MINISTÈRE
CHARGÉ
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

Données de recherche : caractérisation des fonctions clés liées à leur gestion, leur exploitation, leur diffusion et leur utilisation

Table des MATIÈRES

INTRODUCTION

Les données de recherche ont toujours constitué un élément central de la plupart des pratiques scientifiques. Aujourd'hui, la production de données nativement réutilisables - FAIR-by design (Faciles à trouver, Accessibles, Interopérables et Réutilisables) - s'impose comme un objectif majeur afin de faire des données une véritable production scientifique, à la fois citable et réutilisable. La réutilisation des données concerne aussi bien les communautés à l'origine de leur création que, lorsque cela est possible, l'ensemble de la société. La réutilisation des données existantes réduit la duplication des efforts de collecte, d'expérimentation ou de traitement, ce qui permet des gains de temps, de ressources et de financements augmente la capacité à traiter des questions scientifiques complexes. Gérer les données en suivant les principes FAIR, c'est aussi éviter de garder des données inutiles car mal documentées et donc impossibles à utiliser, ce qui permet par la même occasion de réduire l'impact environnemental du numérique dans le domaine de la recherche.

Rendre les données FAIR mobilise un éventail de compétences et de fonctions spécifiques. Ces activités, souvent transversales et en constante évolution, font intervenir une diversité de profils professionnels, depuis la production jusqu'à la valorisation des données, en passant par leur structuration, leur exploitation et leur protection ou partage.

Comprendre et caractériser les fonctions impliquées dans cette dynamique est devenu un enjeu stratégique, tant pour accompagner la transformation des métiers de la recherche que pour développer des formations adaptées, valoriser les compétences mobilisées et garantir l'impact des productions scientifiques. C'est dans cette perspective qu'une étude a été menée, combinant revue de la littérature et enquête auprès de la communauté française, afin de faire émerger un panorama des activités, rôles et expertises structurant l'écosystème des données de recherche.

1 POINTS DE REPÈRE : DÉFINITIONS ET CONCEPTS CLÉS

En préambule, il convient de rappeler que les données de la recherche se définissent comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche » [OCDE, 2007]. Elles constituent l'ensemble des informations collectées ou produites au cours d'un travail scientifique, dans le but de mieux comprendre ou simuler un phénomène, de tester des hypothèses pour produire de nouvelles connaissances. Elles prennent des formes variées, en fonction du domaine disciplinaire, de la méthodologie employée ou de la nature de l'objet étudié.

- Les **données d'expérimentation** sont issues de dispositifs sur lesquels le chercheur intervient de manière active, en contrôlant certaines variables dans le cadre d'un protocole défini. Ces données permettent de tester des hypothèses dans un environnement contrôlé, notamment en laboratoire ou dans le cadre d'essais cliniques.
- Les **données issues de capteurs ou d'instruments** scientifiques sont recueillies automatiquement grâce à des dispositifs techniques : données GPS, images satellites, IRM, relevés de capteurs environnementaux notamment.
- Les **données d'enquête** sont produites par l'intermédiaire de questionnaires ou d'entretiens. Ce type de données est particulièrement utilisé dans les sciences humaines et sociales, car il permet d'accéder directement aux représentations, opinions, comportements ou expériences des personnes interrogées.
- Les **données synthétiques** sont des données artificiellement créées par des algorithmes pour reproduire les caractéristiques statistiques de données réelles, sans en copier les valeurs exactes. Elles sont utilisées pour protéger la confidentialité, tester des outils ou simuler des scénarios, tout en restant proches des conditions réelles.
- Les **données de simulation**, générées par des modèles mathématiques ou informatiques, reproduisent artificiellement un phénomène réel ou théorique. Ces données sont précieuses notamment pour la compréhension des phénomènes climatiques ou économiques.

- Les **données textuelles ou documentaires**, telles que les textes littéraires, les documents administratifs, les contenus numériques, les corpus historiques, des compte-rendus d'entretiens d'enquête…, sont souvent explorées dans des recherches qualitatives ou en sciences humaines.
- Les **données dérivées** ne sont pas collectées directement mais construites à partir de données brutes préexistantes. Ce sont, par exemple, les données issues d'analyse statistique, les visualisations, les données issues du *text and data mining*, les synthèses ou les transcriptions d'entretiens, des données issues de simulation... Ces données sont le fruit d'un travail de traitement, de sélection ou d'analyse, et elles peuvent parfois servir à d'autres recherches que celle pour laquelle elles ont été initialement produites.

Les données de recherche constituent un matériau riche, varié, parfois brut, parfois transformé, dont la diversité reflète la pluralité des approches scientifiques. Pour devenir des jeux de données exploitables, durables et réutilisables, ces données requièrent un travail scientifique de **sélection**, mise en qualité, de structuration et de documentation permettant leur usage non seulement par l'équipe qui les a produites, mais aussi par d'autres chercheurs, des acteurs économiques ou la société dans son ensemble.

Les métadonnées, selon la NISO (2004), sont des informations structurées qui décrivent, expliquent, localisent ou facilitent la gestion des ressources informationnelles.

Dans cette étude dédiée aux données de la recherche, on considère trois types principaux de métadonnées : descriptives, structurelles et administratives.

- Les **métadonnées descriptives ou scientifiques** dans le cadre des données de recherche précisent le protocole et le contexte d'obtention des données, leur origine, leurs transformations et outils associés, les conventions adoptées et les conventions de nommages scientifiques (e.g., nomenclatures, ontologies, terminologies) et standards utilisés pour garantir leur interopérabilité et leur pérennité.
- Les **métadonnées structurelles** décrivent l'organisation et les relations entre les fichiers ou éléments d'un jeu de données de recherche. Elles facilitent la compréhension, le traitement et la réutilisation des données.
- Les **métadonnées documentaires** servent à identifier et à situer un jeu de données dans son contexte et modalités de production et consommation. Elles comprennent notamment le titre, les producteurs, la date de création, la version, l'établissement responsable, ainsi que les modalités d'accès et les licences associées.

2 REVUE DE LITTÉRATURE

La revue de littérature a révélé qu'un flou persiste autour des fonctions liées à la gestion des données de recherche. Il est ainsi fréquent de les voir se confondre dans des intitulés de postes ou des titres disparates, d'être attribuées à tort à des personnes dont les compétences ou le positionnement dans l'organisation ne permettent pas d'agir de manière optimale, et surtout, de ne pas être reconnues à leur juste valeur.

La littérature identifie cinq grands rôles dans la gestion des données, chacun couvrant une partie du cycle de vie.

- Le data steward assure la coordination des actions tout au long du cycle de vie des données, en veillant à leur qualité, leur conformité aux normes et politiques en vigueur, ainsi qu'à leur bonne gestion à chaque étape, depuis la donnée brute jusqu'à sa valorisation.
- Le data engineer conçoit, développe et optimise les systèmes de collecte, de structuration, de stockage et de traitement technique des données, en lien avec les infrastructures.
- Le data curator sélectionne, prépare, enrichit, décrit et préserve les données afin de garantir leur qualité, leur interopérabilité, leur accès pérenne et leur réutilisation.
- Le data scientist explore, analyse, modélise et valorise les données, produisant des résultats exploitables et des données dérivées à forte valeur ajoutée. De son action sont issus les jeux de données validées scientifiquement.
- Le data librarian documente les jeux de données, facilite leur découverte, leur diffusion et leur archivage, et définit leurs modalités d'accès, de citation et leurs licences d'usage.

Cette revue de littérature a révélé un périmètre généraliste, non spécifique aux contextes de recherche et très orienté vers le monde anglo-saxon, nécessitant donc de mener un travail complémentaire et dédié au contexte français de la recherche.

3 ENQUÊTE AUPRÈS DE LA COMMUNAUTÉ SCIENTIFIQUE

3.1. Méthodologie

L'enquête menée auprès de la communauté scientifique avait pour objectif principal de mieux cerner les profils, les activités et les parcours de formation des personnels impliqués dans la gestion des données tout au long de leur cycle de vie. Il s'agissait de comprendre de manière fine la diversité des intervenants, qu'il s'agisse de chercheurs, de personnels de soutien à la recherche ou de spécialistes des données, et d'identifier les pratiques et compétences mobilisées à chaque étape, de la collecte à la diffusion des données. L'enquête quantitative a ainsi structuré son questionnaire en s'appuyant sur le cycle de vie des données de recherche ci-dessous.

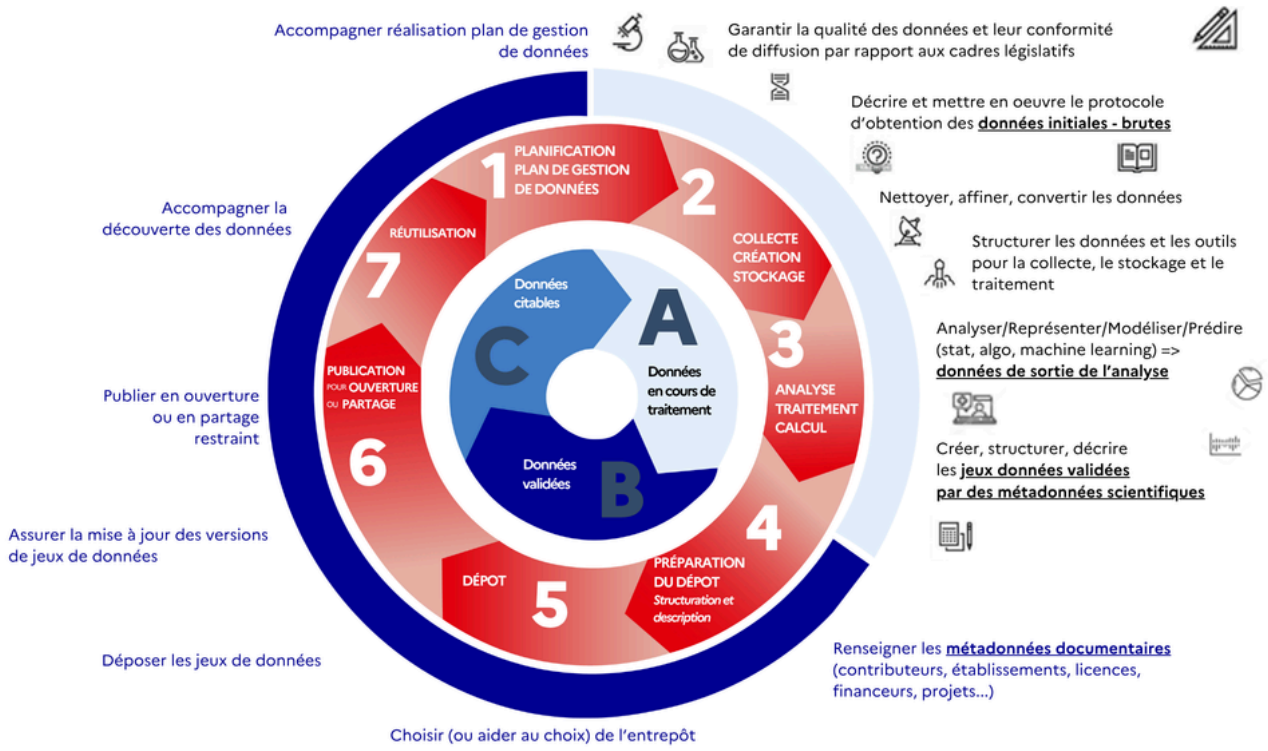

Figure 1 : Cycle de vie de la donnée et activités à chaque étape

Bien que le cycle de vie des données soit souvent présenté comme fermé et linéaire, il est en réalité ouvert, flexible et dépend fortement du contexte. Les données sources peuvent provenir de dispositifs spécialisés, d'entrepôts ou de fournisseurs externes (Open Data, archives, données météo, etc.), déjà partiellement structurées. Dans les faits, le cycle de recherche se caractérise par des discontinuités, des itérations et des chevauchements, témoignant de la pluralité des démarches scientifiques, lesquelles varient selon la nature des questions de recherche et les spécificités disciplinaires. Les producteurs et consommateurs de données mènent des activités couvrant partiellement ce cycle. Malgré cela, ce modèle cyclique reste utile pour illustrer les grandes étapes — nettoyage, transformation, organisation, exploitation — qui permettent de produire des jeux de données structurés et potentiellement diffusables. Il sert de cadre de référence pour situer les interventions, qu'elles soient de nature technique, scientifique ou documentaire, tout au long de l'organisation de la recherche.

Pour enrichir et nuancer les résultats de l'enquête quantitative, une enquête qualitative complémentaire a été menée sous la forme d'entretiens semi-directifs. Ceux-ci ont permis de recueillir des retours d'expérience approfondis, de mieux comprendre les logiques de collaboration entre acteurs, et de mettre en lumière la diversité des trajectoires professionnelles et des contextes institutionnels.
Cette double approche méthodologique, combinant une enquête par questionnaire ayant recueilli 1 565 réponses et 26 entretiens semi-directifs, a permis de croiser des données quantitatives avec des témoignages approfondis, offrant ainsi une vision à la fois globale et contextualisée des pratiques liées à la gestion des données de recherche.

3.2. Résultats de l'enquête : profils, pratiques et enjeux de reconnaissance

Les enquêtes ont révélé que le laboratoire[1] constitue le principal lieu d'ancrage des activités liées aux données de recherche, dans **69 %** des cas. La répartition des répondants entre les catégories professionnelles est relativement équilibrée : **55 %** sont des personnels de soutien à la recherche et **45 %** des (enseignants-)chercheurs, y compris les personnes en doctorat ou post-doctorat. Les participants couvrent l'ensemble des champs disciplinaires, avec une forte représentation dans les sciences humaines et sociales, la biosanté, les sciences de la Terre et de l'Univers, l'environnement, l'agronomie et l'écologie.

L'enquête montre que les compétences mobilisées s'étendent sur l'ensemble du cycle de vie des données : collecte, nettoyage, documentation, stockage, analyse, diffusion, ouverture et préservation. Les activités sont rarement limitées à une seule étape : les répondants interviennent souvent sur plusieurs segments du cycle, ce qui rend les frontières entre fonctions particulièrement poreuses.

Les parcours menant à ces fonctions sont variés. Certains y accèdent par nécessité dans des projets scientifiques ou par une prise de conscience individuelle de l'importance de ces questions. D'autres à travers des réseaux liés à la science ouverte ou à des missions spécifiques dans la recherche. La reconnaissance de ces missions progresse, mais demeure inégale, souvent invisible et peu reconnue au niveau institutionnel. De nombreux répondants se considèrent comme des intermédiaires entre chercheurs, infrastructures techniques et dispositifs de diffusion, jouant un rôle de passeurs de compétences.

[1] *Dans ce document, le terme « laboratoire » s'entend au sens large d'une structure de recherche ou d'appui : plateforme scientifique, laboratoire de recherche, équipe scientifique, …*

L'étude confirme un enjeu fort de reconnaissance et de formation. Les fonctions liées aux données restent encore insuffisamment valorisées, ce qui complique leur structuration et l'adéquation entre besoins et formations disponibles. Si l'auto-formation comble partiellement ce déficit (elle est largement pratiquée – **51%** des répondants de l'enquête), une demande claire émerge pour des formations mieux structurées et pour la reconnaissance des compétences acquises sur le terrain.

Dans le prolongement des constats issus de la synthèse bibliographique, l'enquête met également en lumière le flou persistant autour des fonctions associées à la gestion des données. Celles-ci sont souvent mal définies, dispersées entre des intitulés de postes très variés, et parfois inadaptés. Chaque acteur tend à adopter la dénomination qui lui semble correspondre, sans cadre commun, ce qui entretient la confusion. Cette situation révèle à la fois la porosité entre les fonctions et une confusion entre les activités centrées sur les données de la recherche elles-mêmes et celles portant sur les métadonnées.

Ces constats, croisant analyses bibliographiques et données d'enquête, soulignent la nécessité de mieux définir les fonctions opérationnelles liées à la gestion, au traitement et à la diffusion des données de recherche. La suite du document vise ainsi à proposer des repères concrets pour les acteurs concernés, les encadrants et les institutions, afin de mieux identifier ces fonctions, en faciliter la reconnaissance et accompagner le développement des compétences associées.

4 VERS UN CADRE COMMUN DES FONCTIONS OPÉRATIONNELLES CLÉS DANS LA GESTION DES DONNÉES DE RECHERCHE

Ce document propose des repères pour mieux comprendre sept fonctions clés qui présentent une dimension opérationnelle concrète et effectivement exercées dans les contextes actuels de la recherche.

- Intendance des données,
- Curation des données,
- Ingénierie des données,
- Analyse des données,
- Science des données,
- Curation des métadonnées documentaires,
- Documentation des données.

Ces 7 fonctions ne résument pas toutes les expertises nécessaires pour la gestion et la diffusion des données. Ces 7 fonctions s'appuient selon les contextes et données sur les expertises de délégué à la protection des données, juriste, référent intégrité scientifique, responsable de l'éthique, du RSSI, du FSD, d'acteurs de la valorisation mais aussi des archivistes et des métiers d'informaticiens spécialistes des infrastructures de stockage, archivage, calcul… Nous avons fait le choix de ne pas détailler dans les fonctions clés l'archivage des données, une étude étant en cours. Le cas spécifique de la fonction de gouvernance des données, qui est stratégique et définit un cadre global dans lequel s'inscrivent les 7 fonctions clés, est précisé en fin de document.

Les 7 fonctions décrites jouent chacune un rôle spécifique dans la gestion, le traitement et la diffusion des données de recherche et interviennent à différentes étapes du cycle de vie des données. Une présentation plus détaillée sous forme de fiche en est faite après la présentation globale de chacune de ces fonctions.

Intendance des données ou *data stewardship*

L'intendance des données supervise la qualité, la conformité, la sécurité, la documentation et la gestion des accès aux données. Elle joue un rôle d'interface entre les scientifiques, les informaticiens, les experts de la documentation et d'autres expertises spécifiques selon la nature des données. L'intendance des données (ou *data stewardship*) assure la cohérence et la coordination des actions tout au long du cycle de vie des données, en veillant au respect des normes scientifiques, politiques et aux cadres légaux en vigueur. Elle s'inscrit ainsi dans un rôle de facilitateur, garant du respect des règles, acteur de la sensibilisation et relais de compétences entre les différentes parties prenantes.

Curation des données de recherche ou *research data curation*

La curation des données consiste à sélectionner, organiser, nettoyer, enrichir, décrire et documenter les données afin de garantir leur qualité, cohérence, interopérabilité et pertinence scientifiques. Elle transforme les données brutes en ressources exploitables, bien documentées et partageables, tout en assurant leur accessibilité durable et en facilitant leur réutilisation. Cette fonction s'inscrit dans une approche scientifique visant à valoriser les jeux de données. Les métadonnées utilisées sont de nature scientifique et précisent le protocole et le contexte de collecte, l'origine des données, leurs transformations, les outils associés, ainsi que les conventions et standards adoptés pour garantir leur interopérabilité et pérennité : choix des nomenclatures de nommage des données, références aux ontologies, terminologies et vocabulaires contrôles spécifiques à chaque discipline.

Ingénierie des données ou *data engineering*

La fonction d'ingénierie des données consiste à concevoir, développer et maintenir les infrastructures, outils et flux nécessaires à la collecte, au stockage, au traitement technique et à la mise à disposition des données de recherche. Elle assure la mise en place, la maintenance et l'optimisation des architectures de données, en garantissant la performance, la robustesse et la sécurité des systèmes de traitement. Elle veille à ce que les données soient accessibles, sécurisées, bien structurées et techniquement prêtes à être exploitées par les autres fonctions. Il s'agit d'une fonction à dominante technique, dont l'objectif principal est d'automatiser, fiabiliser et sécuriser les processus de gestion des données, en les adaptant aux besoins spécifiques des projets ou des environnements de recherche.

Analyse des données

La fonction d'analyse des données consiste à explorer, interpréter et exploiter les données de recherche afin de produire des résultats scientifiques, tester des hypothèses et générer de nouvelles connaissances. Elle mobilise des méthodes quantitatives, qualitatives ou mixtes, selon les disciplines, et peut s'appuyer sur des techniques statistiques, des outils de modélisation, de visualisation, ou encore des approches computationnelles. Elle contribue à la valorisation scientifique des données, en les traduisant en résultats interprétables, publiables et potentiellement réutilisables.

Science des données ou *data science*

La science des données est un domaine interdisciplinaire qui mobilise statistiques, apprentissage automatique, modélisation et visualisation pour transformer des données hétérogènes en connaissances exploitables. Elle conçoit et met en œuvre des modèles complexes pour simuler, prédire ou comprendre des phénomènes, de plus en plus souvent à partir de données massives ou non structurées. Au-delà de l'analyse, elle intègre une forte dimension computationnelle et algorithmique. Son application s'étend de la recherche fondamentale à l'aide à la décision, en passant par la production de données dérivées à haute valeur ajoutée. Elle requiert une expertise croisée en mathématiques, programmation, traitement de données et compréhension des problématiques scientifiques. Les résultats sont restitués sous forme de modèles, visualisations interactives ou intégrés à des systèmes intelligents.

Documentation des jeux de données validés

La documentation des données consiste à créer, organiser et maintenir des informations détaillées sur les jeux de données validés afin de garantir leur compréhension, leur accessibilité et leur réutilisabilité. Cette fonction se concentre sur la gestion active des métadonnées documentaires (format, date de création, modalités d'accès, licences…) qui facilitent leur découvrabilité, leur citation et leur réutilisation par des scientifiques et la société dans son ensemble. Au-delà de son rôle d'expertise sur la publication et la citation des données, cette fonction vise à sensibiliser et à accompagner les chercheurs tout au long du cycle de vie des données. Elle intervient à deux étapes clés du cycle de vie des données : en amont, en soutenant les équipes dans la rédaction des plans de gestion des données, et en aval, en les guidant dans le partage ou l'ouverture des jeux de données issus des projets, via des entrepôts accessibles au-delà du cercle restreint de l'équipe de recherche qui les a produites.

Curation des métadonnées documentaires

Cette fonction constitue un sous-ensemble de la précédente, dans la mesure où elle se concentre sur le renseignement et la qualité des métadonnées saisies par les déposants de jeux de données dans un entrepôt, en vue de leur diffusion, qu'elles soient en accès restreint ou ouvertes.

Bien que ces 7 fonctions soient décrites séparément pour des raisons de clarté, elles sont en pratique étroitement imbriquées et souvent assumées simultanément par une seule et même personne, en particulier par les scientifiques eux-mêmes. Ce cumul de responsabilités souligne l'implication directe des scientifiques dans la gestion, la structuration, la documentation et la valorisation des données. Il est donc **essentiel de reconnaître pleinement ce travail autour des données comme une activité scientifique à part entière, nécessitant des compétences spécifiques, et méritant d'être valorisée au même titre que les autres dimensions de la recherche.**

Il est aussi important de noter que ces fonctions ne sont pas exclusivement liées à un domaine de recherche particulier ; elles sont transférables à d'autres contextes organisationnels. Cela favorise la création de passerelles entre les écosystèmes de recherche publics et privés, et même vers d'autres secteurs du milieu académique, voire en dehors.

5 PRÉSENTATION DES FONCTIONS

5.1. Intendance des données

L'intendance des données ou *data stewardship* se positionne comme garante des principes FAIR, des standards de la discipline et cadres règlementaires en veillant à respecter ou faire respecter, s'il en existe une, la politique des données définie par l'établissement. De par son rôle transverse, cette fonction est généralement le premier point de contact entre le chercheur et les problématiques afférentes aux données de recherche. Elle joue fondamentalement un rôle d'interface entre les équipes scientifiques, informatiques et différents experts de la donnée. Connaître les services numériques de stockage, calcul… homologués par l'établissement, savoir qui contacter sur des sujets d'éthique ou de conformité règlementaire, conseiller les équipes de recherche sur les sources et entrepôts de données de confiance sont autant d'exemples de tâches effectuées au niveau de l'intendance des données.

Elle accompagne et joue un rôle de conseil auprès des chercheurs dans la mise en place des principes FAIR (Faciles à trouver, Accessibles, Interopérables, Réutilisables), effectue le suivi de l'évolution des données, identifie les pratiques potentiellement problématiques ou incompatibles avec la politique de l'établissement. L'intendance des données veille à ce que le processus de la collecte jusqu'à la valorisation des données respecte notamment les cadres juridiques, contractuels ou éthiques relatifs à certaines données. En retour, elle fait remonter du terrain les problématiques concrètes auxquelles sont confrontées les équipes de recherche pour permettre une amélioration continue du dispositif de gouvernance ancrée dans la réalité des activités de recherche.

L'intendance des données est le garant de la qualité des données qui repose sur leur exactitude, complétude, cohérence et traçabilité, garantissant leur fiabilité scientifique. En plus de cette robustesse scientifique, les données doivent être bien documentées et conformes aux standards qui permettent leur accessibilité, interopérabilité et réutilisation, afin de favoriser leur partage et leur valorisation.

Activités principales

Quelles sont les 5 activités principales associées à l'intendance des données de recherche ?

1. Coordination du cycle de vie des données

La coordination du cycle de vie des données représente une fonction transversale, articulant toutes les étapes dans un continuum cohérent. Elle implique la supervision de l'ensemble du processus, depuis la collecte initiale des données jusqu'à leur diffusion, qu'elle soit ouverte ou restreinte. Cette fonction s'appuie sur la mobilisation de différents acteurs et compétences, afin d'assurer un enchaînement fluide entre les phases de nettoyage, d'analyse, de structuration, de validation et de dépôt des jeux de données.

2. Accompagnement, médiation et valorisation

Cette fonction vise à soutenir les chercheurs dans la gestion de leurs données, en les accompagnant dans l'acquisition de compétences spécifiques et en favorisant l'appropriation des bonnes pratiques. Elle contribue également à la valorisation du patrimoine informationnel du laboratoire ou de l'établissement, en assurant la gestion du catalogue de données, leur découvrabilité, et leur diffusion selon des modalités claires et encadrées. Elle participe ainsi à la reconnaissance des données comme production scientifique à part entière.

3. Qualité, conformité et traçabilité des données de recherche

Cette fonction recouvre un ensemble d'activités essentielles visant à garantir la qualité, la fiabilité et la structuration cohérente des données produites dans le cadre de la recherche. Elle suppose un travail rigoureux sur la complétude, l'actualisation et la validation des données, en lien avec les exigences disciplinaires et institutionnelles. Elle implique également le respect des cadres juridiques, des politiques internes de gouvernance, et des standards internationaux.
Une attention particulière est portée à la traçabilité des données, depuis leur collecte jusqu'à leur diffusion, afin de garantir leur intégrité scientifique et leur réutilisation dans des conditions maîtrisées. Cette transparence est indispensable à la reproductibilité des résultats et à la confiance dans les processus scientifiques.

4. Documentation et interopérabilité

La documentation et l'interopérabilité des données reposent sur la production de métadonnées complètes et précises, qu'elles soient scientifiques, structurelles ou documentaires. Ce travail s'inscrit pleinement dans le respect des principes FAIR, dont il constitue l'un des fondements opérationnels. Il peut inclure le développement ou l'adaptation de standards de métadonnées propres à chaque domaine disciplinaire. L'objectif est de favoriser la compréhension et la réutilisation des données dans une logique d'ouverture et de partage à l'échelle interdisciplinaire et internationale.

5. Souveraineté et gestion responsable des données

Cette fonction contribue à la souveraineté des données, en veillant à ce que leur traitement, préservation et diffusion soient effectués au sein d'infrastructures maîtrisées, respectueuses des intérêts scientifiques, institutionnels et nationaux. Elle participe activement à une gestion responsable des données de recherche, en conciliant ouverture maîtrisée, sécurité, conformité réglementaire et maîtrise des usages à long terme. Ce positionnement est essentiel pour garantir une autonomie numérique durable au sein des écosystèmes de recherche.

Profil

Quels sont les profils qui peuvent exercer cette fonction d'intendance des données ?

La principale caractéristique associée à l'intendance des données est de ne pas être une fonction de spécialiste nécessitant des connaissances pointues dans des domaines spécifiques. Au contraire, il s'agit d'un rôle généraliste nécessitant une vision large des problématiques associées aux données de recherche.
Agissant comme **interface** et **médiatrice**, il s'agit d'une fonction de facilitation des interactions entre les différents acteurs en lien avec les données de recherche. Les qualités humaines, associées à de bonnes capacités d'organisation et de gestion des priorités, sont donc essentielles.

La fonction d'intendance de la donnée nécessite par conséquent des **compétences scientifiques** (généralistes ou disciplinaires) et des **compétences techniques**, permettant de s'adapter aux besoins de chaque domaine scientifique.

Deux profils types émergent pour tenir ce rôle :

1. Un profil **scientifique** issu d'un champ disciplinaire spécifique développant, au fil de son parcours, une expertise sur les différentes étapes du cycle de vie des données : collecte, nettoyage, structuration, documentation, analyse, préservation et partage. Sa connaissance fine des pratiques de recherche lui permet d'accompagner efficacement ses pairs dans la mise en œuvre de bonnes pratiques en matière de gestion des données.
2. Un profil plus « **spécialiste de la donnée** », formé aux techniques de gestion, de normalisation, de méta documentation et de diffusion des données et/ou des connaissances, se spécialisant progressivement dans un domaine scientifique. Il apporte une expertise technique et méthodologique pour garantir la qualité, la préservation de données souveraines et leur réutilisabilité.

Illustration dans l'Odyssée fantastique

Dans cette odyssée, nous avons transposé la fonction d'intendance des données, sur un personnage, qui, fictivement, accompagne les données dans leur voyage symbolique du cycle de vie des données de recherche. C'est sous les traits d'un personnage dévoué, méticuleux mais aussi excellent communiquant, que nous avons imaginé s'exercer la fonction d'intendant des données et qui tient le rôle de Data Steward. Il a un rôle primordial dans l'histoire, tant il est présent à toutes les étapes du voyage des données. Un rôle augmenté par la connexion permanente qu'il assure avec les scientifiques, à tout instant.

Dans L'Odyssée fantastique des données, Kirk est *data steward*...

Passionné par la gestion des données de recherche, j'accompagne les chercheurs dans leur quête de données d'excellente qualité. Je pilote ce parcours à travers les différentes étapes pour obtenir l'intégrité et l'accessibilité des données selon les principes **FAIR**, et pour optimiser leur exploitation et leur réutilisation.

Dans **L'Odyssée fantastique de la donnée**, vous découvrirez mon engagement aux côtés de **Diana** (les données de recherche). Ma mission est de la guider vers les experts qui vont la rendre exploitable et réutilisable. Je valide l'atteinte des objectifs de chaque étape du voyage, avant de revenir vers **Cristina** (la chercheuse), avec une Diana documentée et exploitable qui pourra être préservée et réutilisée dans différents contextes.

5.2. Curation des données ou research data curation

La curation des données est au cœur du travail scientifique des données. Elle vise à renforcer leur pertinence scientifique et leur interopérabilité, afin de les rendre réutilisables dans une variété de contextes, y compris ceux éloignés de leur domaine ou environnement d'origine. Il s'agit d'une activité préparatoire préfigurant l'exploitation et la valorisation des jeux de données. La curation des données de recherche consiste non seulement à nettoyer, prétraiter, organiser et documenter les données selon les standards du domaine scientifique auquel elles se rattachent dans le but d'une exploitation immédiate, mais également de garder en perspective une potentielle exploitation ultérieure. Il s'agit donc d'un rôle très opérationnel et fortement intriqué avec le processus de recherche en tant que tel.

Les pratiques de curation, si elles doivent être conformes aux principes de gouvernance des données de l'établissement, restent spécifiques aux projets de recherche pour lesquels les données sont mobilisées. Elles peuvent ainsi relever du nettoyage, du formatage, de la normalisation des données pour en garantir la qualité et la cohérence, de l'anonymisation/pseudonymisation, de l'ajout des métadonnées scientifiques (choix des terminologies, description du protocole d'obtention, traitement, matériel, logiciels, …). Par ce travail sur les métadonnées scientifiques, cette fonction assure un accès durable à des données de recherche porteuses de sens et prépare ainsi les données à leur dépôt et leur partage. Ainsi préparées, les données sont plus faciles à enrichir et à apparier avec d'autres données.

La fonction de curation des données de recherche se situe généralement dans une structure de recherche (laboratoire, plateforme, infrastructure de recherche…). **En tant que fonction d'intérêt général scientifique, elle contribue à la conservation durable du capital scientifique et à sa mise à disposition au bénéfice des communautés de recherche, des institutions et, plus largement, de la société.**

Activités principales

Quelles sont les 3 activités principales associées à la curation des données de recherche ?

1. Nettoyage, qualité et préparation des données pour leur exploitation

Cette dimension opérationnelle est essentielle pour **produire des données fiables, cohérentes, et exploitables à court et long terme**. Elle englobe la préparation technique et qualitative des données. Elle comprend le nettoyage, l'affinage, la conversion et la normalisation des jeux de données pour en assurer la qualité, la fiabilité et la cohérence. Cette étape permet également d'assurer leur réutilisation et leur reproductibilité, y compris en dehors de leur contexte initial. Elle englobe la gestion des versions et la mise à jour des données, ainsi que l'anonymisation ou la pseudonymisation si nécessaire. L'objectif est ici de produire des données exploitables, robustes et conformes aux exigences scientifiques, éthiques et réglementaires.

2. Organisation, structuration et documentation des données

Cette activité vise à garantir **la traçabilité et la réutilisabilité des données**, en veillant à leur qualité scientifique et à leur interopérabilité. Elle comprend l'organisation, la structuration et la documentation des données de manière à les rendre compréhensibles, traçables et interopérables. Cela implique de sélectionner les données pertinentes, de les structurer en lien avec l'ingénierie des données et de les décrire à l'aide de métadonnées scientifiques précises (protocoles d'obtention, traitements, outils utilisés, etc.). Cela nécessite d'utiliser des vocabulaires contrôlés au niveau disciplinaire (ontologies, codex, classifications taxonomiques, thésaurus, lexiques, etc.) et de participer à la création et à l'évolution des vocabulaires et standards du domaine de recherche. Elle est l'artisane principale de la **mise en œuvre des principes FAIR**, afin de rendre les données trouvables, accessibles et réutilisables. Cette fonction peut également contribuer au renseignement des métadonnées documentaires (contributeurs, licences, financeurs, etc.).

3. Préservation, valorisation et accès durable aux données

Cette activité consiste **à inscrire les données dans une logique d'ouverture, de partage maîtrisé et de pérennisation**. Elle participe à l'élaboration, veille à la mise en œuvre et à l'actualisation du plan de gestion des données, afin de garantir un cadre structuré pour le cycle de vie des données. Elle intervient dans le choix des entrepôts de dépôt adaptés, facilite le dépôt des jeux de données, et accompagne leur publication, que ce soit en accès ouvert ou restreint selon les règles relatives au projet de recherche. Par ce travail, la curation des données favorise **la valorisation des données et leur accès durable.**

Ces activités concourent à **la reproductibilité** des résultats scientifiques et à poser les fondations d'une éventuelle **réutilisation** des données pour de nouvelles analyses, méta-analyses ou projets interdisciplinaires. Elles sont réalisées en utilisant pour cadre les actions définies dans le plan de gestion des données du projet de recherche, qui peut notamment inclure des exigences liées au domaine de recherche (vocabulaires contrôlés, conditions d'ouverture, période d'embargo…), cadre juridique, contractuel et la conformité aux normes éthiques et légales, notamment en ce qui concerne les données sensibles ou personnelles.

Profil

Quels sont les profils qui peuvent exercer cette fonction de curation de données ?

Dans la mesure où les activités réalisées nécessitent une connaissance très fine du domaine de recherche concerné, la fonction de curation des données est généralement assurée par un **personnel scientifique** ou le chercheur lui-même travaillant dans une structure de recherche (laboratoire, plateforme, infrastructure de recherche …)[2]. En fonction du contexte, des compétences techniques de programmation peuvent également être requises pour rédiger les scripts de transformation des données.

Illustration dans l'Odyssée fantastique

Avant que qui que ce soit n'intervienne dans cette odyssée, il est indispensable de procéder à la curation. Dans le voyage des données, c'est à chaque fois la première étape. Parfois la plus longue, d'autant qu'elle peut se réactiver au fil du cycle de vie. Toute la suite, toute la pertinence du projet de recherche en dépend.

Pour trouver cette donnée qui peut changer le sens de la recherche, il faut être curieux et inventif. C'est ainsi que l'on pourra décrire le caractère du personnage qui a la lourde tâche de parfaire un jeu de données brutes que lui confie une équipe de chercheurs. Il est accompagné dans sa mission, mais demeure seul au moment des choix. De lui dépendent les données qui sont sélectionnées. Un réel pouvoir qui peut changer la trajectoire de n'importe quel projet de recherche. Dans l'Odyssée, Youri accomplit sa mission avec brio !

[2] Le corps d'État des astronomes et physiciens (CNAP) reconnait ces activités comme partie intégrante des missions de recherche. Ils exercent ces activités sur les données, en lien avec leurs missions d'« organisation et réalisation de tâches scientifiques d'intérêt général d'observation ou d'accompagnement de la recherche en astronomie et sciences de la planète ayant un caractère national ou international et labellisées par l'Institut national de sciences de l'univers du Centre National de la Recherche Scientifique ».

Dans L'Odyssée fantastique de la donnée, Youri est data curator…

Toujours en recherche de données nouvelles, je mets ma curiosité et mon goût pour l'innovation au service de la science. Au-delà de la découverte, j'apprécie que les données soient parfaitement formatées et standardisées afin que l'on puisse tout à la fois, les repérer, les traiter, les analyser et bien sûr les partager.

Dans L'Odyssée fantastique de la donnée, je vous montrerai comment j'applique avec une rigueur absolue les différentes méthodes de nettoyage, de tri et de codification, dans le but de **rendre les jeux de données parfaitement propres**. J'interviens dès le début du voyage, pour faciliter le travail de mes collègues *data Engineer* et *data Analyst* ou *data Scientist*. Souvent, j'ajoute des données collectées pour enrichir l'ensemble et le rendre plus cohérent et exploitable par les chercheurs.

5.3. Ingénierie des données

L'ingénierie des données est un domaine de spécialisation informatique qui vise à construire des systèmes de gestion de données nécessitant une forme de "passage à l'échelle". En effet, passés certains seuils, les outils standards mis à disposition des scientifiques ne sont plus suffisants pour lui permettre de travailler de manière efficace : traitements trop lents, capacités de stockage ou de flux insuffisantes, risques d'erreur humaine trop importants, volumes trop massifs pour accéder efficacement aux données, complexité d'ordonnancement des traitements d'acquisition et de retraitement … Le rôle de l'ingénierie des données est de définir et mettre en œuvre les solutions permettant d'apporter des réponses opérationnelles à ces limitations. Elle joue également un rôle clé dans la sécurisation des données à travers la mise en place de ces solutions, sur le plan de la cybersécurité et de la gestion des sauvegardes.

Il s'agit donc d'une fonction très technique, qui peut être pleinement intégrée au sein des équipes de recherche pour les projets nécessitant d'importants moyens techniques ou, plus souvent intervenir ponctuellement en mettant à profit des infrastructures mutualisées déjà en place, à l'échelle d'un établissement, d'un laboratoire ou d'une structure dédiée, que ces structures sont alors en responsabilité de maintenir et de faire évoluer pour répondre aux attentes des équipes de recherche.

Activités principales

Quelles sont les 4 activités principales associées à l'ingénierie des données ?

1. Concevoir et gérer les infrastructures de gestion de données : Cette activité fondamentale consiste à concevoir, construire, structurer et maintenir les systèmes et les infrastructures nécessaires pour collecter, stocker et analyser les données, y compris de grands volumes. L'ingénierie de données est également responsable de la gestion des bases de données, en assurant leur maintenance, leur optimisation, leur performance et leur sécurité. Cela implique aussi de surveiller et maintenir les infrastructures existantes, de concevoir et maintenir des architectures de données robustes et évolutives, d'organiser l'enchaînement des traitements et d'optimiser les performances des systèmes de traitement.

2. Organiser la collecte et l'intégration des données : Il s'agit, lorsque le type de données le nécessite, d'organiser la collecte, la gestion, la préparation des données pour leur traitement. L'ingénierie des données organise la collecte, l'intégration et la préparation des données en s'appuyant sur des sources hétérogènes (bases SQL/NoSQL, APIs, fichiers, etc.), qu'elle agrège et structure selon les besoins des projets. Elle peut également définir les processus d'acquisition des données lorsque ceux-ci n'existent pas.

3. Développer et automatiser les flux de traitement des données : Cette activité est essentielle pour assurer un traitement efficace des informations. Elle consiste à développer des flux de données automatisés, notamment en développant des *pipelines* ETL/ELT (Extract, Transform, Load / Extract, Load, Transform) pour l'ingestion, le nettoyage, la transformation et le chargement des données. Il s'agit aussi d'automatiser et exploiter les procédures de validation et de traitement des données pour garantir leur qualité et la fiabilité.

4. Industrialiser et mettre en production les traitements et modèles : L'ingénierie des données a pour mission d'industrialiser des traitements en mettant en production les algorithmes et les traitements, souvent développés par les *data scientists*. Cela inclut l'industrialisation et l'automatisation du nettoyage des données et le développement de l'industrialisation de modèles statistiques ou de *machine learning*. L'implémentation du suivi de la validité du modèle statistique fait partie de cette tâche, tout comme l'optimisation de la performance des systèmes de traitement des données massives, l'assurance du suivi de production et de la maintenance de ces systèmes.

L'ingénierie des données nécessite la maîtrise de nombreux outils pour gérer, stocker, transformer, et déplacer des données de manière massive et automatique et la capacité à concevoir des architectures à la fois évolutives et pérennes.

Profil

Quels sont les profils qui peuvent exercer cette fonction d'ingénierie des données ?

Cette fonction requiert des compétences en programmation (Python, Java, Scala), en gestion de bases de données (SQL, NoSQL), en outils Big Data (Spark, Kafka…), ainsi qu'une bonne connaissance des systèmes d'exploitation et des pratiques de gestion de la donnée.

Parmi les qualités recherchées pour exercer cette fonction, on retrouvera naturellement rigueur et méthode pour garantir la qualité et la fiabilité des données mais aussi esprit analytique et de synthèse pour résoudre des problèmes complexes et optimiser les flux de données.

Les profils les plus adaptés ont également la capacité de concevoir et gérer des architectures de données tels que des entrepôts et lacs de données.
Du fait de la forte proximité avec des fonctions analogues dans le secteur privé et de la rareté de ce type de compétences, les personnes exerçant ces fonctions sont souvent issues de recrutements externes sur des emplois contractuels.

Pour autant, là encore, deux profils émergent dans les organisations publiques :

- **Des personnels scientifiques** issus de différents champs disciplinaires qui se forment en data science ou informatique, au fil de leur parcours.
- **Des personnels techniques**, diplômés en informatique, ingénierie logicielle par exemple et dont les compétences techniques sont complétées par une formation aux enjeux spécifiques au domaine de recherche de la structure dans laquelle ils travaillent.

Illustration dans l'Odyssée fantastique

Mais notre personnage, dans l'Odyssée, est présenté comme un ingénieur des données chevronné. Fort de ces nombreuses expériences, il s'est spécialisé dans la structure et l'organisation des données de recherche. En parfaite harmonie avec l'équipe orchestrée par l'intendant des données, il prépare le travail des analystes et des scientifiques de la donnée.

Dans L'Odyssée fantastique de la donnée, Janus est data engineer…

Je suis un fervent défenseur de la structure. Si certains sont partisans du foisonnement des données, de leur nécessaire complexité, je suis toujours là pour prôner l'organisation. Pour moi, il n'y a rien de plus beau qu'un système.

Dans L'Odyssée fantastique de la donnée, je prends la suite de Youri, et j'ai le plaisir de travailler sur un jeu de données déjà bien nettoyées. Mon rôle est majeur puisque **mon intervention permettra les analyses**, garantira la performance future des algorithmes comme la justesse des statistiques. Je veille par ailleurs à ce que **les systèmes soient robustes** et adaptés au volume de données en croissance continuelle que nos chercheurs nous envoient.

5.4. Analyse des données

L'analyse des données correspond à un premier type d'exploitation des données à des fins scientifiques. Elle est centrée sur l'exploration et l'utilisation d'outils statistiques et/ou algorithmiques. Sa finalité est de tester des hypothèses scientifiques, en vérifiant si les observations contenues dans les données, concordent avec les modèles que l'on souhaite tester. Il s'agit d'une activité d'exploration, d'analyse, de visualisation et d'interprétation des données de recherche.

Par conséquent, ce travail fait partie intégrante du travail de recherche en tant que tel, et les résultats obtenus alimentent directement les connaissances scientifiques.

La personne qui assure cette fonction d'analyse des données est soit data analyst, soit le scientifique lui-même lorsqu'il maîtrise les méthodes d'analyse de données.

Activités principales

L'analyse de données peut suivre des modèles très différents en fonction des hypothèses à tester, de la nature et de la complétude des données disponibles et des modèles statistiques à mobiliser.

Quelles sont les 3 activités principales associées à l'analyse des données ?

1. Analyse statistique des données

L'analyse des données consiste à appliquer des méthodes statistiques simples (tests d'hypothèses, régressions, etc.) ou informatiques pour identifier des tendances, corrélations ou anomalies dans les données et en tirer des résultats exploitables sur le plan scientifique. La fonction d'analyse des données soutient les scientifiques dans l'interprétation des résultats et contribue à la rédaction de rapports et articles scientifiques. Elle implique une compréhension fine des enjeux de recherche et une capacité à mobiliser les outils statistiques les plus adaptés.

2. Visualisation et communication des résultats

La fonction d'analyse des données crée des supports visuels (graphiques, tableaux, tableaux de bord, …) pour rendre les résultats clairs et accessibles aux chercheurs et parties prenantes du projet. Elle accompagne également les équipes dans l'utilisation des outils d'analyse et l'interprétation des données.

3. Gestion des données et veille technologique

Elle garantit la gestion sécurisée et conforme des données, en veillant à leur stockage approprié. Elle mène également une veille active pour rester à jour sur les innovations, évolutions méthodologiques, logicielles et technologiques dans le domaine de l'analyse de données, et anticiper leur intégration dans les pratiques de recherche.

Profil

Quels sont les profils qui peuvent accomplir cette fonction d'analyse des données ?

Généralement située dans une structure de recherche, la fonction de *data analyst* est très souvent prise en charge par des chercheurs ou enseignants-chercheurs dans le cadre de leur activité de recherche. L'exercice de cette fonction nécessite néanmoins, au-delà des compétences disciplinaires propres à chaque chercheur, une maîtrise des méthodes et outils de traitement statistique. La personne chargée de cette fonction d'analyse de données peut également être un spécialiste des méthodes ou des outils statistiques et se positionner dans un rôle d'accompagnement vis-à-vis des autres chercheurs.

Illustration dans l'Odyssée fantastique

Analyser des données, n'est-ce pas avant tout une aventure numérique ? Certes, on attendra d'une analyse qu'elle éclaire, qu'elle illumine ce qui au départ ressemble davantage à une accumulation d'informations. Trouver du sens, interpréter sans dénaturer, en respectant le contexte scientifique et les lois de la statistique, exige bien plus que de la précision. Il faut voir, là où l'on est le plus souvent dans le noir.

Naturellement celle qui accomplit cette mission dans l'Odyssée est une femme parfaitement à son aise parmi les graphiques, les courbes et les cascades de chiffres. Elle jugule les aléas, réduit les écarts-types et extrait les évidences les plus fines, dans cette jungle numérisée.

Dans L'Odyssée fantastique de la donnée, Anya est data analyst…

On me félicite souvent pour mon esprit synthétique. J'aime utiliser mes compétences statistiques et informatiques pour apporter une vision claire à la valeur des données que l'on me propose. Elles sont alors plus faciles à lire, à étudier et vont permettre aux chercheurs de valider des résultats.

Dans L'Odyssée fantastique de la donnée, j'ai la chance de travailler en symbiose avec Albert, un data scientist génial. Ensemble nous élaborons des modèles prédictifs. Ma mission est de respecter éthique et réglementation en vigueur, et aussi d'éviter tout biais dans les futures analyses extraites des données de recherche.

5.5. Science des données

La science des données est un domaine interdisciplinaire qui combine méthodes scientifiques, techniques statistiques, algorithmes d'intelligence artificielle et outils informatiques afin de transformer les données en connaissances exploitables, en description et modélisation de phénomènes complexes, en prédictions ou en aides à la décision. Elle regroupe un ensemble de pratiques, d'outils et de méthodes visant à extraire, traiter, modéliser, visualiser et interpréter des données, en particulier celles issues de la recherche.

Dépassant le simple cadre de l'analyse de données, la science des données intègre une forte dimension computationnelle et mobilise des approches statistiques avancées. Elle vise à produire des connaissances profondes à travers la modélisation, l'expérimentation numérique et la conception algorithmique, en travaillant souvent à l'interface entre les sciences fondamentales et les technologies numériques.

Appliquée à des jeux de données massifs, parfois non structurés ou collectés sur le long terme, elle permet de représenter des phénomènes complexes au moyen de modèles sophistiqués, notamment issus du machine learning ou du deep learning. Les résultats sont restitués sous forme de visualisations interactives ou intégrés à des systèmes intelligents, facilitant la compréhension et la prise de décision dans des contextes variés.

Activités principales

La science des données consiste principalement à transformer des données en connaissances exploitables, grâce à des techniques avancées d'exploitation de données notamment d'intelligence artificielle.

Quelles sont les 4 principales activités associées à la science des données ?

1. Concevoir, implémenter et évaluer des modèles mathématiques et algorithmiques avancés

Développer, ajuster et comparer des approches de modélisation basées sur des techniques d'apprentissage automatique supervisé et non supervisé, de modélisation probabiliste, ou de deep learning. L'activité inclut le choix des algorithmes adaptés, la sélection des variables pertinentes, la validation croisée des performances, ainsi que l'interprétation des résultats pour répondre à des problématiques scientifiques.

2. Concevoir des stratégies d'exploration computationnelle de données complexes

Mettre en place des approches d'exploration automatique et interactive de données hétérogènes (temps réel, spatiales, textuelles, etc.) afin de révéler des structures sous-jacentes, d'identifier des comportements émergents ou de détecter des anomalies. Cela implique la conception d'outils algorithmiques adaptés aux spécificités des données, ainsi que leur visualisation et interprétation pour favoriser la découverte de connaissances.

3. Développer des modèles scientifiques complexes

Formaliser des problématiques scientifiques dans des cadres computationnels robustes, en intégrant des méthodes statistiques, des systèmes dynamiques, ou encore des modèles hybrides combinant données et théorie. Cette activité vise à produire des modèles explicatifs et prédictifs utiles à la compréhension de phénomènes complexes dans des domaines variés (physique, biologie, sciences sociales, etc.).

4. Optimiser le traitement des données dans des environnements de calcul haute performance ou distribués

Déployer des stratégies d'optimisation algorithmique et d'adaptation des traitements sur des infrastructures de calcul intensif (HPC) ou des environnements de cloud scientifique, dans le but d'améliorer l'efficacité computationnelle. Cette démarche vise à assurer la reproductibilité, la montée en charge (scalabilité) et la robustesse des analyses de données massives, tout en intégrant une approche responsable et sobre en ressources numériques.

Profil

Quels sont les profils qui peuvent exercer cette fonction de science des données ?

La fonction de data scientist peut se trouver dans un laboratoire, une équipe ou une infrastructure de recherche. Cette fonction nécessite une expertise approfondie en statistiques avancées, machine learning, programmation (Python, R, etc.) et une capacité à traiter des données très volumineuses et variées pour générer des connaissances nouvelles et orienter la recherche.

Le data scientist dispose généralement d'une formation approfondie en mathématiques appliquées, statistiques et/ou informatique et dans certains cas d'une formation dans une discipline scientifique complétée par une spécialisation forte en science de données. Si le doctorat n'est pas obligatoire, il correspond très souvent au niveau d'étude des personnes qui exercent cette fonction.

Dans les laboratoires, la fonction est très souvent exercée par un chercheur ou un enseignant-chercheur. Dans une infrastructure de recherche, le rôle du data scientist peut être plus technique et davantage orienté vers la mise à disposition de services au chercheur.

Illustration dans l'Odyssée fantastique

Il y a dans cette fonction les éléments d'une connaissance qui dépasse nos limites communes. C'est sans doute un personnage considéré comme hors normes qui manipule des algorithmes, à l'image d'un alchimiste des temps numériques. Dans l'Odyssée, c'est Albert, qui se charge d'extraire l'or des données de recherche qu'on lui propose de modéliser. Savant, scientifique, technophile, il n'en demeure pas moins un humain au service d'une équipe, démontrant sa patience, son exigence et sa passion pour la science.

Dans L'Odyssée fantastique de la donnée, Albert est data scientist…

Je suis tous les jours plongé dans le monde de demain. Prévoir, anticiper ce que sera notre futur, ne peut se faire sans des données fiables, ni sans une approche claire des algorithmes et de la science. J'apprécie vraiment de travailler sur des données scientifiques, déjà triées, organisées, et porteuses de sens.

Dans L'Odyssée fantastique de la donnée, je cohabite avec d'autres experts des chiffres et des analyses, et nous mettons notre passion pour la science au service des chercheurs pour construire des modèles et des analyses utiles à la validation des résultats de la recherche. J'ai toujours l'espoir que les données seront partagées pour le bien de tous.

5.6. Documentation des données – data librarian

La fonction de documentation des données se concentre sur l'accès aux données, leur documentation et leur préservation et la sensibilisation des équipes de recherche à une gestion raisonnée et responsable des données.
Cette fonction, étroitement liée aux enjeux de partage et à l'ouverture des données, vise à garantir la valorisation et la citation des données de recherche, tout en facilitant leur découverte et leur réutilisation, contribuant ainsi à l'accélération de la découverte scientifique et à l'innovation par les chercheurs, les entreprises et les citoyens. Elle accompagne les chercheurs dans la gestion durable de leurs données, notamment via l'élaboration de plans de gestion, le choix des standards de métadonnées, des entrepôts de dépôt et des licences. Agissant comme un intermédiaire entre producteurs et utilisateurs de données, elle facilite le partage au-delà du cadre du laboratoire, au bénéfice de la communauté scientifique, des entreprises et des citoyens.

Concrètement, cette fonction consiste à classifier et structurer les ensembles de données pour assurer leur accessibilité et leur réutilisation. L'ajout des métadonnées documentaires dans le respect des standards permet ensuite de faciliter le partage et la compréhension des données, en plus du maintien des catalogues de données dans les dépôts institutionnels ou dans les archives ouvertes.

La documentation des données est assurée par une personne qui assure un rôle de sensibilisation et d'accompagnement des scientifiques à la gestion et à la diffusion des données de recherche et plus largement à la science ouverte. Elle exerce le plus souvent en service commun de documentation, de science ouverte ou dans une infrastructure de recherche.

Activités principales

Quelles sont les 2 principales activités associées à la documentation des données ?

1. Accompagner les équipes de recherche dans leur démarche de gestion et de partage des données à long terme, après la publication des études afférentes, en les conseillant sur les standards de métadonnées, les entrepôts de dépôt, ou encore les licences de diffusion

2. Promouvoir les bonnes pratiques de gestion, description, partage et réutilisation des données. Elle joue également un rôle clé dans le fait de sensibiliser le personnel de recherche aux enjeux de la science ouverte, des principes FAIR et de mise en place des plans de gestion des données. Elle intervient à deux moments clés : en amont pour accompagner la rédaction des plans de gestion des données, et en aval pour faciliter leur partage ou ouverture via des entrepôts accessibles à l'ensemble de la communauté.

Profil

Quels sont les profils qui peuvent exercer cette fonction de documentation des données ?

Les data librarians sont aujourd'hui principalement des personnels de bibliothèque qui ont souhaité se spécialiser dans la gestion des données de recherche. Ils peuvent être spécialistes d'un petit nombre de disciplines et se concentrer sur un nombre limité de tâches ou être généraliste, intervenant sur un large éventail de disciplines et un grand nombre de tâches.

Ce métier émerge selon deux trajectoires principales :

- **Des personnels d'appui à la recherche** et notamment des personnels de bibliothèque et de documentation, disposant de connaissances à la fois disciplinaires et techniques. Des documentalistes spécialisés du Centre des Données astronomiques de Strasbourg (CDS) exercent par exemple cette fonction.
- Des **personnels scientifiques** au profil hybride qui exercent à la fois une activité de recherche et prennent en charge une mission d'accompagnement à la gestion des données telle que décrite ici.

Illustration dans l'Odyssée fantastique

Coordonner et accompagner sont les impératifs de la mission accomplie par celui ou celle qui veille à la documentation des données. Il faut être à l'écoute des attentes des chercheurs mais également se projeter dans ce qui permettrait à d'autres, plus tard, d'exploiter leur valeur. Comprendre les uns et anticiper les usages des autres, c'est en quelque sorte une forme de médiation.

Notre personnage dans l'Odyssée est tout à fait dans cette posture, entre ouverture et rappel des exigences de la science. Gaya, puisque dans ce voyage c'est un rôle féminin, fait preuve d'empathie mais aussi de sa passion pour cette fonction. Sans elle, l'Odyssée serait inachevée.

Dans L'Odyssée fantastique de la donnée, Gaya est data librarian…

Faciliter l'accès aux données de recherche est une mission durable et noble. C'est ce que je crois, avec ferveur. Elle est passionnante mais exigeante. C'est un méticuleux travail de coordination pour permettre l'ouverture et le partage des données.

Dans L'Odyssée fantastique de la donnée, j'interviens lorsque les analystes ont terminé leur travail. Il me faut alors documenter et soigneusement ranger ces données. Il s'agit de renseignements mais aussi de cartographie, de choix d'emplacement où l'on retrouvera facilement ce dont la science et les chercheurs ont besoin pour demain.

5.7. Curation des métadonnées documentaires

Cette fonction constitue un sous-ensemble de la précédente, dans la mesure où elle se concentre sur le renseignement et la qualité des métadonnées saisies par les déposants de jeux de données dans un entrepôt, en vue de leur diffusion, qu'elles soient en accès restreint ou ouvertes.

La curation des métadonnées documentaires contribue à faciliter le partage et la citation des jeux de données. Il s'agit également d'assurer la traçabilité et la conformité des métadonnées aux standards internationaux (Dublin Core, DataCite …).

La curation des métadonnées documentaires intervient dans les phases de préparation au dépôt, puis de dépôt et publication pour ouverture ou partage des données.

Elle contribue à la fois à créer, structurer et décrire les jeux de données validés et à renseigner les métadonnées documentaires. Elle assure la mise à jour des différentes versions de jeux de données et publie les données en ouverture ou partage restreint selon les cas.
Cette fonction est assurée par un metadata curator, qui travaille dans des services de soutien ou d'appui à la recherche ou d'infrastructures dédiées à la diffusion des données de recherche.

Activités principales

Quelles sont les 4 principales activités associées à la curation des métadonnées documentaires ?

1. Structurer ou contrôler la structuration des jeux de données de recherche validés en veillant au respect de l'application des règles du domaine scientifique.

2. Renseigner, mettre à jour ou contrôler les métadonnées documentaires sur les jeux de données pour en assurer une description claire et complète en conformité avec les standards reconnus afin de garantir l'interopérabilité et l'accessibilité des données. L'objectif est de faciliter la découverte et le partage des données, notamment via des catalogues associés aux entrepôts de donnée et infrastructures de recherche.

3. Accompagner les déposants dans leurs publications pour en sécuriser l'usage.

4. Élaborer des procédures institutionnelles concernant la gestion des métadonnées et la préservation numérique, en lien avec les standards de métadonnées (Dublin Core, DataCite, normes ISO de certaines données, etc.) en fonction des disciplines scientifiques.

Profil

Quels sont les profils qui peuvent exercer cette fonction de curation des métadonnées ?

Formées le plus souvent en sciences de l'information, en documentation scientifique, les personnes exerçant cette fonction disposent d'une bonne connaissance des normes de métadonnées, de la gestion de l'information et d'une sensibilité ou spécialisation dans le domaine scientifique auquel se rattachent les données dont elles s'occupent. Elles disposent de compétences en structuration de données et en gestion de vocabulaires.

Cette fonction joue un rôle majeur dans la sensibilisation des chercheurs et équipes de recherche à la bonne gestion des métadonnées et les accompagnent dans le renseignement de celles-ci.

Illustration dans l'Odyssée fantastique

Aujourd'hui les données sont de plus en plus nombreuses et leur impressionnante croissance nous incite, plus encore, à les documenter avec rigueur. D'une part, nous devons assurer leur pérennité dans l'usage, d'autre part, nous devons les préparer à la science ouverte, qui est notre avenir. Dans l'Odyssée fantastique de la donnée, cette fonction de curation des métadonnées est confiée à un personnage discret et d'une totale intégrité.

Respectueux des règles, scrupuleux sur les procédures, Markos est un homme de confiance. Il collabore efficacement avec les scientifiques et avec l'intendant des données qui l'instruit sur la gouvernance des données mise en place par l'établissement porteur du projet ou par le domaine de recherche.

Dans L'Odyssée fantastique de la donnée, Markos est metadata curator…

Il y a en moi cet instinct de protection mais aussi de valorisation du travail des chercheurs. Je reconnais aisément toute la valeur créée par les données de recherche, aussi, je tiens à en structurer l'accès et à préserver leur intégrité en leur ajoutant toutes les métadonnées documentaires nécessaires.

Dans L'Odyssée fantastique de la donnée, j'interviens à la fin du voyage, une fois les données validées par notre chercheuse, pour garantir leur usage futur dans le strict respect des règles et des procédures. Mon travail s'effectue en bonne collaboration avec Gaya, notre data librarian. Les données peuvent alors être présentées au monde de la science.

6 LA GOUVERNANCE DES DONNÉES DE LA RECHERCHE : UN RÔLE STRATÉGIQUE AU COEUR DE L'ORGANISATION DE LA RECHERCHE

Contrairement aux 7 fonctions opérationnelles du cycle de vie des données, la gouvernance des données de la recherche constitue une fonction stratégique, transversale et structurante pour les établissements scientifiques. Elle dépasse les étapes opérationnelles du cycle de vie des données pour définir un cadre global de référence, structuré autour de politiques, normes, processus et responsabilités. Cette gouvernance vise à garantir la qualité, la sécurité, la traçabilité et la valorisation des données, dans le respect des réglementations nationales et européennes (RGPD, Data Governance Act, AI Act), des principes de la science ouverte et des priorités scientifiques institutionnelles.

Les principales fonctions exercées par un Chief Data Officer (CDO) dans le domaine des données scientifiques sont listées ci-dessous.

- **Définir et piloter une stratégie de gouvernance des données** de recherche orientée vers l'ouverture, l'interopérabilité, la qualité, la souveraineté numérique et la réutilisation des données scientifiques.
- **Encadrer les politiques d'ouverture, de mutualisation et de valorisation des données**, algorithmes et codes sources, en assurant la conformité juridique et éthique des usages.
- **Identifier, évaluer et anticiper les risques et opportunités** associés à la gestion des données (perte, non-conformité réglementaire, fuite de données sensibles, mais aussi mutualisation, valorisation ou innovation ouverte).
- **Superviser la cartographie des données de la recherche**, leurs conditions d'accès, de conservation, d'archivage et de destruction.
- Assurer une **veille stratégique sur les standards, outils et cadres réglementaires** en constante évolution, tout en pilotant l'adaptation des politiques de l'établissement à ces évolutions.
- **Coordonner l'écosystème des parties prenantes** impliquées dans la chaîne de valeur des données (chercheurs, ingénieurs, documentalistes, informaticiens, valorisateurs, juristes, partenaires externes, etc.).

- **Conseiller les directions sur les arbitrages stratégiques** liés aux données : ouverture, confidentialité, souveraineté, innovation etc.
- **Représenter la fonction « données de recherche »** dans les instances stratégiques de l'établissement, et contribuer à la coordination nationale ou internationale des politiques de gestion des données scientifiques.

En croisant les dimensions scientifiques, juridiques, techniques et organisationnelles, le CDO contribue à inscrire les données de la recherche dans une dynamique d'**excellence, de conformité, de transparence et d'impact**, tout en **minimisant les risques** et en **exploitant les opportunités** offertes par une gestion maîtrisée et stratégique des données. Il est ainsi un acteur clé de la transformation numérique de la recherche, garantissant un usage responsable, durable et stratégique des données produites par la communauté scientifique.

Ces fonctions sont généralement assurées par des membres expérimentés de la gouvernance académique (VP recherche, chargé de mission science ouverte, directeur délégué à la recherche, etc.), issus de la recherche et en lien étroit avec les unités et services de l'établissement.

S'il peut exister, au sein des établissements, un rôle analogue pour les données de gestion et de pilotage de la recherche, ces 2 périmètres méritent d'être dissociés du fait du caractère spécifique des données de recherche.

NOTES

Auteurs

Isabelle Blanc, Administratrice ministérielle des données, des algorithmes et des codes sources au Ministère chargé de l'Enseignement Supérieur et de la Recherche

Anne Laurent, Professeure à l'Université de Montpellier, Vice-Présidente déléguée à la science ouverte et aux données de la recherche, Directrice de l'Institut de Science des Données de Montpellier (ISDM), Directrice de l'Antenne Inria de l'Université de Montpellier

Remerciements pour leur expertise au cours de l'étude

Zoé Ancion, ANR

Pierre-Yves Arnould, Collège données, CNRS, Université de Lorraine

Johann Berti, ADBU

Fabien Borget, Aix-Marseille Université

Stéphanie Cheviron, Université de Strasbourg

Sylvie Damy, Université Marie et Louis Pasteur

Frédéric De Lamotte, Collège données, INRAE

Juliette Dibie-Barthelemy, INRAE

Gabriel Gallezot, GIS URFIST

Sandrine Gropp, Université de Montpellier

Christine Hadrossek, Collège données, DDOR, CNRS

Céline Hernandez, Collège données, CNRS

Cyril Heude, Sciences Po

Joanna Janik, DDOR, CNRS

Mariannig Le Bechec, GIS URFIST

Soizick Lesteven, Centre de Données astronomiques de Strasbourg

Grégoire Rey, Inserm, France Cohortes

Sylvie Rousset, DDOR, CNRS

Véronique Stoll, Collège données, Observatoire de Paris

Accompagnement de la mission

Datactivist

Simone et les Robots

Direction de la publication

Ministère chargé de l'Enseignement Supérieur et de la Recherche

2025

DOI : 10.52949/83

MINISTÈRE
CHARGÉ
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

Données de recherche : caractérisation des fonctions clés liées à leur gestion, leur exploitation, leur diffusion et leur utilisation