MÉTIERS D'AVENIR

DATA SCIENTIST (H/F)

Septembre 2017
Le Forem, Service de veille, analyse et prospective du marché de l'emploi

DATA SCIENTIST, UN MÉTIER D'AVENIR ?

Anticiper les évolutions, l'émergence de métiers ou la transformation de métiers actuels constitue un axe majeur de la mission d'analyse et d'information sur le marché du travail du Forem. Une première étude exploratoire réalisée en 2013 a permis de dégager les grandes tendances d'évolution des secteurs. En 2016, Le Forem a poursuivi sa démarche en publiant des rapports sur les effets de la transition numérique sur les secteurs en termes d'activités, métiers et compétences. Des métiers d'avenir ont ainsi été identifiés. Il peut s'agir de :

- nouveaux métiers ;
- métiers actuels qui évoluent considérablement ;
- métiers avec un potentiel de croissance en effectifs.

Partant de cette base, une analyse en profondeur, « métier par métier » est mise en œuvre. Elle permet de mieux cerner les évolutions des métiers et d'adapter, après l'analyse de grands domaines de transformation attendus, l'offre de prestations. Cette analyse prospective se fonde sur la méthode *Abilitic2Perform.*

Abilitic2Perfom est une méthode d'anticipation des compétences basée sur l'animation de groupes d'experts lors d'ateliers successifs et éprouvée sur une quinzaine de métiers lors de son développement dans le cadre de projets européens « Interreg IV ». Cette méthode est inspirée des études relatives à la prospective stratégique[1], dont certains outils sont mobilisés comme l'analyse structurelle ou morphologique. Les rapports d'analyse font l'objet d'une publication régulière sur le site Internet du Forem.

Le présent rapport, réalisé en partenariat avec le Centre de compétence « Technofutur TIC », porte sur le métier de *data scientist*.

Le *data scientist* trouve son origine dans la digitalisation de la société et l'avènement du *Big data*. Ce sont J. Hammerbacher et D. Patil[2], deux ingénieurs travaillant pour *LinkedIn* et *Facebook* qui inventèrent ce terme pour qualifier le métier des personnes confrontées aux problématiques liées au traitement d'énormes masses de données numériques.

[1] Voir notamment, Godet M., Manuel de Prospective stratégique - Tome 1 : *Une indiscipline intellectuelle*, Paris, Dunod, 2007 et Godet M., Manuel de Prospective stratégique - Tome 2 : L'art et la méthode, Paris, Dunod, 2007.

[2] Dhanurjay Patil est un Data Scientist américain, nommé en 2015 comme le "First First U.S. Chief Data Scientist ». Il est actuellement Chief Data Scientist à l'Office of Science and Technology Policy. Après avoir mené l'équipe « data » de Facebook, Jeff Hammerbacher a co-fondé la société « Cloudera » où il travaille comme Data Scientist.

TABLE DES MATIÈRES

DATA SCIENTIST, UN MÉTIER D'AVENIR ? 2

Partie 1 – Synthèse des résultats 5

Quelles sont les grandes tendances qui détermineront le plus l'évolution du métier de data scientist dans les prochaines années. 5

Partie 2 – La démarche et les résultats pas à pas 8

1. Le périmètre du métier 9
2. Les facteurs les plus influents 10
3. Les évolutions probables et souhaitables 11
4. Le profil d'évolution 11
5. Les impacts sur les activités et les besoins en compétences 16

Le *data scientist* a pour mission de traiter une grande quantité de données brutes pour en retirer des informations précieuses et de valeur pour les entreprises. Ces informations, issues de bases de données internes et externes des entreprises, permettent de mettre en avant et de comprendre des tendances et des comportements complexes pour améliorer leurs performances marketing, financières ou de production.

Le champ des possibles est très important. UPS est par exemple couramment cité pour illustrer le potentiel d'utilisation des données. Cette société, grâce à la mise en place d'algorithmes, a optimisé les itinéraires des livraisons en limitant les virages à gauche, ce qui a permis d'économiser une grande quantité de carburant et de réduire considérablement l'émission de CO2.

Le métier de *data scientist* requiert des profils hautement qualifiés. Il doit posséder des connaissances solides en mathématiques et en informatique (programmation et bases de données). Il doit, pour être efficace, développer beaucoup de compétences non techniques telles que l'empathie pour le secteur qu'il analyse et se montrer créatif, innovant et curieux.

Le *data scientist* a été élu par la *Harvard Business Review* en 2012 comme « *le métier le plus sexy du XXIe siècle »*[3] et ce grâce à la plus-value de la connaissance qu'il génère. En 2017, le site de recherche d'emploi Glassdoor[4] a élu le *data scientist* pour la deuxième année consécutive en première position de son top vingt-cinq des meilleurs métiers aux États-Unis, sur base du nombre de postes à pourvoir, de la rémunération et des possibilités d'évolution de carrière.

Si l'engouement est bien présent aux États-Unis, qu'en est-il en Europe ? Glassdoor n'effectue pas l'analyse pour l'Europe entière mais uniquement pour le Royaume-Uni. Pour ce pays, le *data scientist* arrive en sixième position alors que l'année précédente, il n'apparaissait pas du tout dans ce top. Cette différence peut s'expliquer par différents facteurs :

- L'Europe a une politique de protection de la vie privée et des données plus stricte qu'aux États-Unis.[5]
- Les États-Unis investissent des sommes importantes dans la recherche dans ce domaine, et notamment depuis les attentats du 11 septembre 2001 ; avec parfois des dérives importantes au niveau de la protection de la vie privée.[6] Faute d'une communautarisation des budgets de la défense, l'Europe n'a pas les capacités d'investissement des États-Unis.
- Enfin, la *Silicon Valley* (USA) concentre les industries mondiales de l'informatique et du numérique à la pointe de l'innovation.

En Belgique, le profil de *data scientist* est de plus en plus recherché. D'une part des startup spécialisées dans ce domaine voient le jour avec une mission de consultance pour les entreprises. D'autre part, les entreprises elles-mêmes, tous secteurs confondus, sont intéressées par ce profil.

Cependant, il semblerait que les offres d'emplois sous l'appellation *data scientist*, concernent majoritairement, des *data analyst* très qualifiés. La différence fondamentale entre ces deux fonctions se situe au niveau du caractère de l'analyse. Celle du *data analyst* reste descriptive tandis que celle du *data scientist* est prédictive.

Le *data analyst* ne crée pas d'algorithmes, ne réalise pas de modélisation et le niveau requis en mathématiques est moindre que celui du *data scientist*. Il doit être capable de réaliser une analyse prédictive en utilisant des outils standards, parfois automatisés et basés sur les algorithmes existants. Cette tendance est

[3] Cf. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century, consulté le 2/03/2017.
[4] Glassdoor est un site de recrutement et de recherche d'emploi américain utilisant une base de donnée croissante d'avis d'entreprise, de notes d'approbation de PDG et de rapports de salaires, d'avis et de questions d'entretien, d'avis sur les avantages sociaux ainsi que des photos des bureaux et bien plus encore ces informations sont partagées par les employés de l'entreprise. Glassdoor se targue de permettre de voir quels employeurs embauchent, les conditions de travail réelles dans l'entreprise ou comment sont les entretiens d'embauche selon les employés, et les niveaux de salaires possibles.
[5] Cf. http://www.europarl.europa.eu/news/fr/headlines/priorities/20130901TST18405, consulté le 23/06/17.
[6] Le programme PRISM est un programme américain de surveillance électronique par la collecte de renseignements à partir d'Internet et d'autres fournisseurs de services électroniques.

également annoncée par l'entreprise de conseil *Gartner inc*[7], qui parle du « *citizen data scientist* ».

Les profils de *data scientist* sont aujourd'hui assez rares. Cette dénomination n'existant pas en tant que telle dans les bases de données du Forem, il est difficile de donner un chiffre exact et précis. Néanmoins il semblerait que moins de 100 personnes dans la réserve de main d'œuvre soient dans les conditions pour accéder au métier ou à la formation. Pour certaines entreprises, *le citizen data scientist* sera une bonne alternative et pourra déjà réaliser des analyses avancées, laissant au *data scientist* les problématiques les plus complexes.

Le métier de *data scientist* étant récent, l'offre de formation est à l'heure actuelle assez limitée mais va s'étoffer dans les mois à venir. Actuellement les *data scientist* se forment de manière autodidacte grâce à des plateformes, des MOOC[8] ou encore en participant à des hackathons.[9] Au niveau académique, la majorité des universités francophones propose un master en data science pour la rentrée académique 2017-2018. En fonction de l'université, les prérequis sont parfois différents. L'université de Namur ouvre par exemple ce master aux ingénieurs de gestion, aux informaticiens et aux mathématiciens. Au niveau des centres de compétences, Technofutur Tic a mis en place depuis janvier 2017 *la data academy*, qui forme à différents métiers liés à la filière data et notamment le *data scientist*. Cette formation est relativement courte par rapport aux formations académiques et a pour vocation d'être plus professionnalisante. Elle s'adresse principalement à des demandeurs d'emploi ayant déjà un bagage professionnel adéquat pour la fonction et souhaitant se réorienter.

Les participants à la démarche soulignent le fait que le niveau de compétence à acquérir à l'université ou en centre de compétence ne doit pas être le même. Le premier offre un apprentissage en profondeur alors que le second est plus professionnalisant. Afin que les cursus de formations soient complémentaires, le groupe de travail souligne l'importance de positionner les programmes de formation en distinguant les compétences visées par chacun d'eux, les méthodologies et le public ciblé. Le profil issu des centres de compétence pourrait être assimilé au *citizen data scientist*.

Enfin, à la question de savoir quels sont les besoins en *data scientist*, il est, selon le groupe, très difficile d'avoir des chiffres précis et objectifs. Selon les experts, la demande est supérieure à l'offre et les perspectives d'emploi sont bonnes. Il semblerait toutefois que 70% des offres d'emplois concerne plutôt des *data analyst,* voire des *citizen data scientist* pour 30% de *data scientist.* Dans la base de données des offres d'emploi du Forem, seules deux offres d'emploi ont été trouvées. Cependant, il semblerait que ce canal ne soit pas privilégié par les entreprises. Fin septembre 2017, le site LinkedIn proposait quatre-vingt-deux offres pour les régions de Bruxelles et la Wallonie. On peut également noter que les entreprises recrutent pas mal via la communauté.

L'enjeu pour la Wallonie concernant ce métier est de soutenir l'innovation et de favoriser la création et le développement des startups offrant des produits et services basés sur l'exploitation du *Big data.*

Ce rapport comprend deux parties. La première présente une synthèse des résultats reprenant l'ensemble du profil d'évolution et les activités clés pour l'avenir. La seconde reprend dans le détail l'ensemble du processus d'analyse selon l'ordre chronologique de son déroulement.

Le lecteur y retrouvera notamment la liste (non exhaustive) des compétences pointées comme importantes par les experts pour la réalisation des activités clés.

---

[7] http://www.gartner.com/newsroom/id/3570917, consulté le 23/06/2017.

[8] MOOC signifie Massive Open Online Course (formation en ligne ouverte à tous).

[9] Le mot hackathon désigne un événement où un groupe de développeurs volontaires se réunissent pour faire de la programmation informatique collaborative sur plusieurs jours. Cf. http://www.creativeworkslondon.org.uk/wp-content/uploads/2013/11/Digital-Innovation-The-Hackathon-Phenomenon1.pdf, consulté le 11/09/17.

Partie 1 – Synthèse des résultats

Quelles sont les grandes tendances qui détermineront le plus l'évolution du métier de *data scientist* dans les prochaines années ?

Dans le cadre de cette étude, dix facteurs d'évolution ont été retenus comme étant clés pour le métier de *data scientist* dans les trois à cinq prochaines années. La filière data au sens large ayant fait l'objet d'une analyse publiée en octobre 2015[10], la présente analyse s'est basée sur ces résultats. Les facteurs déterminés à ce moment-là ont été mis à jour : certains ont été supprimés et d'autres se sont ajoutés. Parmi les facteurs retenus, certains impactent directement les compétences du *data scientist*, d'autres représentent plutôt des possibilités de développement de celui-ci en termes d'opportunités à saisir ou encore de développement du potentiel d'emploi.

L'accessibilité et la maturation des technologies du *Big data* est, sans surprise, un facteur déterminant pour la filière data de manière générale et particulièrement pour le métier de *data scientist*. Selon *Gartner inc.*, le *Big data* est exploité par une famille d'outils qui répondent à une triple problématique : un volume de données important à traiter, une grande variété d'informations en provenance de sources multiples et structurées ou pas, ainsi qu'un certain niveau de vélocité à atteindre. [11]

Concernant la Belgique et particulièrement la Wallonie, c'est la variété des sources qui est à mettre en avant plutôt que le volume qui est relativement restreint en raison de la taille du territoire.

Un certain nombre d'outils « open source » existent sur le marché et les possibilités de stockage se démocratisent. Cependant il manque des environnements où ils peuvent être mis en œuvre et exploités. De plus un effort d'adaptation est requis pour les rendre efficaces et les intégrer à l'entreprise car les outils d'architecture ne sont pas encore tout à fait prêts à combiner les différents types de données.

Le développement de l'internet des objets est une évolution technologique en cours de diffusion. [12] L'exploitation des données issues de ces objets permet non seulement d'améliorer la gestion des entreprises mais aussi d'améliorer certains produits ou services. La quantité de données générées par cette technologie est considérable. Selon le cabinet de recherche américain *Gartner inc.*, il y aurait fin 2017, 8,4 milliards d'objets connectés et plus de 20 milliards en 2020.

Si l'Europe est à la traine par rapport à la conception des objets eux-mêmes, c'est au niveau de l'analyse des données que vont générer les capteurs, particulièrement au niveau industriel, que se situent les enjeux pour le métier de *data scientist*.

Les données deviennent un vecteur d'innovation. Depuis peu on assiste à la création de services uniquement basés sur l'exploitation des données. À titre d'exemple, l'ONG *Bayes Impact* a notamment pour vocation de créer des services sociaux grâce aux algorithmes. Ils ont ainsi mis en place le service « Bob Emploi »[13] en partenariat avec Pôle Emploi. Ce service permet, grâce à des algorithmes, de mettre en place un accompagnement sur mesure pour les chercheurs d'emploi. Si dans ce cas, les données ne sont plus uniquement génératrices de profit, d'autres industriels pourraient, grâce aux données, améliorer la chaine de production et vendre la méthode à d'autres industries. Le profit de la démarche se situe donc à plusieurs niveaux.

Les entreprises sont de plus en plus conscientes des opportunités offertes par le traitement des données et de plus en plus de managers poussent à **la promotion et à la réflexion sur le data.** Certaines entreprises sont proactives, d'autres plutôt attentistes, mais elles

[10] Cf. https://www.leforem.be/MungoBlobs/465/148/20151001_Rapport_A2P_LaFiliereData_Final.pdf.
[11] Par vélocité on entend la fréquence de création, de collecte et de partage des données.
[12] Cf. le communiqué de presse de *Gartner inc*. concernant l'IOT : http://www.gartner.com/newsroom/id/2684616, consulté le 26/06/17.
[13] Cf. https://www.bob-emploi.fr/ consulté le 20/03/17.

semblent être de moins en moins complétement ignorantes par rapport aux traitements des données y compris dans les TPE-PME, qui sont déjà actuellement ouvertes à « l'analyse des affaires ».

L'informatique industrielle et l'informatique de gestion ont tendance à se rapprocher. Les logiciels de type ERP[14] permettent actuellement de réaliser des analyses de contrôle. L'objectif principal dans les trois à cinq ans serait d'optimiser les processus et les chaines de production grâce à la *data science.*

Le traitement des données n'est pas seulement orienté vers l'entreprise. **Le consommateur est aussi directement concerné par le partage de ses données.** Il est difficile de dégager une tendance claire à cet égard. Il semblerait toutefois que celui-ci est d'accord de léguer ses données à partir du moment où il y trouve un intérêt comme de bénéficier de services personnalisés et ce, même dans des domaines sensibles tels que la santé.

L'ouverture des données (open data) représente un enjeu important. Il est aujourd'hui question que certaines bases de données issues des services publics soient ouvertes. En Flandre plusieurs fichiers sont déjà disponibles. En Wallonie un certain nombre d'initiatives ont vu le jour suite au décret « open data » adopté par le gouvernement Wallon et sont répertoriées sur Digital Wallonia.[15] Cependant ces fichiers sont aujourd'hui encore difficilement exploitables, certains restent encore assez figés. Pour certaines administrations, des solutions d'anonymisation doivent être trouvées avant de pouvoir livrer les données.

La politique de protection et de sécurisation des données est un des facteurs retenus comme clé par les experts. Depuis les révélations sur le programme Prism[16], l'Union Européenne a décidé de revoir la réglementation sur la protection des données en place depuis 1995.[17] Après des mois de concertation, le règlement général sur la protection des données[18]a été adopté par le Parlement Européen en avril 2016 et sera d'application à partir de mai 2018. Les trois grands éléments en sont :

- La logique de responsabilisation
- La coresponsabilité des sous-traitants
- Le « privacy by design »[19]

Ces changements de réglementation risquent de brider les travaux du *data scientist* qui devra trouver d'autres sources de données. La recherche devrait par ailleurs s'en trouver stimulée, par exemple en matière d'anonymisation des données.

Le contexte politique influencera sans aucun doute les demandes adressées aux *data scientist*. Les gouvernements sont de plus en plus friands d'outils capables de détecter la menace terroriste.[20]

L'analyse des données est également utilisée dans les campagnes électorales les plus importantes et ce depuis les élections présidentielles américaines de 2012 qui a vu le camp du candidat démocrate utiliser les données pour orienter la campagne. Ces technologies peuvent aujourd'hui être utilisées, quel que soit le type d'élection et les budgets de campagne. De nombreuses opportunités existent mais l'Europe est en retrait par rapport aux États-Unis qui sont prêts à faire de gros investissements dans le domaine.

Enfin, **les formations de data scientist sont de plus en plus accessibles.** Des MOOC[21] réalisés par des universités réputées telles que Stanford ou Harvard existent sur le marché bien qu'il s'agit surtout d'y apprendre les fondamentaux et d'acquérir les prérequis pour aller plus en avant ou de se tenir à jour de spécificités. Si l'offre de formations augmente, elle n'est cependant pas accessible à tous et demande des prérequis, notamment en mathématiques et en informatique.

[14] Entreprise Resource Planning.
[15] cf. https://www.digitalwallonia.be/decret-open-data/, consulté le 4/07/17.
[16] Edward Snowden, cf. http://www.lemonde.fr/international/infographie/2013/06/11/le-programme-prism-en-une-infographie_3427774_3210.html, consulté le 4/07/17.
[17] Cf. la Directive 95/46/CE. http://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A31995L0046, consulté 5/07/17.
[18] Cf. Règlement (UE) 2016/679. http://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679, consulté le 5/07/17.
[19] La protection de la vie privée est intégrée dès la conception.
[20] À titre d'exemple, c'est au lendemain des attentats du 11 septembre 2001 que la startup « Palantir Technologies » a été créée, son objectif étant de mettre à profit la puissance des algorithmes informatiques pour contrer les actes de terrorisme. Cette entreprise est aujourd'hui présente sur pratiquement tous les continents et a pour client de nombreux services de renseignements.
[21] Massive Open Online Course peut être traduit par « formation en ligne ouverte à tous ».

Impacts des évolutions : quels besoins en compétences ?

Cette analyse a permis de mettre en évidence les compétences déterminantes pour le métier de data scientist. Les hypothèses d'évolution déterminées lors des premiers ateliers révèlent sans surprise que ce sont les activités d'identification des besoins, d'identification des sources de données et d'exploitation des données qui sont les plus impactées. Les participants soulignent qu'à l'horizon 2020-2022, le *data scientist*, de manière générale, au-delà des compétences en statistiques, devra développer encore plus ses compétences informatiques.

Comme énoncé, les outils du *big data* manquent d'environnements où ils peuvent être mis en œuvre et exploités. S'il ne revient pas dans les tâches premières du *data scientist* de s'occuper de créer les *data flow*[22] ou encore de structurer l'environnement (missions respectivement dédiées au data engineer et data architect), il doit néanmoins être **conscient des contraintes** techniques liées aux outils utilisés. Les modèles qu'il proposera doivent tenir compte de ce paramètre.

L'internet des objets va devenir une source de données importante à l'avenir. Le data scientist devra donc **comprendre les fondements et les contraintes** spécifiques de celles-ci, c'est-à-dire des données distribuées et en streaming. Comprendre les impacts de la **logique temporelle** sur les méthodes d'accès et de traitement des données semble également incontournable.

Les entreprises étant de plus en plus conscientes du potentiel de leurs données, le *data scientist* doit avoir une **vision « métier » de l'organisation**. Il doit comprendre les enjeux du business, être dans une **logique diagnostique** et poser les questions pertinentes. Il a également dans ce contexte un **rôle d'évangélisation**, il doit pouvoir expliquer la valeur potentielle des données et les opportunités business qu'elles proposent.

Comme déjà considéré, la nouvelle règlementation européenne sur la protection des données impacte également le métier. Le *data scientist* doit **connaitre ces contraintes légales** et règlementaires au moins dans les grandes lignes. Ce facteur couplé avec la difficulté pour la Wallonie à ouvrir ses bases de données, demande au *data scientist* de faire preuve d'imagination[23] pour pouvoir accéder et extraire les données là où elles se trouvent.[24]

Le *data scientist* doit accorder de l'importance au **temps de veille**. Il doit veiller d'une part les outils disponibles : ceux-ci évoluant constamment, il doit impérativement se tenir informé, au niveau des outils de stockage, d'extraction et d'accès aux données. D'autre part, il doit pouvoir veiller les méthodes et les outils de traitement existants. Avec R[25] par exemple, tous les mois, de nouveaux packages sortent. Pour y arriver, il doit notamment être actif dans la communauté, afin d'échanger avec d'autres *data scientist*. L'accessibilité des formations (MOOC, formations en ligne, etc..) facilite le développement de ces compétences.

Cette compétence de veille est particulièrement à développer pour les *citizen data scientist*, comme déjà évoqué, qui utiliseront plus souvent des modèles standardisés dans leurs pratiques.

[22] Flux de données.
[23] Ex : on ne peut utiliser tel type de données pour des raison légales donc.
[24] Les experts parlent de web scraping : il s'agit d'une technique d'extraction de contenu de site web.
[25] R est un langage de programmation, disponible en « open source » et utilisé de plus en plus par les universités.

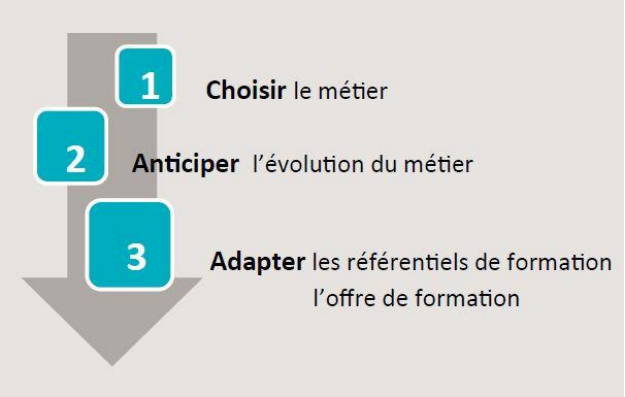
Partie 2 – La démarche et les résultats pas à pas

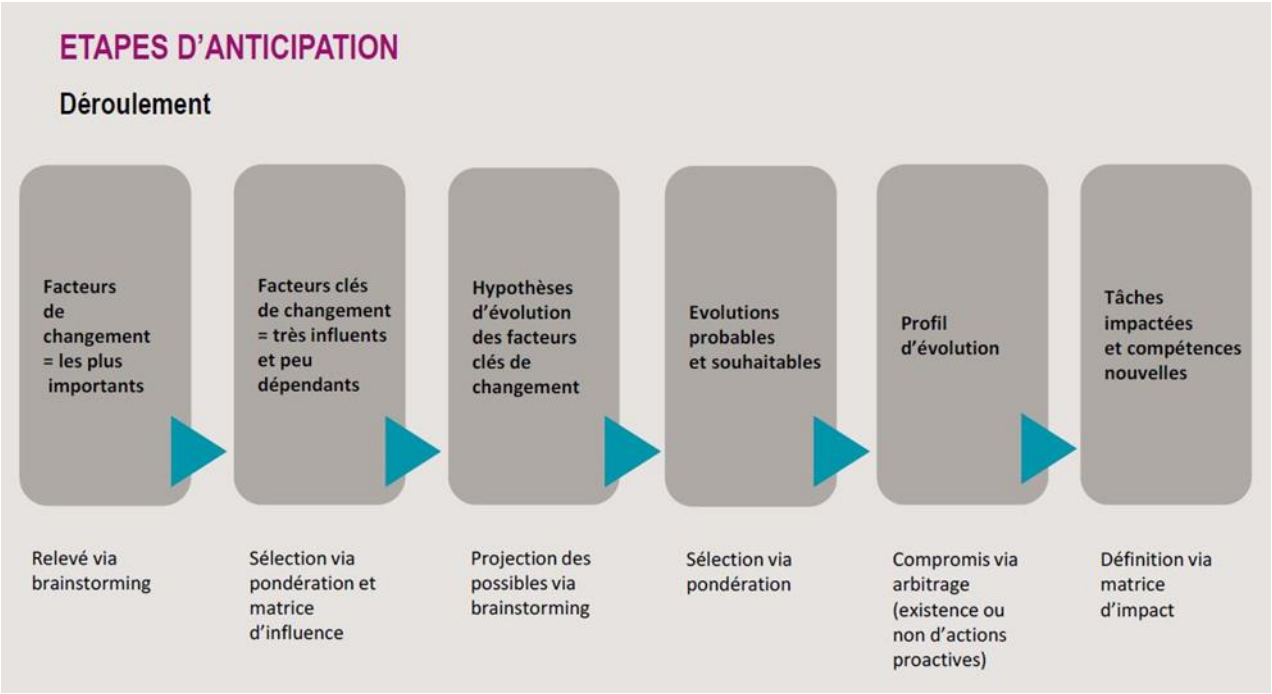
Cette partie du document décrit l'ensemble du processus suivi dans le cadre du déploiement de la méthode *Abilitic2Perform* appliquée au métier de *data scientist.*

La démarche se base sur la participation d'un panel d'experts à une série d'ateliers encadrés par un animateur qui conduit les réunions et par un back officer qui prend note des éléments cités en séance.

La méthode alterne, d'une part, des phases de réflexions créatives et collectives de type brainstorming et, d'autre part, des phases individuelles destinées à noter la pertinence ou l'impact des idées précédemment émises. Le traitement de ces notes par le back officer et l'animateur permet d'objectiver les éléments récoltés. Les résultats obtenus au terme de chaque phase servent de matière première à la phase suivante.

Trois grandes étapes doivent être parcourues : choisir un métier, anticiper les évolutions et leurs impacts sur le métier, puis adapter les prestations. Le présent rapport se focalise essentiellement sur la deuxième phase consacrée à l'anticipation.





L'analyse de ce métier a été réalisée lors de trois ateliers du 9 février au 7 mars 2017 qui ont rassemblé une dizaine de personnes issues de différents milieux : entreprises, centre de compétence, opérateurs de formation, représentants du secteur et Le Forem (cf. colophon).

Le métier de *data scientist* a été sélectionné pour faire l'objet d'un exercice détaillé d'anticipation sur base de l'analyse de grandes tendances d'évolution des secteurs.

La suite du document reprend, étape par étape, la procédure d'analyse :

1. Périmètre du métier
2. Recensement des facteurs de changement
3. Hypothèses d'évolution des facteurs clés de changement
4. Évolutions probables et souhaitables
5. Impacts sur les activités et les besoins en compétences

1. LE PÉRIMÈTRE DU MÉTIER

La définition du métier a été réalisée sur base de références françaises (OPIIEC[26]) et rediscutée en séance avec les participants lors du premier atelier.

« Le data scientist est un expert de la gestion et de l'analyse pointue de données complexes. Il vise à extraire de la connaissance de celles-ci en s'appuyant sur des modèles mathématiques et algorithmiques.
Il possède donc des connaissances pointues en mathématiques et en informatique (développement) tout en faisant preuve d'empathie pour le secteur d'application des données analysées. »

Les activités et tâches du métier sont basées sur le site de référence *Data Science Central* et les commentaires du groupe afin que celles-ci correspondent au mieux avec la réalité de terrain.

Activités	Tâches
1. Identifier les besoins et la problématique	- Identifier le type de problème. - Comprendre l'infrastructure IT et analytique en place. - Identifier les modèles mathématiques et les algorithmes à utiliser. - Benchmarking.
2. Identifier les sources de données disponibles	- Identifier les sources de données supplémentaires non disponible. - Extraire et vérifier les données de l'échantillon. - Effectuer l'analyse exploratoire. - Évaluer la qualité des données et prendre des mesures afin de les rendre exploitables. - Identifier et sélectionner les outils nécessaires. - Déterminer comment stocker et accéder aux données.
3. Exploiter et analyser les données	- Utiliser les méthodes d'imputation (traitement éventuel des valeurs manquantes) selon les besoins. - Détecter, expliquer et éventuellement supprimer les valeurs aberrantes. - Sélection des variables (réduction des variables). - Analyser les corrélations croisées. - Modéliser mathématiquement les données. - Analyser la sensibilité des données. - Validation croisée, adaptation de modèle. - Mesurer l'exactitude, fournir un intervalle de confiance. - Exploiter les données distribuées ou en streaming.
4. Mise en œuvre et développement	- Support et déploiement. - Mise en œuvre de données.
5. Communiquer les résultats	- Déterminer la manière adéquate de communiquer. - Réaliser une interface de visualisation des résultats. - Évaluer la valeur business. - Avancer des pistes d'amélioration. - Rédiger un rapport. - Éventuellement former le personnel aux outils mis en place.

***Tableau 1** : Activités / tâches du data scientist.*

[26] OPIIEC : observatoire Paritaire des métiers de l'Informatique, de l'Ingénierie des Études et du Conseil.
Référentiels métiers de la branche du numérique, de l'ingénierie, des études et du conseil et de l'évènement.

2. LES FACTEURS LES PLUS INFLUENTS

L'anticipation des facteurs de changement, c'est-à-dire la détermination des facteurs de l'évolution du métier de *data scientist* s'est effectuée sur base de l'analyse prospective, selon la méthodologie *Abilitic2Perform*, sur les métiers de la filière data réalisée fin 2015.[27]

Elle s'est déroulée en deux étapes. La première consistait à présenter de manière détaillée aux experts les douze facteurs les plus influents retenus lors de l'analyse prospective relative aux métiers de la filière data.

Suite à cette présentation, et pour réaliser la mise à jour des facteurs, la question suivante a été posée aux experts : *Quels sont, dans un horizon de trois à cinq ans (2020-2022), les facteurs qui détermineront/influenceront le métier du data scientist ?*

Après un temps de réflexion individuelle, chaque expert a commenté et réagi aux facteurs présentés. Certains ont été reformulés, d'autres supprimés.

Au total, les participants ont ainsi recensé dix facteurs de changement qui relevaient de différentes dimensions : politique, économique, socioculturelle, technologique, légale.

Ces deux étapes sont réalisées lors du premier atelier.

A1	Accessibilité technologique des outils du Big data.
A2	Maturation des technologies informatiques du *Big data*.
A3	Émergence de standards (plus d'interopérabilité).
A4	Rapprochement de l'informatique industrielle et de l'informatique de gestion.
A5	Obligation / impulsion d'ouvrir les données (*open data*).
A6	Développement et établissement de l'Internet des Objets.
A7	Comportement de l'utilisateur en demande d'individualisation et d'immédiateté.
A8	Diversification des types de données (textes, images, vidéos, sons, …).
A9	Développement dans les entreprises d'ERP/CRM/robotisation... qui ouvre à la prise en compte de la réflexion en matière de data.
A10	Perception des data comme ressources valorisables.
A11	Évolution de politiques de protection et de sécurisation des données.
A12	Ajout de service / d'intelligence sur un produit.

***Tableau 2** : Facteurs dominants retenus lors de l'analyse « filière data ».*

A1	Accessibilité technologique des outils du *Big data*.
A2	Développement de l'internet des objets.
A3	L'ajout de service/d'intelligence sur les produits : les données comme vecteur d'innovation.
A4	Rapprochement de l'informatique industrielle et de l'informatique de gestion.
A5	Ouverture à la réflexion data.
A6	L'ouverture des données.
A7	Attitude du consommateur face au partage des données.
A8	Politiques efficaces de protection et de sécurisation des données.
A9	Contexte politique.
A10	Accessibilité des formations.

***Tableau 3** : Facteurs les plus influents retenus pour le métier de data scientist.*

[27] https://www.leforem.be/chiffres-et-analyses/metiers-d-avenir-prospectives-abilitic2perform.html.

3. LES ÉVOLUTIONS PROBABLES ET SOUHAITABLES

Une fois ces dix facteurs déterminés, il s'agissait d'envisager leur évolution possible. Pour ce faire, il a été demandé aux experts, lors du second atelier, de décrire les situations actuelles et futures (dans un horizon de trois à cinq ans) pour chaque facteur de changement. Il leur a été proposé dans cet ordre ; un temps de réflexion individuelle puis en plénière, en repartant des scénarios élaborés lors des ateliers de la filière data, d'actualiser et d'adapter ceux-ci aux réalités du *data scientist*. Ils devaient décrire trois types d'évolution potentielle : une pessimiste, une intermédiaire et une optimiste. Chaque scénario a été débattu et reformulé en séance, afin qu'il soit validé par le groupe.
Ils ont ensuite été soumis au vote des participants qui étaient invités à exprimer, d'une part, une estimation du caractère probable du scénario, d'autre part, une appréciation de son caractère souhaitable.

4. LE PROFIL D'ÉVOLUTION

Le tableau des pages suivantes a servi d'input au troisième atelier, dont le premier objectif était, pour chaque facteur, de retenir le scénario à considérer pour la suite du travail : le scénario le plus probable a été confronté au scénario le plus souhaitable. Lorsque le scénario le plus probable était différent du scénario le plus souhaitable, un arbitrage était réalisé entre les deux scénarios. Si le groupe estimait qu'il était possible de mettre en œuvre des actions permettant d'atteindre le scénario le plus souhaitable, c'est celui-ci qui était retenu. Dans le cas inverse, on retenait le scénario le plus probable.
La formulation de certains des scénarios retenus a été légèrement précisée ou enrichie à l'occasion de cette discussion.

Note de lecture du tableau 4 :
Les hypothèses d'évolution ayant été identifiées comme les plus probables sont sur fond bleu et en italique.
Les hypothèses d'évolution identifiées comme les plus souhaitables sont sur fond jaune et soulignées.
Lorsque l'hypothèse d'évolution la plus probable est identique à la plus souhaitable, elle apparaît sur fond rose en italique et soulignée. Les **hypothèses d'évolution retenues**, parce que probables et souhaitables, ou après arbitrage, sont **en gras**.

Facteurs de changement	Hypothèses d'évolution des facteurs clés à l'horizon 2020-2022			
	A	B	C	D
F1. Accessibilité et maturation des outils technologiques du big data.	En 2022, l'usage d'outils de gestion de bases de données distribuées permet la mise en œuvre de projets de collecte et de stockage de données dans des grandes entreprises de la grande distribution. Cela reste cependant cantonné à certains outils "emblématiques" tel que Hadoop. L'offre technologique n'est pas arrivée à maturité. L'utilisation s'apparente encore à du bricolage.	***En 2022, le succès d'outils open source a fait émerger quelques standards qui rendent des outils accessibles et plus simples à l'usage. Ils sont souvent proposés sous forme de « services spécialisés ». Les technologies du Big data sont stables, adaptées aux besoins et reposent sur une interopérabilité entre les différents outils*.**	En 2022, l'offre technologique s'est développée et des outils automatiques, flexibles, multiplateformes sont disponibles. Quelques grandes industries intègrent dans leur business plan des projets d'exploitation de *Big data* en s'appuyant sur des outils de stockage et quelques applications d'analyse customisées. Les technologies du *Big data* sont matures et permettent un développement exponentiel de la filière.	
F2. Développement de l'internet des objets.	En 2022, les débouchés industriels sont nombreux pour les objets connectés qui se multiplient, mais restent cantonnés dans cette seule sphère. La concurrence ralentit l'émergence de standards. Le sens et la valeur des milliards de données récoltables et récoltées sont encore peu clairs.	***En 2022, la technologie des objets connectés concerne de plus en plus les particuliers (voiture, énergie, domotique, santé…). Les objets connectés deviennent ainsi de plus en plus nombreux, générant une quantité importante de données.***	En 2022, les objets connectés sont répandus tant au niveau industriel que privé. De grandes quantités de données sont produites. Leur gestion et leur exploitation sont le plus souvent maîtrisés.	
F3. L'ajout de services / d'intelligence sur les produits : les données comme vecteur d'innovation.	En 2022, la recherche commence à se développer, des startups se lancent sur des produits et des services innovants. L'e-commerce est le terreau de cette exploitation.	***En 2022, l'offre de gestion à distance (maintenance, services annexes, …) se développe, intégrée aux produits. Les secteurs technologiques sont bien sûr les premiers concernés***	En 2022, le "connected / data driven by design" est rentré dans les processus de conception et de fabrication. C'est un axe d'innovation, une opportunité pour les PME et leur emploi, mais qui nécessite des compétences nouvelles et spécifiques.	

Bleu italique : plus probable – Jaune souligné : plus souhaitable – *Rose italique souligné : probable et souhaitable* – **Gras : hypothèse retenue**

Facteurs de changement	Hypothèses d'évolution des facteurs clés à l'horizon 2020-2022			
	A	B	C	D
F4. Rapprochement de l'informatique industrielle et de l'informatique de gestion.	En 2022, le rapprochement entre les informatiques industrielles et de gestion se concentre sur des portions du cycle de production et essentiellement dans des grandes entreprises.	***En 2022, les outils de l'informatique de gestion "se diffusent" dans l'ensemble du cycle industriel (évolution numérique des métiers). Des formations de niche à haute valeur ajoutée voient le jour.***	En 2022, dans l'entreprise 4.0, informatique de gestion et informatique industrielle sont intégrées. Le marché de l'emploi est en demande de personnes avec des connaissances dans les domaines des "données", des "métiers" et de la "production".	
F5. Ouverture-Perception des data comme ressources valorisables (en entreprise).	En 2022, les perceptions n'ont pas beaucoup changé, et les réticences au développement du management des données demeurent. Les discours et les incitants n'intéressent que les convaincus.	En 2022, les politiques d'information et d'incitation soutiennent une progression lente, mais régulière de la prise de conscience de la valeur des données. La qualité des données est un enjeu majeur dans cette optique, pour renforcer le mouvement.	*En 2022, des produits et services ciblés sur la valorisation et l'exploitation des données se sont progressivement développés à destination de l'ensemble des PME dans tous les secteurs. La collecte et l'analyse des données au sein des entreprises est plus systématique.*	En 2022, le data management fait partie du business plan de 50 % des PME. Les données constituent un "patrimoine" qu'il faut gérer, négocier, développer. Cette perception est étendue aux secteurs industriels.
F6. L'ouverture des données/open data.	**En 2022, des initiatives politiques dispersées et limitées ouvrent l'accès à des bases de données proches de l'accessibilité, mais les investissements nécessaires restent très faibles. La Wallonie est à la traîne par rapport aux autres régions**	*En 2022, notamment sous les impulsions européennes, l'ouverture des données publiques commence à se répandre. Toutefois, certains "propriétaires de données" sont encore récalcitrants. De plus, l'ouverture des données nécessite des connaissances techniques et "métier" difficiles à trouver sur le marché.*	En 2022, l'ouverture des données est stimulée par la stratégie européenne et les politiques régionales. Les flux dérivés font l'objet d'exploitation. Les données publiques sont utilisées pour compléter les données "privées", permettant d'obtenir de meilleurs résultats.	
F7. Attitude du consommateur face au partage des données.	En 2022, les préoccupations liées à la protection de la vie privée tendent à réduire l'explosion de la collecte de données individuelles, le particulier ayant à cœur une protection accrue de sa sphère privée.	*En 2022, le modèle dominant est l'individualisation des données. L'utilisateur en est demandeur sans toutefois être conscient ou sans pouvoir contrôler ses données. Le besoin d'immédiateté limite l'esprit critique sur la nécessité.*	**En 2022, la demande d'individualisation de la part de l'utilisateur pour certains types d'applications est un soutien déterminant du développement du big data. Toutefois, cette individualisation est réalisée de manière plus transparente et plus contrôlée avec par ex. davantage de modèles "à la carte".**	En 2022, les services apportés par l'exploitation des méga données sont vécus par la majorité des utilisateurs comme synonyme d'amélioration de confort, de qualité de vie, de satisfaction. Le consommateur trouve un réel intérêt à partager ses données.

Bleu italique : plus probable – Jaune souligné : plus souhaitable – *Rose italique souligné : probable et souhaitable* – **Gras : hypothèse retenue**

Facteurs de changement	Hypothèses d'évolution des facteurs clés à l'horizon 2020-2022			
	A	B	C	D
F8. Politiques efficaces de protection et de sécurisation des données.	En 2022, les règlementations sur la protection des données sont fortes et contraignantes. Certains projets n'aboutissent pas et sont annulés. La quantité de données exploitables ayant diminué fortement.	***En 2022, la définition de cadres de protection applicables est en train de se faire, notamment sous l'impulsion européenne. Certains projets prennent du retard. Les sociétés s'adaptent difficilement à ces contraintes.***	En 2022, des réglementations fortes sont sorties, qui ont force de loi et doivent être prises en compte. Protection et sécurité des données ont pris toute leur place dans le data management. Les sociétés se sont adaptées en développant des compétences en "droit" et "techniques données".	
F9. Contexte politique et sociétal (terrorisme, fraude fiscale, mobilité, smart-cities services publics…).	En 2022, le contexte politique tendu bride fortement l'accès aux données exploitables, de peur que celles-ci soient utilisées à mauvais escient.	***En 2022, le pouvoir politique comprend l'intérêt d'utiliser les données afin d'optimiser leur gestion.***	En 2022, d'importants financements publics sont déployés pour favoriser l'utilisation des données pour une meilleure gouvernance.	

Bleu italique : plus probable – Jaune souligné : plus souhaitable – *Rose italique souligné : probable et souhaitable* – **Gras : hypothèse retenue**

Facteurs de changement	Hypothèses d'évolution des facteurs clés à l'horizon 2020-2022			
	A	B	C	D
F10. Accessibilité des formations (cursus scolaire, formation qualifiante, MOOC, formations en ligne, etc.).	En 2022, l'offre de formation de haut niveau stagne, les filières en place n'attirent pas assez de candidats, étant donné les prérequis demandés. Les formations en ligne de qualité disponibles sont coûteuses et non « certifiantes » (ou peu valorisables sur le marché du travail).	***En 2022, l'offre de formation s'est développée. Les masters en data science se sont développés au départ des filières scientifiques et technologiques classiques. Des formations BI, à l'analyse et à la visualisation des données, etc. enrichissent les formations dans différents secteurs (marketing, informatique, management…). En ligne, l'offre de formation de base est assez riche et offre une initiation de bonne qualité pour entrer dans un processus de (ré)orientation vers des métiers ou des « grappes de compétences » data. Elle offre également de bonnes bases de formations de mise à jour aux nouveaux outils et développements.***	En 2022, les formations initiales en sciences des données se sont multipliées. Toutes les filières scientifiques et technologiques traditionnelles (informatique, mathématiques, statistiques…) ont inclus des enseignements et des certifications «data » sur des thèmes (visualisation, analyse de données, science des données…) et à des niveaux variés de qualification. L'offre de formation en ligne est très importante et couvre l'ensemble des besoins, de l'initiation aux certifications spécialisées en passant par des masters généralistes. Cette offre en ligne est accessible sur différentes bases : intégrée dans des formations initiales ou indépendantes, gratuites ou (partiellement ou totalement) payantes, avec des certifications reconnues et valorisables sur le marché du travail ou à seul titre d'autoformation initiale et continue. Les nouvelles tendances et les nouveaux développements (plateformes, langages, domaines d'application…) sont rapidement intégrés dans l'offre en ligne, soutenant tout à la fois la mise à jour des formations initiales que facilitant les reconversions professionnelles vers les data sciences.	

Tableau 4 : Hypothèses d'évolution.

Bleu italique : plus probable – Jaune souligné : plus souhaitable – *Rose italique souligné : probable et souhaitable* – **Gras : hypothèse retenue**

5. LES IMPACTS SUR LES ACTIVITÉS ET LES BESOINS EN COMPÉTENCES

La dernière étape du travail réalisé avec les participants a porté sur l'identification des compétences que le *data scientist* devrait développer pour mener à bien ses tâches d'ici 2022. L'objectif de ce recensement de compétences est d'éclairer sur les futurs besoins en compétences.

Le groupe a donc été invité à proposer des ressources nécessaires à l'exercice du métier de *data scientist*. Pour alimenter les réflexions, le groupe s'est appuyé sur le chemin d'évolution (soit les dix scénarios) construit durant les deux premiers ateliers ainsi que sur les activités de base telles qu'elles ont été proposées lors du premier atelier.

Cet exercice a fait ressortir une liste non exhaustive de compétences clés à approfondir et/ou à développer pour les *data scientist* dans un horizon de trois à cinq ans.

Activités	Hypothèses impactantes	COMPÉTENCES À MOBILISER (il faut …)
Identifier les besoins et la problématique	5-9	- Comprendre les enjeux de la problématique traitée. - Comprendre l'environnement suffisamment que pour traduire la réalité et les besoins en un ensemble de règles mathématiques, les affiner, les actualiser. - Identifier le potentiel de valeur des données, à quoi les données pourraient être utiles. - Expliquer la valeur potentielle des données et les opportunités business (évangélisation). - Avoir des attitudes positives envers la recherche : créativité, pensée prédictive. - Mobiliser une logique diagnostique (Formuler des hypothèses afin de pointer le problème et pouvoir le résoudre).
Identifier et extraire les données disponibles	2-6-10	- Comprendre les fondements et les contraintes spécifiques de l'IOT28 : données distribuées, données en streaming, etc. - Accorder de l'importance et du temps à une veille des outils d'extraction et de stockage, d'accès. - Déterminer les données à extraire des flux disponibles. - Accéder aux données et les extraire là où elles se trouvent (web scraping[29]). - Comprendre les impacts de la logique temporelle sur les méthodes d'accès et de traitement des données. - Connaitre les contraintes légales et règlementaires relatives aux données (GDPR…).
	1-2-10	- Développer, mettre à jour ses compétences en développement informatique (méthode, langage).

[28] IOT : Internet Of Thing (internet des objets).
[29] Technique d'extraction de contenu de site web.

Exploiter et analyser les données		- Adapter la complexité des algorithmes aux besoins et objectifs des traitements et analyses (quantités de données, rapidité attendue de résultats, finesse et précision des résultats attendus).
		  o Estimer des ordres de grandeur : temps de traitement et de réponse, marges d'erreur, validité des résultats (de contenu, prédictive…), tailles d'échantillon…
		  o Dans l'ensemble des méthodes existantes choisir la plus adaptée à la problématique.
		- Accorder de l'importance et du temps à une veille des méthodes et outils de traitement (Participer à la communauté).
		  o Rechercher des algorithmes existants proches de ses besoins.
		  o Adapter les algorithmes existants à ses besoins.
		- Avoir conscience des contraintes liées à l'architecture du Big data.
		- Connaitre et utiliser des outils de text mining (NLP)[30].
		- Connaitre et utiliser des outils de gestion et de traitement des données distribuées (Hadoop, Spark, Flink...).

Communiquer les résultats	5	- Utiliser les outils de reporting et de visualisation des données (« data viz »). - Utiliser les outils analytiques interactifs (notebooks, shiny, pages web+D3JS…)

***Tableau 5** : Détails des compétences clés.*

[30] Ensemble de techniques permettant de créer de l'information à partir de texte.

NOUS REMERCIONS POUR LEUR PARTICIPATION AU PROCESSUS EN QUALITÉ D'EXPERTS

Hervé BATH, Founder of Eura Nova
Stéphane FAULKNER, Directeur du département des sciences de gestion, Université de Namur
Pierre GEURTS, Institut Montefiore, ULG
Eric LECOUTRE, Data Science Advisor, WeLoveDataScience
Pierre LELONG, Manager Formations Entreprises et Enseignement, Technofutur TIC
Simon PETIT, Associate Partner, Dataroots
Martine VANCAMBERG, Manager formations qualifiantes pour demandeurs d'emploi, Technofutur TIC
Elizabeth VERSAILLES, Academics Relationship Manager, SAS
Michel VERSTREPEN, Responsable ligne de produits transition numérique, Forem
Jef WIJSEN, professeur département informatique, UMons

ENCADREMENT MÉTHODOLOGIQUE DE LA DÉMARCHE ET RÉDACTION DU RAPPORT FINAL

Le Forem, Service de veille, analyse et prospective du marché de l'emploi :
Jean-Claude CHALON, Direction
David Pieroux, Coordination du projet
Michel Orban, Back Officer
Aurélie LELUPE, Animation et rédaction

ÉDITEUR RESPONSABLE

Marie-Kristine VANBOCKESTAL, Administratrice générale, Le Forem