CNSD

— RAPPORT D'ACTIVITÉ

2024

SOMMAIRE

Direction de la publication
Kamel Gadouche

Réalisation/rédaction
Roxane Silberman

Conception graphique
Guillaume Léonardy — boreaone@gmail.com
Crédits photos : AdobeStock

LE MOT DU *DIRECTEUR*

Dans un contexte où les défis de tout ordre s'accumulent à l'échelle nationale, européenne et internationale, pouvoir utiliser l'immense ressource des grandes bases de données individuelles de la statistique publique, des administrations ou d'autres détenteurs publics ou privés est un enjeu majeur pour la recherche comme pour l'évaluation des politiques publiques. Le CASD espère pouvoir contribuer à y répondre en continuant à développer ses activités en 2024 avec l'appui des producteurs de données pour permettre le traitement en toute sécurité de ces sources, sans silos entre grands domaines de l'économie, de l'environnement et de la santé.

En 2024, le nombre d'utilisateurs du CASD en France comme à l'étranger a continué à croître. Soucieux de fournir à tous un service toujours plus fluide tout en maintenant au plus haut niveau la sécurité indispensable au maintien de la confiance des producteurs comme de la société sur la protection de la confidentialité, le CASD a déployé de nouvelles fonctionnalités intégrées dans le portail CDAP (Confidential Data Access Portal) qui fédère l'ensemble des parties prenantes, producteurs et utilisateurs en collaboration avec le Comité du secret statistique et la Banque de France, tous les deux en charge des autorisations.

Les investissements réalisés sur la sécurité, l'ergonomie et la performance ont permis la mise en place d'une collaboration étroite avec la DREES et la DARES pour développer des environnements sécurisés de datascience dédiés au data management et la création de nouvelles sources. Comme les années précédentes, le CASD a également, comme tiers de confiance pour des appariements, participé directement à la création de nouvelles sources.

Acteur présent à l'international, avec 20% de ses SD-Box hébergées dans les grandes universités et centres de recherche de pays étrangers, coordinateur du réseau de centres sécurisés IDAN (International Data Access Network), le CASD a été choisi par META pour un second contrat pilote pour des projets de recherche américains désireux d'utiliser les données américaines d'Instagram. Le CASD est enfin particulièrement heureux d'avoir contribué comme partenaire à l'édition 2024 de la Conference of European Statistical Stakeholders (CESS), « The beyond GDP agenda : past, present, visions for the Future » organisée à Paris en octobre par l'INSEE et la Banque de France, sous l'égide de l'ESAC en partenariat avec Eurostat, la Banque Centrale Européenne, PSE et le CNIS.

Kamel Gadouche
Directeur

© Photo : Mathieu Boutang

01.

DES DONNÉES POUR LA RECHERCHE, L'ÉVALUATION DES POLITIQUES PUBLIQUES ET LA DATASCIENCE

Le CASD propose des environnements de calculs sécurisés dédiés au traitement de données de gros volumes, utilisés principalement par des chercheurs et data scientists. Le CASD agit comme tiers de confiance entre les producteurs de données et les utilisateurs pour le traitement sécurisé de données confidentielles.

Les données confiées au CASD couvrent la plus grande partie des données de la statistique publique, en particulier celles de l'INSEE et d'un grand nombre de services statistiques des ministères (SSM), un nombre croissant de bases de données de grandes agences publiques et administrations dont les données médico-administratives ainsi que des données issues de la recherche telles que celles de grandes cohortes épidémiologiques. Depuis 2022 la Banque de France met également ses données à disposition de la recherche via le CASD. Ces sources qui peuvent être utilisées conjointement et, dans certains cas, appariées, permettent ainsi le développement de travaux multi-domaines sans silos, ce qui est assez unique dans le paysage européen et international.

Le référentiel de sources de données interrogeable permet de lier sources, documentées au format DDI, projets de recherche et publications.

Les utilisateurs, une fois accrédités après avis du Comité du secret statistique et/ou des producteurs, et enrôlés auprès du CASD, travaillent à distance avec un accès direct aux données (remote access) dans l'environnement sécurisé du CASD via un terminal sécurisé (la SD-Box) depuis leur bureau dans leur organisme d'appartenance.

Les données sont utilisées tant pour les travaux de recherche menés, souvent en collaboration, par les universités de toute l'Union européenne (également de pays hors Union européenne disposant d'une décision d'adéquation de la Commission européenne pour le RGPD) que pour des travaux d'évaluation des politiques publiques par des instances institutionnelles ou pour des travaux de datascience.

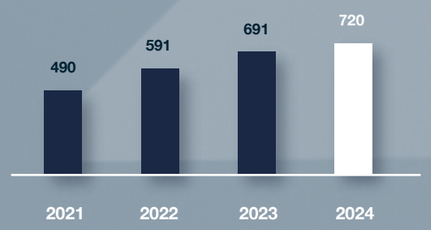
Le CASD accompagne les utilisateurs en organisant avec les producteurs des présentations de sources.

—NOS GRANDS *INDICATEURS*

720

En 2024, 720 projets ont été menés au CASD. Cela représente une augmentation de 4% par rapport à 2023.

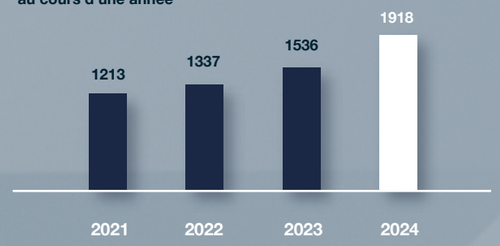
Nombre de projets menés au CASD
au cours de l'année



1 918

En 2024, 1 918 comptes utilisateurs étaient actifs en 2024. Cela représente une augmentation de 25% par rapport à 2023.

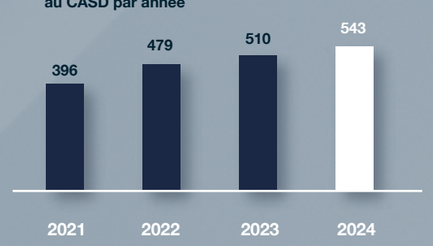
Comptes utilisateurs actifs
au cours d'une année



33

En 2024, 33 nouvelles sources ont été mises à disposition par le CASD. Cela a porté le total des sources mises à disposition à 543 sources.

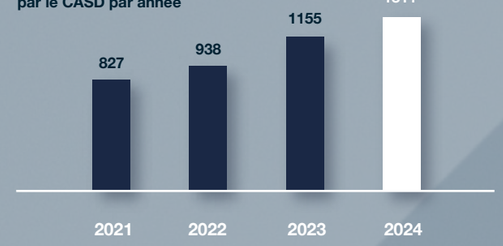
Nombre de sources de données hébergées
au CASD par année



1 311

En 2024, le CASD a déployé 1 311 boîtiers, cela représente une évolution de 14% par rapport à 2023. Depuis 2020, le CASD a donc doublé le nombre de SD-box déployées pour passer de 652 à 1 311 box.

Nombre de box déployées
par le CASD par année



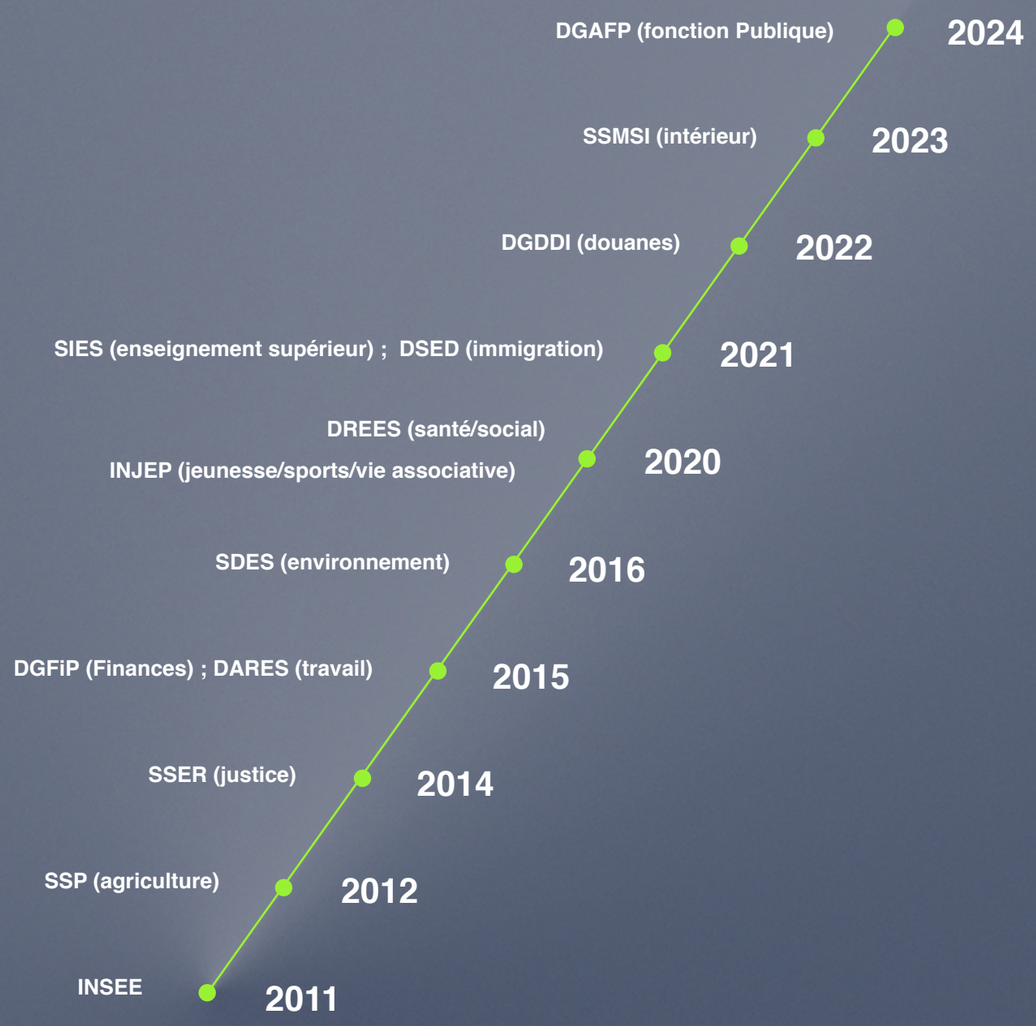
// UNE LARGE OUVERTURE DES DONNÉES DE LA STATISTIQUE PUBLIQUE *POUR LA RECHERCHE*

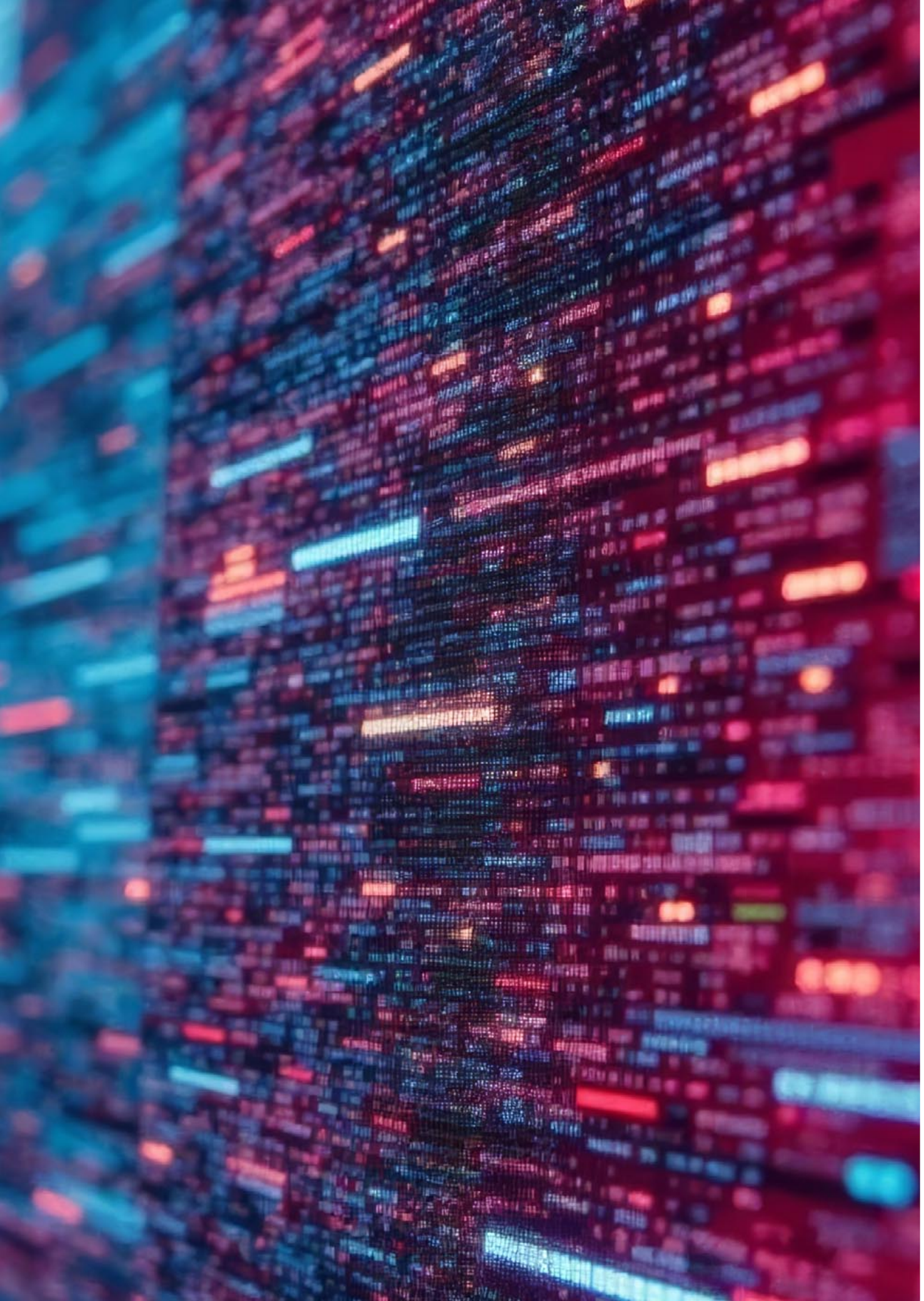
En 2024, la Sous-direction des Études, des Statistiques et des Systèmes d'Information (SDESSI) de la Direction Générale de l'Administration et de la Fonction Publique (DGAFP) a rejoint les 37 producteurs de données qui confient une copie de leurs bases de données individuelles au CASD pour les mettre à disposition des utilisateurs.

Une première source de données a été déposée au CASD

La Base statistique concours IRA (BSC IRA) fournit des informations administratives et sociodémographiques sur les candidats inscrits à un concours de la fonction publique. Issue de l'enquête Concours, qui interroge l'ensemble des candidats inscrits à un concours ciblé, et d'une base de données administratives (la BAC) sur laquelle l'enquête est adossée et qui contient des renseignements fournis par le candidat à son inscription, ses notes, son parcours dans le concours, et des informations sur le concours, la BSC IRA permet de disposer d'une vision complète du parcours dans le concours des candidats, allant de l'inscription aux résultats, et de l'analyser selon leurs caractéristiques sociodémographiques.

Ce sont ainsi maintenant les données de 12 SSM sur les 16 pour lesquelles les accès sont autorisés après avis du Comité du secret statistique et accord de leur part et de l'Administration des archives. Les données au CASD couvrent ainsi en 2024 une grande partie du périmètre de la statistique publique, INSEE et SSM.





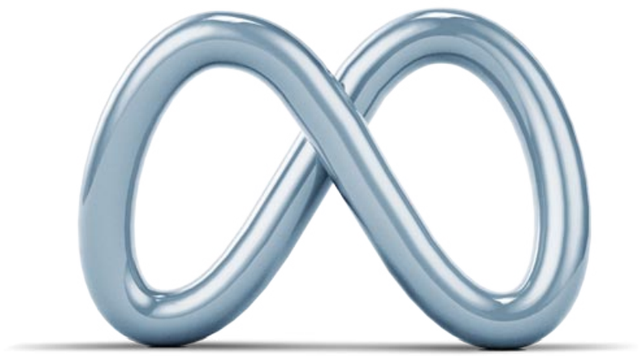
// LA TECHNOLOGIE ET LES SERVICES DU CASD UTILISÉS PAR META POUR *FOURNIR UN ACCÈS AUX DONNÉES POUR LES CHERCHEURS*

Très riches et sensibles pour nombre d'entre elles, les données des grandes plateformes suscitent l'intérêt quant à la façon dont elles peuvent être utilisées. Dans un objectif de régulation de leur usage par les grandes plateformes et de lutte contre les risques systémiques éventuels, le Digital Service Act « DSA » exige ainsi qu'elles mettent à disposition des chercheurs leurs données. Les chercheurs sont par ailleurs intéressés depuis longtemps à utiliser ces données pour des travaux de recherche dans différents domaines.

Dans le cadre de pilotes en amont de la mise en œuvre du DSA, de grandes plateformes se sont tournées vers le CASD pour fournir des accès aux données pour les chercheurs.

- META qui recherchait en Europe un organisme indépendant et expérimenté, avait signé un contrat pour des données de Facebook en 2023 avec le CASD dans le cadre des discussions en cours à l'EDMO (European Digital Media Organisation) visant à mettre en place un Organisme Intermédiaire Indépendant (OII) pour examiner les demandes d'accès, le CASD étant le tiers de confiance offrant la solution technique d'accès sécurisé aux données.

- En 2024, en dehors du contexte du DSA, META a signé un contrat pour des données d'Instagram avec un programme pouvant aller jusqu'à 7 projets qui sera développé en 2025.

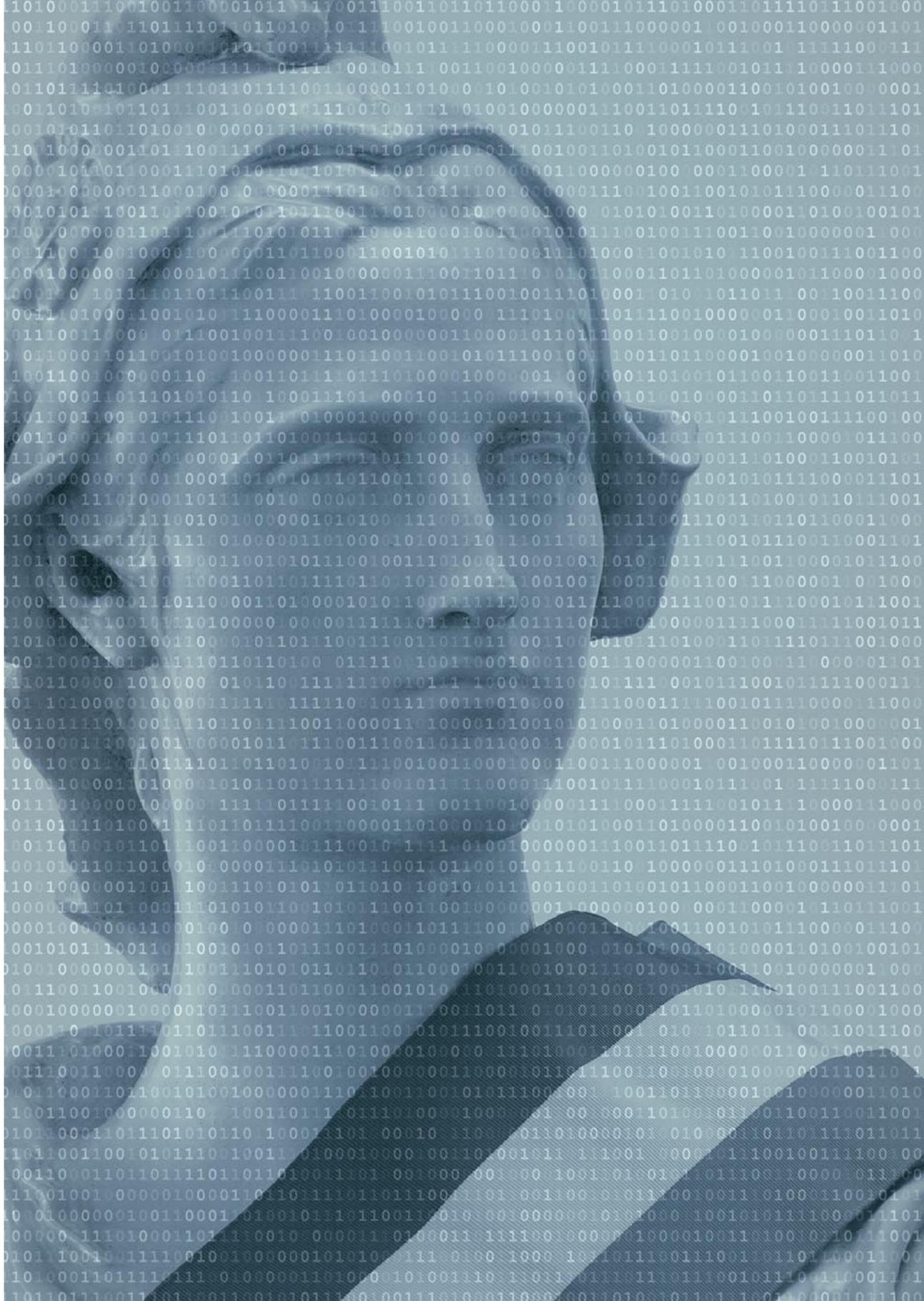


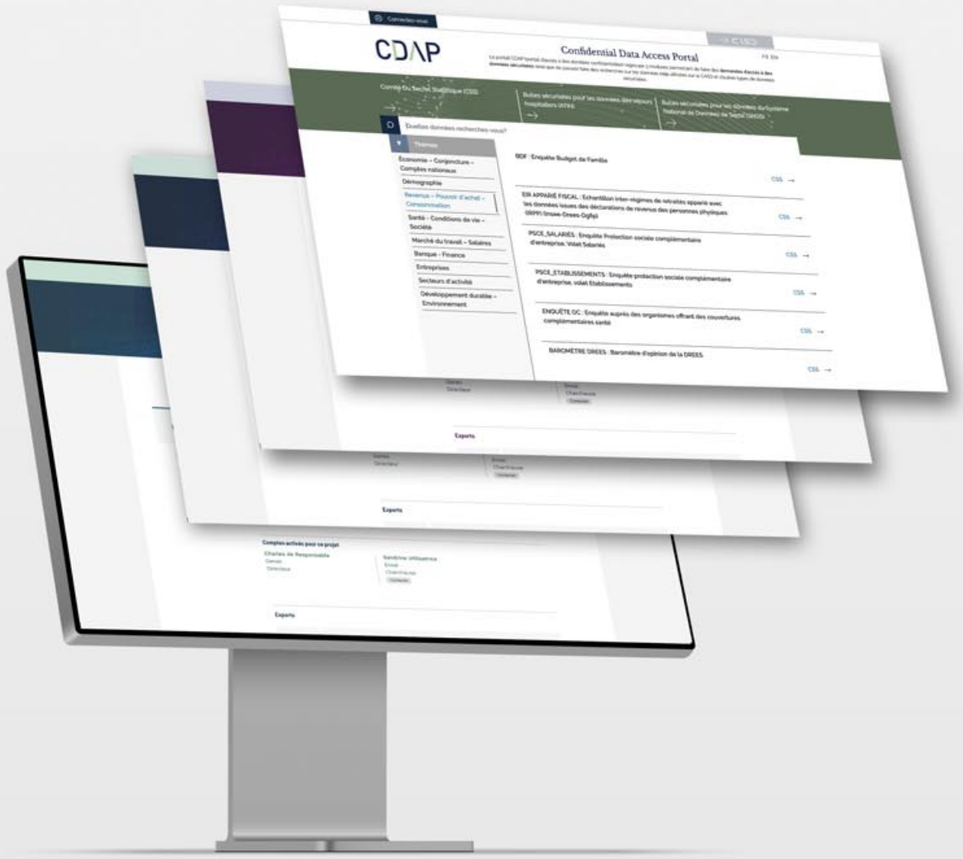
// DE PLUS EN PLUS *D'UTILISATEURS INSTITUTIONNELS*

Pour répondre à leurs besoins, les administrations publiques bénéficient d'une procédure « administration » auprès du Comité du Secret Statistique : cette procédure leur permet d'ajouter des sources de données et/ou des membres à leurs projets en dehors du calendrier des séances régulières du Comité. Elles peuvent alors travailler sur le CASD.

En 2024, 14 administrations publiques avaient un accès actif au CASD pour leurs travaux internes.

- Direction Générale des Entreprises - Ministère de l'Économie (DGE)
- Direction Générale du Trésor - Ministère de l'Économie (DG Trésor)
- Conseil d'Analyse Économique - Commissariat Général à la Stratégie et à la Prospective
- Commissariat Général à la Stratégie et à la Prospective
- Inspection Générale des Affaires Sociales
- Cour des Comptes
- Inspection Générale de l'Environnement et du Développement Durable
- Inspection Générale des Finances
- Direction Générale des Outre-Mer - Ministère des outre-mer
- Sénat
- Unédic
- Direction de l'Habitat, de l'Urbanisme et des Paysages
- Centre d'Études et d'Expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement
- Direction Générale de la Cohésion Sociale





// CDAP : DE NOUVELLES FONCTIONNALITÉS *POUR UN PORTAIL FÉDÉRATEUR*

Les demandes d'accès aux données de la statistique publique, à celles couvertes par le secret fiscal et aux données de la Banque de France se font sur le module du Comité au sein du portail CDAP (Confidential Data Access Portal) développé par le CASD pour le Comité du secret statistique. Les utilisateurs y trouvent la liste des données, y déposent leurs demandes, peuvent y contacter les producteurs concernés par leurs demandes. Après un premier développement en collaboration avec le Comité du secret statistique, le CASD a mis en place une interface dédiée à ses utilisateurs : l'interface CASD du portail CDAP. Cette interface permettait déjà à tous les utilisateurs du CASD d'avoir une vision d'ensemble de leurs différents projets avec le suivi de leurs habilitations et de leurs abonnements.

- En 2024, l'interface CASD du portail CDAP https://cdap.casd.eu/ a poursuivi son enrichissement avec la mise en place d'une nouvelle fonctionnalité de récupération des sorties de résultats. La demande de « sortie des résultats » depuis la bulle sécurisée du projet, suit une nouvelle procédure de transmission plus rapide et fluide, ne nécessitant plus de passer par des étapes par mail et envoi d'un code.

- Une fois le contrôle d'anonymisation effectué par le CASD, les résultats sont directement récupérables pendant 15 jours depuis l'interface CASD du projet concerné dans CDAP où l'utilisateur disposera de l'historique de l'ensemble des sorties réalisées. Des notifications automatiques informent de l'avancement des demandes.

- En 2024, le déploiement de cette nouvelle fonctionnalité a été effectué pour les nouveaux utilisateurs avant sa généralisation progressive pour 2025.

- CDAP devient ainsi un point fédérateur pour toutes les parties prenantes en fluidifiant les processus tout en maintenant le même niveau de sécurité dédié aux travaux des chercheurs.

CDAP

// UN WEBINAIRE TRÈS SUIVI *SUR LES DONNÉES DE LA DGFIP*

Le CASD s'attache à accompagner au mieux les utilisateurs avec l'aide des producteurs et autres parties prenantes afin qu'ils puissent se saisir au mieux des sources de données disponibles. Ces webinaires favorisent aussi le retour des chercheurs vers les producteurs, une interaction importante qui permet aussi d'améliorer la qualité et la documentation des données.

Le 21 mai 2024, afin de permettre aux utilisateurs de mieux appréhender les données DGFIP mise à disposition au CASD, une session de présentation concernant certaines sources fiscales a été organisée par le CASD et la DGFIP. Cette session a porté sur les sources suivantes :

BIC-IS : Bénéfices industriels et commerciaux - tous régimes

BIC-RN : Bénéfices industriels et commerciaux - régime normal

BIC-RS : Bénéfices industriels et commerciaux - régime simplifié

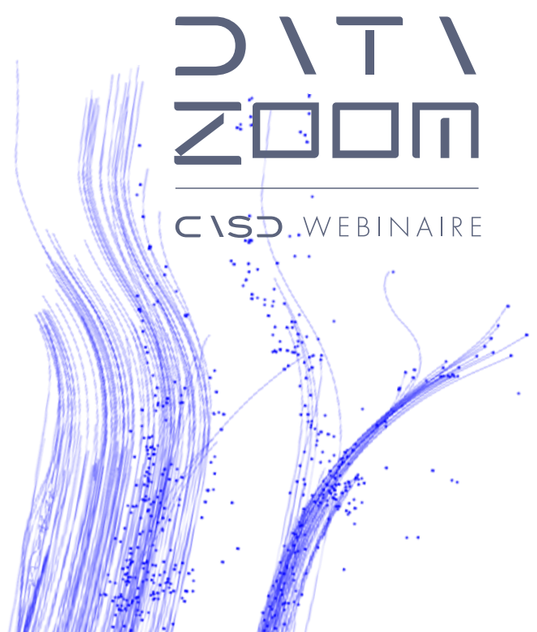
BA : Bénéfices agricoles - régime normal et simplifié

PERIM : Périmètres des groupes fiscaux

ISGROUPE : Groupes fiscaux des entreprises à l'Impôt sur les Sociétés (IS)

Liasses fiscales pro : Liasses fiscales Entreprises

Elle a été présentée par Monsieur Gérard Forgeot, Chef de la Section Production statistique, diffusion et qualité et Monsieur Roddy Caccialupi, Direction Générale des Finances Publiques.



// DES *DONNÉES DE SANTÉ* TOUJOURS TRÈS UTILISÉES

L'utilisation des données de santé très sensibles est soumise à des prérequis spécifiques. Le CASD héberge de nombreux projets de santé, utilisant notamment les données médico-administrative comme les données PMSI de l'ATIH sur les séjours hospitaliers, les données de l'Assurance Maladie (SNDS) et les données des cohortes INSERM (Constances…).

L'infrastructure du CASD a été mise à jour aux deux référentiels de sécurité des données de santé : pour la certification « Hébergeur de données de santé » et pour l'homologation au référentiel de sécurité des données de santé (relatif à l'exploitation statistique des données du Système National des Données de Santé, SNDS)

Plus de 100 projets ont traité sur le CASD des données santé

PMSI	47
INSERM	35
CNAM	17
IRDES	11

// LE CASD *PARTENAIRE DU PROJET* GRAPH4HEALTH

Le Projet Graph4Health, sélectionné en 2023 par l'ANR, coordonné par le Centre de recherche en économie et statistique (CREST), en collaboration avec l'Unité Mixte de Service Constances de l'Inserm, l'ESSEC Business School et, pour l'infrastructure, avec le CASD, a commencé ses premiers travaux. Le projet, qui étudie la formation des liens entre patients et professionnels de santé ainsi que la structure et l'évolution de ces liens, utilise les données quasi-exhaustives du SNDS sur 11 années de 2008 à 2018 avec les données du référentiel des professionnels de santé (RPPS), ce qui représente plusieurs dizaines de Teraoctets.

En 2024 le CASD a fourni une expertise technique cruciale pour le projet Graph4Health. Il apporte notamment :

- **l'infrastructure et l'environnement de calculs avec les logiciels de data science** (Python, R…) et des ressources matérielles spécifiques (double GPGPU NVIDIA H100 NVL)
- **l'expertise à la mise en œuvre de « frameworks » de calculs :**
  - Spark pour le calcul distribué
  - SEDONA pour travailler sur les données géo-spatiales. Les données géo-spatiales sont notamment utilisées pour détecter les déserts médicaux. En effet, l'équipe Datascience du CASD a développé des algorithmes de calculs pour prendre en entrée la localisation des patients et des docteurs/hôpitaux (que ce soit en coordonnées ou en localisation GPS) et calculer la distance moyenne entre les deux localisations afin de détecter la présence des déserts médicaux
- **le catalogue de données au sein de l'outil Openmetadata** où sont listées toutes les tables du projet, leur localisation, le nombre de colonnes et le « data lineage » (à partir de quelles tables de données brutes ont été construites). Ce catalogue de données comporte notamment les données SNDS et les données de géolocalisation et est constitué d'environ 300 tables
- **un appariement entre le code postal et le code commune INSEE** des établissements hospitaliers et la correction des éventuelles erreurs

02.

DES ENVIRONNEMENTS SÉCURISÉS POUR LA *CRÉATION DE NOUVELLES SOURCES*

Par essence, les environnements du CASD permettent le traitement de données confidentielles : large échelle, sensibilité des informations, etc. Il est ainsi techniquement possible, dans ces environnements hautement sécurisés, d'apparier diverses sources de données, en particulier sur la base d'identifiants directs (SIRET des entreprises, état civil des personnes voire NIR). Ces appariements, le plus souvent réalisés à large échelle (plusieurs centaines de milliers, voire plusieurs millions, de personnes), permettent la création de nouvelles sources de données décuplant les possibilités d'exploitation pour la recherche.

La Statistique publique (avec un grand S et l'article défini) réalise depuis plusieurs années déjà des appariements de ce type, on peut penser notamment au dispositif RESIL de l'INSEE pour les individus et les logements, mais également à d'autres types tels que les enquêtes relatives à l'autonomie appariées aux données de santé ou la constitution d'échantillons inter-régimes pour le suivi des cotisants au système de retraite français des ministères sociaux.

Pour disposer d'un environnement performant et sécurisé, certains producteurs de données font appel au CASD, au travers des offres INFRASEC, permettant de disposer d'environnements sécurisés hautement personnalisables pour réaliser leurs travaux de Data Management et d'exploitation statistique, ou encore au titre d'une prestation de réalisation d'appariement, dans le cadre de laquelle le CASD joue le rôle de tiers de confiance chargé des appariements. (cf. infra). Il en est de même pour certains projets portés par des chercheurs qui, associés à une ou plusieurs administrations, peuvent mener à la création d'une nouvelle source.

Le but est in fine de mettre à disposition des sources enrichies, une fois pseudonymisées, à disposition de la communauté de la recherche.

// *LES PRODUCTEURS DE DONNÉES* DEVIENNENT UTILISATEURS DES ENVIRONNEMENTS CASD

Le projet ESTRADD de la DREES et de la DARES (ministères sociaux), vise à offrir aux chargés d'études/data scientists des deux SSM des environnements de calcul pour leurs travaux de production statistique : les deux directions ont choisi de faire un partenariat avec le CASD pour les parties qui concernent les traitements les plus sensibles ou requérant une puissance de calcul supérieure associée à des outils ad hoc et à une grande stabilité.

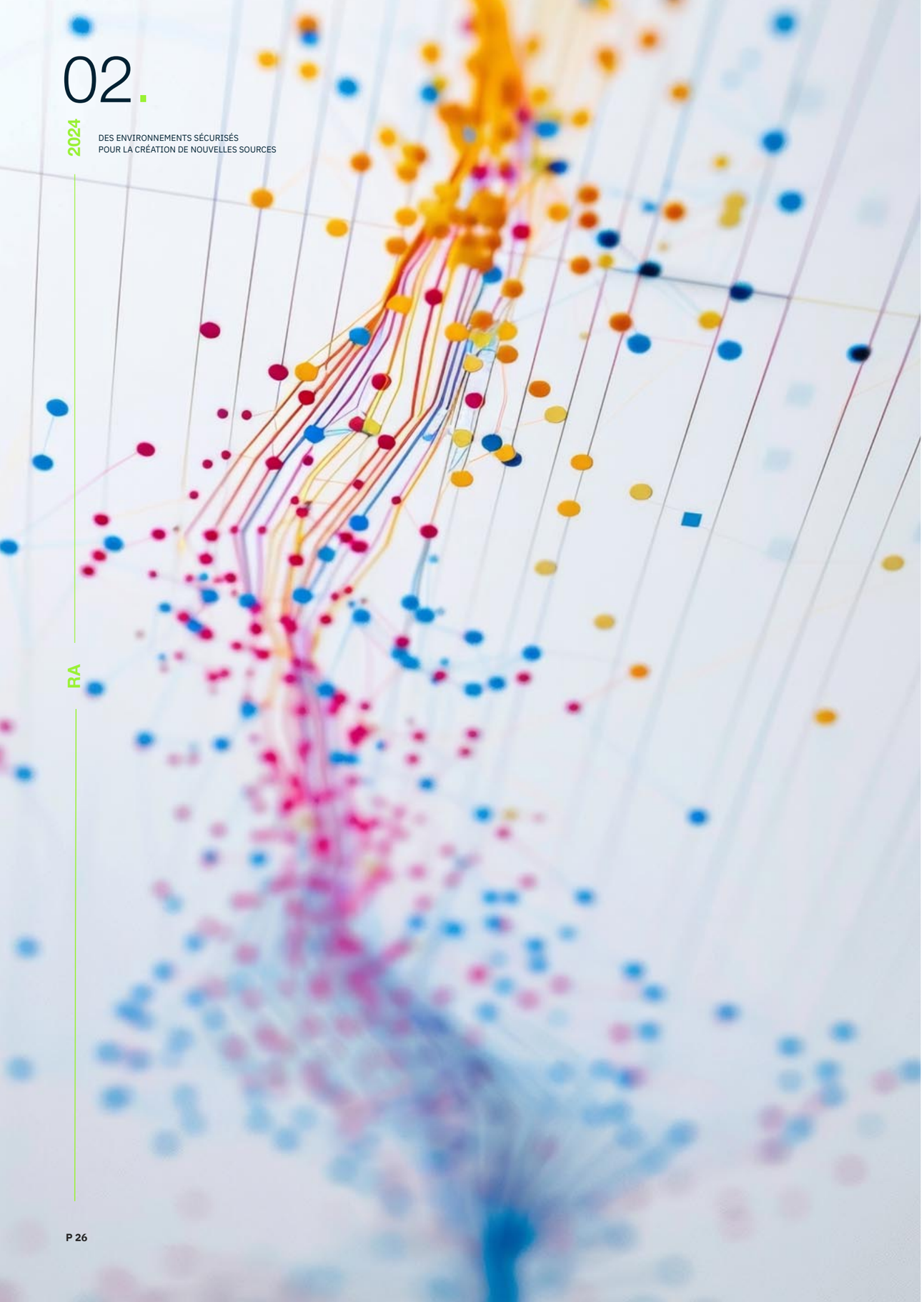
Initié en 2023, ce projet s'est élargi pendant toute l'année 2024 et continue de monter en ampleur. Chaque direction dispose d'une bulle sécurisée dédiée à l'administration des sources et d'une autre mutualisée permettant un partage facilité d'informations. D'autres bulles sécurisées sont installées à la demande. Fin 2024, la DREES dispose de 17 bulles actives ; la DARES 7, dont celle permettant l'accès au cluster Spark de calculs dédié à MIDAS (Minima sociaux, droits d'assurance chômage et parcours salariés) qui permet de reconstituer les trajectoires professionnelles des allocataires de l'assurance chômage et des minima sociaux.

Les SSM du Ministère de l'Intérieur (SSMSI) et du Ministère de la Justice (SSER) ont fait appel au CASD pour créer deux environnements dédiés à la création d'une nouvelle source de données reproduisant la chaîne pénale complète des individus présents dans leurs fichiers. Ce traitement nécessite un très haut niveau de sécurité, en particulier en termes d'étanchéité des environnements vis-à-vis de l'extérieur, de gestion des identifiants permettant les appariements et de traçabilité des accès.

Dans le secteur de la santé, le CASD opère les infrastructures de calcul de l'IRDES (3 bulles dédiées en fonction des thématiques d'études et de production) et de l'INSERM pour la cohorte Constances (5 bulles dédiées à diverses étapes de construction de la cohorte et de son exploitation). Constances permet par ailleurs à d'autres équipes de recherche d'utiliser leurs données via une gouvernance dédiée. Le CASD est également partie prenante pour développer une application de recette des données du SNDS reçues par l'équipe Constances au sein d'une bulle dédiée. Le développement du premier module s'est poursuivi sur toute l'année 2024 et devrait aboutir à l'été 2025.

estradd
Environnement sécurisé
de travail **DREES DARES**

Envirronnement sécurisé de travail
DREES DARES - ESTRADD



// LE CASD, *UN ACTEUR* POUR LES OPÉRATIONS DE TIERS DE CONFIANCE

Outre l'hébergement direct d'environnements de production pour plusieurs services statistiques ministériels, le CASD prend part à la création de nouvelles sources de données par la réalisation d'appariements de sources existantes. Ces appariements, reposant la plupart du temps sur des sources administratives à large échelle (plusieurs centaines de milliers d'entreprises ou plusieurs millions de personnes) et sur des données confidentielles, nécessitent à la fois un accompagnement solide en termes organisationnels et relativement à la réglementation, des compétences pointues en data management, une sécurité à l'état de l'art et une confiance de la part des organismes confiant les données et des autorités de contrôle. A ce titre, le CASD joue le rôle de tiers de confiance pour la réalisation des appariements ou pour assurer certaines opérations de dés-identification des personnes avant mise à disposition des données aux équipes de chercheurs.

En 2024, le CASD a continué les appariements semestriels des opérations FORCE et MIDAS de la DARES. La première opération apparie les données de France Travail (ex-Pôle Emploi) et des mouvements de main d'œuvre (MMO produits par la DARES), tandis que la seconde associe en plus les données de la CNAF sur les allocataires de minima sociaux. Ces deux sources de données sont mises à disposition des équipes de recherche qui en font la demande, toujours en passant par le processus du Comité du secret statistique.

L'enquête CARE (Capacités et Ressources des Séniors) de la DREES, pour son volet institutionnel (résidents des EHPAD en majorité), continue le suivi de la mortalité des personnes interrogées, pour lequel le CASD réalise les opérations en lien avec l'INSEE, sur la base du NIR.

Pour certains projets de santé, le CASD permet un dialogue entre les chercheurs et les services de la CNAM en vue de l'appariement de données avec celles du Système national des données de santé (SNDS). Le CASD accompagne les équipes de recherche dans la gestion des flux de données, ce qui nécessite plusieurs environnements cloisonnés. Des flux de données de complexité variable ont été détaillés lors d'une présentation conjointe du CASD et de la CNIL en juin 2024 pour garantir la sécurité de bout en bout de ces traitements d'appariement.

Ce type d'opérations vise à ce qu'aucun des acteurs d'un appariement (producteur des données et chercheurs) ne dispose des mêmes identifiants, renforçant les mesures de pseudonymisation des données.

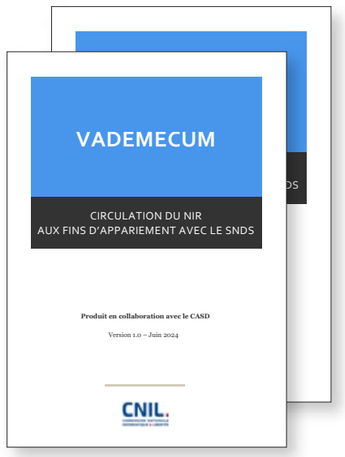
// SANTÉ ET *APPARIEMENTS*

Les appariements constituent une opportunité très importante pour de très nombreux projets de recherches d'enrichir les données de santé.

Voir le replay

25 juin 2024 **Webinaire CNIL-CASD :** La CNIL, en partenariat avec le CASD, a organisé un webinaire dédié aux appariements de données pour les recherches et études en santé. Ce webinaire, qui a réuni près de 300 personnes, a été l'occasion de présenter les différents cas d'usage et a permis également de disposer d'illustrations étape par étape des schémas de circulation des données qu'il est nécessaire d'anticiper lorsque l'on souhaite apparier plusieurs tables de données en s'appuyant sur le NIR dans le domaine de la santé. Le replay est disponible en ligne.

Le CASD a également participé à la rédaction avec la CNIL de fiches pratiques pour les circuits d'appariement avec le SNDS.



03.

SÉCURITÉ ET TECHNOLOGIES AVANCÉES DE DATA SCIENCE

La sécurité des données est au cœur des missions du CASD : concrètement, plus de 700 mesures de sécurité (dont les 114 issues de la normes ISO27002) sont mises en œuvre ; qu'elles soient organisationnelles, contractuelles, physiques ou informatiques. L'ensemble du système « bulles sécurisées » du CASD fait l'objet d'audits techniques réguliers par des prestataires certifiés par l'ANSSI. Tout ajout de fonctionnalité fait l'objet au préalable d'une analyse de sécurité dès la conception jusqu'à la mise en production. Le CASD ne faisant appel à aucun sous-traitant qui pourrait avoir accès aux données confiées, c'est toute la chaîne de traitement qui est maîtrisée de bout en bout.

L'homologation au référentiel de sécurité des données de santé et les certifications garantissent le maintien d'une sécurité optimale des données hébergées et des traitements associés. Le CASD assure une veille constante de la technologie et de la réglementation européenne et nationale afin de se conformer à ces exigences. L'intégralité des procédures est documentée, ce qui assure également la résilience des équipes. Les certifications font l'objet d'audits de surveillance réguliers (deux par an). L'objectif est également l'amélioration continue des systèmes de management de la sécurité et de la protection des données.

Si la sécurité – notamment selon les critères de disponibilité, intégrité, confidentialité et traçabilité - est absolument primordiale pour tous les acteurs de l'écosystème, des producteurs de données aux utilisateurs finaux, l'environnement de travail des bulles sécurisées se doit d'être adapté aux travaux d'analyse des utilisateurs, tout en répondant aux plus hauts niveaux d'exigence pour les méthodes de calcul innovantes et particulièrement gourmandes en capacités machine : ainsi, outre l'environnement standard et les outils dédiés aux travaux statistiques (R, Python, Stata, Julia, QGIS), de data management (HeidiSQL, DuckDB…) et de valorisation des résultats (suite Office et OpenOffice, éditeur LaTeX…), le CASD s'attache à construire des architectures dédiées à des projets spécifiques. A chaque fois, l'équipe Datascience du CASD accompagne les utilisateurs dans la définition précise des besoins et dans l'utilisation des outils mis à disposition.

A ce titre, la quasi intégralité du dépôt de packages R (CRAN) et des ado Stata (REPEC) est disponible nativement dans l'environnement de l'utilisateur, ainsi que les packages Python (et leurs dépendances) les plus utilisés. Si d'autres packages sont apportés par l'utilisateur, le CASD peut les introduire dans la bulle sécurisée. Enfin, l'équipe IT peut réaliser, à la demande, l'installation de logiciels spécifiques nécessaires à certains projets en datascience. Comme toujours, cette installation se fait à la suite d'une analyse préalable de sécurité pour garantir l'intégrité de l'environnement.

— LA SD-BOX® *AU CŒUR* DU DISPOSITIF

Conçue en 2010, la SD-Box s'est améliorée au fil des générations. Elle est devenue plus facile à utiliser, plus économe en énergie et la plupart de ses composants sont réutilisables et recyclables.

04

ÉVOLUTION DE LA SD-BOX

01

02

03

— *DÉPLOIEMENT* DE LA CARTE SSO BIO

Il y a quelques années, une carte à puce biométrique nécessaire pour se connecter était dédiée à chaque projet. Un utilisateur qui avait accès à plusieurs projets, et donc autant de bulles sécurisées, devait disposer de plusieurs cartes et changer de carte à chaque fois qu'il voulait passer d'un projet à l'autre. Le temps de connexion, qui intégrait de nombreuses transactions pour les vérifications de sécurité (validation du certificat…) était un peu long.

Le CASD a développé une interface sécurisée d'authentification permettant à la fois :

- de cloisonner les autorisations aux bulles sécurisées sur la carte à puce et ainsi obtenir une carte multi-projets et fluidifier la transition entre bulles sécurisées,
- d'optimiser le processus de validation de la chaîne d'authentification et le processus de validation des certificats (en réduisant les temps de transaction par exemple) et ainsi réduire le temps de connexion.

Ces développements ont été réalisés pour apporter un plus grand confort pour les utilisateurs tout en maintenant le haut niveau de sécurité exigé pour le traitement de données confidentielles.

1 172

Le déploiement se fait progressivement et en 2024 le nombre de carte SSO BIO déployée était de 1 172

LA SÉCURITÉ *UNE PRÉOCCUPATION CENTRALE :* DES AUDITS RÉGULIERS ET UNE DÉMARCHE POUR LA CERTIFICATION EUROPRIVACY

Avec la sécurité au cœur, les certifications et les audits sont essentiels. Le CASD qui dispose des certifications ISO 27001, ISO 27701, HDS y attache une attention toute particulière.

En 2024

- Le CASD a commencé la démarche pour le nouveau référentiel RGDP – Europrivacy : certification de service officielle prévue à l'article 42 du RGPD ce qui lui permettra d'être la 1re hébergeur de données en Europe certifié sur ce référentiel.

- Le CASD a lancé par ailleurs une démarche pour la mise en place d'un système de management de la qualité ISO 9001.

- Le CASD a effectué la migration vers la nouvelle méthodologie d'analyse de risque EBIOS RM (version modifié d'EBIOS).

- Le CASD réalise régulièrement des audits. L'audit de sécurité technique réalisé par la société Orange Cyberdefense en juillet 2024 a confirmé le très bon niveau de sécurité du dispositif.

// UN TEMPS DE RÉPONSE TOUJOURS RAPIDE À DES *DEMANDES D'EXPORTS EN FORTE CROISSANCE*

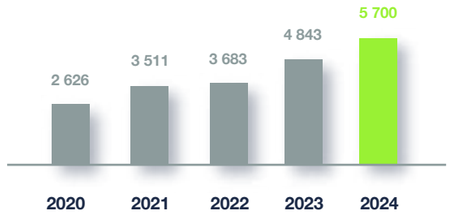
Le choix fait par l'INSEE et le Genes, à l'origine du projet CASD, d'un système en « remote access » qui permet à l'utilisateur de travailler dans la bulle sécurisée du projet directement sur les données jusqu'à obtention des résultats souhaités est un grand avantage, encore peu développé dans de nombreux autres pays. Pour autant, les exports de résultats finaux doivent respecter les règles pour l'anonymisation des données par les producteurs. Les producteurs de données choisissent le mode de contrôle des exports de résultats (contrôle manuel systématique par le CASD qui est majoritaire) ou sorties automatiques.

Avec 5700 sorties manuelles vérifiées en 2024, l'augmentation annuelle moyenne dépasse les 20%, ce qui reflète la croissance du nombre de projets. Le temps de traitement est néanmoins resté court .

2024 a vu l'implémentation d'un outil interne d'aide aux traitements des sorties Permettant un gain de temps très important aux équipes du CASD : concrètement, le flux des exports de résultats est bien plus rapide et nécessite moins d'étapes aux équipes. Lorsque la vérification de l'anonymat des sorties est manuelle, la plupart des autres étapes de traitement de la sortie sont, elles, automatiques, de la demande à la livraison de la sortie à l'utilisateur sur le portail CDAP, qui étend ses fonctionnalités pour les utilisateurs du CASD. Le niveau de sécurité reste le même, toujours aussi exigeant.

5 732

Nombre de sorties manuelles vérifiées par année



// *FLUIDIFIER LES EXPORTS ET LES IMPORTS :* DES AVANCÉES MAJEURES

Les exports comme les imports que l'utilisateur peut avoir besoin de faire pour son analyse ne doivent pas compromettre la sécurité du système. Sur ces deux points essentiels, le CASD s'efforce de faciliter le travail des utilisateurs.

Développement d'algorithmes de machine learning et LLM pour la vérification de la confidentialité des sorties de résultats à partir du CASD

A l'issue de leurs travaux sur le CASD, lorsque les chercheurs souhaitent récupérer les résultats de leur analyse, ils doivent soumettre leurs fichiers au service Data management du CASD pour vérifier que les fichiers respectent bien les règles de confidentialité.

Pour assister les vérificateurs, le CASD développe un système d'analyse des fichiers de résultats et de leur contenu, basée sur la génération de caractéristiques à partir de groupes de fichiers exportés et sur l'entraînement d'un modèle de renforcement. Le système s'appuie sur des données historiques provenant d'examens antérieurs effectués 'manuellement', où les décisions (Accepté/Refusé) ont servi d'étiquettes, et des caractéristiques structurées ont été extraites des exportations examinées. L'objectif premier était d'identifier les situations susceptibles de présenter un risque pour les données confidentielles et de déclencher des alertes en vue d'un examen par un expert.

En 2024, le CASD a commencé à réaliser une nouvelle amélioration pour accroître la fiabilité du système. En tirant parti des grands modèles de langage (LLM), il a été ajouté de nouvelles caractéristiques prédictives qui enrichissent le modèle d'évaluation des risques. En particulier, les LLM permettent la détection d'attributs clés précédemment absents du système, tels que l'identification des colonnes contenant des effectifs et la garantie de la conformité avec les exigences de valeur minimale.

Au-delà de la validation numérique, les LLM améliorent également la capacité du système à analyser le contenu textuel, en déterminant si les données exportées consistent en du code, des ensembles de données structurées ou des documents en texte libre. En améliorant la transparence et la traçabilité, cette intégration, au stade de prototype actuellement, devrait renforcer à la fois la fiabilité et l'interprétabilité des contrôles de conformité du respect de la confidentialité des données.

Import sécurisé automatique de scripts

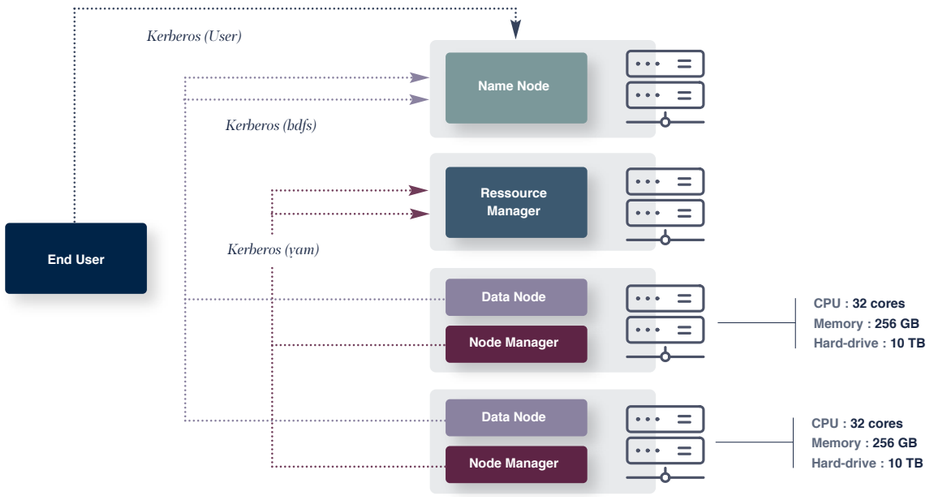
Pour faciliter les imports, le CASD a développé la possibilité pour l'utilisateur de pouvoir de lui-même depuis sa bulle démarrer une procédure pour importer un bout de texte.

// *DES ENVIRONNEMENTS DÉDIÉS* POUR DES PROJETS EN DATASCIENCE

L'équipe IT peut à la demande réaliser l'installation d'environnements de travail spécifiques nécessaires à certains projets en datascience. Comme toujours cette installation se fait à la suite d'une analyse préalable de sécurité pour garantir l'intégrité de l'environnement.

Le CASD a mis en place, dans le cadre d'un projet MIDAS de la DARES, un cluster SPARK/HDFS, permettant de distribuer les calculs au plus près des données réparties sur 15 serveurs, et qui rassemblent :

- 150 vCPU,
- 2,8 To de RAM,
- 30 To de disque brut



// UN ENVIRONNEMENT TRÈS SPÉCIFIQUE POUR LES *DONNÉES GÉNOMIQUES*

Des données génétiques sur le CASD

En 2024, dans le cadre du programme transversal Variabilité génomique de l'Inserm, également appelé GOLD pour GenOmics variability in heaLth & Disease, le projet GOLD-GENOPHENOMET s'appuie sur les données génétiques des volontaires de la cohorte Constances, générées dans le cadre du projet pilote du plan France Médecine Génomique 2025 POPGEN, mises à disposition de façon sécurisée au CASD.

L'objectif de ce projet est d'évaluer des méthodes statistiques et bioinformatiques afin de mieux comprendre l'impact des variations génétiques sur la santé et ainsi pouvoir, si possible, mettre en place des mesures plus efficaces dans la prise en charge de la population.

Pour ce faire, un environnement de travail spécifique, permettant de programmer dans un large spectre de langages (C/C++, Java, Perl, Python, Bash, Awk) et avec des logiciels spécifiques du domaine (Plink, LDPred2, snptest, bolt-lmm, etc), a été installé dans une bulle sécurisée dédiée sous Linux, accessible après authentification sur une SD-Box depuis l'environnement sécurisé mis à disposition au CASD.

04.

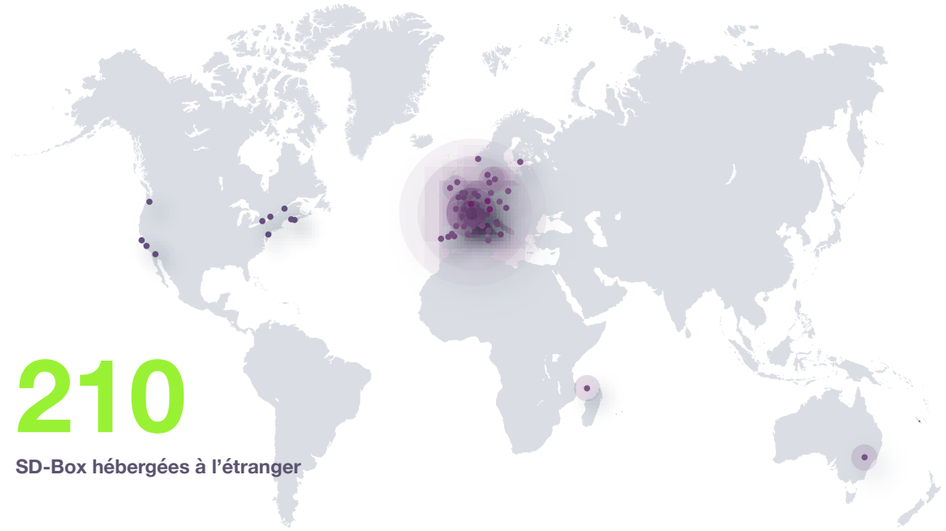
UNE PRÉSENCE FORTE À *L'INTERNATIONAL*

Dès sa création et grâce à l'accord des producteurs de données, le CASD s'est engagé dans l'ouverture des données confidentielles pour la recherche à l'international. Le CASD est particulièrement en pointe sur cette question.

L'accès sécurisé à distance est possible depuis les pays de l'Union européenne et les pays associés à l'UE dans les mêmes conditions que pour les chercheurs des universités et centres de recherche en France. C'est également le cas pour les pays disposant d'une décision d'adéquation pour les données personnelles de la Commission européenne et pour les pays d'Amérique du Nord (USA et Canada) sous certaines conditions. Ces accès depuis l'étranger sont autant utilisés par des chercheurs français en mobilité dans des universités à l'étranger que par des chercheurs de ces pays, avec de nombreuses coopérations impliquant souvent plusieurs institutions de différents pays qui peuvent travailler ensemble dans l'environnement de recherche commun pour le projet que met à disposition le CASD.

Le CASD participe également à des projets visant à faciliter l'utilisation des données confidentielles par delà les frontières nationales. Avec le réseau IDAN (International Data Access Network) qu'il coordonne, il s'attache à bâtir des coopérations en ce sens avec plusieurs centres sécurisés d'autres pays.

// NOS UTILISATEURS *À L'ÉTRANGER*



210

SD-Box hébergées à l'étranger

Les SD-Box à l'étranger représentent maintenant environ 20 % du total des SD-Box, un taux élevé et constant

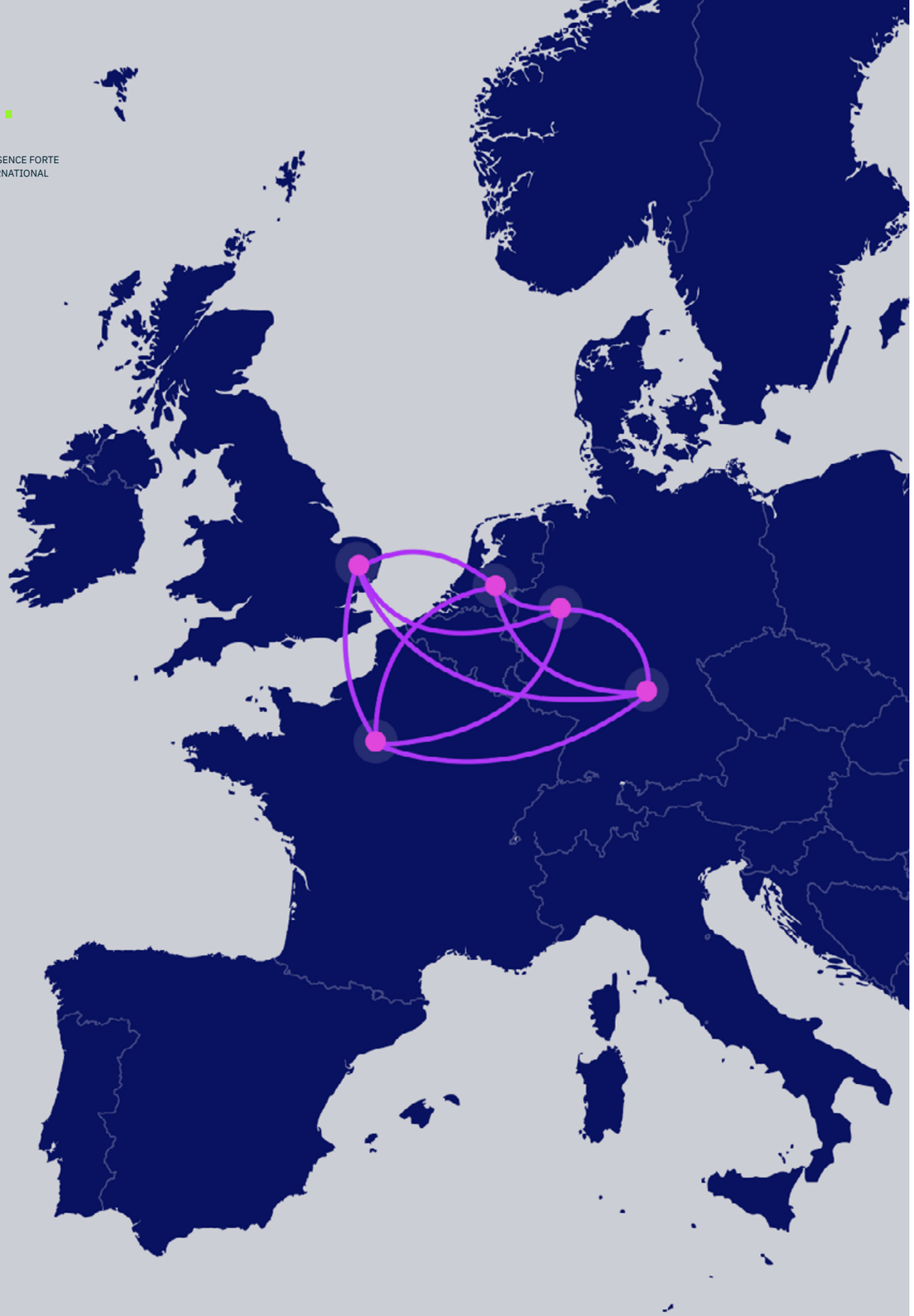
En décembre 2024, 210 SD-Box étaient hébergées à l'étranger dans 18 pays depuis lesquels des chercheurs travaillent sur les données françaises.

Les pays en tête en 2024 pour le nombre de SD-Box hébergées étaient :

Pays-Bas, Italie, Belgique, Royaume-Uni, Allemagne, Etats-Unis, Suisse, Canada, Espagne, Autriche, Suède, Luxembourg, Danemark, Norvège, Irlande, Portugal, Finlande, Australie.

En 2024 un projet depuis un nouveau pays, l'Australie a été autorisé.

L'enrôlement sur place au CASD est indispensable pour pouvoir travailler sur la SD-Box. Si pour nombre de chercheurs, le déplacement est aussi une opportunité de venir voir des collègues en France, discuter avec eux de projets pour certains menés en commun sur les données françaises, pour certains ce peut être une contrainte d'autant plus forte que le voyage est long et coûteux. La pandémie a rendu cette contrainte bien visible. Le Comité du secret statistique avait, pour la procédure d'accréditation, progressivement levé la contrainte d'une présence sur place. Après la validation opérationnelle de sa technologie d'enrôlement à distance expérimentée en 2023, le CASD a poursuivi en 2024 ses travaux d'élaboration du protocole encadrant l'opération afin de lui conférer le même niveau de sécurité que lorsqu'elle se déroule dans ses locaux.



// IDAN : *LES DONNÉES ALLEMANDES ET BRITANNIQUES* ACCESSIBLES DEPUIS LE CASD

Faciliter l'utilisation sécurisé de données de plusieurs pays est l'objectif de IDAN (International Data Access Network) https://idan.network/ le réseau de centres sécurisés que coordonne le CASD. Dans une première étape, des accords sont passés pour permettre depuis les sites physiques de chacun des partenaires l'accès à distance aux données de tous les partenaires.

- Comme les années précédentes, plusieurs cher cheurs ont ainsi travaillé depuis la safe room IDAN du CASD sur les données allemandes d'IAB fdz.iab.de/en/data-access/, cependant que d'autres utilisaient les données françaises depuis le GESIS à Mannheim www.gesis.org/en/gml/safe-room-mannheim
- En 2024, c'est l'accès sécurisé distant aux données très détaillées ou sensibles disponibles au UKDA (UK Data Archive) www.ukdataservice.ac.uk qui a été ouvert depuis le CASD en France.
- Il est ainsi maintenant possible au CASD de travailler sur des données allemandes, britanniques et françaises.
- Toutes les informations sur les catalogues, les procédures d'autorisation et les procédures de réservation une fois les autorisations obtenues sur le site IDAN et sur le site du CASD.

https://idan.network

// LE CASD *SOLLICITÉ* À L'INTERNATIONAL

Acteur reconnu à l'international, où il est sollicité pour son expertise, le CASD intervient régulièrement dans des conférences ou des séminaires.

Le CASD a été partenaire de plusieurs conferences internationales

La conférence COSMOS (11-12 avril 2024) organisée par l'INSEE sur les metadonnées qui a réuni plus d'une centaine de participants d'une vingtaine de pays. Le CASD y est intervenu avec l'INSEE pour presenter leurs travaux d'échanges de métadonnées qui visent à faciliter et améliorer la mise à disposition de metadonnées pour les utilisateurs http://cosmos-conference.org/2024/

CESS (Paris 15-16 octobre 2024) sur le thème « L'agenda au-delà du PIB : passé, présent, visions pour l'avenir ». L'édition 2024 était organisée conjointement, sous l'égide de l'Esac, par l'INSEE, la Banque de France, Eurostat et la BCE, l'École d'économie de Paris (PSE) ainsi que le CNIS et le CASD. Le CASD était partenaire de la conférence. Kamel Gadouche, directeur du CASD et Olivier de Bandt de la Banque de France y ont présenté les nouvelles possibilités offertes en termes de données grâce au partenariat qu'ils ont établi. Les avancées sur l'accès aux données administratives dans le cadre du réseau de centres sécurisés IDAN coordonné par le CASD y ont été également présentées. Les supports de présentation sont disponibles en ligne.

Le CASD est également intervenu en 2024 avec des présentations dans plusieurs autres conférences

Qualidata : Le CASD a participé à la 11ème Conférence européenne sur la qualité des statistiques officielles, Q2024 https://www.q2024.pt/ organisée par Statistics Portugal et Eurostat du 4 au 7 juin 2024, lors de laquelle Halima Bakia, Ifaliana Rakotoarisoa de l'équipe Data management du CASD ont présenté avec Thomas Dubois de l'INSEE l'expérimentation que le CASD mène avec l'INSEE sur l'échange de métadonnées de documentation au standard DDI (Data Documentation Initiative), permettant des gains significatifs en termes de fiabilité de la documentation des données, en accélérant l'affichage de la documentation sur le site du CASD

Le CASD a participé au **séminaire Mapineq** « Improving Accessibility, Harmonisation and Data Linkage in Europe »[Le séminaire sur YouTube]

Le CASD accueille régulièrement des visiteurs de centres étrangers intéressés par le développement du CASD, sa technologie et son positionnement. En 2024, il a accueilli à deux reprises l'instance britannique en charge des développements des infrastructures de recherche (UKRI), et notamment son directeur Richard Welpton.

— " CASD is leading the way in secure data access to confidential and sensitive data sources for research through its innovation and client-based approach. By develping technologies and makint research support interesting an rewarding CASD is showing us how to scale up research capacity in trusted research environment "

Richard Welpton - Blog

05.

GOUVERNANCE

Créé en 2010 par l'INSEE, le CASD, après avoir bénéficié d'un financement dans le cadre du Plan Investissements d'Avenir (dispositif « Equipements d'Excellence » ou Equipex de 2011 à 2019), a pris la forme d'un Groupement d'intérêt public (GIP) créé par un arrêté interministériel du 29 décembre 2018. Le GIP rassemble l'État représenté par l'INSEE, le Genes, le CNRS, l'École polytechnique, HEC Paris et la Banque de France.

Il a pour objet principal d'organiser et de mettre en œuvre des services d'accès sécurisé pour les données confidentielles à des fins non lucratives de recherche, d'étude, d'évaluation ou d'innovation. Il a également pour mission de valoriser la technologie développée pour sécuriser l'accès aux données dans le secteur public et dans le secteur privé.

Ses différentes instances sont :

L'assemblée générale

Le conseil scientifique

Le comité des producteurs

Le CASD, dirigé par Kamel Gadouche, est organisé autour de plusieurs services : PMS (Project Management Service) ; DMS (Data Management Service) ; IT-DS (IT-Datascience) ; R&D (Recherche & Développement)

// LES DIFFÉRENTES INSTANCES DU CASD *SE RÉUNISSENT RÉGULIÈREMENT — 2024*

Catherine GAUDY
Directrice générale du GENES

- **L'Assemblée générale,** présidée par Catherine Gaudy, Directrice Générale du Genes, a tenu ses réunions le 21 juin 2024 et le 02 décembre 2024

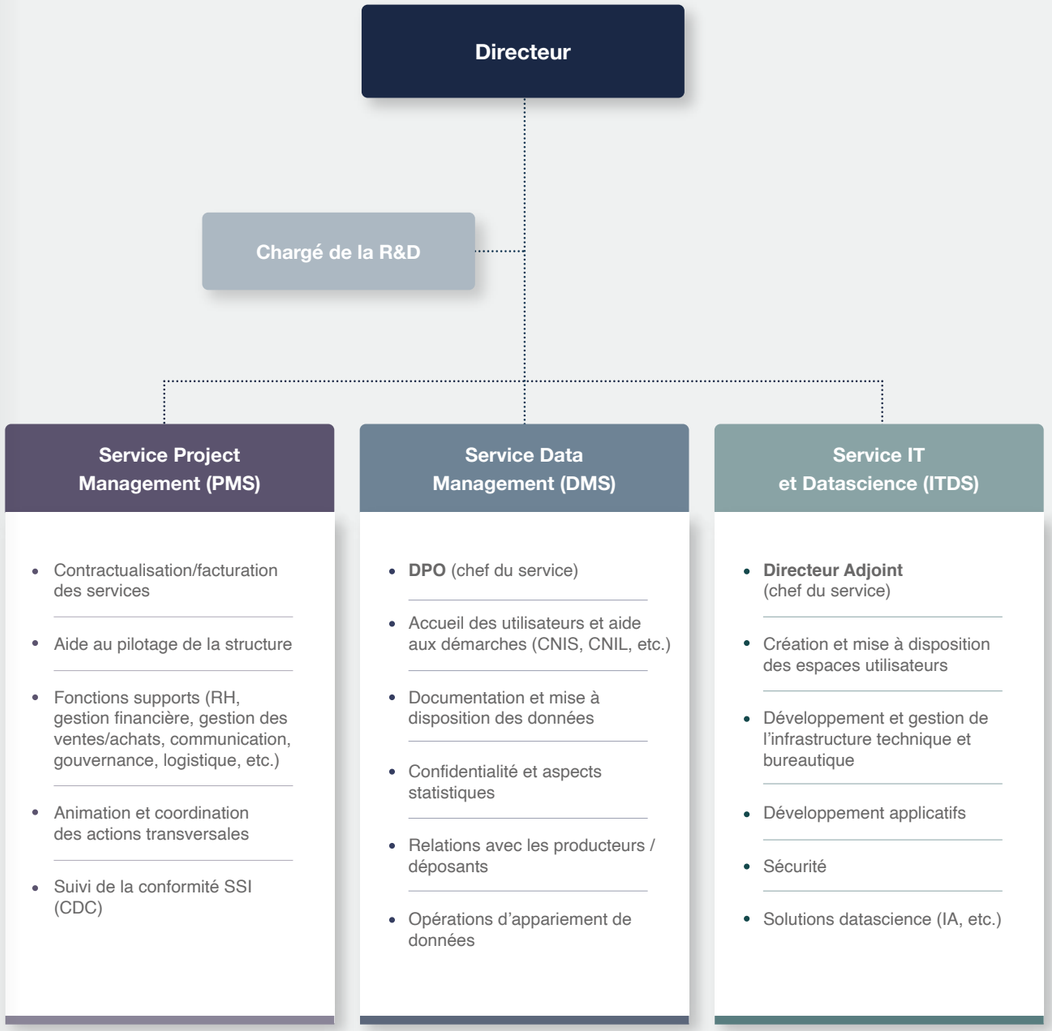
- **Le Conseil scientifique** a accueilli deux nouveaux membres :
  - Romain Lesur, administrateur hors classe de l'INSEE et chef du SSP Lab de l'INSEE, il représente l'INSEE au Conseil scientifique du CASD.
  - Paola Tubaro, directrice de recherche au Centre national de la recherche scientifique (CNRS) et membre du Centre de Recherche en Economie et Statistique (CREST). Ses recherches portent notamment sur l'économie des plateformes numériques, les réseaux de production mondiaux de l'industrie de l'intelligence artificielle, le rôle du travail humain dans le développement de l'automatisation et les inégalités numériques.

  Il a tenu deux réunions les 31 janvier et 10 octobre 2024 sous la présidence de Lars Vilhuber (Cornell University). Lors de ces séances, les présentations et les échanges ont porté sur :
  - l'avancement des axes stratégiques de développement du CASD,
  - le contrôle d'anonymisation des exports de résultat et le projet Colysée,
  - les derniers développements du CASD (ergonomie, outils, interface, confidentialité…),
  - les usages à venir des chercheurs.

- **Le Comité des producteurs,** présidé par Christel Colin, Directrice des statistiques démographiques à l'INSEE, s'est réuni le 22 novembre 2024. Les discussions ont porté autour :
  - des macro indicateurs de l'étude de coût CASD,
  - l'animation d'une communauté de chercheurs utilisant les données confidentielles,
  - la documentation des données,
  - le format de stockage des données sources.

ANNEXES

ORGANIGRAMME *FONCTIONNEL* DU CASD



— 2024 INDICATEURS *CLEFS*

+ 19,5%

Augmentation du nombre d'utilisateurs par rapport à décembre 2023

+ 26%

Augmentation du nombre de SD-Box déployées et suivies par rapport à décembre 2023 [boitiers connectés]

+ 20%

Augmentation du nombre d'organisations co-contractantes par rapport à décembre 2023

+ 5%

Augmentation du nombre de projets gérés et facturés : + 5 % par rapport à décembre 2023

+ 7 ETP

Evolution de l'effectif géré en fin d'année par rapport à décembre 2023

+ 5%

Budget géré par rapport au budget 2023 (augmentation du Chiffre d'affaires : + 29% par rapport à décembre 2023)

- **Pour accompagner la dynamique de son activité et faire face à la montée en charge, le CASD a renforcé ses équipes en accueillant 7 nouveaux collaborateurs en 2024**

- **le GIP a engagé une réflexion sur son organisation interne qui permettrait d'offrir un meilleur accompagnement à ses utilisateurs en mutualisant certaines missions tout en sanctuarisant les fonctions support :**
  - **dans une optique d'amélioration de l'existant**
  - **pour répondre à des exigences de qualité de service et d'accountability toujours plus élevées**
  - **dans un souci de respect des engagements du CASD à l'égard de ses diverses parties prenantes (membres du GIP, producteurs de données, usagers, administrations et pouvoirs publics…)**

- **Axes de développement 2022-2025 : 2024, l'année de la concrétisation et de la préparation d'un nouveau cycle stratégique en vue de répondre à de nouveaux enjeux de passage à l'échelle et d'usages diversifiés (montée à l'échelle de l'infrastructure, datascience / IA, traitements de données massives issues des grandes plateformes de données, démarche qualité, structuration d'une communauté d'utilisateurs des données, etc.)**

— *L'AVANCEMENT* DES AXES STRATÉGIQUES

AXE 01

Développement et renforcement de l'offre technologique

- **Développement technologie cloud privé datascience (Spark, ML) :** Développement cluster SPARK / HDFS à grande échelle pour la Dares (MIDAS)
- **Développement entrées automatiques de codes :** Réalisé (cofinancé par DREES/DARES)
- **Développement enrôlement biométrique à distance :** Réalisé, testé, à valider avec la CNIL puis mettre en production
- **Développement machine learning sur les sorties de fichiers (Colysée) :** Réalisé, testé, en production
- **Analyse des vidéos de session (computer vision) :** Proof of concept en cours
- **Amélioration du temps de connexion et switch entre bulle :** en production (passage de 45 secondes à 15)
- **Sécurité :** certification avec succès pour ISO 27001, ISO 27701, HDS

85%

AXE 02

Consolidation et développement de l'offre data

- **Documentation DDI des données / traduction de la documentation**
- **Développement d'une nouvelle interface pour la présentation de la documentation sur le site du CASD**
- **Appariements :** FORCE, MIDAS, BADS2 (PSE), SAMU
- **Développement des collaborations avec les producteurs** (Infrasec-DMA)
- **Formation sur les données entreprises,** Fideli, DGFiP….
- **VTL** (Validation et Transformation Language) intégré dans l'infrastructure CASD et intégration DuckDB/parquet

75%

AXE 03

Automatisation fonctionnement interne

- **Développement du front office (CDAP) et du Back Office (ROME)**
- **Interface utilisateur CDAP** (suivi projets/abonnement), **Intégration des données de la BDF**
- **Intégration du module facturation dans le SI interne**

80%

AXE 04

International et relations avec les autres HUBS

- **Développement collaborations avec autres centres (IDAN, etc.)**
- **Début d'élargissement du réseau IDAN** (International data access network) au-delà des membres fondateurs

75%

AXE 05

Valorisation de la technologie

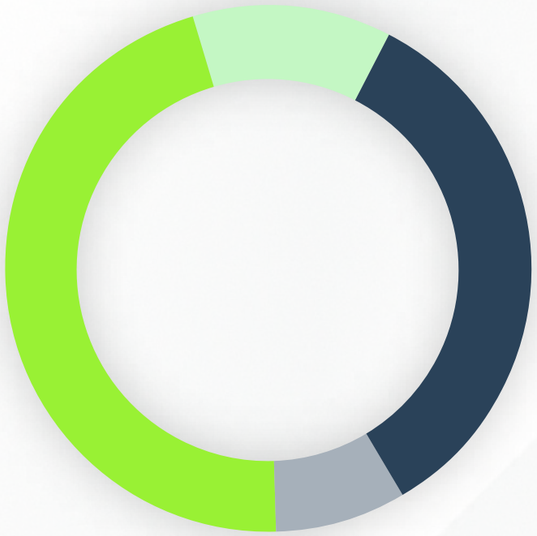
- **Développement Commercial**
- Vente Meta, TikTok pour des pilotes, contact avec Google et Microsoft Linkedin

55%

— ASPECTS *FINANCIERS*

6M€

En 2024, un budget de près de 6 M€ réparti comme suit :



Investissements :
Amortissement :
12% | En volume annuel d'achat : ~ 500K€

Charges de fonctionnement :
34%

Charges de personnel
46%

Impôts et taxes
8%

L'ACTIVITÉ DES *DIFFÉRENTS SERVICES*

le service

PROJECT MANAGEMENT SERVICE — *(PMS)*

Missions

- **Administration des ventes : Relation et conseils aux utilisateurs sur les aspects contractuels, de financement et d'expression de leur besoin ; suivi du déploiement des services et de leur facturation**
- **Administration des Ressources Humaines**
- **Gestion budgétaire / comptabilité**
- **Facility management / logistique**
- **Soutien à la gouvernance et secrétariat de direction**
- **SMSI (Système de Management de la Sécurité de l'Information) et conformité aux certifications**
- **Communication**
- **Soutien au pilotage des projets transversaux de la structure**

Chef de service :
Tanguy Libes

L'année 2024 pour le PMS

- Recrutements : Afin de faire face à l'évolution de l'effectif, le pôle RH du PMS s'est vu renforcé d'un second poste

- Début des travaux sur l'automatisation des principaux processus métiers (onboarding des projets utilisateurs, réflexion sur l'automatisation de l'activation / désactivation des services, etc.) afin de pouvoir concentrer l'effort sur les tâches à plus forte valeur ajoutée et le conseil aux utilisateurs ; en préfiguration d'un « front office » web.

- Coordination de la mise en place de nouvelles modalités de gestion et d'accueil de projets spécifiques, dont les enjeux diffèrent de ceux des projets de recherche et d'étude classiques :

  - Coordination des modalités d'accueil et de déploiement du pilote META sur les données Instagram pour la recherche sur le bien-être

  - Prise en charge des projets portés par les organisations internationales utilisatrices, avec un encadrement juridique répondant à leurs spécificités (OCDE, FMI, etc.)

  - Mise en œuvre des nouvelles modalités de prise en charge et de gestion des infrastructures sécurisées de production et d'analyse statistique des SSM des ministère sociaux (DREES / DARES)

- Contribution à l'accompagnement de projets transverses (enrôlement distant, SSO biométrique, évolution du clausier contractuel encadrant le déploiement des services, pilotage et amélioration du SMSI et des documents sous-jacents, agrandissement et sécurité des locaux occupés, etc.)

le service

RECHERCHE & DÉVELOPPEMENT — *(R&D)*

Missions

Créée en 2015, la R&D du CASD voit ses missions évoluer au cours du temps selon les besoins du GIP.

- **Elle mène des études, généralement mais pas exclusivement, dans des domaines technologiques.**
- **Elle réalise des projets complets destinés à la production.**
- **Elle assure des fonctions de veille technologique.**
- **Elle est chargée de la conception, des évolutions et du suivi de la fabrication des SD-Box. Elle assure pour cette fonction le choix et l'approvisionnement des composants électroniques, du système d'exploitation et des cartes à puce ainsi que la veille associée à ces composants, y compris de leurs logiciels associés.**

La R&D peut être saisie d'une problématique par la direction ou les différents services du CASD, généralement autour de la SD-Box ; sur ses aspects matériels et logiciels mais également sur ses interactions avec l'environnement sécurisé auquel elle se connecte et certains aspects légaux comme les règles du transport aérien, la douane et parfois même en dehors de sa sphère habituelle comme lors de la création de salles serveurs en free-cooling.

La R&D peut aussi s'autosaisir de tous sujets, généralement liés à l'environnement informatique du CASD ; afin de faire des propositions au comité de direction ou aux services concernés.

La croissance du CASD et l'accumulation des générations des composants matériels et logiciels complexifient la gestion et l'évolution du parc. Par exemple le CASD supporte actuellement 9 variantes de carte à puce. L'anticipation des changements en devient d'autant plus importante.

Chef de service :
Philippe Donnay

Quelques études menées par la R&D en 2024

- Évaluation des changements à apporter aux certificats et à leurs conteneurs sur les cartes à puce.
- Évaluation des protections du microcode des processeurs de la SD-Box.
- Protocoles d'authentification système et leur utilisation par le CASD.
- Renouvellement à distance des certificats des cartes à puce.
- Attestation matériel pour les SD-Box.
- Évolution du système de connexion des SD-Box vers l'infrastructure CASD.
- Évolution du système d'exploitation des SD-Box.

le service

DATA MANAGEMENT SERVICE *(DMS)*

Missions

Le Data Management Service remplit principalement une partie des tâches dites d'exploitation : ouverture des droits pour les utilisateurs, réception et mise à disposition des données, relations avec les producteurs et les utilisateurs, vérifications des sorties de résultats. Outre cette activité principale, le DMS assure les missions suivantes :

- Veille réglementaire sur la protection des données, le secret statistique et le droit en cybersécurité : RGPD, loi 51-511, Code pénal, référentiels de la CNIL notamment. Le DMS est à ce titre en relation régulière avec la CNIL et participe au comité du secret statistique (CSS)
- Dialogue et suivi des conventions avec les producteurs de données ou les porteurs de projets
- Réalisation d'opérations de Tiers de confiance en charge des appariements, nécessitant l'accès restreint à des données directement identifiantes
- Documentation des sources de données mises à disposition, y compris au niveau des variables et disponible en libre accès sur le site du CASD
- Analyse de l'activité du CASD vis-à-vis des projets des utilisateurs : nombres de projets, nombre de sorties, etc. et des profils d'utilisation
- Mise en place et exploitation de l'enquête de satisfaction des utilisateurs du CASD

Chef de service :
Rémy Marquier

En 2024, le DMS a réalisé d'autres tâches importantes :

- La réalisation d'appariements au titre du rôle de « Tiers de confiance » du CASD : les dispositifs FORCE et MIDAS de la DARES continuent et le DMS continue d'accompagner les projets d'appariements d'utilisateurs.
- La veille réglementaire et juridique : le dialogue avec la CNIL est fourni, en particulier sur les référentiels relatifs au traitement des données de santé. Le DMS accompagne également les porteurs de projets dans leurs démarches de mise en conformité, notamment pour la mise en place d'entrepôts des données de santé.
- L'exploitation de la base de données de gestion du CASD : la mise en forme statistique de cette base, regroupant toute l'activité d'exploitation, permettra de réaliser des analyses approfondies sur les utilisations du CASD : typologies de projets, d'utilisateurs, etc.
- L'enrichissement de la documentation des données des producteurs, ainsi que les travaux sur la refonte de la page documentaire du site casd.eu. L'équipe DMS a également participé à deux conférences internationales relatives aux standards internationaux de métadonnées et au partage de connaissances dans le domaine : COSMOS (Conference on Smart Metadata for Official Statistics) et QSTAT (European Conference on Quality in Official Statistics)
- L'organisation d'une séance de présentation des données entreprises de la DGFiP
- Le rôle de MOA pour les outils Colysée (vérification semi-automatique des sorties de certains projets) et Vespa (amélioration du flux des sorties de résultats des utilisateurs)

le service

IT & DATASCIENCE *(IT&DS)*

Missions

Le service ITDS est composé de 3 services complémentaires, ses missions principales sont la conception, le déploiement, la maintenance et l'optimisation de l'infrastructure

Trois analystes-développeurs de l'équipe DEV-WEB assurent les évolutions, la documentation et la maintenance des applicatifs web du CASD : son site casd.eu, l'ensemble des modules du portail CDAP, l'application interne ROME supportant l'exploitation des bulles sécurisées et les sites webs externes (fiche documentation des DOIs, enquête de satisfaction, quiz pour la sensibilisation etc).

En 2024 leurs travaux ont majoritairement porté sur l'interface CASD de CDAP, notamment l'intégration des nouvelles fonctionnalités INFRASEC (import de texte, changement d'IP depuis le domicile), l'automatisation de l'activation de services PMS (déclenchement d'actions à une date programmée ou suite à la complétion de pré-requis contractuels : activation des comptes, création des espaces projets, changement de configuration matérielle, envoi de facture) et la prise en compte de besoins spécifiques de la Banque de France pour les autorisations d'accès à ses données : intégration souple dans le module CSS, gestion d'une contractualisation supplémentaire à plusieurs niveaux comportant des signatures électroniques, gestion sur mesure de son référentiel de source.

L'équipe Datascience compte 3 datascientists qui forment un pool de compétences de haut niveau sur un ensemble de domaines très étendus, à destination des utilisateurs du CASD : mise en place cluster Spark, l'optimisation de code Python, assistance au choix de configuration matérielle, réalisation de comparaisons de performances entre diverses piles logicielles & matérielles voire de format de fichiers. L'équipe Datascience est aussi mobilisée pour assurer des besoins internes du CASD : utilisation de modèles entrainés sur l'historique des exports acceptés et refusés pour prioriser les contrôles (Colysée), modèle de classification pour les entrées au format texte, détection avancée de comportements anormaux via exploitation des journaux de sécurité, programme de conversion de format de fichier de très gros volume, documentation de l'intégration d'outils datascience/IA au CASD)

Leurs travaux en 2024 ont permis de renforcer les capacités de Colysée (OCR sur les fichiers images, amélioration continue du modèle sous-jacent, travaux exploratoires sur l'utilisation de LLM), de tester des outils basés sur des modèles de langage de grande taille (LLM) au sein des bulles sécurisées, de proposer une infrastructure tant matérielle (Hyperviseur dédié 2To de RAM, 2 NVIDIA H100 NVL) que logicielle (passage au format parquet, gestion performante de graphes distribués sous Spark)

Chef de service :
Éric Debonnel

l'année 2024 pour le service IT & Datascience

- Leurs missions principales sont la conception, le déploiement, la maintenance et l'optimisation de l'infrastructure informatique du CASD (120 serveurs physiques, 1500 serveurs virtuels, 1300 SD-Box, 2Po de capacité de sauvegarde). Ils assurent aussi l'assistance aux utilisateurs, la mise en œuvre des mesures permettant les certifications techniques du CASD renforçant son auditabilité.

- En 2024, ils ont concentré leurs efforts sur la disponibilité de l'infrastructure (4 liens Internet, redondance (régulièrement testée) des services critiques, transformation du site secondaire de secours en équivalent du site principal et croisement des sauvegardes)), une automatisation croissante des processus d'exploitation courante, l'offre de nouveaux services (INFRASEC, VESPA, environnements sous Linux) et la constitution d'une base de connaissances pour faciliter l'assistance aux utilisateurs.

casd.eu