Guide pour la mise en œuvre et le fonctionnement d'un lac de données dans un établissement de santé et de services sociaux

CITADEL
CENTRE D'INTÉGRATION ET D'ANALYSE DE DONNÉES MÉDICALES

CRCHUM
CENTRE DE RECHERCHE

CONSORTIUM SANTÉ NUMÉRIQUE

Fonds de recherche Santé Québec

CHU Sainte-Justine
Le centre hospitalier universitaire mère-enfant

Université de Montréal

INTRODUCTION

Lancée en 2018, l'**initiative « Accès aux données » de la Table nationale des directeurs de la recherche (TNDR)**[1] a pour mission centrale d'accroître de façon majeure la capacité de recherche et la génération de nouvelles connaissances en organisant et facilitant le traitement des données en santé et services sociaux. Cofinancée par le ministère de l'Économie et de l'Innovation dans le cadre de la Stratégie québécoise des sciences de la vie (SQSV), l'initiative « Accès aux données » est copilotée par le Centre de recherche du Centre hospitalier de l'Université de Montréal (CRCHUM) et le Fonds de recherche du Québec Santé (FRQS). Ses actions s'orientent autour de trois volets principaux :

- Soutien aux établissements du réseau de la santé et des services sociaux dans l'organisation et la structuration de leurs données pour favoriser leur utilisation secondaire pour la recherche et l'amélioration des services aux populations;
- Établissement de règles et de cadres de gouvernance en partenariat avec les établissements du réseau pour assurer une gestion responsable des données de santé selon les meilleures pratiques et en conformité avec les lois, règlements et politiques en vigueur;
- Construction et recensement d'outils et de ressources pour soutenir les établissements et les communautés de recherche et d'innovation dans la conduite de projets favorisant l'acceptabilité sociale de l'accès et de l'utilisation des données de santé.

Fruit de la collaboration entre les équipes du **Centre d'IntégraTion et d'Analyse en DonnÉes médicaLes (CITADEL) du CHUM**[2], du **Consortium Santé Numérique de l'Université de Montréal**[3] et de l'**Univers informationnel du CHU Sainte-Justine (UniC)**[4], le présent guide propose de venir soutenir les établissements de santé et de services sociaux dans la mise en œuvre et le fonctionnement d'un lac de données.[5]

Le guide s'adresse à l'ensemble des acteurs du réseau qui souhaitent favoriser la valorisation des données de santé au service de la recherche, de l'évaluation et de l'amélioration continue des soins et services, dans un esprit de collaboration entre les équipes et les directions d'une institution.

[1] Table nationale des directeurs de la recherche, Initiative Accès aux données : https://tndr-donnees.ca/
[2] Centre d'intégration et d'analyse des données médicales (CITADEL) du CHUM : https://citadel-chum.com/
[3] Consortium Santé Numérique de l'Université de Montréal : https://santenumerique.umontreal.ca/accueil/
[4] Univers informationnel du CHU Ste-Justine : https://recherche.chusj.org/fr/Projet-UnIC
[5] Pour une explication de ce qu'est un lac de données, ses avantages et ses limites, nous vous référons à l'Annexe 2.

Les modèles de lacs de données à la source du Guide

Le guide s'appuie tout d'abord sur le modèle du lac de données **CITADEL du CHUM**.

CITADEL est une plateforme constituée d'un centre d'expertise en science des données et d'une infrastructure de lac de données dans laquelle sont versées les données cliniques, administratives et de recherche du CHUM. La mission de CITADEL est de promouvoir l'innovation en sciences des données dans le domaine de la santé. CITADEL organise et analyse les données clinico-administratives et de recherche en plus d'offrir un appui à l'organisation du CHUM. L'équipe est composée d'experts hautement qualifiés (ingénieurs de données, scientifiques de données, bio-informaticiens, biostatisticiens et autres experts) pour soutenir les chercheurs dans leurs projets de recherche, notamment ceux requérant l'accès et l'analyse de données en santé. Tous les travaux effectués à CITADEL se font sous le cadre réglementaire de la Loi sur les Services de Santé et les Services Sociaux, en conformité avec les normes et règles en vigueur en matière de gestion des données de santé.

De plus, le présent guide contient des éléments complémentaires en lien avec des travaux du **Consortium Santé Numérique de l'Université de Montréal** qui agit en soutien à l'**Univers Informationnel du CHU Sainte-Justine (UniC)**.

L'Univers informationnel est une initiative structurante dont l'objectif est de créer une plateforme centralisée pour fournir aux chercheurs, cliniciens, administrateurs et partenaires des données cliniques et administratives complètes, bien documentées, organisées et mises à jour, afin de promouvoir et faciliter la recherche, l'évaluation et l'innovation. Cette initiative est centrale pour le développement de la recherche clinique, de l'intelligence d'affaires et de l'intelligence artificielle au CHU Sainte-Justine et sera la pierre angulaire du nouveau Centre de valorisation des données mère-enfants.

TABLE DES MATIÈRES

1. PLANIFICATION

1.1. Formation de l'équipe-projet

La planification du lac de données implique la mise en place d'une **équipe-projet multidisciplinaire** qui est chargée de la gestion, du développement et de l'implantation du lac de données de l'établissement. L'équipe-projet mobilisée se compose d'individus portant les rôles suivants :

1. **Porteur de projet**[6] : il va notamment défendre et promouvoir le projet auprès des différentes instances et participer aux discussions concernant les principales décisions stratégiques à prendre. Il peut s'agir par exemple d'un chercheur-clinicien qui est familier avec l'environnement et les pratiques en recherche.
2. **Gestionnaire de projet** : il orchestre et centralise les demandes pour le développement et l'implantation du lac de données dans l'établissement. Un profil mixte à l'interface de la santé et de l'informatique permettra d'assurer une excellente communication avec l'ensemble des parties prenantes.
3. **Expert en technologie de l'information (TI)** : il assure la connexion au réseau du lac, le versement sécurisé des données hospitalières et la communication avec les différents pilotes des systèmes[7].
4. **Expert en gestion, manipulation et analyse de données** : il rend utilisable les données pour les cliniciens ou gestionnaires. Ce profil nécessite une excellente connaissance de la science des données et des problématiques de santé. Il peut s'agir d'un individu ou d'une équipe réunissant l'ensemble de ces compétences. Cette personne est différente d'un responsable qualité, qui peut être mandaté pour faire des audits.
5. **Expert des questions éthiques et juridiques** en lien avec l'accès et l'utilisation des données : il s'assure que le cadre de gestion est en accord avec les politiques d'accès aux données et des cadres éthiques qui peuvent évoluer. Il peut faire partie de l'équipe de gouvernance, selon l'institution visée.
6. **Patient ou citoyen partenaire** : il apporte à l'équipe technique et aux cliniciens le point de vue essentiel de l'utilisateur du système de santé.
7. **Expert en sécurité** : il a pour rôle de veiller à la sécurité de l'information et des ressources.

[6] Pour des fins de concision, le genre masculin est adopté à travers ce document. Il n'est toutefois pas exclusif mais vise à simplifier l'expression des concepts et processus présentés.
[7] Voir Glossaire en fin de document.

1.2. Définition du but poursuivi

Dans le cadre de la planification, il est essentiel d'élaborer **la vision et la mission du lac et de l'utilisation secondaire des données** qui y seront entreposées. Ceci influencera les décisions prises par rapport aux permissions d'accès aux données (cadre de gouvernance, zones de sécurité des données, par exemple) et à l'infrastructure du lac (type de machine et technologies). Dans le cas d'un lac en milieu hospitalier, les objectifs d'utilisation envisageables mais non exhaustifs sont les suivants :

- Acquérir et lier les données pour améliorer les soins aux patients et augmenter la performance du système;
- Fournir un accès facile, sécurisé, approprié, dans un temps raisonnable aux données, pour faciliter la recherche, l'évaluation, l'innovation et l'enseignement;
- Appuyer les décisions par l'analyse et la visualisation de données de haute qualité obtenues en temps réel.

Par ailleurs, pour spécifier le but poursuivi par le lac de données, les éléments suivants sont à considérer:

- les **types de données** qui seront rendues accessibles à travers le lac (p. ex. données clinico-administratives, données d'enquêtes colligées à des fins de recherche, données financières, de gestion, systèmes d'informations et entrepôts de l'établissement, etc.);
- les **modes d'utilisation** qui seront prévus (ex. : utilisation secondaire des données pour la recherche, la gestion, l'amélioration de la prestation des soins et services, etc.);
- les **utilisateurs** qui seront éligibles à accéder aux données provenant du lac (ex. : chercheurs académiques, gestionnaires, chercheurs privés, etc.);
- les **conditions d'accès** de chaque type d'utilisateur (ex : administrateurs de système qui ont un accès élargi incluant les données brutes, utilisateurs avec accès limité aux données dé-identifiées uniquement, etc.);
- les **rôles et responsabilités** des membres de l'équipe du lac de données et des tiers vis-à-vis des données accessibles dans le lac : qui en est le fiduciaire, qui en est l'intendant ou le mandataire (dans le cas d'une délégation de gestion de données à un tiers, etc.).[8]

[8] Voir Glossaire en fin de document.

1.3. Mobilisation des parties prenantes

Si l'équipe-projet du lac de données assume le leadership de la mise en œuvre, elle doit aussi s'assurer du **soutien continu de l'ensemble des acteurs, services et directions** en matière d'accès et d'utilisation des données. Ceux-ci constituent les parties prenantes clés du projet de lac de données.

Avant d'entamer la mobilisation des parties prenantes, il est important d'avoir en sa possession toutes les autorisations nécessaires en vue de l'organisation des données de l'établissement. En général, une entente doit être conclue entre la direction générale de l'établissement et l'équipe-projet afin d'autoriser la structuration du lac de données et son opérationnalisation. Le soutien de la direction générale permet une meilleure communication avec les parties prenantes et assure la mobilisation des différents départements de l'établissement autour du projet et de ses enjeux.

Liste indicative des parties prenantes pouvant être mobilisées dans le cadre de la planification[9] :

- Direction de la qualité, de l'évaluation de la performance et de l'éthique;
- Direction des technologies de l'information/Ressources informationnelles;
- Direction des services professionnels;
- Direction de la recherche;
- Direction des ressources humaines et des affaires juridiques;
- Bureau de la convenance institutionnelle;
- Comité d'éthique de la recherche;
- Comité représentatif des usagers;
- Directions de départements impliquées (pharmacie, laboratoire, etc.) et différents pilotes de systèmes[10].

1.4. Connaissance du contexte législatif, réglementaire et normatif

Soutenue par les parties prenantes, l'équipe projet a pour mission de s'informer des cadres législatifs, réglementaires et normatifs dans lesquels le lac de données va opérer. Ces cadres définissent les modalités de l'accès aux données et les modes d'utilisation possibles.

[9] Selon l'établissement, il est possible que les fournisseurs de système source soient aussi impliqués.
[10] Voir Glossaire en fin de document.

Liste indicative des cadres légaux, réglementaires et normatifs de référence :

- Loi sur les services de santé et les services sociaux (LSSSS);
- Loi sur l'accès aux documents des organismes publics et sur la protection de renseignements personnels (« Loi sur l'accès »);
- Loi concernant le partage de certains renseignements de santé (LPRS);
- Loi modernisant des dispositions législatives en matière de protection des renseignements personnels (Loi 25 ou Projet de loi 64);
- Politiques et procédures de l'établissement dans lequel sera mis en place le lac de données;
- Politique des Trois-Conseils sur la gestion des données de recherche (qui recommande notamment (sans obligation) les principes FAIR et les principes de « science ouverte »);
- Normes et politiques des organismes de financement tels que le Fonds de recherche du Québec Santé (FRQS);
- Normes ISO/EIC en matière de gestion et de sécurité des données;
- Normes en vigueur pour conduire les « Évaluations des facteurs relatifs à la vie privée »;
- Contrats existants avec des fournisseurs des systèmes d'information de l'établissement.

2. GOUVERNANCE

La gouvernance du lac de données se définit en amont dans le processus de planification. Celle-ci fournit un cadre pour que des politiques, procédures et meilleures pratiques soient mises en place afin d'assurer la protection de la vie privée et de réduire les risques liés à la confidentialité et la sécurité des données.

2.1. Organigramme

La structure de gouvernance et les différentes instances assurant le fonctionnement et la gestion du lac peuvent être schématisées sous la forme d'un organigramme. Le modèle de gouvernance présenté à la figure 1 ci-dessous constitue un exemple de ce qui a été mis en place dans les lacs de données à la source du présent guide. Ce modèle peut être utilisé par un établissement et être ensuite adapté aux spécificités de fonctionnement et aux

objectifs de chacune des structures. Par exemple, le projet UniC n'a pas de comité de pilotage, ces fonctions étant assurées par le responsable du lac.

Figure 1 : Proposition d'organigramme pour l'organisation de la gouvernance d'un lac de données

2.2. Définition des rôles et responsabilités

Le **comité de gouvernance** définit les grandes orientations stratégiques du lac de données et de ses utilisations secondaires. Il reçoit notamment les rapports de performance et de sécurité produits par le comité de pilotage afin de s'assurer de la conformité des pratiques vis-à-vis des politiques et procédures instituées. Le comité comprend les représentants des directions impliquées dans l'orientation stratégique des utilisations du lac de données (TI, qualité et performance, recherche, éthique et juridique, services professionnels, enseignement, etc.)

Le **comité de pilotage** est responsable de coordonner les opérations et le développement du lac de données (intégration des systèmes, structuration du lac, transformation des données pour utilisation secondaire) tout en mettant en œuvre les orientations stratégiques décidées par le *comité de gouvernance.* Le comité comprend le *responsable du lac,* ainsi que les représentants des autres directions impliquées dans l'opérationnalisation du lac de données (TI, qualité et performance, recherche et services professionnels, minimalement).

Le **responsable du lac**, avec le soutien du **responsable des opérations**, dirige les équipes opérationnelles (Infrastructure, Scientifiques des données et Bureau d'accès) et veille à ce que tous les membres respectent les normes et standards de sécurité, de qualité et de rigueur scientifique. Il a également le mandat de s'assurer que les politiques, procédures et meilleures pratiques mises en place dans les équipes respectent le cadre de gestion établi, ainsi que les lois et règlements en vigueur, Les demandes d'accès aux données du lac sont traitées par le *bureau d'accès* et autorisées par le *responsable du lac* suite à l'obtention de toutes les approbations requises.

L'**équipe de gestion des infrastructures** réunit des expertises en termes de maintenance technique (serveurs, machines virtuelles, etc.) et de gestion des outils et méthodes de structuration et de protection des données. L'équipe travaille en collaboration avec la direction des ressources informationnelles et la direction de la qualité, de l'évaluation, de la performance et de l'éthique.

L'**équipe des scientifiques des données** possède un volet « intégration » et un volet « analyse des données ». Elle est chargée de l'évaluation de la faisabilité des demandes d'accès aux données, de la gestion des données incluant le contrôle qualité, de la préparation et de l'extraction des sous-ensembles et de l'analyse des données (volets statistiques et bio-informatique). Veuillez noter que ce modèle peut varier selon les centres.

Le **bureau d'accès** est chargé de la réception, de l'évaluation et du traitement des demandes d'accès et d'utilisation des données du lac. Ce bureau assume des fonctions de gestion de projets, d'administration financière et logistique, de gestion des ententes et contrats, de maintenance des registres de suivi et d'analyse éthique et juridique en lien avec le cadre de gestion du lac de données.

2.3. Cadre de gestion, politiques et procédures

Chaque établissement désirant mettre en place un lac de données doit se doter d'un cadre de gestion balisant l'organisation, la gestion et l'utilisation des données.

Le document *Principes directeurs pour assurer le fonctionnement et la gestion optimale d'un centre d'accès aux données* élaboré dans le cadre de l'initiative « Accès aux données » de la TNDR peut servir de guide pour l'élaboration d'un cadre de gestion et donner des exemples de cadres existants.[11]

[11] Table nationale des directeurs de recherche (TNDR), Sous-groupe Gouvernance et cadre de gestion. (2021). Principes directeurs pour assurer le fonctionnement et la gestion optimale d'un centre d'accès aux données de santé. Document en ligne, version 2, 14 octobre 2021 https://tndr-donnees.ca/wp-content/uploads/2021/10/tndr-cadre-principes-directeurs_2021_v2.0-1.pdf

Le cadre de gestion du lac de données s'appuie sur les politiques et procédures existantes au sein de l'établissement, ainsi que sur l'environnement législatif et normatif qui vient réglementer la protection des renseignements personnels. Le cadre de gestion du lac **se décline ensuite en un ensemble de politiques, procédures et autres documentations complémentaires** qui permettent son opérationnalisation, tel que présenté à titre d'exemple à la figure 2 ci-dessous.

Les procédures doivent couvrir l'ensemble du cycle de vie des données incluant leurs utilisations secondaires. Par exemple, une procédure clé à développer est celle qui a trait à la mise en place d'un registre des demandes d'accès aux données et à la journalisation des utilisations faites des données.

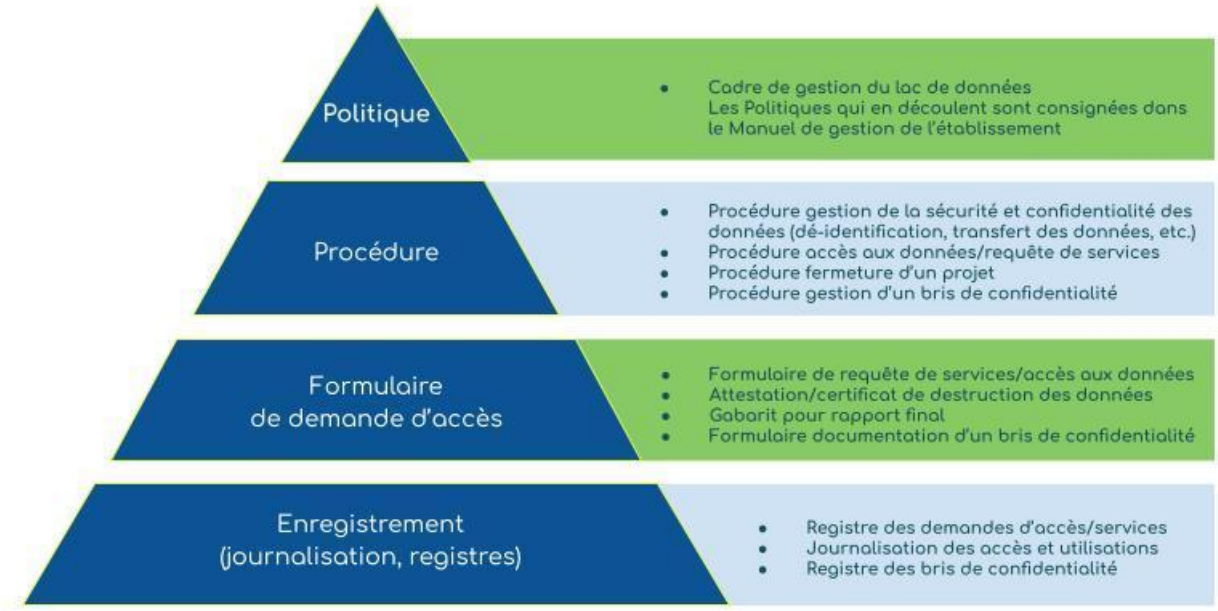


Figure 2 : Représentation de l'architecture des documents de gouvernance d'un lac de données

3. MISE EN OEUVRE

3.1. Réunion des expertises et compétences

Une fois les différents organes de gouvernance mis en place, la structuration du lac de données peut débuter. Mandatée par le responsable du lac, l'équipe-projet commence alors l'organisation et la structuration des données de l'établissement.

La mise en œuvre opérationnelle du lac de données est ainsi assurée par la mobilisation des expertises suivantes :

- Technologies de l'information (TI) et interopérabilité
- Architecture des données
- Science des données
- Gestion de projet

Un **expert en TI et interopérabilité** est indispensable pour établir des connexions avec les données sources, s'assurer que les protocoles de sécurité sont appropriés, configurer les serveurs, mettre en place les environnements, etc.

Un **architecte de données** doit connaître les principes de conception des bases de données et avoir une expérience substantielle dans la conception et l'optimisation de requêtes.

Un **scientifique des données** doit avoir de l'expérience dans la création de jeux de données prêts pour la recherche (y compris le nettoyage et la manipulation des données) et dans la réalisation d'analyses statistiques, d'apprentissage automatique ou les deux. Un gestionnaire de projet va également s'avérer indispensable pour guider l'exécution des processus de structuration du lac de données.

De plus, la **connaissance des principes de base de la recherche en santé**, tels que la conception des études et la sélection des cohortes de patients, est aussi pertinente pour concevoir des systèmes qui seront les plus utiles pour l'établissement et ses utilisateurs.

Un **conseiller ayant une connaissance à la fois des processus cliniques et des systèmes d'information** est aussi utile pour identifier les tables de données et les variables importantes.

D'autres **spécialistes en intégration et analyse des données** peuvent aussi être consultés pour s'assurer que la structure du lac réponde bien aux besoins des utilisateurs. Ces spécialistes peuvent être : des chercheurs, biostatisticiens, spécialistes en intégration de données de format particulier, etc. Ils peuvent être des employés de l'établissement ou des consultants mandatés par l'organisation pour un temps déterminé.

À noter : Le processus de recrutement des membres de l'équipe peut souvent prendre plus de temps qu'anticipé puisqu'un mélange de compétences spécialisées est nécessaire pour mettre en œuvre avec succès un lac de données.

3.2. Implantation du lac de données

Afin d'assurer le bon déroulement de l'implantation du lac de données, différentes étapes au niveau de l'élaboration de l'infrastructure et de la structuration des données elles-mêmes devraient être réalisées. Une vision globale des éléments à considérer est présentée à la figure 3 ci-dessous.

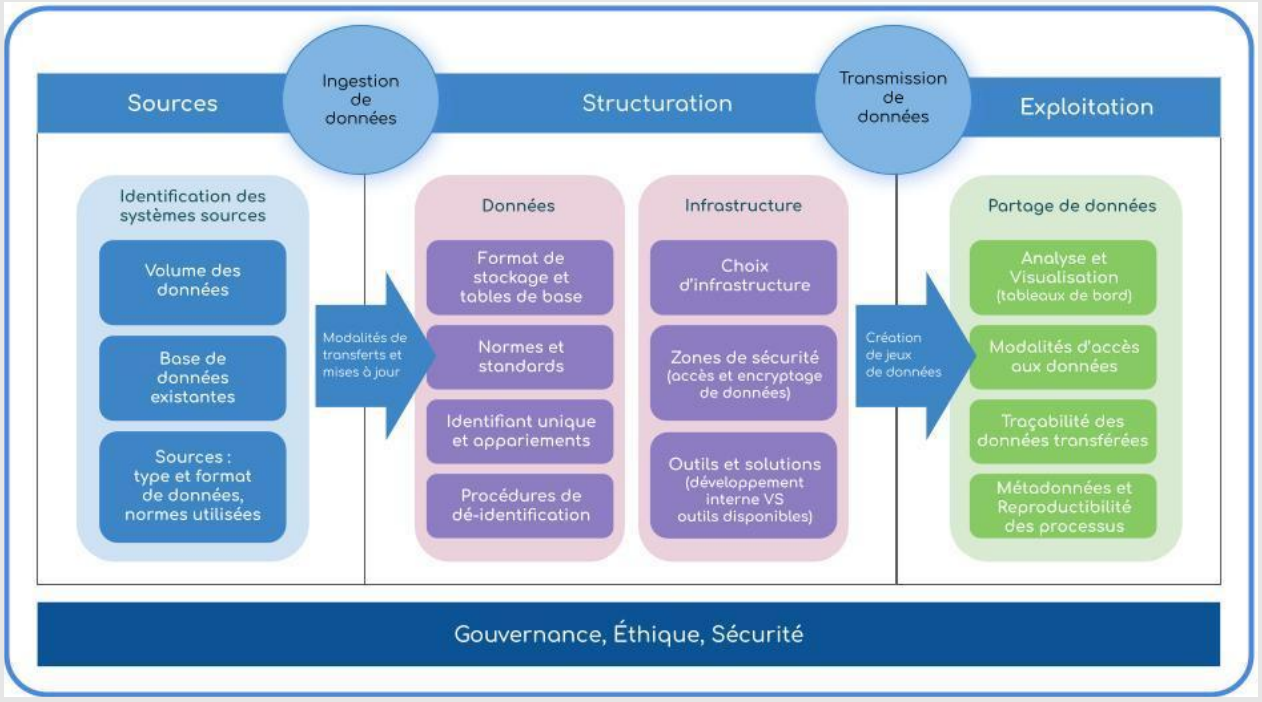


Figure 3 : Éléments structurants de la mise en place d'un lac de données
(pour voir la figure en grand format, se référer à l'annexe 3)

3.2.1. Élaboration de l'infrastructure

L'élaboration de l'infrastructure concerne les sources des données, la planification de l'ingestion et l'architecture interne du lac.

a. Identification des sources de données

Au niveau de la collecte des données destinées à être conservées dans le lac, les démarches suivantes sont fortement recommandées :

- **Recenser les différentes sources de données à travers l'établissement** : Plusieurs systèmes sources peuvent exister selon les départements concernés (ex : système d'admission, départ et transfert (ex. ADT), dossier clinique informatisé (ex. OACIS), laboratoires, pharmacie, entres autres). Il est donc nécessaire d'identifier les

responsables des données de chacun de ces systèmes sources (aussi appelés « pilotes des systèmes »).

- **Assurer l'accessibilité aux systèmes sources** : Il est préférable de se munir d'un accès à toutes les tables de données des systèmes sources. Cependant, selon les modalités d'accès ou la lourdeur des systèmes sources, il est possible qu'une sélection préliminaire de tables soit nécessaire pour peupler le lac de données initialement. À ce titre, l'équipe-projet peut s'appuyer sur des initiatives existantes (telles que des tableaux de bord institutionnels) pour identifier les systèmes et les tables les plus pertinents.
- **Identifier le formatage et schéma des données pour chaque système source** : Il est préférable de répertorier les dictionnaires de données disponibles auprès des différents fournisseurs afin d'obtenir les structures de bases de données, c'est-à-dire la nature et le format des données. Toute information permettant de comprendre l'univers de la donnée (ou méta-donnée) est utile pour bien situer la donnée dans son contexte.

L'implantation d'un lac de données et la description détaillée des différents systèmes sources peuvent nécessiter l'intervention des **experts informatiques** de chacun de ces systèmes (les pilotes ou fournisseurs des systèmes). Si l'intervention de ces experts n'est pas toujours nécessaire pour le versement des données dans le lac, leur expertise pourrait en revanche être déterminante au moment de la structuration des données dans l'infrastructure du lac.

b. Planification de l'ingestion des données

Il est essentiel à ce stade de déterminer comment les données des systèmes sources de l'établissement seront réceptionnées dans le lac. Il s'agit de l'**ingestion des données**. Une base de données tampon, où séjournent les données de manière temporaire, peut être créée sous la responsabilité des administrateurs du lac de données (cas de CITADEL au CHUM) ou de la direction des TI (cas de l'Univers Informationnel au CHUSJ).

Dans tous les cas, il est important de privilégier le transfert des données des systèmes sources vers cette zone tampon ou vers le lac lui-même en dehors des périodes d'utilisation des systèmes dans l'établissement. Il est essentiel de planifier le transfert adéquatement afin de ne pas perturber le fonctionnement opérationnel ni de compromettre l'utilisation des systèmes cliniques correspondants.

Pour créer un système de mise à jour du lac de données :

- **La fréquence et la stratégie (incrémentale ou complète) de chargement des données vers la zone tampon et vers le lac devraient être déterminées.** La

fréquence des mises à jour des données provenant des systèmes sources dépendra de plusieurs facteurs (niveau d'accès accordé, taille des tables, format des données, etc.). Il est conseillé de débuter par des mises à jour moins fréquentes et de développer ultérieurement des solutions augmentant cette fréquence.

- **Le format et l'accessibilité des données nécessaires aux mises à jour devraient être évalués.** Par exemple, les sources de données opérationnelles sont souvent sous la forme de bases de données relationnelles (ex. SQL). Il est alors nécessaire d'assurer la disponibilité des messages transmis en flux continu sous format standard HL7 pour une mise à jour plus fréquente.

c. Définition de l'architecture interne du lac de données

Définir les étapes et les stratégies de transformation des données permet d'établir une architecture globale du lac de données où plusieurs zones de stockage peuvent être considérées. Les différentes zones sont liées au degré de transformation des données et au niveau de confidentialité/sécurité requis. Il est recommandé de prévoir une zone initiale où les données seront déposées dans leur état brut. Ainsi, même si les données sont dé-identifiées ou transformées par la suite, il demeure possible de retrouver leur état initial au cas où il y aurait un besoin de ré-identification, de correction ou tout simplement de validation d'un processus de transformation.

Voici quelques considérations importantes :

- **Prévoir l'infrastructure nécessaire à l'entreposage des données.** Par exemple, les données d'imagerie nécessitent des coûts d'entreposage supplémentaires, des procédures distinctes et des considérations spéciales pour la confidentialité des patients. Une expertise en la matière est essentielle pour assurer la conformité de l'utilisation de ce type de données.
- **Identifier différentes zones de niveau de sécurité selon l'architecture établie.** Les modalités d'implémentation et de séparation de ces zones (utilisation des machines virtuelles par exemple) doivent être déterminées.

3.2.2. Structuration et traitement des données

a. Création et validation des identifiants uniques

Un identifiant unique doit être créé pour lier les données provenant de différentes sources pour un même individu. Un identifiant exploitable peut parfois déjà être fourni par un système source de l'établissement (ex. identifiant du système d'admission, départs et transferts). Une étape de validation de cet identifiant est toutefois nécessaire.

Une stratégie de validation doit ainsi être déterminée en particulier dans le contexte d'utilisation à des fins de recherche où la qualité des données initiales utilisées peut avoir un impact sur les résultats d'une étude. L'étape de validation s'effectue entre autres en croisant plusieurs informations permettant d'identifier des duplicatas d'individus ou de compléter des informations manquantes.

b. Établissement des tables de base

Les tables de base[12] incluent communément la table d'index des patients (identifiants uniques validés) et la table des épisodes de soins. Ce sont les premières tables qui sont créées au moment de la mise en œuvre du lac. Il convient ensuite d'identifier les autres types d'informations qui sont le plus souvent utilisées (un noyau d'information minimum) et d'établir une façon de procéder pour optimiser la liaison de ces informations avec l'identifiant unique du patient. Il est ainsi possible de commencer à créer des tables clés reliées à l'identifiant unique du patient dans un format de schéma SQL par exemple.

Toutes les informations contenues dans les tables du lac de données n'ont pas besoin d'être identifiées de prime abord ; certaines informations peuvent être ajoutées en fonction des besoins des projets de recherche. À chaque ajout de nouvelles informations/tables, les procédures pour l'intégration et la validation de ces nouvelles données, incluant leur format dans le nouvel environnement, doivent être développées dans un souci d'optimisation ou de réutilisation.

Le lac de données et les méthodes d'intégration et de validation devraient viser la flexibilité et l'agilité pour rendre les tables de données prêtes à l'utilisation en fonction des besoins. Un exemple de stratégie pour l'ajout de tables dans un lac de données est présenté dans la figure 4 ci-après. D'autres établissements pourront définir leurs propres tables clés et tables supplémentaires, selon les structures de données disponibles.

À noter : Afin d'optimiser le processus d'extraction des jeux de données et la ré-utilisabilité des résultats intermédiaires, le pipeline de transformation de données peut inclure la création d'une zone où des données nettoyées, inter-croisées et bien documentées sont entreposées (entrepôt de données). Initialement, l'entrepôt peut être composé d'une table par système ou par concept clinique. Il est possible par la suite de l'enrichir avec des variables dérivées qui peuvent servir plusieurs projets de recherche. Ainsi, ces tables d'entrepôt de données peuvent être utilisées pour la plupart des requêtes génériques quotidiennes (création de jeux de données pour les utilisateurs).

[12] Voir Glossaire en fin de document.

c. Tables transactionnelles et mises à jour en temps réel

La mise à jour en temps réel des tables, devenant ainsi des tables transactionnelles (*online transaction processing*, en anglais), apporte sans contredit une valeur ajoutée pour la réalisation de certains projets. Ceci peut être réfléchi dès la mise en place de l'infrastructure, même si les tables transactionnelles peuvent être intégrées dans un second temps. **Avec les tables transactionnelles, il est possible de faire passer les mises à jour hebdomadaires ou quotidiennes de tables à des mises à jour plus fréquentes.** Les tables clés du lac de données ainsi mises à jour en temps réel permettent d'obtenir des données supplémentaires sur l'évolution clinique qui pourraient autrement être perdues dans une base de données opérationnelle. De plus, pour les tables de grande taille, où les approches de mise à jour incrémentale usuelles prendraient un temps considérable, comme les laboratoires, la collecte des flux HL7 en temps réel pour obtenir un accès plus fréquent aux mises à jour des données peut être intéressante à envisager.

Tables de base	Tables clés	Tables supplémentaires
Nécessaire pour la liaison des tables clés	Représente le noyau d'information minimum avec les tables de base	Tables ajoutées selon les besoins
• Identification et données démographiques des patients (index patient) • Épisodes (unités de soins, salle d'urgence, clinique) - données d'admission, de sortie et de transfert (ADT-épisode de soins)	• Laboratoires (sang et microbiologie) • Archives (diagnostics, interventions au cours d'un épisode hospitalier) • Données structurées sur les signes vitaux • Prescriptions	• Données d'imagerie • Signes vitaux en direct et données des capteurs • Tables spécifiques à une spécialité (oncologie, cardiologie, etc.) • Données du bloc opératoire • Données relatives aux transfusions • Données relatives aux ressources humaines • Données financières • Données relatives aux approvisionnement et ressources matérielles • Autres

Figure 4 : Stratégie pour l'ajout de tables dans un lac de données
À noter : les tables de base et tables clés pourront varier selon l'établissement ou la disponibilité des données pour la recherche

3.2.3. Exploitation et valorisation des données

Afin de favoriser l'exploitation optimale des données du lac, plusieurs outils d'exploration, d'analyse et de visualisation peuvent être mis en place et adaptés en fonction du profil et des besoins des différents utilisateurs. Par exemple, au niveau des équipes de gestion et d'évaluation de la qualité et de la performance, des **tableaux de bords, alimentés en continu par les analyses tirées des données réelles**, sont utiles pour favoriser la prise de décision basées sur les données probantes. Ces tableaux de bord devraient être développés à travers un partenariat entre les équipes des directions et les experts du lac de façon à y intégrer les indicateurs de performance et de suivi pertinents pour les utilisateurs de données. Les solutions informatiques et logicielles adaptées pour développer de tels tableaux peuvent être développées à l'interne ou en collaboration avec un partenaire externe.

Au niveau de la recherche, les données du lac peuvent servir à générer de nouveaux projets qui reçoivent alors le soutien des équipes d'experts de l'infrastructure pour accélérer l'accès, la préparation et l'analyse des données pour les projets autorisés. Basé sur des mécanismes de traçabilité et de suivi des accès aux données, **un registre des projets faisant usage des données du lac** peut être mis à jour en continu et publié pour informer les communautés et le public des projets de valorisation en cours dans l'établissement. De surcroît, le lac de données peut faire office de dépôt institutionnel de données pour venir héberger les données de recherche, de manière à assurer un entreposage sécurisé et mettre à disposition des utilisateurs des outils d'analyse à la fine pointe.

3.3. Documentation et transparence au service du partage de connaissances

Finalement, lors de l'implantation du lac de données, il est important de mettre au point des mécanismes et des outils pour **documenter l'évolution des systèmes d'information et des modalités de structuration et d'exploitation des données**. Une grande partie de l'infrastructure générale de programmation du lac devrait être centralisée sous la forme de fonctions personnalisées, de paquets (des ensembles de fonctions standardisés) et de documentation sur les extractions de données. Ainsi, la gestion des paquets de programmation statistique est importante pour s'assurer que les extractions de données sont cohérentes, sans erreur et reproductibles.

La documentation est aussi essentielle pour garantir un **transfert des connaissances** indépendamment du personnel qui effectue une tâche au sein de l'équipe du lac de données. À ce titre, pour valider la qualité et la cohérence des données, des relations étroites avec les experts de contenu terrain et les pilotes de systèmes peuvent être établies

afin de s'assurer que les données du lac de données et celles des systèmes sources correspondent. Ces experts terrain sont essentiels notamment pour aider l'équipe du lac à déterminer les étiquettes, les codes et les définitions à utiliser qui font sens d'un point de vue clinique lors du filtrage et de la sélection des tables de données.

Enfin, l'équipe du lac de données peut activement contribuer au partage des expertises et des savoirs avec les communautés de valorisation des données. Ceci passe notamment par le recours à des plateformes ouvertes de partage et de coconstruction telles que GitHub qui permettent de rendre disponibles les codes et analyses, et de recevoir les apports et rétroactions d'experts internes et externes à l'établissement. Ces contributions sont utiles pour bonifier sans cesse la rigueur et la sécurité des pratiques en cours et permettre la reproductibilité des analyses réalisées. À ce titre, **l'endossement des principes FAIR** pour rendre les données **F**acilement trouvables, **A**ccessibles, **I**nteropérables et **R**éutilisables devrait constituer une valeur centrale de l'implantation et du fonctionnement d'un lac de données.

Le document *Recommandations pour le bon fonctionnement d'un centre d'accès aux données* élaboré dans le cadre des activités des Fonds de recherche du Québec peut servir de guide pour l'intégration des meilleures pratiques en matière de valorisation des résultats de la recherche conduite à partir des données d'un organisme public.[13]

4. FONCTIONNEMENT

4.1. Gestion des requêtes de services et/ou d'accès aux données

4.1.1. Réception des requêtes

La gestion des requêtes de services et/ou d'accès aux données du lac se veut en adéquation avec les modalités d'accès décrites dans le cadre de gestion (voir section 2.3). Conformément aux politiques et procédures en vigueur, la gestion des requêtes suit un modèle d'accès déterminé dont le processus est initié par l'**envoi d'un formulaire de requête** au bureau d'accès.

[13] Fonds de recherche du Québec (2022). Recommandations pour assurer le bon fonctionnement d'un centre d'accès aux données pour la recherche au sein d'un organisme public. Document en ligne, version 2, juin 2022. https://frq.gouv.qc.ca/app/uploads/2022/06/recommandations-pour-les-centres-dacces-aux-donnees_comite-frq_juin2022.pdf

Idéalement, ce formulaire devrait être accessible via une plateforme centralisée de gestion et de suivi des demandes. Les utilisateurs sont encouragés à contacter le bureau d'accès via une adresse courriel qui lui est propre pour obtenir du soutien pour la formulation des requêtes et/ou valider qu'ils ont les autorisations et documents nécessaires pour obtenir l'accès aux données.

La figure 5 ci-dessous présente un exemple de processus d'accès aux données pour exposer les principales étapes parcourues par les utilisateurs à travers leur requête.

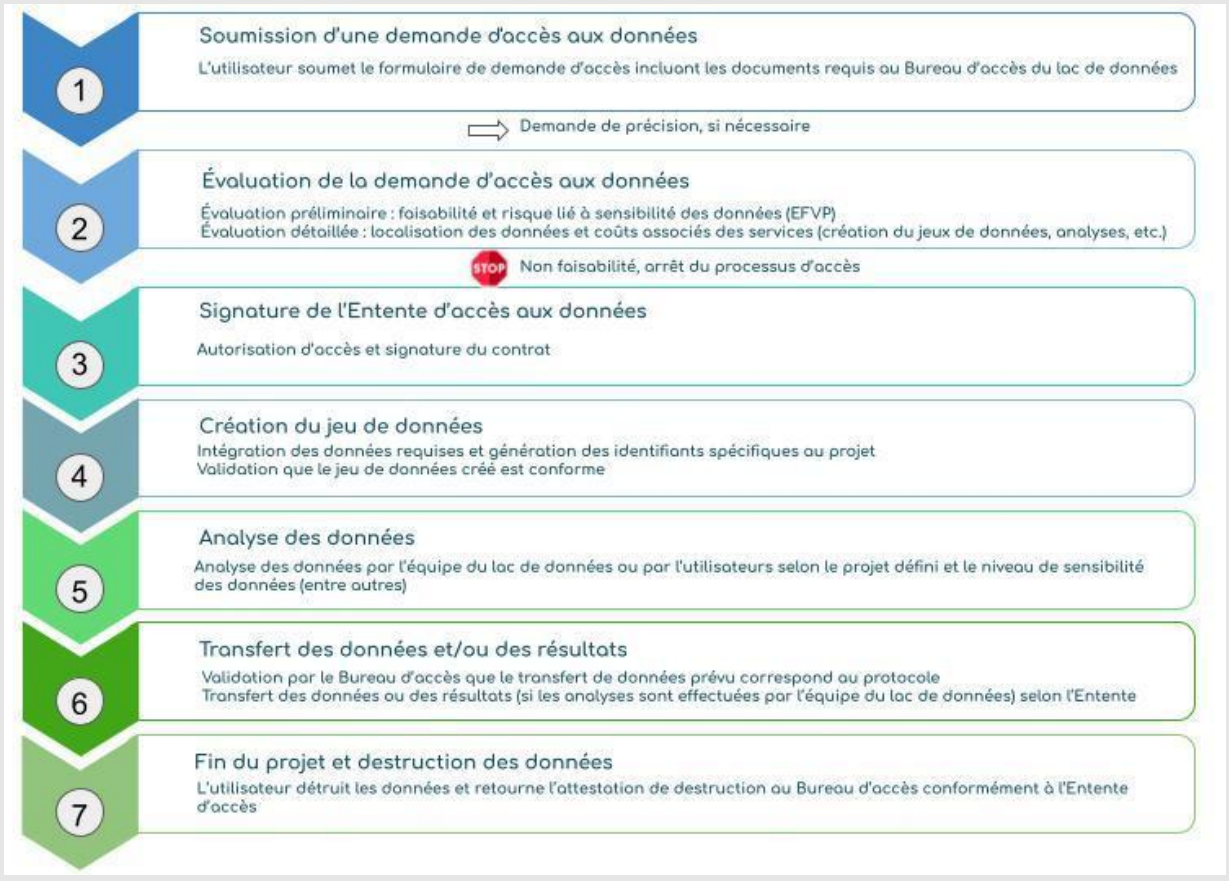


Figure 5 : Exemple de modèle d'accès aux données (source pour inspiration : CITADEL)

Les requêtes d'accès aux données sont réceptionnées par le coordonnateur du bureau d'accès qui se réfère au **cadre de gestion** pour valider que les conditions d'accès aux données sont remplies et vérifier que les utilisateurs possèdent les autorisations requises ainsi que les fonds nécessaires pour la réalisation du projet.

Conditions d'accès aux données de CITADEL (volet recherche)

Les projets de recherche pour lesquels toutes les autorisations requises ont été obtenues (approbation du comité d'éthique de la recherche, approbation du directeur des services professionnels, autorisation d'effectuer de la recherche par le directeur de la recherche, autres approbations d'organismes détenteurs de données si applicable) peuvent faire l'objet d'une demande d'accès aux données de CITADEL. L'instigateur de la demande d'accès doit être un utilisateur interne CHUM (i.e. individu ayant un statut de chercheur au CHUM ou un résident).

Pour les projets provenant du milieu académique externe au CHUM ou les projets provenant de l'industrie, une collaboration doit être établie avec un chercheur ou un département du CHUM préalablement à la demande d'accès aux données. L'accès aux données de CITADEL à des fins d'enseignement (ex. étudiant réalisant un stage nécessitant des données de CITADEL ne s'inscrivant pas dans le cadre d'un projet de recherche) est possible sous certaines conditions et évalué au cas par cas.

4.1.2. Priorisation des requêtes

Après réception de la requête par le bureau d'accès, un gestionnaire de projet peut effectuer une priorisation selon des critères préétablis et la communiquer aux utilisateurs.

Outre la date de réception du formulaire de requête, les éléments suivants peuvent influencer la priorité de traitement des requêtes :

- Il s'agit d'une étude de faisabilité qui précède une requête d'accès aux données ou de services;
- La requête est liée à une demande de subvention (avec une date limite de soumission aux organismes subventionnaires);
- Il s'agit d'une question en lien avec une revue de document (protocole, article, publication, question des réviseurs, analyses);
- Il existe une entente préalable signée entre l'utilisateur et la plateforme de services du lac de données pour la réalisation de projets contractuels.

4.1.3. Évaluation des requêtes

Après réception de la requête et son évaluation primaire (documents complets et valides), une consultation initiale est prévue avec l'utilisateur (et/ou les membres désignés de son équipe), le gestionnaire de projet et le membre de l'équipe du lac de données assigné au projet (analyste TI, biostatisticien, etc.) afin de préciser le mandat et les attentes.

L'utilisateur doit avoir en sa possession pour la rencontre toute information relative au projet, par exemple son protocole de recherche, une liste de participants ou les critères d'identification de sa population cible, une liste de variables d'intérêts incluant les codes standards ou internes utilisés pour créer ces variables lorsque disponible.

À la suite de cette rencontre, le gestionnaire de projet du bureau d'accès produit une estimation des coûts (soumission) et une évaluation du délai d'exécution en fonction de la nature de la requête, qui sera validée par le responsable des opérations et le responsable du lac avant son envoi à l'utilisateur.

Il est possible que, suite à l'étape de soumission, un projet ne soit pas pris en charge par la plateforme de services. Cela peut être le cas lorsque le délai d'exécution demandé est trop court par exemple. L'utilisateur peut alors être redirigé vers d'autres ressources.

À noter : la priorisation des projets et l'évaluation des requêtes sont des étapes à considérer par les autres centres hospitaliers et de recherche qui pourront en adapter les modalités. Par exemple, les rencontres avec les utilisateurs peuvent être précédées par des évaluations préliminaires permettant l'établissement d'un plan d'extraction de données et la détermination des coûts associés. Lors de la rencontre, des ajustements de l'énoncé des travaux et/ou des prix peuvent être apportés.

4.2. Offre de services

4.2.1. Services en lien avec les requêtes d'accès aux données

L'équipe du lac de données peut offrir un ensemble de services en vue de préparer les jeux de données pour optimiser leur utilisation secondaire :

- **Étude de faisabilité** : Identification de la population ciblée par l'équipe de recherche et/ou extraction de données préliminaires pour estimer la faisabilité d'un projet;
- **Demande impliquant la recherche ciblée d'un ou de plusieurs profils patients** dont le résultat correspond à des critères précis nécessaires à l'élaboration d'un projet de recherche;

- **Préparation des jeux de données clinico-administratives** : extraction, nettoyage et transformation des données pour obtenir un jeu de données prêt à l'analyse;
- **Gestion des données en prospectif** : analyse des besoins pour colliger des données en temps réel et mise en place de l'infrastructure de collecte et d'analyse;
- **Intégration de nouvelles variables/données à l'infrastructure d'analyse**;
- **Gestion et entreposage des données de recherche** : construction de schémas de données spécifiques à un projet et services de maintenance;
- **Vérification algorithmique** : Développement, validation interne ou externe de solutions et/ou algorithmes sur des données de recherche ou clinico-administratives.

Des collaborations avec d'autres départements de l'établissement peuvent être établies pour mettre en place une offre de services plus complète.

4.2.2. Services complémentaires en statistiques et méthodologie

Outre les services en lien avec les requêtes d'accès aux données formulées par les utilisateurs, l'équipe responsable du lac de données pourrait aussi proposer des services complémentaires en méthodologie et statistiques selon l'expertise disponible.

- **Consultation méthodologique** : soutien quant au choix de la meilleure méthodologie de recherche, d'analyse et de modélisation;
- **Analyse des données et interprétation des résultats** : réalisation d'analyses par l'équipe d'experts du lac de données et rédaction de rapports d'analyse (analyses statistiques et bio-informatique);
- **Soutien à la production scientifique** : relecture d'articles ou de demandes de subvention, participation à l'écriture de certaines sections méthodes d'écrits scientifiques pour fins de subventions ou publications, réponses aux questions des réviseurs.

4.2.3 Établissement du modèle de coûts

Le modèle de coûts des services proposés doit être établi en parallèle à la définition des services offerts. Selon les types de collaborations établies (académique, publique-privée, etc.), les coûts définis peuvent être différents. **La grille tarifaire des services offerts indique les éléments qui font varier les coûts et est partagée de façon transparente avec les utilisateurs.**

4.3. Entente d'accès et date de début du projet

Suite à l'obtention de toutes les autorisations requises et l'acceptation de la soumission de l'équipe du lac de données, l'utilisateur doit signer une entente d'accès aux données ou de services avec l'institution d'attache du lac.

En plus de contenir une brève description du projet et un aperçu des données requises, l'entente d'accès, dans un cadre de recherche, devrait couvrir au minimum les points suivants :

- **Responsabilités et obligations de chacune des parties**
- **Transfert des connaissances, diffusion des résultats et reconnaissance des auteurs** : Toute publication ou présentation utilisant les données du lac de données devrait inclure une mention reconnaissant l'apport de l'équipe du lac de données. Il est recommandé que l'utilisateur s'engage à reconnaître la contribution d'un membre de l'équipe du lac de données lorsque celle-ci respecte les règles et meilleures pratiques en matière de reconnaissance d'auteurs sur les publications.
- **Période d'usage et de conservation des données utilisées** : il s'agit dans cette section de prévoir le devenir du jeu de données confié à l'utilisateur, lorsque le projet de recherche est terminé. La conservation des données est encouragée à condition de pouvoir en démontrer la pertinence et que les meilleurs standards de sécurité soient assurés. Si les garanties ne sont pas adéquates, une date et un processus de destruction des données doivent être proposés. À noter que tout règlement décrivant des conditions de conservation de données de santé et de données administratives doit être respecté.
- **Découverte d'une erreur fortuite dans le sous-ensemble des données auquel l'utilisateur a obtenu accès** : L'utilisateur doit aviser les responsables du lac de données immédiatement afin que puissent se déployer les interventions nécessaires pour corriger l'erreur le plus rapidement possible.
- **Découverte d'une information permettant d'identifier un individu ou bris de confidentialité** : L'utilisateur doit aviser les responsables du lac de données immédiatement afin que puissent se déployer les interventions nécessaires pour remédier à la situation le plus rapidement possible.
- Selon le niveau de risque établi, **l'entente stipule également si l'utilisateur est autorisé à recevoir une copie des données** extraites ou s'il devra les travailler à l'intérieur des espaces sécurisés du lac de données.
- **La soumission comprenant le détail des services entendus et les coûts** est aussi intégrée à l'entente.

Le projet débute lorsque la documentation minimale requise est reçue et que l'entente est dûment complétée, signée et validée par le Bureau d'accès. Advenant une modification du

mandat en cours de projet, un amendement à la soumission peut être effectué et devrait être à nouveau endossé par le responsable du projet. Aucune dépense ne devrait être encourue sans l'autorisation explicite de la personne responsable du projet.

Modèle Entente d'accès aux données de CITADEL

Au sein de CITADEL, l'entente d'accès aux données est dûment signée par l'utilisateur de données, le Responsable de CITADEL et le Directeur de la recherche du CHUM pour permettre le transfert et l'utilisation des données de CITADEL. Le responsable de CITADEL se réserve le droit de suggérer que l'instigateur de la demande d'accès soit accompagné par l'équipe scientifique de CITADEL pour obtenir du soutien méthodologique lorsque jugé nécessaire. Ceci est alors stipulé dans l'entente.

4.4. Transfert des données ou résultats

Une fois la demande d'accès aux données ou de prestation de services complétée, le bureau d'accès revoit la documentation du projet dans son ensemble (incluant les documents éthiques et réglementaires requis) et valide que tout est conforme avant le transfert de données et/ou de résultats à l'utilisateur. Une procédure de transfert des données et/ou résultats permet de s'assurer que les méthodes de transfert utilisées sont sécuritaires (ex. dépôt dans répertoire sécurisé interne, fichier encryptée transmis par canal sécurisé). La politique spécifie, entre autres, le niveau de confidentialité pour l'envoi des données et/ou des résultats, ainsi que le mode d'expédition autorisé selon que l'envoi s'effectue intra ou extra-établissement.

Procédure de transfert des données et/ou résultats de CITADEL

Procédure de transfert des données et/ou résultats de CITADEL À titre d'exemple, pour le transfert de résultats à l'interne, CITADEL dépose les résultats directement dans le répertoire sécurisé du chercheur-utilisateur. Dans le cas de transfert à l'externe l'encryptage du fichier est effectué avec l'utilisation d'un mot de passe transmis verbalement au chercheur-utilisateur.

4.5. Formations

Un guide de formations et des mécanismes de suivi pour le rappel de ces formations devront être développés avant la mise en exploitation des données du lac. Les formations couvrent les formations obligatoires dans l'établissement auxquelles sont ajoutées les formations spécifiques à l'exploitation des données, y compris les formations éthiques et techniques qui peuvent être différentes d'un établissement à l'autre.

Le guide de formation spécifie les formations pour l'équipe opérationnelle du lac de données mais également les formations requises pour les utilisateurs. Ces dernières devraient suivre le modèle de l'établissement et sont en adéquation avec le statut de l'utilisateur et les utilisations prévues des données (recherche, gestion, évaluation, etc.).

Formations de l'équipe opérationnelle CITADEL

Au sein de CITADEL, tous les membres de l'équipe, sans exception, doivent compléter les formations suivantes et fournir un certificat de complétion ou une attestation de lecture selon la formation :

- Canada GCP 1 Good Clinical Practice (GCP) – Canada
- Canada GCP 2 Health Canada Division 5 - Drugs For Clinical Trials Involving Human Subjects
- Formation en éthique de la recherche (MSSS) : les 3 niveaux
- Attestation de conduite responsable en recherche du CRCHUM
- MON's du CRCHUM (1 à 30)

Des formations complémentaires peuvent aussi s'avérer utiles dans le cadre de la gestion des données de santé.

Il s'agit par exemple des formations à l'utilisation du formatage de données FHIR qui sont organisées par le Consortium Health Level Seven International (HL7), ou encore les formations en IA et valorisation des données proposées par IVADO et MILA, ainsi que les formation à la gouvernance et la gestion des données des Premières Nations, Métis et Inuits dispensées par le Centre sur la gouvernance de l'information des Premières Nations.

Conclusion

Le guide dont vous achevez la lecture a pour ambition principale de fournir un ensemble de pistes de réflexion et d'actions en vue de mettre en œuvre un lac de données dans un établissement de santé et de services sociaux.

À dessein, le guide aborde ce projet d'envergure en partant de la phase de la planification, pour passer à celle de la gouvernance et de la mise en œuvre pour aboutir à celle de la gestion du fonctionnement des demandes d'accès et d'utilisation des données. Il est en effet essentiel de bien considérer l'ensemble de ces étapes avec attention et d'y dédier les efforts et les ressources suffisantes. Le changement organisationnel amené par cette nouvelle infrastructure nécessite un leadership actif et continu des instances dirigeantes, un alignement des visions des différentes parties prenantes, la mobilisation des expertises et des ressources, et le soutien des acteurs tout au long du projet.

Grâce à cela, il est possible de parvenir à l'établissement d'un consensus durable autour de la valeur ajoutée que représente le lac de données pour l'organisation de santé et la rencontre de ses finalités de recherche et d'amélioration continue des soins et services.

GLOSSAIRE

Base de données : Collection de données structurées ou non pour permettre des opérations, parfois très complexes, de lecture, de suppression, de déplacement, de tri, de comparaison ou autres opérations. Lorsque plusieurs bases de données sont constituées sous forme de collection, on parle alors d'une banque de données.

Chercheur-utilisateur : Correspond à la personne physique qui possède un statut en règle au sein de l'établissement ou du centre de recherche et qui en tant que tel est autorisé à accéder à la donnée afin d'accomplir une tâche donnée dans le cadre du programme de recherche dont il est responsable. De manière plus générale, l'utilisateur de la donnée est responsable de suivre les lois, politiques, procédures et standards associés à la donnée qu'il utilise et d'utiliser cette dernière uniquement aux fins auxquelles il a été autorisé. L'utilisateur de la donnée a également la responsabilité de signaler à l'intendant de la donnée tout accès non autorisé, utilisation abusive ou problème de qualité de la donnée.

Dé-identification (aussi nommé dépersonnalisation ou pseudo-anonymisation) : Procédure qui consiste à remplacer les informations nominatives contenues dans un document par un code d'identification, de manière à empêcher l'identification des individus auprès desquels elles ont été recueillies.

Fiduciaire de la donnée : Correspond à la personne physique ou morale qui est responsable de la planification et de l'élaboration des politiques en lien avec la gestion de la donnée (notamment en ce qui concerne son accès et son utilisation). Le fiduciaire de la donnée applique les exigences légales et réglementaires en lien avec la donnée et supervise la mise en place de politiques et de processus de gestion des données. À ce titre, le fiduciaire de la donnée assure notamment les orientations stratégiques et financières liées à la donnée.

Intendant de la donnée : Correspond à la personne physique ou morale qui est directement responsable de la gestion de la donnée au niveau opérationnel. L'intendant de la donnée gère et maintient la donnée dont il a la charge et à ce titre, assure la qualité, la sécurité et la confidentialité de ladite donnée. L'intendant de la donnée est notamment responsable d'implanter et de gérer l'application de toutes les politiques liées à la qualité de la donnée, la standardisation de la donnée et l'accès à la donnée.

Jeu de données : Ensemble de valeurs « organisées » ou « contextualisées » (alias « données »), où chaque valeur est associée à une variable (ou attribut) et à une observation. Une variable décrit l'ensemble des valeurs décrivant le même attribut et une observation contient l'ensemble des valeurs décrivant les attributs d'une unité (ou individu statistique).

Mandataire de la donnée : Correspond à la personne physique ou morale qui est mandatée pour accomplir le rôle d'intendant de la donnée.

Pilote des systèmes : Correspond à la personne responsable de l'exploitation technique d'un système source et des bases de données qui y sont associées. Les systèmes sources peuvent correspondre à des départements du système hospitalier (pharmacie, imagerie, obstétrique par exemple) ou couvrir plusieurs départements.

Ré-identification : Tout processus rétablissant le lien entre l'information et l'identité d'un individu.

Renseignements personnels : Renseignements portant sur un individu et permettant d'établir son identité.

Table de données : Dans les bases de données relationnelles, une table est un ensemble de données organisées sous forme d'un tableau où les colonnes correspondent à des catégories d'information (une colonne peut stocker des identifiants uniques, une autre des informations (âge, sexe, poids, etc.…) et les lignes à des enregistrements, également appelés entrées.

RÉFÉRENCES ET LIENS PERTINENTS

Table nationale des directeurs de recherche (TNDR), Sous-groupe Gouvernance et cadre de gestion. (2021). Principes directeurs pour assurer le fonctionnement et la gestion optimale d'un centre d'accès aux données de santé. Document en ligne, version 2, 14 octobre 2021. https://tndr-donnees.ca/wp-content/uploads/2021/10/tndr-cadre-principes-directeurs_2021_v2.0-1.pdf

Fonds de recherche du Québec (2022). Recommandations pour assurer le bon fonctionnement d'un centre d'accès aux données pour la recherche au sein d'un organisme public. Document en ligne, version 2, juin 2022. https://frq.gouv.qc.ca/app/uploads/2022/06/recommandations-pour-les-centres-dacces-aux-donnees_comite-frq_juin2022.pdf

Site Internet de CITADEL : https://citadel-chum.com/

Site Internet de la TNDR, Initiative Accès aux données : https://tndr-donnees.ca/

Informations sur la communauté de pratique Accès aux Données : https://tndr-donnees.ca/communaute/

ANNEXE 1 : LES AUTEURS ET AUTRICES DE CE GUIDE

Ce guide est le fruit du travail assidu et collaboratif des personnes suivantes, expertes de domaines de connaissances divers et complémentaires en gestion et valorisation des données en santé et services sociaux.

Liste des auteurs et autrices par ordre alphabétique :

- Camille Craig, Adjointe à la direction au CRCHUM
- Cécile Petitgand, Ph.D., Coordonnatrice de l'initiative Accès aux données de la TNDR et Conseillère en gestion de données au CRCHUM et au FRQS
- Kamran Afzali, Ph.D., Conseiller principal, scientifique de données, Consortium Santé Numérique
- Khedidja Seridi, Ph.D., Conseillère principale, scientifique de données, Consortium Santé Numérique
- Kip Brown, Directeur des opérations de CITADEL, CRCHUM
- Michaël Chassé, intensiviste au CHUM, chercheur au CRCHUM, Directeur scientifique de CITADEL et Directeur adjoint Sciences des données au CRCHUM
- Marie-Ève Cantin, gestionnaire de projets à CITADEL, CRCHUM
- Marie-Noël Nadeau, responsable du Bureau d'accès de CITADEL, CRCHUM
- Pascale Beliveau, Ph.D., Conseillère principale, scientifique de données, Consortium Santé Numérique
- Yan Terrat, Ph.D., Conseiller principal, scientifique de données, Consortium Santé Numérique

Nous tenons à remercier tout particulièrement Catherine Boileau, épidémiologiste et gestionnaire de projet au Centre de recherche du CHU Ste-Justine, pour la richesse de ces apports et commentaires au cours de la conception de ce guide.

Pour citer ce document :

Initiative « Accès aux données » de la Table nationale des directeurs de la recherche (CRCHUM et FRQS), Centre d'IntégraTion et d'Analyse en DonnÉes médicaLes (CITADEL) du Centre hospitalier de l'Université de Montréal (CHUM), Consortium Santé Numérique de l'Université de Montréal, & Univers informationnel du CHU Ste-Justine (UniC). (2022, September 15). Guide pour la mise en œuvre et le fonctionnement d'un lac de données dans un établissement de santé et de services sociaux. Zenodo. DOI: 10.5281/zenodo.7218119

ANNEXE 2 : QU'EST-CE QU'UN LAC DE DONNÉES ?

Comme les autres secteurs d'activités, le milieu hospitalier n'échappe pas à l'augmentation importante du volume de données générées par les activités administratives et de soins prodiguées. En 2019, le rapport « Dell Technologies Global Data Protection Index » mentionnait que la quantité de données générées dans le secteur hospitalier avait augmenté de 878% de 2016 à 2018 (Donovan, 2019). En plus d'avoir un volume accru, les données hospitalières répondent également aux caractéristiques des données massives que sont les 3V du « Big Data », à savoir :

- La Variété : ce sont des données multidimensionnelles et de sources très variées. C'est le cas de plus de 80% des données produites dans le domaine de la santé (Hyoun-Joong, 2019);
- Le Volume : les hôpitaux hébergent aujourd'hui des pétabytes de données (il y a 1024 térabytes dans un pétabyte);
- La Vélocité : un flux continu de données sont produites et demandent à être analysées en temps réel.

Les lacs sont des infrastructures qui permettent l'hébergement et l'analyse croisée de données massives et hétérogènes. Cette option de stockage apparaît donc comme une solution viable pour l'exploitation des données dans le domaine de la santé (Zagan et al., 2019). L'appellation de « lac » est d'ailleurs appropriée pour résumer les propriétés d'une telle infrastructure. Elle reçoit en effet des flux continus de données de nombreuses sources aux formats variés. Ces données ne sont pas transformées et restent donc sous forme « brutes », en contraste avec les pratiques antérieures de stockage de données structurées et formatées. Le stockage des données dans leur format natif comporte d'ailleurs plusieurs avantages dont celui de faciliter leur réutilisation pour un ensemble de finalités (Sawadogo & Darmont, 2021).

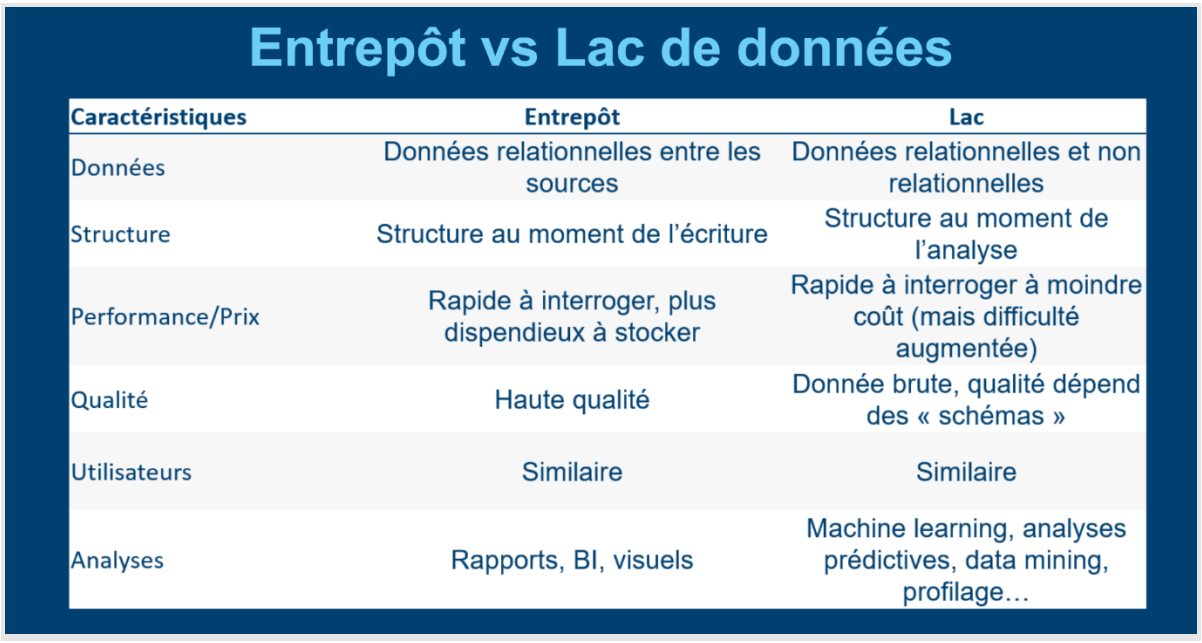
Le lac de données peut ainsi soutenir la valorisation des données au service de la recherche, de la gestion, de l'innovation et des soins au service des patients, soignants et populations. Un accès plus optimal aux données via un lac permet notamment de :

- Favoriser la mise en œuvre de projets de recherche fondées sur des données massives et hétérogènes;
- Faciliter la mise en œuvre de la médecine personnalisée et des systèmes d'intelligence artificielle;
- Améliorer la qualité et la sécurité des soins et services grâce aux suivis en temps réel;
- Optimiser l'utilisation des ressources et des processus de gestion;
- Mieux anticiper le futur et planifier la croissance des unités de soins.

Quels sont les avantages et les limites des lacs de données ?

Les avantages

À l'origine, les lacs ont été mis en place pour pallier les limitations des bases de données relationnelles de type « entrepôts de données ». Ces dernières imposent en effet un format de dépôt défini, sous forme de tableaux, qui est peu susceptible d'héberger la grande variété des types de données rencontrées dans un système de santé. Il s'agit par exemple d'images, de notes cliniques ou d'enregistrements de signes vitaux. Au sein d'une infrastructure agile comme celle d'un lac des données, les données sont organisées dans leur format natif et ne seront transformées que lors d'étapes subséquentes à leur entreposage (Zagan et al., 2020).



Caractéristiques	Entrepôt	Lac
Données	Données relationnelles entre les sources	Données relationnelles et non relationnelles
Structure	Structure au moment de l'écriture	Structure au moment de l'analyse
Performance/Prix	Rapide à interroger, plus dispendieux à stocker	Rapide à interroger à moindre coût (mais difficulté augmentée)
Qualité	Haute qualité	Donnée brute, qualité dépend des « schémas »
Utilisateurs	Similaire	Similaire
Analyses	Rapports, BI, visuels	Machine learning, analyses prédictives, data mining, profilage…

Figure 6 : Entrepôt versus Lac de données (source pour inspiration : CITADEL)

Un autre avantage d'un lac de données est de pouvoir conserver l'intégralité de l'information qui y est contenue. Au sein d'un lac, il n'est pas nécessaire d'intégrer ou de relier toutes les sources de données potentiellement pertinentes pour les recherches futures et les besoins institutionnels. La philosophie du lac de données est de permettre une certaine flexibilité pour mieux s'adapter aux demandes à venir en lien avec la valorisation des données.

Finalement, les données d'un lac peuvent être organisées par la suite dans un format commun (*common data model*), tels que FHIR et OMOP qui facilitent l'interopérabilité et les analyses entre différents établissements de santé.

Les limites

Tout d'abord, la taille de l'infrastructure peut limiter la capacité de fonctionnement d'un lac de données. Bien que beaucoup de données soient entreposées au sein d'un lac, ces données peuvent ne pas être suffisantes pour répondre à une question scientifique spécifique (Yu et al. 2022). Dans ce cas, l'approche d'apprentissage dite « fédérée » permet de mettre en commun des entrepôts de données décentralisées de manière à maximiser le potentiel de valorisation des données. C'est ce que démontre le projet CODA-19[14], fondé sur la connexion de bases de données en lien avec la COVID-19 dans neuf établissements du Québec.

Yu et collaborateurs (2022) notent également que la mise en place d'un lac n'assure pas nécessairement le déploiement fructueux de solutions technologiques fondées sur la valorisation de données massives telles que les solutions d'IA. Malgré l'existence d'un lac, les défis de développement et de déploiement de l'IA au chevet des patients sont nombreux et comprennent par exemple :

- La faible qualité de données ainsi que le manque de standards de formatage;
- L'existence de systèmes sources de données hétérogènes;
- Le faible nombre de données à l'échelle d'une même institution, rendant par exemple difficile l'entraînement et la mise en place d'outils diagnostics fondées sur l'IA;
- La spécificité des données hébergées dans un établissement. Les modèles d'IA entraînés dans une institution donnée peuvent être difficilement transférables à d'autres institutions où les populations sont différentes.

Ainsi, pour des projets complexes comme ceux fondés sur l'IA et les données massives, l'existence d'un lac est un facteur moteur mais n'est pas en soi suffisante. Des collaborations stratégiques peuvent alors être établies pour avoir plus de données et obtenir la puissance statistique nécessaire pour répondre à une question de recherche spécifique ou garantir la fiabilité des modèles d'IA dans plusieurs contextes.

Finalement, pour reprendre l'analogie du lac, il est important que ces structures ne se transforment pas en « marais » au cours du temps. Il est ainsi primordial de constituer des registres de documentation des données conservées dans le lac et des actions entreprises, de manière à produire et conserver en tout temps des métadonnées standardisées et à jour (Sawadogo et al., 2021). Ceci permet ainsi de garantir que l'information et les connaissances extraites des données du lac soient les plus fiables et rigoureuses possibles, de façon à servir les besoins de l'institution, des soignants, gestionnaires et patients.

[14] https://www.coda19.com/

Références :

Donovan (2019). Organizations See 878% Health Data Growth Rate Since 2016. HIT Infrastructure (May 8, 2019). Online: https://hitinfrastructure.com/news/organizations-see-878-health-data-growth-rate-since-2016

Kong, H. J. (2019). Managing unstructured big data in healthcare system. Healthcare informatics research, 25(1), 1-2.

Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. Journal of Intelligent Information Systems, 56(1), 97-120.

Yu, M., Tang, A., Brown, K., Bouchakri, R., St-Onge, P., Wu, S., ... & Chassé, M. (2021). Integrating artificial intelligence in bedside care for covid-19 and future pandemics. bmj, 375.

Zagan, E., & Danubianu, M. (2020). Data Lake Approaches: A Survey. In 2020 International Conference on Development and Application Systems (DAS) (pp. 189-193). IEEE.

Zagan, E., & Danubianu, M. (2019). From Data Warehouse to a New Trend in Data Architectures–Data Lake. IJCSNS International Journal of Computer Science and Network Security, 19(3).

ANNEXE 3 : ÉLÉMENTS STRUCTURANTS DE LA MISE EN PLACE D'UN LAC DE DONNÉES