

CépiDc

 Centre d'épidémiologie sur
les causes médicales de décès

Livrable Phase 1 – Analyse de l'Existant

Document C2-Restreint

Table des matières

1.	Introduction.....	4
1.1.	Contexte du document et démarche	4
1.2.	Rappels sur l'activité du CépiDc.....	4
1.3.	Acteurs en jeu	4
1.4.	Organisation de l'équipe.....	6
2.	Formalisation des processus CépiDc	7
2.1.	Accueil.....	7
2.2.	Traitement	8
2.3.	Exploitation	9
2.4.	Pilotage	9
3.	Système d'Information	11
3.1.	Définitions et méthode	11
3.2.	Inventaire applicatif	11
3.2.1.	Inventaire par bloc de processus.....	12
3.2.2.	Inventaire des bases de données	20
3.3.	Plan d'occupation des sols applicatif	21
3.4.	Infrastructures	22
3.5.	Exploitation	23
4.	Irritants	23
4.1.	Disponibilité	23
4.1.1.	Incapacité à accéder en simultané aux ressources.....	23
4.1.2.	Avenir des postes fixes du Kremlin Bicêtre	24
4.1.3.	Suppression de droits.....	24
4.1.4.	Synchronisation annuelle INSEE	24
4.2.	Infrastructure.....	24
4.2.1.	Manque de serveurs de calcul pour soutenir les processus de production. 25	
4.2.2.	Manque d'intégration dans le SI des processus d'IA.....	25
4.2.3.	Manque d'environnement de tests / préproduction.	25
4.2.4.	Besoin d'homologation du SI CépiDc.....	25
4.2.5.	Sécurité à renforcer	25

4.3.	Gestion des droits	25
4.3.1.	Différence de groupes Active Directory CépiDc et INSERM	25
4.3.2.	Découplage entre les AD régionales et nationales	25
4.3.3.	Mauvaises applications de la gestion des droits	25
4.4.	Résolution des incidents	26
4.4.1.	Gestion du ticketing	26
4.4.2.	Difficulté d'investigation	26
4.4.3.	Documentation obsolète	26
4.4.4.	Amélioration continue	26
4.5.	Supervision.....	26
4.5.1.	Pas de levée d'alertes via une supervision des flux et traitements de production.....	26
4.5.2.	Lisibilité sur les processus d'Accueil et Traitement.....	26
4.6.	Organisation	26
4.6.1.	Sous-effectif DSI	27
4.6.2.	Rôles et responsabilités TMI, TMA	27
4.6.3.	Gestion des sources / des versions	27
4.6.4.	Prolifération technologique et cadre de cohérence technique	27

1. Introduction

1.1. Contexte du document et démarche

Ce document s'inscrit dans la première phase de l'étude précédant la refonte du système d'information du CépiDc. Cette première phase vise à comprendre comment est réalisé le processus d'accueil de données et de production de la base statistique sur les causes de décès, les systèmes d'information mis en œuvre et les irritants rencontrés.

Pour rappel, cette intervention fait suite à un contexte de mise en place de moyens dédiés pour rattraper le stock des données à traiter et améliorer le processus de production du CépiDc.

Ce document se fonde sur une démarche d'entretiens menés avec différents acteurs du CépiDc entre mars et mai 2024. Les acteurs suivants ont été sollicités.

#	Fonction	Contact	Date
1	Directrice CépiDc	Elise Coudin	21-mars
2	Pôle Production + Responsable d'automatisation	Diane Martin Aude Robert	28-mars
3	Pôle Diffusion	Cecilia Rivera	13-mai
4	DSI INSERM	Sammy Sahnoune	23-avr
5	Délégué DSI CépiDc	Julio Martins	11-avr
6	RSSI	Vincent Archer	25-mars
7	DSI SPI (TMA) + Atos	Shaka Pouth Wang Olivier Duparc	10-avr

1.2. Rappels sur l'activité du CépiDc

La mission principale du CépiDc est la **production de la base statistique des causes de décès en France**. Cette base permet ensuite de :

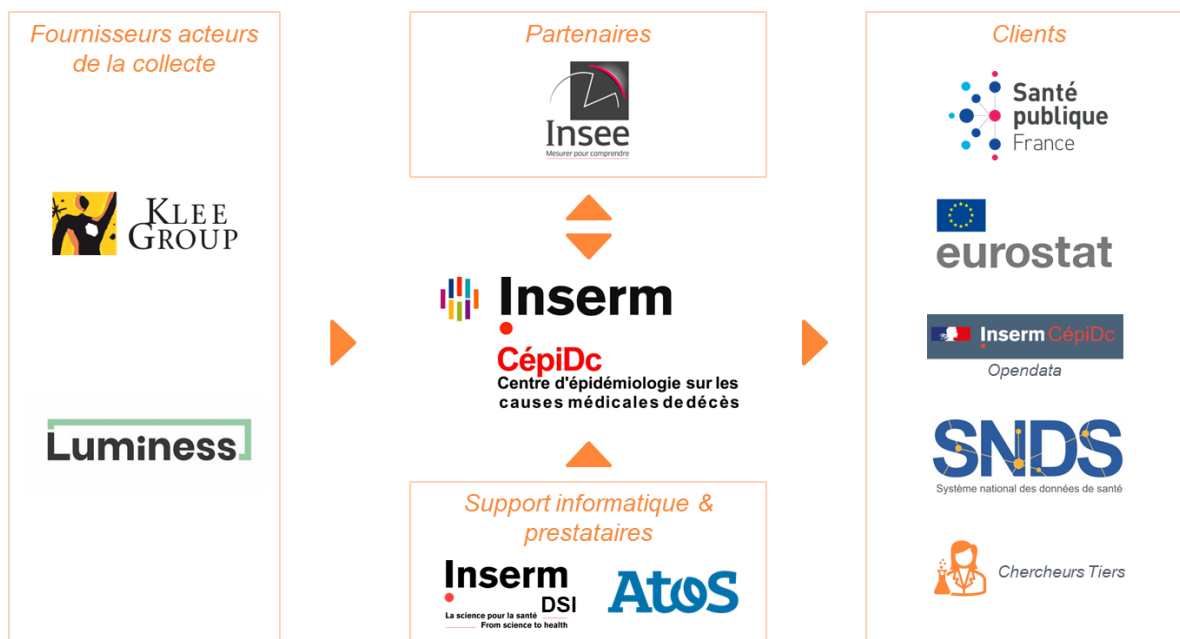
- alimenter la veille sanitaire
- produire les statistiques fournies annuellement à l'OMS et Eurostat
- fournir la base de données individuelles qui est mise à disposition dans le système national des données de santé (SNDS).

Le CépiDc doit pour ce faire accueillir les volets médicaux des environ 650 000 décès par an, et transformer les textes des certificats en codes de la norme CIM10, supprimer les doublons, compléter les trous de collecte (synchroniser avec les décès enregistrés à l'Insee) et construire les différentes variables de la base statistique.

Le processus de production de la base statistique est outillé par de nombreux applicatifs et traitements (décrits ci-dessous). Parmi eux : l'application européenne IRIS/Muse permettant un codage des causes de 63% des certificats en automatique, des algorithmes de *deep learning* développés adhoc pour environ 25 % des codages, pour ne laisser que 14% de codage manuel pour les cas les plus difficiles, requérant une attention particulière, ou permettant de réentraîner et surveiller les algorithmes (~100 000 par an). S'y ajoutent divers traitements et applications permettant de contrôler/corriger, synchroniser, compléter les données et les diffuser.

1.3. Acteurs en jeu

Le CépiDc s'appuie sur plusieurs partenaires pour réaliser ces missions. Le schéma suivant permet d'identifier les acteurs majeurs avec lesquelles le CépiDc est en contact pour réaliser ses missions.



Fournisseurs acteurs de la collecte :

- **KLEE Group** : est le prestataire de certification électronique des décès. Il transmet au CépiDc les volets médicaux des certificats de décès « électroniques » (aujourd’hui environ 45% des cas). La saisie par le médecin certificateur et le transfert d’un flux vers le CépiDc se fait via la plateforme CertDc. La DGS est MOA de cette plateforme avec appui du CépiDc sur la partie volet médical. Cette plateforme n’est pas accessible aux membres du CépiDc (sauf en recette dans le cadre de la mission de MOA du CépiDc pour le volet médical).

L’application CertDc ne fait pas partie du périmètre de « système d’information » du CépiDc tel qu’il est décrit plus bas.

LUMINESS : (anciennement Jouve) est le prestataire de saisie et de numérisation pour les certificats de décès papier (aujourd’hui 55% de cas). De même, elle met à disposition les données des certificats de décès au travers d’un flux informatisé. Ces deux acteurs peuvent être considérés comme les points d’entrée principal des flux du système d’information, et masquent une multitude de « Fournisseurs » que constituent les médecins certificateurs, les ARS...

NB : La DSI de l’Inserm est maîtrise d’œuvre de l’application CertDc. La maîtrise d’ouvrage est assurée par la DGS avec l’appui du CépiDc. Pour les volets médicaux des certificats papiers, le CépiDc est maîtrise d’œuvre de la numérisation (avec appui sur Luminess).

Support informatique & prestataires :

- **Atos** : Responsable de la Tierce Maintenance Applicative des systèmes d’information du CépiDc.
- **Le DSI de l’Inserm** : fourniture de la TMI, relais avec la TMA, migration, réseaux, bureautique...

Partenaires :

- **L’INSEE** n’est pas à proprement parler un fournisseur, mais un partenaire dans la mesure où les données d’état civil (côté INSEE) sont réconciliées avec les données du CépiDc pour s’assurer d’une cohérence mutuelle. Ce processus dit de « synchronisation » est décrit plus bas. Il est générateur de flux d’entrée et de sortie.

Clients :

Le terme client est à comprendre comme « consommateur » des données de l’INSERM. Les principaux sont :

- **Santé publique France**, qui assure une veille sanitaire. Peuvent s’y ajouter d’autres acteurs de la veille comme l’ARS IdF.

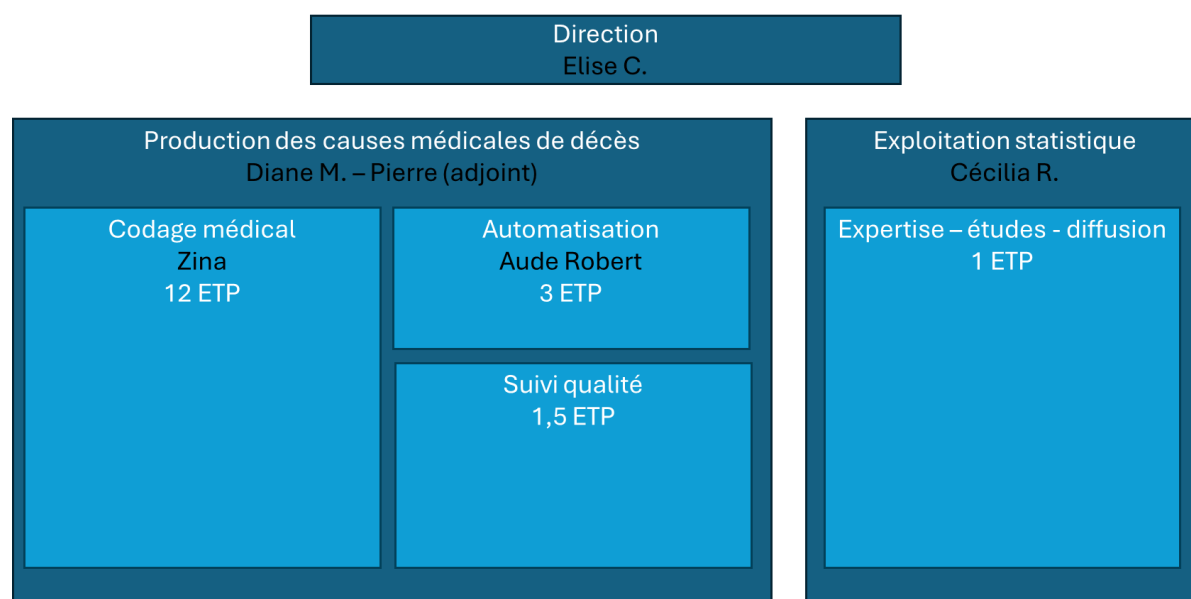
- **Eurostat**, organisme communautaire de statistique, qui consolide les données de décès à l'échelle européenne à des fins d'études et benchmark.
- **Le site OPENDATA de l'INSERM CépiDc**, c'est-à-dire tous citoyens, décideurs ou autre souhaitant consulter des agrégats statistiques non réidentifiants.
- **Le SNDS**, Système National des Données de Santé, utilisé pour centraliser les données individuelles sources à des fins de recherche et de travaux statistique de santé publique. Elles couvrent les causes de décès, les données de parcours de soins hospitalier – PMSI et de médecine de ville - SNIRAM, des données de handicap (MDPH, CNSA), ainsi que des données liées à la Covid-19. Les causes de décès (données individuelles) sont une donnée « source principale » du SNDS.

1.4. Organisation de l'équipe

Le CépiDc est constitué de 2 pôles majeurs :

- Le pôle Production, qui prend en charge la chaîne de valeur visant à créer une base de données finale et complète sur les décès en France.
- Le pôle Diffusion, qui prend en charge la diffusion des données aux différents acteurs de l'écosystème, ainsi que leur valorisation par des études statistiques réalisées sur les bases.

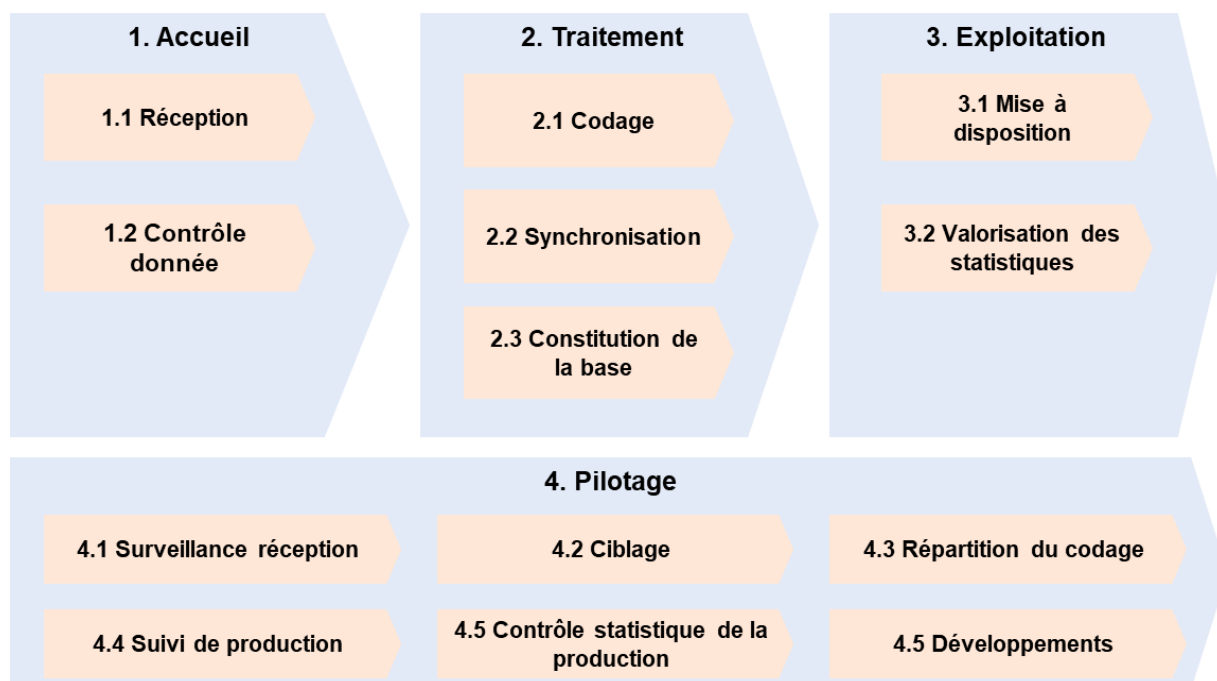
Le schéma informatif ci-dessous présente les effectifs des pôles :



2. Formalisation des processus CépiDc

La démarche a permis de proposer un découpage en 4 processus fonctionnels et 11 sous-processus qui soutiennent l'activité du CépiDc.

L'objectif de ce découpage est de clarifier le vocabulaire commun, ainsi qu'offrir un prisme d'analyse pour le système d'information.



Les processus sont les suivants :

- **L'Accueil** des données, qui a pour but d'accueillir les données des certificats, de manière exhaustive et juste, ce qui suppose certaines activités de contrôle et échange avec les prestataires situés en amont du processus et une disponibilité du système d'accueil.
- **Le Traitement**, qui comprend notamment la partie de codage de la chaîne morbide et la cause initiale, et a pour but de fournir une base de données finale aux standards internationaux, exploitable par les acteurs compétents.
- **L'Exploitation**, qui a pour but de mettre à disposition toute ou partie des données à des acteurs externes au CépiDc, en assurant des opérations d'agrégation, filtrage ou anonymisation, ainsi que la valorisation de la base pour produire des connaissances prenant la forme d'études ou publications.
- **Le Pilotage**, qui regroupe les activités support des processus décrits plus haut : l'objectif est de donner de la visibilité et d'améliorer le fonctionnement de la chaîne de valeur de manière réactive, ainsi que piloter la qualité des processus dans une démarche d'amélioration continue. On y ajoute aussi les missions de développements, tests, bacs à sable, recherche, indispensables à l'évolution et amélioration continue de la chaîne de production.

Sont décrits ci-dessous les sous-processus pour chaque processus.

2.1. Accueil

La réception consiste à l'accueil des informations des volets médicaux des certificats de décès. L'objectif est d'acquérir une description informatique de tous les volets médicaux des certificats de décès, qu'ils soient papier ou électroniques, quel que soit le modèle du formulaire utilisé (2017, 1997), avec des informations spécifiques ou non

selon les cas (ex. selon l'état de grossesse). Le livrable de ce processus est une base de données contenant la base des informations relativement brute, avant codage et avant synchronisation avec l'INSEE (bases différentes selon papier/elec). Notons qu'à ce stade, la base peut encore avoir des informations erronées, avant de procéder aux traitements, une activité de contrôle actif des données est nécessaire. Sont également accueillies les images des documents papier saisis par le prestataire (volets médicaux / B7).

Le **contrôle des données**, qui consiste à faire des vérifications sur les données numérisées issues des certificats de décès, comme des contrôles de complétude des champs et de cohérence, sur les champs individuellement et entre les différents champs (ex. « Etat de grossesse » = Oui alors que « Sexe » = masculin).

Si des erreurs massives de saisie par le prestataire sont observées, un retour lui est fait pour mise en place de correctif sur les saisies suivantes. Les corrections d'erreurs de saisie ou de cohérence de données sont faites directement par le CépiDc. En certification électronique, un contrôle des données est fait par des règles de gestion en cours de saisie du certificat et des messages remontent au certificateur. Des corrections a posteriori peuvent aussi être faite par le CépiDc.

2.2. Traitement

Le codage des causes médicales de décès, qui consiste à un double objectif pour chaque décès :

- Transformer chaque cause de décès (ou entité nosologique) intervenant dans l'enchaînement morbide en code(s) de la classification internationale des maladies actuellement dans sa version CIM10.
- Identifier la cause initiale de décès et lui attribuer un code CIM10, en appliquant les règles de décision internationales qui prennent en compte l'enchaînement des causes entre elles.

Ce processus est en partie automatisé via des traitements batchs du système expert IRIS/Muse, ou par un recours à des prédictions de réseaux de neurones profonds (supposant un entraînement préalable). Il est aussi en partie réalisé par codage assisté interactif de l'équipe de codage (composée de « codeurs » et nosologistes). Le livrable du processus de codage est la base dite « codée ».

Remarques sur la partie codage via IRIS : Historiquement, IRIS traite des sous-tables regroupant les certificats par ordre d'arrivée (le département n'est pas pris en compte dans la constitution des lots, mais de fait, les certificats papiers sont envoyés par les départements donc l'ordre d'arrivée correspond à des « paquets » de certificats par département). La stratégie par échantillon n'est plus compatible avec le codage directement dans les lots qui arrivent au fil d'arrivée des certificats en base de données. De nouveaux lots sont créés manuellement par l'équipe automatisation : Il y a toujours la notion de lot pour permettre le codage (auto mais aussi l'organisation du codage manuel). Mais avec l'introduction de l'IA, le codage manuel ne concerne plus tous les certificats non codés automatiquement. Ce codage manuel est ciblé et donc pour organiser le codage manuel, il a fallu définir différents types de certificats à coder manuellement (ciblages « d'échantillons ») qu'on répartit ensuite en lots pour l'organisation du codage et la répartition des lots entre codeurs.

La synchronisation, qui est un processus d'appariement entre la base de données de mortalité de l'état civil de l'INSEE et les décès reçus au CépiDc, afin de s'assurer de la cohérence des données. Pour rappel, le CépiDc est garant du codage de la cause de décès (celle-ci n'est pas reçue par l'INSEE) et l'Insee est garant du statut vital de chaque individu (le CépiDc ne reçoit pas l'identité de la personne décédée).

Ce processus permet, pour le CépiDc, de supprimer les doublons, de compléter et de corriger les données de l'état civil liés aux certificats (sans aller jusqu'à l'identité de la personne décédée), et d'assurer l'exhaustivité des décès. Pour l'INSEE, la réception des volets administratifs directement depuis la plateforme de certification électronique permet de déclencher une présomption de décès en cas de non-déclaration du décès au répertoire des personnes dont il a la gestion.

L'appariement entre les deux bases est essentiellement réalisé à l'INSEE, de manière mensuelle : les décès CépiDc sont envoyés et l'INSEE renvoie les décès CépiDc appariés.

En revanche, la gestion des doublons, des volets médicaux complémentaires et des décès manquants côté CépiDc se fait au CépiDc lors d'une synchronisation annuelle qui permet de finaliser la base avec les données civiles d'une année de décès consolidées par l'Insee.

On distingue donc bien une synchronisation « mensuelle » en tant que temporalité de réception et une synchronisation annuelle cette fois sur une année de décès en particulier.

La constitution de la base, qui consiste en tous les traitements de création des variables et des différentes tables de données constituant la base de données de décès codées et synchronisées avec l'INSEE. Ce processus a pour objectif de consolider la base SRef (base de production, c'est-à-dire la base « vivante » en cours de production) et mettre à jour la base SrefFull, qui est la base de diffusion, c'est-à-dire figée une fois la production terminée. Elle sert d'entrant principal aux travaux d'Exploitation. NB : Ces bases sont individuelles, c'est-à-dire qu'elles contiennent les données à la maille la plus fine (individus) et ne contiennent pas d'agrégats.

2.3. Exploitation

La mise à disposition, aussi appelée Diffusion, qui consiste en 3 objectifs principaux :

- Mettre à disposition sur une base récurrente les données de décès aux consommateurs classiques (Eurostat, OMS, SNDS). Ces mises à disposition se font via des flux automatisés ou des actions manuelles.
- Gérer l'OpenData, c'est-à-dire les statistiques mises à disposition du grand public via des bases ne contenant pas de données de détail, au travers de sites web.
- Répondre aux demandes de données ponctuelles d'acteurs externes, c'est-à-dire cadrer le besoin, préparer les extractions, et mettre à disposition des bénéficiaires les données via des media sécurisés.

La valorisation des statistiques, qui consiste à exploiter la base de données des décès à des fins d'étude et de publication. Parmi les grandes publications, celle de la mise à disposition de la nouvelle édition de la base est prépondérante, avec une attention à expliquer les changements par rapport à la production de l'année précédente et souligner les tendances. Cette activité couvre également les collaborations avec des chercheurs externes, afin d'apporter un éclairage métier sur les données et leur processus de création.

2.4. Pilotage

La surveillance de la réception, qui consiste en l'ensemble des activités pour atteindre l'exhaustivité de la collecte.

Cette activité de surveillance est réalisée via l'interface mise à disposition par Luminess pour suivre la transmission effective des colis transmis par les ARS ainsi que certains des développements R et SQL effectués par le métier, permettant des calculs d'exhaustivité (ex. rapports de volets médicaux reçus par rapport au nbre de décès déclarés à l'Insee) et piloter des relances auprès des départements.

Les problèmes rencontrés au cours de la transmission du papier au fil de l'eau via le portail Luminess (contrôler le fait qu'il y ait bien 1 envoi par mois qui ait fonctionné sans anomalies) font l'objet de communications avec les référents en ARS.

Le ciblage, qui consiste à sélectionner quels certificats de décès doivent faire l'objet d'un codage humain par les codeurs et nosologues, afin de respecter plusieurs objectifs :

- Assurer la qualité et la cohérence du codage de la base finale par comparaison avec une campagne traditionnelle sans intervention du codage par algorithme de *deep learning*, et ce pour chaque catégorie de la shortlist d'Eurostat. Les résultats sont entre autres priorisés par un indicateur de confiance (les certificats obtenant un indice faible sont envoyés à l'équipe de codage).
- Permettre d'enrichir une base d'apprentissage des algorithmes de *deep learning* en mélangeant échantillon aléatoire et reprises ciblées.
- S'assurer que les décès sensibles ou les certificats à vérifier (faisant l'objet d'incohérences manifestes ou dans le cadre d'anomalies connues du batch de codage automatique) sont bien revus par des humains.

Le ciblage des certificats à reprendre en lien avec l'intelligence artificielle intervient en amont de la répartition et est réalisé par l'équipe automatisation du pôle production, via des algorithmes de *deep learning*. Les autres types de ciblage sont faits par tirage aléatoire et des programmes d'identification des décès sensibles ou des certificats à vérifier.

NB : Parmi les traitements liés à l'activité de codage, certains ont pour objectif de faire basculer en codage manuel les certificats dont le codage risque d'être incohérent.

La répartition du codage humain, est une étape nécessaire de gestion entre les membres de l'équipe de codage pour qu'ils se répartissent les lots de données à coder. Celle-ci est constituée d'un premier niveau : codeur, puis d'un second niveau : nosologiste et enfin le niveau le plus expérimenté : expert. Il y a attribution du lot à un codeur nosologiste ou expert et une fois qu'ils ont codé leur lot de données, il y a restitution, ce qui permet de mettre à jour le suivi de la répartition du codage. Les certificats non codés lors de la restitution par les codeurs ou nosologistes le sont par le niveau supérieur.

Le suivi de production, qui consiste à surveiller les statistiques de réception et de codage des certificats, selon différents axes (provenance, temporalité) afin de s'assurer que le codage sera réalisé dans les temps et qu'aucun incident ne nuit à l'accueil ou au traitement des données (anomalie dans les flux d'intégration des certificats, anomalie de transmission des données à l'INSEE ou lors du retour des appariements INSEE, ...)

Le contrôle statistique de la production, qui consiste à explorer grâce à des traitements statistiques les résultats de la base en cours de codage, afin d'identifier de potentielles erreurs et en améliorer la qualité. Ce processus recourt typiquement à des échantillonnages, des contrôles statistiques de moyennes segmenté par populations, géographies, typologies de décès, etc ...

En cas de détection de biais ou anomalie statistiques, des actions correctives sont prises en fonction de l'erreur et de l'explication (par exemple, des vérifications par des codeurs d'échantillons).

Le développement, recherche consiste à élaborer, tester de nouveaux programmes statistiques ou applicatifs en vue des évolutions futures à mettre en œuvre : changement d'algorithmes de *deep learning*, nouveaux indicateurs de suivi statistique, nouvelle classification etc...

3. Système d'Information

3.1. Définitions et méthode

Cette partie permet d'avoir une vision d'ensemble du système d'information du CépiDc qui sert à opérer les besoins internes.

Les entretiens ont permis de décrire un certain nombre d'objets du Système d'Information, souvent dénommés « applications ». Nous proposons de distinguer les objets selon la typologie suivante pour faciliter la compréhension :

- **Applications** : qui consistent en un logiciel, hébergé sur un environnement (VM, conteneur ou autre), doté d'une interface graphique, de fonctionnalités propres et d'un stockage de données persistant.
- **Traitements applicatifs** : traitement appliqué à un lot de données, exécuté par un programme informatique. Ce programme peut être un script, ou une tâche exécutée par une application. Ce programme peut être planifié, déclenché automatiquement sur événement, ou déclenché manuellement par un opérateur. NB : Les flux d'intégration et les traitements statistiques sont à part. Certains peuvent être industrialisés et délégués à une TMA, mais d'autres non.
- **Traitement statistique** : Traitement algorithmique écrit dans un langage de programmation dédié aux statistiques et à l'IA (R, Python, ...) caractérisé par une forte évolutivité. Ils sont indissociables d'actions humaines, tantôt de lancement, de vérification, de correction et/ou exploration. Ceci limite l'intérêt d'une industrialisation, et rend ces traitements incompatibles avec une délégation à une TMA.
- **Flux d'intégration** : Traitement qui permet de déplacer des données d'un système à un autre (y compris base de données et serveurs), en interne CépiDc. Les données intégrées peuvent (ou non) écraser les anciennes.
- **Flux externes** : Traitement qui permet de déplacer des données d'un système à un autre, impliquant au moins 1 système externe au CépiDc.
- **Bases de données** : Instanciation d'un logiciel de base de données, indépendante d'une application (ne sont pas comptées comme bases de données les bases « internes » des applications)
- **Environnements** : espace informatique mettant à disposition des ressources de calcul et de stockage accessibles via un système d'exploitation (OS). Dans ce document, un serveur est considéré comme un environnement.
- **Hébergement** : Site qui met à disposition des infrastructures informatiques, c'est-à-dire les machines physiques qui permettent de mettre à disposition des environnements.

Par la suite, nous exposerons des inventaires ainsi que des projections cartographiques, apportant chacun un point de vue sur le système d'information. Parmi eux :

- Un inventaire applicatif, permettant de comprendre succinctement les raisons d'être des applications.
- Une cartographie dite « Infrastructure » des environnements et hébergements utilisés, pour comprendre la multiplicité des environnements.
- Un plan d'occupation des sols (POS) applicatif sur les différents processus, pour comprendre quels applicatifs pourraient être mutualisés.

3.2. Inventaire applicatif

L'inventaire suivant est essentiellement tiré des interactions avec le pôle production, qui a une maîtrise fonctionnelle aboutie de l'essentiel de ces applications. Il est présenté par grand processus. Les bases de données sont abordées dans un second temps.

3.2.1. Inventaire par bloc de processus

3.2.1.1. Accueil

Type	Processus associé	Application	Rôle
Traitement applicatif (TMA)	1.1 Réception	TraitBPO	Transfert – décryptage des fichiers de données livrées par le prestataire de saisie (données issues des certificats de décès papiers, leurs scans, mais aussi les données issues des certificats électroniques envoyés au prestataire pour correction de saisie par le CépiDc). L'OCRisation est réalisée en amont chez le prestataire. TraitBPO sert à déchiffrer les données puis à les déplacer du serveur FTP vers le serveur DTA.
Traitement applicatif (TMA)	1.1 Réception	Quarantaine	Batch de traitement permettant l'exclusion des fichiers reçus incorrects (certificats électroniques ou données issues de l'INSEE)
Flux d'intégration (TMA)	1.1 Réception	C1	Intégration des données électroniques issues de CertDc dans SreF
Flux d'intégration (TMA)	1.1 Réception	Job SSIS – ImportationAR	Tâche exécutée automatiquement pour prendre en charge les fichiers PDF hebdo des scans des bordereaux d'envoi des certificats de décès papier après leur traitement par TraitBPO. Ces PDF sont générés par Luminess et archivés dans un répertoire dédié.
Traitement applicatif (TMA)	1.1 Réception	Job SSIS – NettoyageFichierCauses	Tâche exécutée automatiquement pour nettoyer les données des causes des certificats de décès papier livrées par Luminess (fichiers txt). Les fichiers ainsi nettoyés sont dans un répertoire dédié.
Flux d'intégration (TMA)	1.1 Réception	Job SSIS – ImportationCause	Tâche exécutée automatiquement pour intégrer les données des causes des certificats de décès nettoyées par le SSIS - nettoyageFichierCauses (fichiers txt) dans la base CorrectionNumerisation
Flux d'intégration (TMA)	1.1 Réception	Job SSIS – RepriseCausesElec	Tâche exécutée automatiquement pour intégrer les données des causes des certificats de décès électroniques corrigées et livrées par Luminess (fichiers txt) dans la base CorrectionNumerisation
Flux d'intégration (TMA)	1.1 Réception	Job SSIS – ImportationFichierTxt « IN2T »	Intégration des fichiers mensuels des données brutes des B7bis fournis par l'Insee dans la base SreF (<i>DonnéesB7bis</i>)
Traitement applicatif (TMA)	1.1 Réception & 1.2 Contrôle donnée	Job SSIS – ImportationIdentNMC	Tâche exécutée automatiquement pour intégrer les données structurées des certificats de décès papier livrées par Luminess (fichiers txt) dans la base CorrectionNumerisation et leur affecter un

			niveau de contrôle selon les incohérences observées ou non.
Traitement applicatif (TMA + Métier CépiDc)	1.1 Réception & 1.2 Contrôle donnée	VMCIMLParis	<ul style="list-style-type: none"> Depuis 2018 des volets médicaux complémentaires sont générés par les médecins ayant conduit des recherches complémentaires pour connaître la cause de décès (notamment en cas d'obstacle médico-légal) Le décret indique qu'ils devraient arriver via CertDc mais l'IML de Paris les transmet actuellement par mail <p>Il a donc fallu développer un nouveau programme d'intégration vers SREF avec un csv en entrée. Le programme identifie les incohérences et permet à un agent du CépiDc de nettoyer les données puis il y a génération de la requête d'intégration à exécuter manuellement.</p>
Traitement applicatif (TMA)	1.1 Réception & 1.2 Contrôle donnée	Purge	Batch de traitement permettant le nettoyage des vieux fichiers dans les répertoires des serveurs du CépiDc
Application (TMA)	1.2 Contrôle donnée	CorrectionNumérisation	Logiciel de correction/Vérif données papiers Contrôle qualité sur les données structurées via une interface métier, le certificat, s'il fait l'objet de contrôle, est bloqué tant qu'il n'est pas validé.
Application (TMA)	1.2 Contrôle donnée	FinalisationDc	Logiciel de visualisation des certificats Les certificats sont chargés via un fichier csv à la demande mais cette application est également connectée à SREF et IRIS pour la récupération des images
Traitement applicatif (Métier CépiDc)	1.2 Contrôle donnée & 2.3 Constitution de la base	Anonymisation	Traitement qui permet de s'assurer qu'aucune adresse, nom, prénom, numéro de téléphone ou identifiant CertDc ou numéro de sécurité sociale n'est présent en base de données. Tourne à la demande et essentiellement en période de validation de la base de codage. Il remplace le texte identifié par une balise <NOM> ou <ADRESSE> ou <NUM> Langages : Python + SQL

3.2.1.2. Traitement

Type	Processus associé	Application	Rôle
Application (métier CépiDc)	2.1 Codage	IRIS	Logiciel de Codage => Codage à la norme internationale (OMS) Il s'agit d'un logiciel développé par un consortium international (Iris Core Group)

			<ul style="list-style-type: none"> • Il présente quelques limitations en termes de sécurité (refus des US de l'utiliser en l'état) • Ce logiciel est assez contraignant et très normé mais reste la référence pour le codage pour beaucoup de pays dans le monde et utilisé depuis 2011. • Les codeurs ont accès aux images et aux données via l'interface pour réaliser le codage semi-assisté • Permet le codage automatique par batch • Fonctionne par lot : chaque certificat doit être dans une table ident (données état civil + code de cause initiale + statut du certificat (codé ou non) ; ainsi que dans une table medcod (texte de causes + codage des causes associées) <p>A grosse maille</p> <ul style="list-style-type: none"> • 63% de codage en auto sur les certificats papier • 45% uniquement pour les certificats électroniques jusqu'en 2017. <ul style="list-style-type: none"> ○ Mais depuis 2018 dépôt des certificats électroniques non codés automatiquement sur le serveur du prestataire de saisie des certificats papiers pour qu'il corrige les données saisies et les renvoie au CépiDc <p>Ce qui permet au final d'atteindre 63% au global</p>
Traitement applicatif (Métier CépiDc)	2.1 Codage	IRIS Batch	Pour le codage : Batch automatique lancé en ligne de commande à la demande sur les certificats avec une statut Initial ou nécessitant un rebatch suite à des décisions de codage.
Traitement applicatif (TMA)	2.1 Codage	MAJ Priorisation Lots	MAJ Priorisation LOTS permet à partir d'une liste d'échantillon donnée pour une année de décès donnée, de ré-injecter le codage réalisé dans les lots créés pour convenir à la stratégie par échantillon, dans les petits lots de production créés au fil d'arrivée des certificats.
Flux d'intégration (TMA)	2.1 Codage	IDPI	Intégration données papier dans Iris / Envoi des données à l'Insee Il s'agit d'un batch d'intégration de la base CorrectionNumerisation d'accueil +/- corrections des données certificats papier

			vers la base de codage (IRIS), mais qui produit également un fichier à destination de l'INSEE pour leur envoyer les données indirectement identifiantes des certificats papiers reçus au fil de l'eau en vue de la synchronisation.
Flux d'intégration (TMA)	2.1 Codage	IDEI	Intégration données élec de SreF dans Iris
Flux d'intégration (TMA)	2.1 Codage	C1C	<p>Mise à jour des données Iris avec les données Insee post-synchro mensuelle</p> <ul style="list-style-type: none"> • Il s'agit d'un traitement interne non activé • Il permettrait de mettre à jour les données IRIS avec les données corrigées de SreF par le retour de l'INSEE • Mais il n'est pas encore activé car tous les appariements ne sont pas validés (certains sont trop permissifs et souvent erronés) • Des évolutions sont nécessaires avant son activation (à minima la suppression de 2 traitements d'appariement trop permissifs côté INSEE + la compatibilité à revoir avec la gestion des lots) <p>A l'avenir il faut viser à avoir des données propres le plus tôt possible et avoir un fonctionnement simple de correction de données au fil de l'eau sans avoir à se soucier des impacts en « dominos » sur le reste du système : refonte de fond à envisager dans le nouveau SI</p>
Flux d'intégration (TMA)	2.1 Codage	MajRepriseIrisElec	Mise à jour des nouvelles causes de décès reprises par le prestataire de saisie pour les certificats Elec dans Iris, mise à jour également du Statut Iris pour la prise en compte du batch automatique.
Traitement applicatif (Métier CépiDc)	2.1 Codage	Vérifications	À partir d'une liste de certificats (ou échantillons) permet d'extraire les certificats nécessaires à une revue codage. Langage : SQL
Traitement applicatif (Métier CépiDc)	2.1 Codage	Traitement des illisibles	Ce traitement réalisé après codage du niveau nosologiste, permet de modifier le « ! » présent sur un certificat papier et qui traduit une problématique de lecture (du nosologiste, du codeur et de l'agent de saisie chez le prestataire). Cette méthode vise à coder le certificat en « ignorant » le « ! » ou en l'isolant pour garantir un codage de qualité. Langages : Python + SQL
Traitement statistique (Métier CépiDc)	2.1 Codage	Entraîner les modèles IA	Utilisation d'algorithmes de deep learning pour entraîner et prédire des données. Actuellement réalisé « en dehors » du système d'information du CépiDc à partir de

			tables extraites à partir du logiciel R et Python NB : traitements sur les machines dédiées à l'IA.
Traitement statistique (Métier CépiDc)	2.1 Codage	Prédire par des méthodes IA	Utilisation d'algorithmes de deep learning pour entraîner et prédire des données. Actuellement réalisé « en dehors » du système d'information du CépiDc à partir de tables extraites à partir du logiciel R et Python NB : traitements sur les machines dédiées à l'IA.
Traitement statistique (Métier CépiDc)	2.1 Codage	Prédiction de la CI par combinaison système de règle et IA	Passage Iris sur les causes prédites par IA, puis choix par IA entre différentes propositions de causes initiales Utilisation d'algo IA pour entraînement et inférence. Actuellement réalisé « en dehors » du système d'information du CépiDc à partir de tables extraites à partir du logiciel R et Python NB : traitements sur les machines dédiées à l'IA.
Application (TMA)	2.2 Synchronisation	ARBI	Logiciel d'arbitrage de la synchro annuelle Il est utilisé lorsqu'il y a des doublons entre les certificats et donc plusieurs possibilités d'appariement entre les décès INSEE et les certificats reçus au CépiDc. L'interface permet d'avoir toutes les images des certificats, les infos INSERM et INSEE afin de choisir l'appariement à retenir. L'interface permet aussi de modifier les données SREF en cas d'erreur de saisie. La base de donnée utilisée est Sref (Tables spécifiques Arbi).
Traitement applicatif et intégration (TMA)	2.2 Synchronisation	Synchro BRPP	Batch avec plusieurs options possibles : La synchronisation annuelle Insee est réalisée en trois étapes. La première consiste à récupérer les fichiers d'entrée fournis par l'INSEE (option pour vérifier la conformité des fichiers XML) et préparer ces derniers pour les intégrer dans une table SynchroINSEE. La deuxième étape réalise une analyse entre les fichiers INSEE et la base SREF. Cette étape peut être lancée autant de fois que l'on souhaite, elle produit juste des rapports et permet de constituer les fichiers d'arbitrage. Enfin, la troisième étape est la phase de synchronisation de la base SREF. Elle opère des mises à jour sur les enregistrements et n'est donc lancée qu'une seule fois. Langage : Java
Traitement applicatif (Métier CépiDc)	2.2 Synchronisation	Synchro (partie interne CépiDc)	Ensemble de programmes pour synchro annuelle entre les données Insee et les données CépiDc : Vérification des fichiers INSEE et modifications

			<p>Génération de groupes de certificats pour arbitrage puis mis à jour des tables nécessaires à l'application ARBI</p> <p>Automatisation de cas d'arbitrage</p> <p>Appariement « désespoir »</p> <p>NB : traitements en local non présents sur le serveur APP.</p> <p>Langages : Python + SQL et 1 macro VBA</p>
Flux externe (TMA)	2.2 Synchronisation	IN2	<p>Réception mensuelle des données Insee post-appariement. (cf. description ci-dessous)</p> <p>L'IN2 mensuel renvoie chaque mois le résultat de l'appariement entre les données du CépiDc (électronique ou papier) et les données Insee. L'appariement étant souple, il permet de « corriger » les données du CépiDc si l'Insee n'a pas la même date de naissance par exemple. C'est la donnée Insee qui est retenue.</p> <p>Extraction sur le serveur de l'INSEE des fichiers .txt des décès reçus le mois précédent avec des données indirectement identifiantes issues des avis B7 bis.</p>
Flux externe (TMA)	2.2 Synchronisation	IN1B*	<p>Transfert de fichiers vers l'INSEE des données indirectement identifiantes issues des certificats papiers (pas de nom / prénom ou cause de décès) générés par IDPI afin de permettre la synchronisation avec les données de l'Insee pour le papier.</p> <p>*Pour les certificats électroniques, ces données (avec le nom et le prénom) sont transmises directement depuis CertDc vers l'Insee via la plateforme d'état HubEE.</p>
Flux d'intégration (TMA)	2.3 Constitution de la base	C1B	<p>Intégration des nouvelles données papiers dans SreF</p> <p>Mise à jour journalière des données SreF relatives au codage à partir des bases actives de codage Iris (Elec + papier)</p>
Traitement applicatif (Métier CépiDc)	2.3 Constitution de la base	Intégrer le codage IA des causes initiales	<p>Traitement qui permet d'intégrer dans les bases de données de codage et SREF et SREFFull le résultat de l'IA pour la cause initiale de décès en particulier. (attention, ce n'est pas l'inférence). Lors de cette intégration, on renseigne le champ 'TypeCodage' indiquant que le codage s'est fait par l'IA (modalités indiquant le détail du modèle utilisé)</p> <p>Langages : Python + SQL</p> <p>NB : traitements en local non présents sur le serveur APP.</p>
Traitement applicatif (Métier CépiDc)	2.3 Constitution de la base	Intégration des causes associées	<p>Traitement qui permet de générer un texte propre diffusable pour chaque certificat codé par l'IA. Il assure également la validité</p>

		pour un certificat codé par l'IA	des codes de causes associées et permet son intégration dans une base de données intermédiaire. Langages : Python + SQL
--	--	---	--

3.2.1.3. Exploitation

Type	Processus associé	Application	Rôle
Application (TMA et DSI)	3.1 Mise à disposition	GenerationIdDeces	Logiciel de génération des fichiers pour le SNDS (Identifiant Crypté) <ul style="list-style-type: none"> En entrée : année, destination (CNAM ou INSEE) puis édition de la requête à lancer pour sélectionner les certificats, les données structurées et les causes de décès. S'appuie sur un WS qui génère un identifiant crypté pour chaque certificat dans le fichier de sortie
Application	3.1 Mise à disposition	Logiciel Serveur Opendata	Applicatif installé sur le serveur OpenData pour héberger les données et permettre un requêtage en R.
Traitement statistique (Métier CépiDc)	3.1 Mise à disposition	MAD externe Extraction / Contrôle / Calculs	Traitements de mise à disposition pour la diffusion de données : Extraction de données, contrôle de données, mises au format, calcul statistiques génération de tables pour mise à disposition externes. Destinataires : Eurostat, OMS, SNDS (qui mobilise aussi generation iddeces).
Traitement statistique (Métier CépiDc + DSI)	3.1 Mise à disposition	MAD OpenData	Traitements de mise à disposition des données sur l'OpenData : extractions, mise au format, anonymisation, mise sur serveur dédié, avec calcul statistiques.
Flux externe (TMA)	3.1 Mise à disposition	C2-C3	Génération de fichier et envoi des données du CépiDc vers SpFrance et l'ARS IDF (pour de la veille sanitaire (fichiers XML chiffrés). <ul style="list-style-type: none"> C2 : SREF vers SPFRANCE & ARS IDF. Il s'agit d'un flux de diffusion qui tourne toutes les 5 minutes pour générer des fichiers puis les envoyer avec les données ayant fait l'objet de tout mouvement réalisé sur la base SREF du CépiDc pour des données non finalisées en cours de production Envoi toutes les 5 minutes pour SpFrance 2 fois par jour pour l'ARS. C3 (=Histo C2) : fichier d'historisation qui accompagne le C2 (liste des certificats envoyés sur les dernières 24h avec quelques

			métadonnées comme les dates de naissance et de décès et le type d'opération)
Traitement applicatif (TMA)	3.1 Mise à disposition	C2R	Envoi ponctuel de données massives finalisées Rattrapage de données par période donnée (fichiers générés au même format que le C2 mais avec un transfert manuel via FileSender)
Traitement statistique (Métier CépiDc)	3.2 Valorisation des statistiques	Valorisation ad hoc Calculs de statistique	Exploitation, statistiques, études

3.2.1.4. Pilotage

Type	Processus associé	Application	Rôle
Traitements statistiques (Métier CépiDc + DSI)	4.1 Surveillance réception & 4.4 Suivi de production	Pilotage Indicateurs réguliers de suivi production	Dsi pour le serveur de mise à disposition de ces indicateurs. Self pour le calcul
Traitement applicatif (Métier CépiDc)	4.2 Ciblage	Création d'un nouvel échantillon	À partir d'une liste de certificats, créé de nouveaux lots de codage au format lisible par IRIS et les intègre dans la table ADMINLOTS. Réalisé par l'équipe automatisation, ce traitement permet également d'alimenter le descriptif des échantillons et d'alimenter une table générique du nouvel échantillon. Langages : Python + SQL
Traitement statistique (métier CépiDc)	4.2 Ciblage	Ciblage IA	A partir de simulation d'une précision à atteindre sur la base de test, et des prédictions d'un premier modèle IA, calcul d'un indicateur de confiance, simulation du volume de certificats à inclure dans les échantillons de reprise IA et identification de ces certificats
Traitement applicatif	4.2 Ciblage	Identification des échantillons	Batches permettant de constituer les listes de certificats à envoyer en reprise manuelle (décès sensibles, ciblage IA, aléatoires, EDP, vérifs...)
Application (TMA)	4.3 Répartition du codage	ChoixLotIris	Logiciel de gestion du codage – attribution des lots aux codeurs, nosologistes, experts (différents profils). Les bases de données utilisées sont les bases annuelles (table AdminLot)
Traitement applicatif (TMA)	4.3 Répartition du codage	FusionLotIris	Concaténation des données partagées en petits lots (contrainte du logiciel Iris) en une seule table. En effet, IRIS fonctionne par lots (10 000 certificats maximum par lot)

			<p>FusionLotIris permet simplement d'interroger les données facilement en concaténant l'ensemble des données des lots au lieu d'aller interroger les milliers de lots pour avoir une vue globale de l'année de codage.</p> <p>Elle n'a aucun impact sur la nouvelle organisation, au contraire elle fonctionne toujours sur les lots historiques toujours constitués au fil de l'eau et qui sont justement à mettre à jour avec les travaux fait dans les nouveaux lots avec la nouvelle organisation.</p>
Traitements statistiques (Métier CépiDc)	4.4 Suivi de production	Dama – Data management de routine en cours de campagne	Maintenance des tables de codages, correction de données, intégration manuelle de données, extractions adhoc,..erreurs d'attribution de lots,
Traitements statistiques (Métier CépiDc)	4.5 Contrôle statistique de la production	Pilotage Indicateurs adhoc de suivi	Extractions, calculs d'indicateurs statistiques spécifiques à la demande, contrôles statistiques / monitoring IA permettant le pilotage de la campagne
Traitement statistique (métier CépiDc)	4.6 Développements	Bac à sable /développement modèles	Bac à sable de développement test de modèles IA, modèles statistiques, prise en compte des évolutions de la CIM, du dictionnaire de codage....

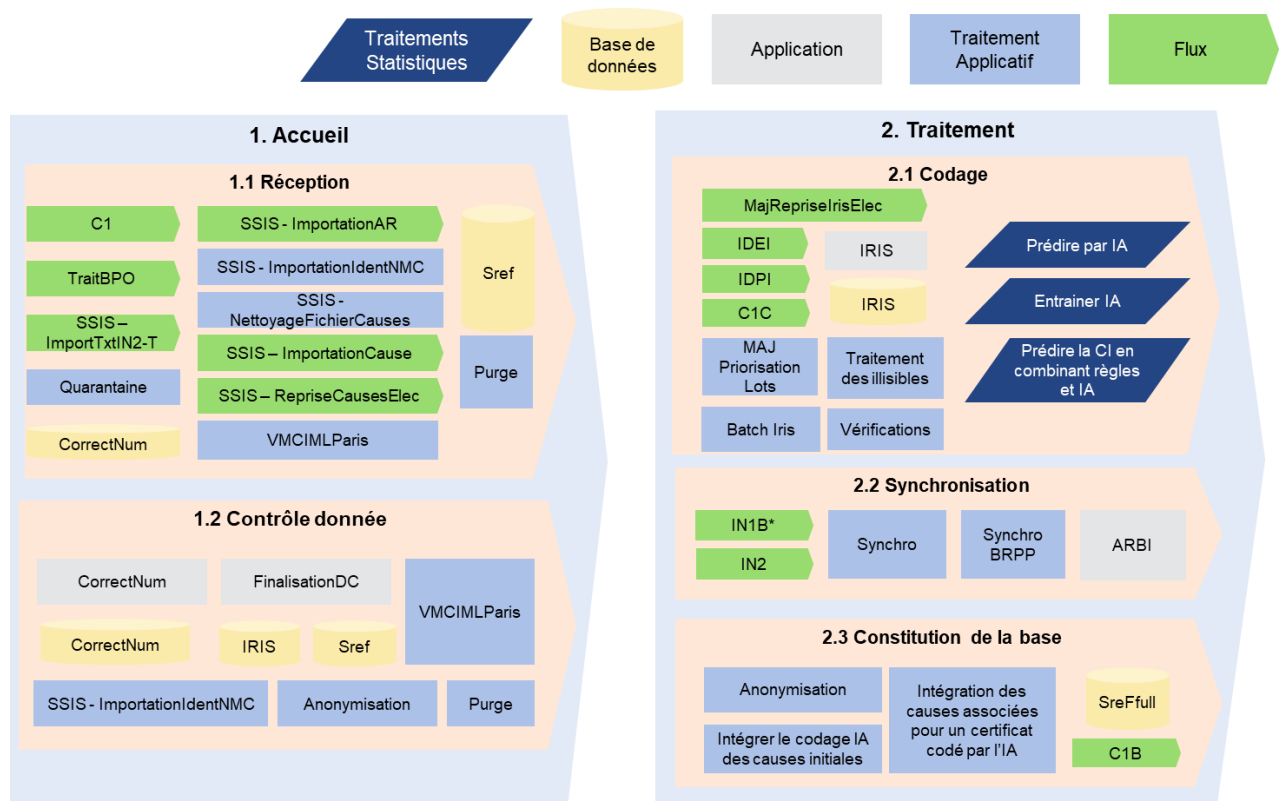
3.2.2. Inventaire des bases de données

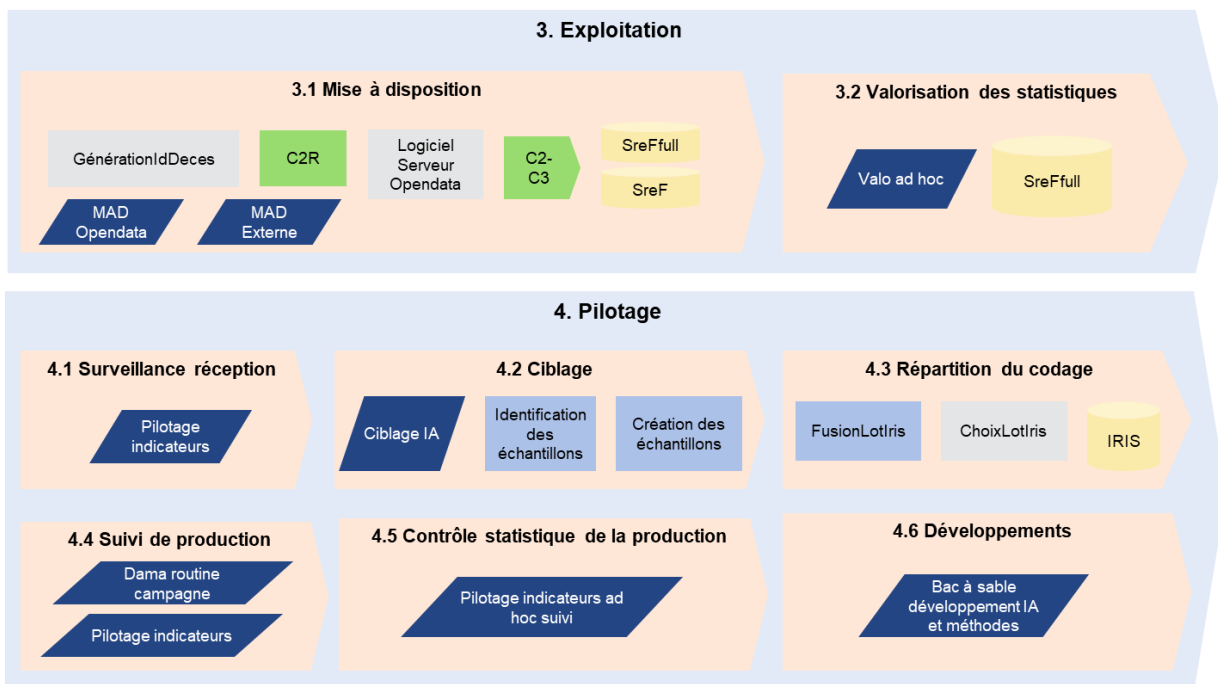
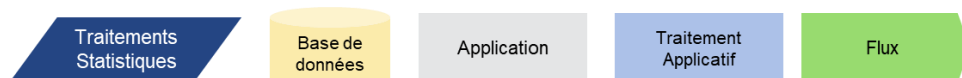
Type	Base de données	Rôle
Base de données (TMA)	BDD CorrectionNumérisation	Base de données liée à CorrectionNumérisation
Base de données (TMA)	BDD IRIS (bases annuelles et bases 'TablesFrAAAA...')	Bases de données liées à IRIS. Elles sont impactées par les batch IDEI, IDPI, FusionLotIris, C1B, C1C. Il s'agit en réalité de plusieurs Bases annualisées pour les données et pour les tables nécessaires au codage (dictionnaire, standardisation)
Base de données (Métier CépiDc)	SreF	Base de référence du CépiDc portant les décès (dont codage) et données Insee brutes et de synchronisation
Base de données (Métier CépiDc)	SreFfull	Base finalisée du CépiDc (copie de SreF), utilisée pour l'exploitation et la diffusion.
Base de données (Métier CépiDc)	TAL	Base servant d'intermédiaire dans le cadre des traitement statistiques métiers

3.3. Plan d'occupation des sols applicatif

A date, le PoS vise à faire apparaître un maximum de traitements pour viser une description exhaustive de l'existant. Ceci ne préjuge pas d'une répartition cible entre les traitements à industrialiser et ceux à garder complètement à la main des métiers (que l'on pourrait appeler « Self-Service »).

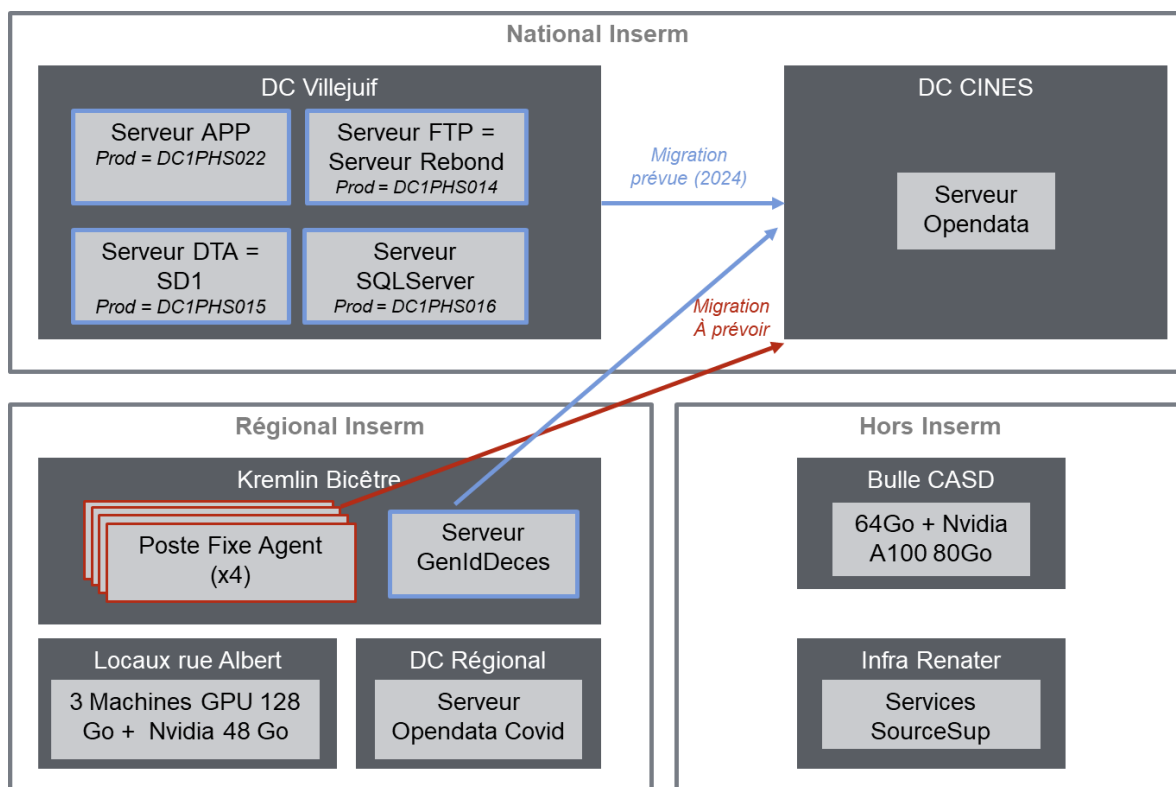
Pour rappel, les « traitements statistiques » dans le cadre de ce document sont les traitements fait par le métier en R, Python et SQL, qui répondent notamment aux besoins qui ne peuvent pas être couverts par les applications industrialisées en production (évolutivité forte, besoin de réactivité...). Il en est de même pour certains traitements applicatifs.





3.4. Infrastructures

Le CépiDc possède une multiplicité d'infrastructures. La migration de certains environnements est en cours depuis le data center (DC) de Villejuif, vers le data center Centre Informatique National de l'Enseignement Supérieur (CINES).



3.5. Exploitation

La supervision du SI du CépîDc est réalisée à plusieurs niveaux :

- La DSI réalise sa propre supervision de l'infrastructure sur la partie centrale (bases de données, serveurs, ...)
- La TMI du reste du parc applicatif, réalisée par Econocom, est garante du maintien en condition opérationnelle de l'infrastructure (CPU, RAM, ...)
- La TMA, réalisée par ATOS, est en charge des applications en charge du contenu (développement, évolution)
- Les métiers CépîDc par leur propre supervision des flux via les logs ou le contrôle des données en base.

Globalement la supervision de l'infrastructure, que ce soit au niveau TMI ou DSI, est opérationnelle. Mais la supervision applicative est à renforcer (Le chantier est identifié et initié mais pas officiellement lancé car non prioritaire) :

- Toutes les applications génèrent des fichiers de logs
- Les applications historiques de Jouve sont développées en java de manière assez homogène, mais les autres applications ont été développées et C# de manière plus éclectique.
- Trois types de logs cohabitent dans des répertoires différents : logs liés aux transferts de données situés sur le FTP, ceux liés à l'intégration des données dans SREF, ceux liés aux applications situées sur le DTA. Leur structure n'est pas homogène. En revanche, les logs des applications historiques de Jouve sont centralisés dans un même répertoire.
- Certains (cas d'erreur seulement, ne couvre pas le cas où le flux n'est pas joué) sont envoyés par mail, mais pour la plupart le métier doit directement consulter de lui-même les fichiers de log car ces derniers ne sont pas intégrés dans un logiciel de supervision.

Pour finir il existe une comitologie dédiée :

- Comité de suivi du CépîDc bimensuel : suivi opérationnel des sujets du moment et de leur résolution (infra + applicative)
- Comité de suivi mensuel : comité plus global en termes de maintenance (TMA, évolution application, projet, migration, benchmark)

4. Irritants

21 irritants ont été soulevés durant les ateliers. Ils ont été groupés en 6 thématiques.

- Disponibilité
- Infrastructure
- Gestion des droits
- Résolution des incidents
- Supervision
- Organisation

Chaque thématique abrite plusieurs irritants, qui peuvent être traités selon les cas au sein d'une refonte du SI ou hors de la refonte. Le présent document s'attache plus à la description des irritants qu'à des plans de remédiations. Toutefois, certaines pistes ont d'ores et déjà été identifiées lors des interactions. Le cas échéant, elles sont retranscrites.

4.1. Disponibilité

La thématique de la disponibilité recouvre l'ensemble des problèmes liés à l'accessibilité et le caractère fonctionnel des outils utilisés, par exemple la capacité à utiliser sans délais les outils mis à disposition. Les irritants sont les suivants :

4.1.1. Incapacité à accéder en simultané aux ressources

La connexion simultanée à plusieurs serveurs n'a pas été rendue possible sur tous les postes et reste instable.

A titre d'exemple, la connexion au serveur de fichiers partagés, géré au niveau régional (accès via VPN régional) et au

serveur de rebond (permettant d'accéder à IRIS), géré au niveau national (accès via VPN national) n'est pas aisé, car la connexion aux 2 VPN en simultané n'est pas possible. Il en était de même pour l'accès à la base de données, hébergée sur un serveur national. Une solution a été développée en juillet 2023 : mettre en place une redondance pour que le serveur de fichiers soit aussi hébergé en national avec un système de synchronisation et à terme la suppression du serveur régional. Depuis novembre 2023, l'accès à ce serveur de fichier depuis le VPN National (et donc simultané au serveur de rebond) est possible mais pas sur tous les postes et de façon instable ce qui peut être très lourd au quotidien.

Il est nécessaire de poursuivre la généralisation et stabiliser ces connexions.

NB : L'instabilité de la connexion est caractérisée par des micro-coupures qui incite les métiers à basculer sur un mode dégradé de téléchargement de fichiers en local pour ensuite recharger les fichiers une fois le travail effectué.

Une autre piste évoquée dans les entretiens consisterait à faire migrer totalement les usages hébergés sur les environnements régionaux sur le national (ex. le serveur de fichiers).

4.1.2. Avenir des postes fixes du Kremlin Bicêtre

Les 4 postes fixes du KB ne sont pas inclus dans le projet de migration/virtualisation des serveurs. Pourtant, des traitements supportés par ces postes sont indispensables à la production. Pour le moment il n'y a pas de projet de reprise/virtualisation/ migration les concernant. A titre informatif, les processus de production hébergés sur les ordinateurs fixes sont les suivants :

- Batch automatique IRIS
- Synchronisation annuelle
- Génération des échantillons
- Traitement des illisibles
- Maintenance du dictionnaire et des contrôles associés
- Indicateur de suivi de production

Le risque opérationnel de continuité d'exploitation pour les usages précédents est donc réel si une solution d'hébergement cible n'est pas trouvée. Ce sujet se recoupe en partie avec l'absence de serveur de calcul (4.2.1).

4.1.3. Suppression de droits

Une discontinuité des droits est observée par plusieurs interlocuteurs. Certaines autorisations sont supprimées de manière intempestive (exemple de cause racine : des migrations de règles de sécurité).

4.1.4. Synchronisation annuelle INSEE

La dernière étape de mise à jour des données dans la base de données SREF est un traitement qui nécessite d'isoler la base de données et d'interrompre l'ensemble des flux qui gravitent autour d'elle. Ce batch dure 25 heures en moyenne.

Ce processus de synchronisation manque de fluidité : avant la première migration du système d'information du CépiDc réalisée, il n'était pas nécessaire d'isoler le SI du CépiDc, mais depuis il est nécessaire de couper l'ensemble des flux touchant au serveur des bases de données (C.SSIS, TraitBPO, IDPI, FusionLotIris, C1, C2, IDEI, C1C, C1B) pendant tout un week-end sous peine de faire planter le batch et de contraindre l'équipe automatisation à de lourdes reprises manuelles en base.

Aussi, la refonte doit viser un objectif de fluidification de ce processus : ne pas nécessiter de bloquer les flux pendant un WE, « nettoyer » au fur et à mesure Sref grâce aux synchros mensuelles.

4.2. Infrastructure

La thématique de l'infrastructure recouvre l'ensemble des problèmes liés aux environnements et hébergements utilisés par le CépiDc. Les irritants sont les suivants :

4.2.1. Manque de serveurs de calcul pour soutenir les processus de production.

Il n'y a pas de serveur de calcul dédié permettant de faire les calculs de manière indépendante des postes des agents. L'ensemble des programmes statistiques en R, mais également les développements en Python, tournent sur la Ram des ordinateurs portables, la RAM des postes fixes au Kremlin Bicêtre ou encore sur les trois « grosses machines » avec GPU achetées dernièrement par le CépiDc pour l'IA. Il en est de même pour les traitements de synchronisation ou d'autres traitements lourds ou transverses comme les batch automatiques et le calcul automatisé d'indicateurs de suivi de la production. Autant, pour les grosses machines adaptées à l'IA cette situation relève d'un choix de grande souplesse, réactivité, etc. autant pour les autres traitements il s'agit d'un manque.

4.2.2. Manque d'intégration dans le SI des processus d'IA.

L'écosystème IA a été mis en place à côté du processus de production sans avoir pris en compte son intégration avec les processus industriels. Il s'agit d'un état de fait assumé par le CépiDc. Le prototypage, l'entraînement et l'inférence est faite sur 3 machines séparées du SI de production. Le processus d'inférence sur les données de production est donc encore artisanal et manuel. L'intégration dans le SI des usages IA suppose une clarification des évolutions à industrialiser, celles à garder sur un mode pérenne de « Self-Service » par les métiers et celles qui sont encore en cours de recherche développement. Cette clarification devra tenir compte du couple flexibilité/risque propre à chaque usage.

4.2.3. Manque d'environnement de tests / préproduction.

L'environnement de PREPROD n'a pas pu être mis en place (des machines étaient mêmes dédiées pour cela). Aussi il n'existe qu'un environnement de recette en plus de la production ce qui est très pénalisant.

Pour l'Opendata par exemple, les résolutions d'incident ont fait l'objet de patch directement appliqués en production sans test préalable car il n'y a pas d'environnement de recette pour cet outil.

4.2.4. Besoin d'homologation du SI CépiDc

Le CépiDc est identifié dans la cartographie des SI devant déclencher une homologation (sécurité, disponibilité, etc ...) Il serait cependant intéressant de disposer à court terme d'une liste de recommandations sur le second semestre sur les volets disponibilité, intégrité et confidentialité.

4.2.5. Sécurité à renforcer

- Le Plan de Continuité d'Activité sur le site de Lognes n'est pas finalisé.
- Les logiciels anti-malware doivent être modernisés (EDR & XDR).
- L'instruction sur les volets disponibilité, intégrité et confidentialité est nécessaire.

4.3. Gestion des droits

La thématique de la gestion des droits recouvre l'ensemble des problèmes liés à la gestion des profils et habilitations informatiques des utilisateurs du CépiDc. Les irritants sont les suivants :

4.3.1. Différence de groupes Active Directory CépiDc et INSERM

Il faudrait sûrement moderniser la gestion actuelle en intégrant toute la structure de l'active Directory local du CépiDc dans l'active Directory national de l'INSERM (8 000 agents + 7 000 chercheurs). Ce projet est porté par le bureau d'ingénierie (globalement il s'agit des mêmes personnes que la SECOP)

4.3.2. Découplage entre les AD régionales et nationales

La responsabilité partagée des AD régionales et nationale est complexe à gérer. Il en découle que la gestion des droits est particulièrement compliquée et qu'au quotidien les problèmes d'accès constituent l'irritant le plus marquant.

4.3.3. Mauvaises applications de la gestion des droits

Certains droits ne sont pas ouverts alors que demandés, certaines fermetures intempestives de droits sont constatées, sans communication préalable, ou calendrier prévisionnel.

4.4. Résolution des incidents

La thématique de la résolution des incidents recouvre l'ensemble des problèmes liés à la prise en charge, gestion et résolution des incidents sur les systèmes d'information qui impactent les utilisateurs du CépiDc. La détection est couverte par la thématique « Surveillance ». Les irritants sont les suivants :

4.4.1. Gestion du ticketing

Le système de ticketing n'est pas réactif et les attributions des tickets sont souvent erronées (ressenti depuis le transfert au national en 2017).

4.4.2. Difficulté d'investigation

L'architecture du Système d'information est très difficilement intelligible, car résulte d'une implémentation au fil de l'eau et souffre de plusieurs sources de complexité (national vs régional par exemple). Il est souvent difficile d'identifier précisément la cause d'un incident.

4.4.3. Documentation obsolète

La documentation n'a pas été mise à jour depuis 2019 (date à laquelle il n'y a plus eu de référent de la DSI au CépiDc). Les mises à jour intermédiaire n'ont pas abouti. En particulier, sur le serveur opendata, la documentation fait défaut de l'implémentation technique aux calculs « haut niveau » d'agrégation en R.

4.4.4. Amélioration continue

Les incidents rencontrés, une fois résolus, font peu l'objet d'un processus de capitalisation, comme une mise à jour de documentation ou une nouvelle procédure de contrôle.

4.5. Supervision

La thématique de la Supervision recouvre l'ensemble des problèmes liés à la détection et déclenchement de mesures suite à des incidents, compris comme anomalie de la qualité de service des systèmes utilisés par les acteurs du CépiDc. Les irritants sont les suivants :

4.5.1. Pas de levée d'alertes via une supervision des flux et traitements de production

Les principaux soucis rencontrés sont d'ordre réseau mais relèvent plutôt d'une gestion des flux qui n'est pas complètement maîtrisée.

Et comme il n'existe pas de supervision technique ou d'alerte (par exemple sur l'expiration des certificats de sécurité, le reporting du serveur en charge d'assurer la pseudonymisation, les VM Opendata ou WS IDDC) la détection des incidents est réalisée directement par le CépiDc ou même parfois directement par les partenaires.

Cet irritant est majeur étant donné que les métiers constatent régulièrement des traitements batch KO.

Un des symptômes les plus flagrant de ce manque de supervision (à connecter avec la résolution des incidents) est le fait qu'un serveur OpenData Covid est en panne depuis ~10 mois (juillet 2023), sans piste de résolution apparente.

4.5.2. Lisibilité sur les processus d'Accueil et Traitement.

La logique de stratification et de croissance organique des développements résulte en un SI abondant en traitements et flux autour de SRef, qui complexifient la lecture séquentielle des processus de contrôle, codage et synchronisation. Ces processus perdent en fluidité à cause de ces nombreux allers retours sur les mêmes lots de données : il est nécessaire de faire transiter des données par plusieurs applicatifs avant de progresser dans le processus métier. Cet irritant plaide pour une remise à plat du processus de bout en bout.

4.6. Organisation

La thématique Organisation recouvre l'ensemble des problèmes liés aux moyens humains, aux procédures et aux rôles et responsabilités relatifs à la gestion des systèmes d'information utilisés par les acteurs du CépiDc. Les irritants sont les suivants :

4.6.1. Sous-effectif DSI

Jusqu'en 2020, une équipe technique dédiée de 2-3 personnes (Côté DSI) faisait la coordination entre le métier et les équipes techniques. Elle a été dissolue et reportée sur la TMA sans accompagnement ni renfort. La coordination globale peut être améliorée entre les intervenants nationaux et régionaux.

Depuis près d'un an, Julio M. est dédié à 10% au suivi du CépiDc, ce qui reste largement insuffisant pour mener le déménagement et assister le métier CépiDc sur les incidents.

4.6.2. Rôles et responsabilités TMI, TMA

Plusieurs problèmes rencontrés proviennent d'un problème de maîtrise de bout-en-bout des systèmes. Certains services ont une gestion à cheval entre la TMA et TMI (par exemple les certificats de sécurité ou les bases de données), ce qui introduit une zone de flottement sur la prise en charge des problématiques. La gestion d'incidents sur l'OpenData constitue un exemple récent.

De manière générale, les rôles et responsabilités des métiers, TMI, TMA et de la Cellule SécOPS doivent être reposés en prenant en compte les problématiques de supervision (détection d'anomalie, investigation, application de correctifs, amélioration continue via mise en place d'une nouvelle fonctionnalité/action/procédure), sur tous les niveaux d'anomalies (de l'infrastructure « bas niveau » à l'erreur métier « dans les données »).

4.6.3. Gestion des sources / des versions

Le partage du code source n'est que peu outillé. Il existe un répertoire Git utilisé uniquement par l'équipe Diffusion à des fins d'archivage. Il n'y a pas d'outil de versionning pour l'équipe Production. Il est hébergé en externe (sur Renater), ce qui le rend dépendant à une connexion internet et indépendant de l'administration des habilitations centralisée de l'INSERM. Une des conséquences est le stockage des codes sources dans le serveur de fichier SD1 sans politique de classement clair (en effet, l'outillage git est trop peu connu et son administration n'est que peu maîtrisée).

Avoir un GitLab interne à l'INSERM (dédié ou non à l'équipe CépiDc), connecté au système de gestion des droits et faisant l'objet d'une administration serait une avancée.

4.6.4. Prolifération technologique et cadre de cohérence technique

Il existe une grande disparité au niveau des modèles de développements (utilisation de frameworks différents) qui complexifient la maintenance (ex : deux applications qui tournent sur SharePoint en C#, versions de Java différentes, ...). Cette prolifération résulte de libertés prises sur les choix d'architecture, de manière historique. Il y aurait un gain à rationaliser certains développements (exemple au niveau gestion des packages SSIS).

Pourraient constituer des avancées :

- La connaissance et mise en place d'un cadre de cohérence technique, validant les choix technologiques des développements.
- Une cartographie des technologies et versions pour repérer les obsolescences et concevoir une cible rationalisée.