXPLORATION OF THE HR DATA CONSTRUCTION PROCESS – QUALIFICATION, CAPITALIZATION AND REQUALIFICATION

This doctoral thesis examines the role of data in the design of artificial intelligence (AI)-enabled tools for human resource management (HRM), with a specific focus on developing a tool to analyze absenteeism. The research investigates how *data scientists* engage both calculative and non-calculative practices in constructing HR data. While a substantial body of academic literature addresses positivist and evidence-based approaches in this field, the internal mechanisms through which these practices are operationalized remain underexplored.

This study employs a qualitative, sequential methodology that bridges action research and ethnography, supported by an immersive three years and seven months in a professional service firm (PSF) specializing in *data science*. It contributes to the sub-fields of electronic human resource management (e-HRM) and HRM instrumentation by shedding light on the socio-digital dimensions of HR data. The thesis challenges the dominant narrative of technical determinism in the field by proposing a new conceptualization of the HR data construction process, articulated in three phases: Qualification, Capitalization and Requalification. This reconceptualization views data as economic assets functioning as epistemic instruments, whose qualities are continually debated, contested and negotiated, thus raising crucial questions about what constitutes 'good' data in the context of absenteeism analysis.

Keywords: HR Data, Human Resource Management, Information Systems, Actor-Network Theory