CASD

— ACTIVITY REPORT 2024

CASD IS A PUBLIC INTEREST GROUP (CONSORTIUM) BRINGING TOGETHER:

CONTENTS

Director of publication
Kamel Gadouche

Production/editing
Roxane Silberman

Graphic design
Guillaume Léonardy — boreaone@gmail.com
Crédits photos : AdobeStock

CASD

A WORD FROM THE *DIRECTOR*

Being able to use the immense resources of large individual databases, from public statistics, government agencies, or other public or private organisations, is a major concern for research and for assessing public policies. CASD strives to contribute positively to this issue by continuing to develop our activities in 2024, with the support of data producers, to enable the secure processing of these data sources, while avoiding silos between the major areas of the economy, the environment, and health.

In 2024, the number of CASD users both in France and abroad continued to increase. CASD is committed to providing everyone with an increasingly seamless service while maintaining the highest level of security necessary to preserve the trust of data producers and society generally regarding the protection of confidentiality. In this vein, we have rolled out new features integrated into the Confidential Data Access Portal (CDAP) which brings together all stakeholders, data producers, and users in collaboration with the Statistical Confidentiality Committee and Banque de France, both of which are responsible for authorising data access requests.

Investments in security, ergonomics, and performance have enabled close collaboration with DREES and DARES to develop secure data science environments dedicated to data management and to create new data sources. As in previous years, CASD has also participated directly in the creation of new data sources as a trusted third party for data matching operations.

As an international actor, with 20% of its SD-Boxes hosted in major universities and research centres abroad, coordinator of the network of secure data access centres (IDAN, International Data Access Network), CASD was chosen by META for a second pilot contract for American research projects seeking to use American Instagram data. CASD is also particularly proud to have been a partner in the 2024 edition of the Conference of European Statistical Stakeholders (CESS), "The Agenda Beyond the GDP: Past, Present and Visions for the Future", hosted in October in Paris by INSEE and Banque de France, under the patronage of ESAC and in partnership with Eurostat, the European Central Bank, PSE and CNIS.

Kamel Gadouche
Director

CASD

DATA FOR RESEARCH, ASSESSMENT OF PUBLIC POLICY AND DATA SCIENCE

01.

CASD provides secure computing environments dedicated to processing large volumes of data, mainly used by researchers and data scientists. CASD acts as a trusted third party between data producers and users for the secure processing of confidential data.

Data entrusted to CASD cover most of the data in public statistics, particularly those of INSEE and a large number of statistical services within ministries (MSS). It also hosts – and provide access to – a growing number of databases from large public agencies and administrations, including medical and administrative data, as well as research data including from large epidemiological cohorts. Since 2022, Banque de France has also made its data available for research through CASD. These sources, which can be used jointly and, in some cases, matched together, enable the development of multi-field work without silos, which is unique in Europe and internationally.

The searchable data source repository enables links to be made between sources, documented in DDI format, research projects, and publications.

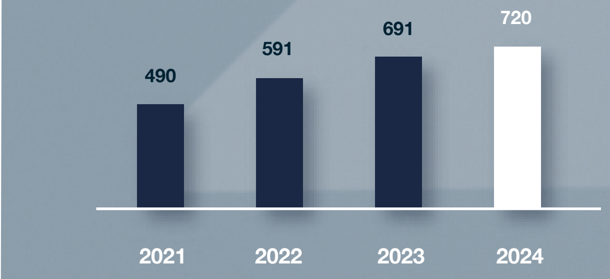
Once accredited, upon assent from the Statistical Confidentiality Committee and/or producers, and registered with CASD, users work remotely with direct access to data in the CASD secure environment via a secure terminal (the SD-Box) from their office within their organisation.

— KEY *INDICATORS*

720

In 2024, 720 projects were conducted with CASD, representing an increase of 4% compared to 2023.

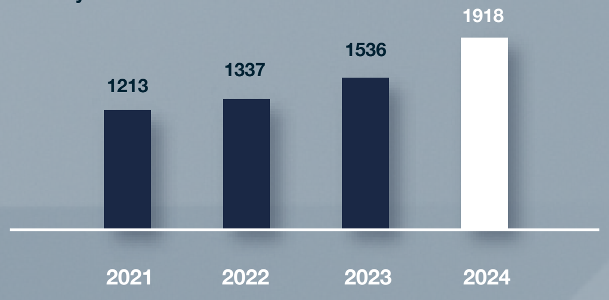
Number of projects conducted through CASD per year



1 918

In 2024, 1918 user accounts were active, representing a 23% increase compared to 2023.

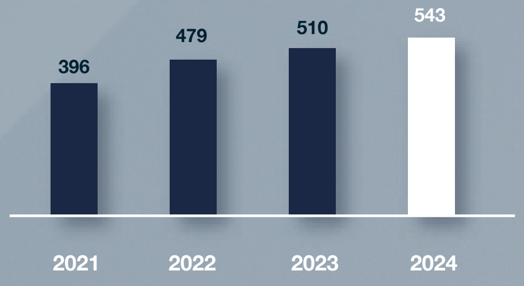
User accounts active in one year



33

In 2024, 33 new data sources were made available through CASD, bringing the total number of sources available to 543.

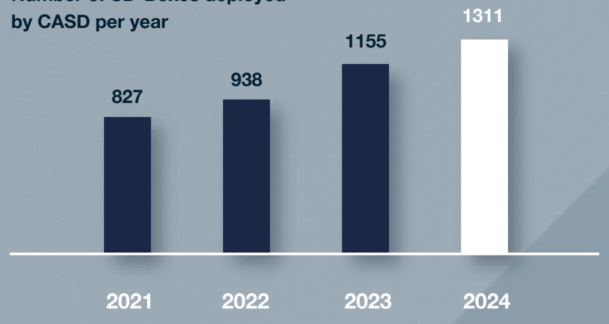
Number of data sources hosted by CASD per year



1 311

In 2024, CASD deployed 1311 SD-Boxes, representing an increase of 14% compared to 2023. Since 2020, CASD has thus doubled the number of SD-Boxes in use, from 652 to 1,311 boxes.

Number of SD-Boxes deployed by CASD per year



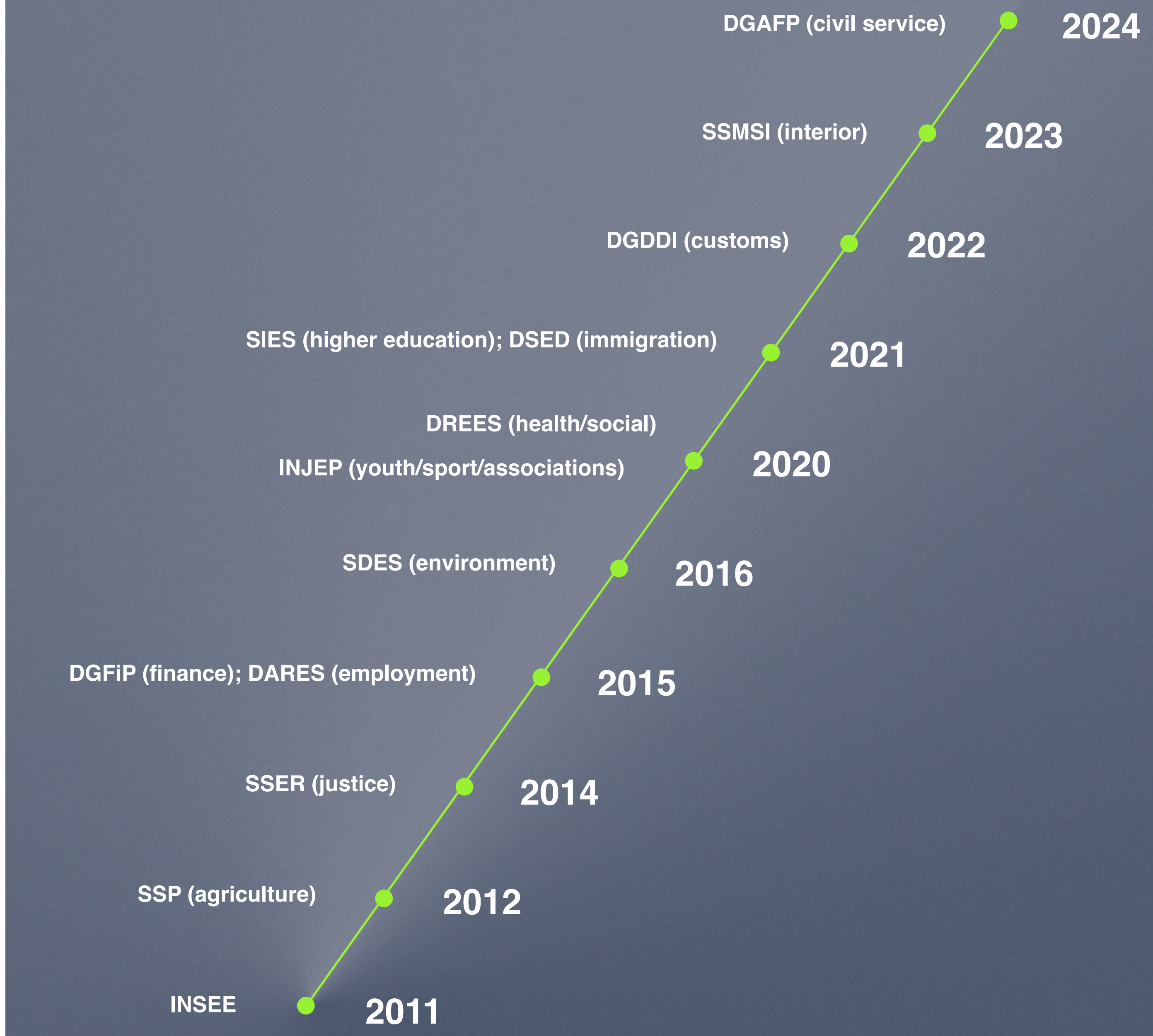
// WIDE ACCESS TO OFFICIAL MICRODATA *FOR RESEARCH*

In 2024, The Sub-Directorate for Studies, Statistics, and Information Systems (SDESSI) of the French Directorate-General for Administration and the Civil Service (DGAFP) joined the 37 data producers that have entrusted a copy of their individual databases to CASD to make them available to users.

An initial dataset was deposited at CASD

The IRA competitive entry process statistical database (BSC IRA) provides administrative and sociodemographic information on candidates registered for civil service competitive entry examinations. Based on the competitive entry process survey, which polls all candidates registered for a specific competitive entry examination, and an administrative database (the BAC) on which the survey is based and which contains information provided by the candidate upon registration, their scores, their progress in the competitive entry process, and information about the competitive entrance process, the BSC IRA provides a comprehensive overview of candidates' progress in the exam, from registration to results, enabling analyses based on their sociodemographic characteristics.

This brings the total to 12 MSS out of 16 for which access is authorised, following approval by the Statistical Confidentiality Committee and agreement from both the MSS and the Archives Administration. In 2024, CASD data thus covered a large part of the scope of public statistics, INSEE, and MSS.



DGAFP (civil service) 2024

SSMSI (interior) 2023

DGDDI (customs) 2022

SIES (higher education); DSED (immigration) 2021

DREES (health/social)
INJEP (youth/sport/associations) 2020

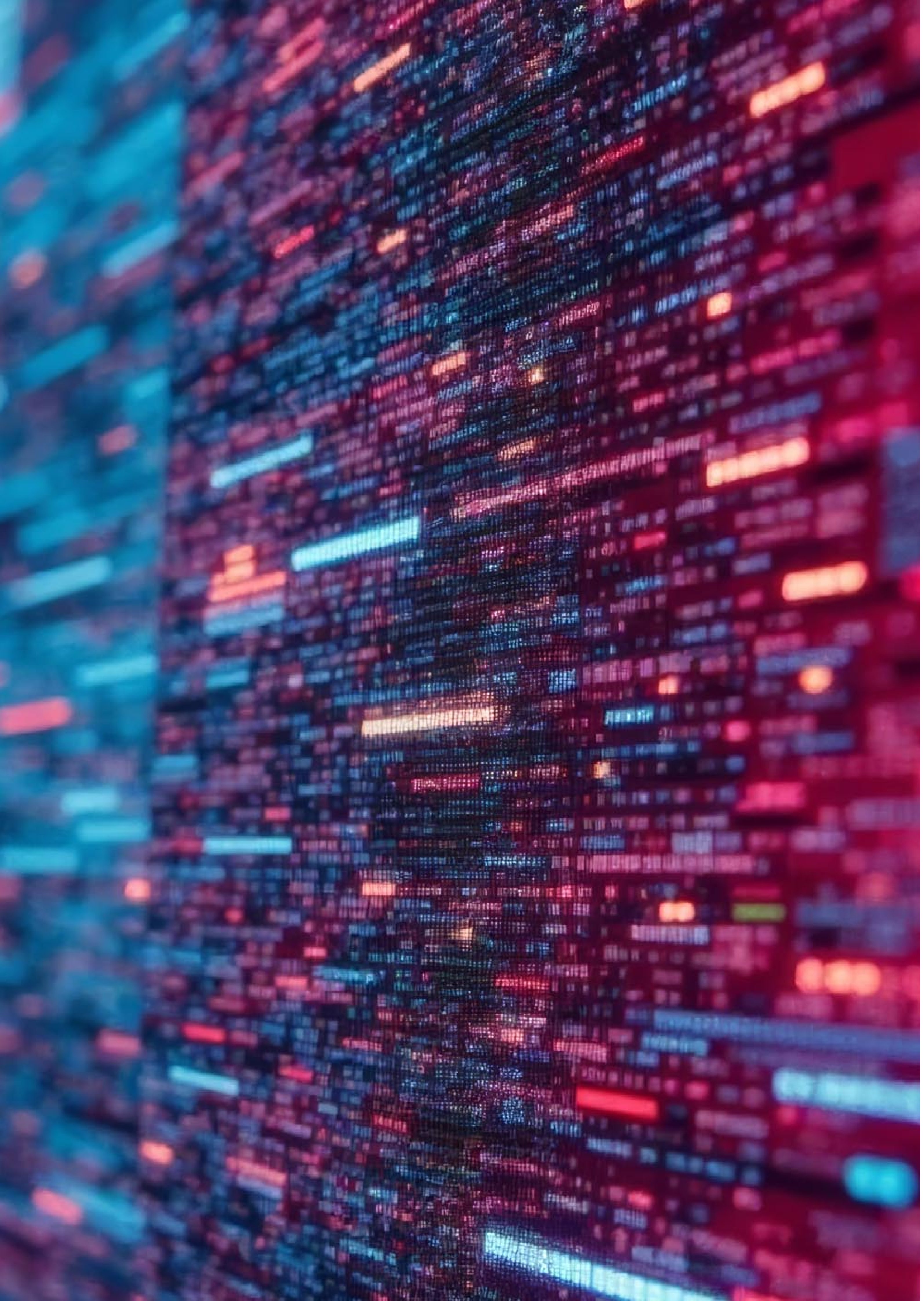
SDES (environment) 2016

DGFiP (finance); DARES (employment) 2015

SSER (justice) 2014

SSP (agriculture) 2012

INSEE 2011



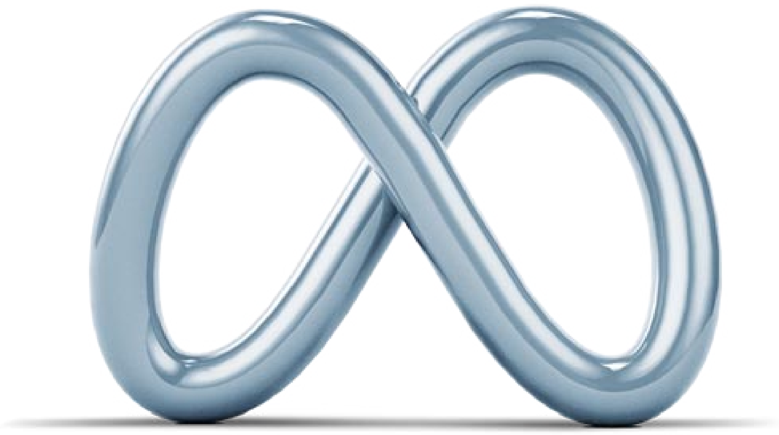
// META IS USING CASD TECHNOLOGY AND *SERVICES TO PROVIDE RESEARCHERS WITH ACCESS TO DATA*

The data gathered by large platforms is often both very rich and sensitive, raising questions about how it can and should be used. With the aim of regulating such data use and addressing potential systemic risks, the Digital Services Act (DSA) requires such large platforms to make their data available to researchers. Researchers have for a long time been interested in using this data for research in various fields.

Prior to the rollout of the DSA, major platforms were already turning to CASD to provide access to data for researchers for pilot projects.

META was seeking an independent and experienced organisation in Europe and signed a contract with CASD in 2023 regarding Facebook data. This contract was drafted in the context of ongoing discussions with EDMO (European Digital Media Organisation) that aimed at establishing an Independent Intermediary Organisation to review access requests, with CASD acting as the trusted third party providing a technical solution for secure access to data.

In 2024, beyond the context of the DSA, META signed a contract for Instagram data with a program enabling up to 7 projects, which will be developed in 2025.

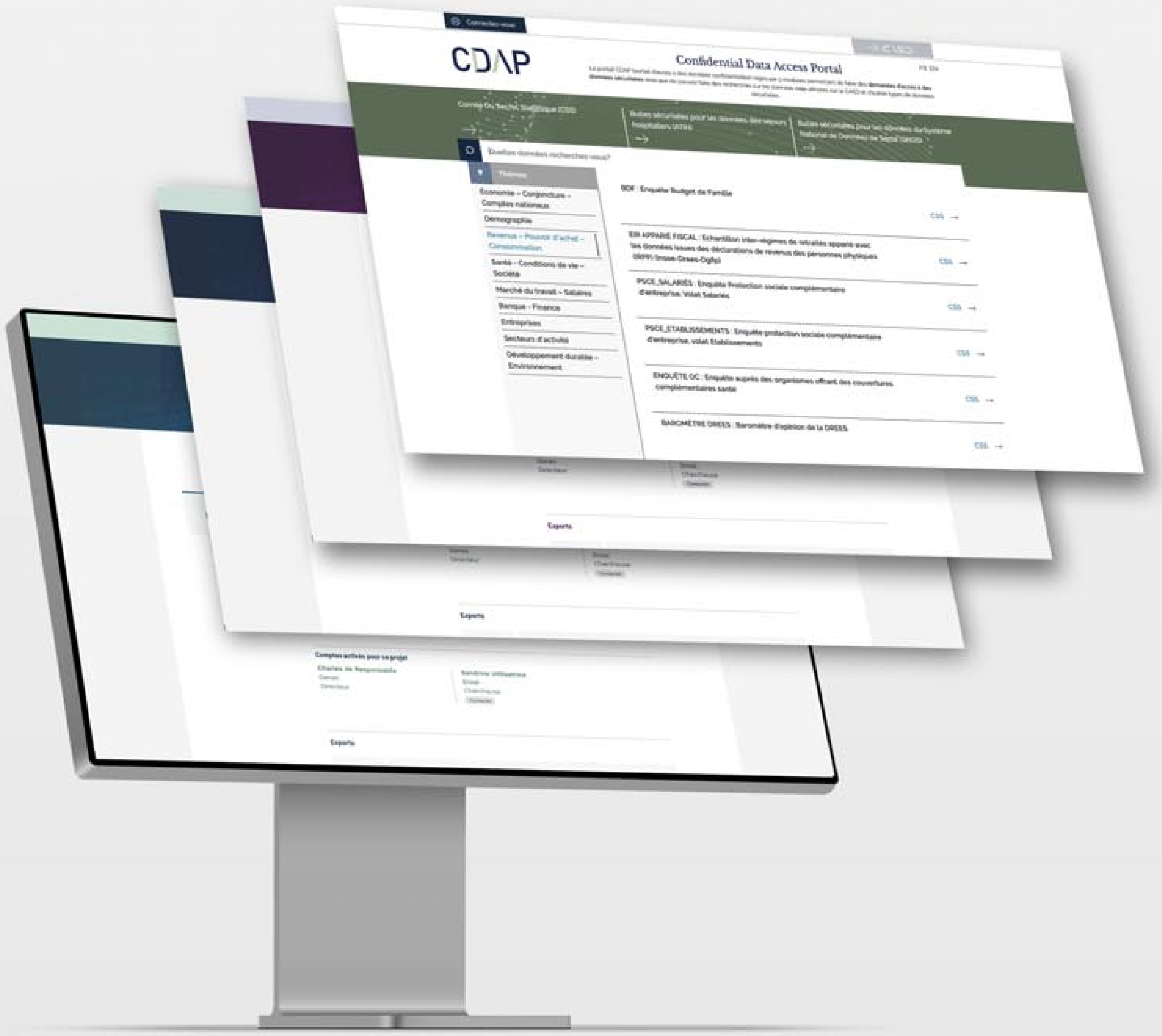


// MORE *INSTITUTIONAL USERS*

To meet the needs of institutional users, the Statistical Confidentiality Committee provides public administrations with a specific and simplified “administrative” procedure, which enables them to add data sources and/or members to their projects outside of the schedule of regular Committee meetings. These members can then work on the CASD platform.

In 2024, 14 public administrations had active access to CASD for their internal work.

- General Directorate of Companies - Ministry of the Economy (DGE)
- General Directorate of the Treasury - Ministry of the Economy (DG Trésor)
- Economic Analysis Council - General Commission for Strategy and Forecasting
- General Commission for Strategy and Forecasting
- General Inspectorate for Social Affairs
- Court of Auditors
- General Inspectorate for the Environment and Sustainable Development
- General Inspectorate for Finance
- Overseas Territories Directorate General - Ministry of Overseas Territories
- Senate
- Unédic
- Department of Housing, Urban Planning, and Landscaping
- Centre for Studies and Expertise on Risks, the Environment, Mobility, and Urban Planning
- General Directorate for Social Cohesion



Connectez-vous
CDAP
Confidential Data Access Portal
FR EN
Comité Du Secret Statistique (CSS)
Bulles sécurisées pour les données des séjours hospitaliers (ATIH)
Bulles sécurisées pour les données du Système National de Données de Santé (SNDS)
Quelles données recherchez-vous?
Thèmes
Économie – Conjoncture – Comptes nationaux
Démographie
Revenus – Pouvoir d'achat – Consommation
Santé – Conditions de vie – Société
Marché du travail – Salaires
Banque – Finance
Entreprises
Secteurs d'activité
Développement durable – Environnement
CSS →
Exports

BDF : Enquête Budget de Famille

EIR APPARIÉ FISCAL : Échantillon inter-régimes de retraités apparié avec les données issues des déclarations de revenus des personnes physiques (IRPP) (Insee-Drees-DgFip)

PSCE_SALARIÉS : Enquête Protection sociale complémentaire d'entreprise, Volet Salariés

PSCE_ETABLISSEMENTS : Enquête protection sociale complémentaire d'entreprise, volet Établissements

ENQUÊTE OC : Enquête auprès des organismes offrant des couvertures complémentaires santé

BAROMÈTRE DREES : Baromètre d'opinion de la DREES

// CDAP: NEW FEATURES *FOR THE CONFIDENTIIAL ACCESS PORTAL*

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.

- In 2024, the CASD interface on the CDAP portal https://cdap.casd.eu/ continued to expand with the introduction of a new feature for retrieving output results. Any request to "export results" from the secure project bubble now follows a new, faster and smoother transmission procedure, no longer requiring email exchanges and the dispatch of a code.
- Once the anonymisation check has been performed by CASD, the results can be retrieved directly for 15 days from the CASD interface of the corresponding project in CDAP, where users will have access to the records of all outputs produced.
- Automatic notifications provide updates on the progress of such requests.
- In 2024, this new feature was rolled out to new users before being gradually extended to all users in 2025.
- CDAP has thereby become a hub for all stakeholders, streamlining processes while maintaining the same level of security dedicated to researchers' work.

// A WIDELY ATTENDED WEBINAR *ON DETAILED TAX DATA*

CASD strives to provide users with the best possible support, with the help of data producers and other stakeholders, to ensure they can take full advantage of the available data sources. These webinars also encourage researchers to provide feedback to data producers, a valuable interaction which also helps to improve data quality and documentation.

On 21st May 2024, in a bid to help users to better understand the data produced by DGFiP (Directorate General for Public Finance) and made available through CASD, CASD and DGFIP co-organised a presentation session on certain tax sources. This session focused on the following data sources:

BIC-IS : Industrial and commercial profits - all regimes

BIC-RN : Industrial and commercial profits - standard regime

BIC-RS : Industrial and commercial profits - standard regime – simplified regime

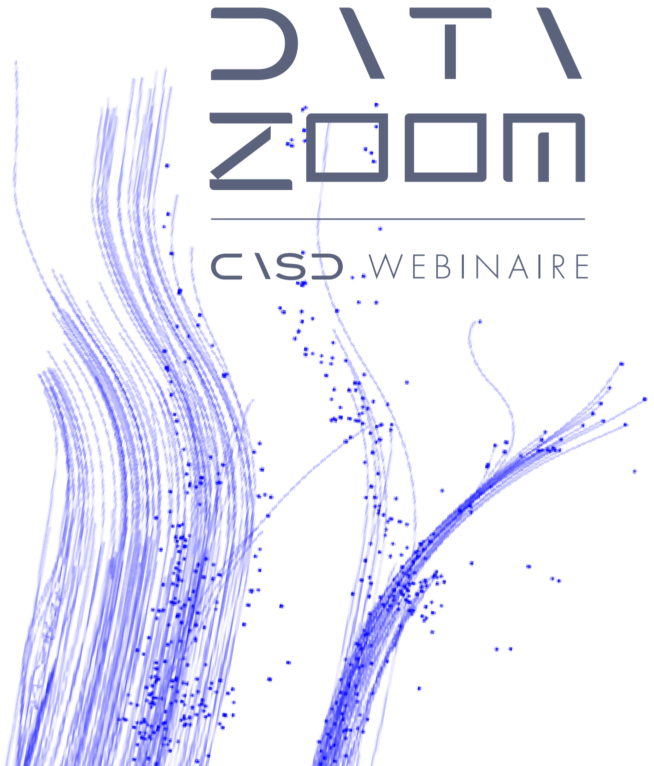
BA : Agricultural profits - standard and simplified regimes

PERIM : Tax group structures

ISGROUPE : Corporate tax groups subject to corporate income tax (CIT)

Professional tax returns: Corporate tax returns

This was presented by Gérard Forgeot, Head of the Statistical Production, Dissemination, and Quality department, and Roddy Caccialupi, of the General Directorate for Public Finances.



// INCREASING USE OF *HEALTH DATA*

The use of highly sensitive health data is subject to specific conditions. CASD hosts a large number of projects using medical and administrative data such as ATIH PMSI data on hospital stays, national health data (SNDS), and data from INSERM cohorts (Constances...).

The CASD infrastructure has been updated to incorporate the two security standards applying to health data, ensuring compliance with (1) the “Health Data Hosting” certification and (2) the health data security reference framework (relating to the statistical use of data from the National Health Data System, SNDS).

More than 100 projects have drawn on health data through CASD.

PMSI (hospital data)	47
INSERM	35
CNAM	17
IRDES	11

// CASD, *A PARTNER* IN THE GRAPH4HEALTH PROJECT

The Graph4Health project was selected in 2023 by the French National Agency for Research (ANR). Coordinated by the Centre for Research in Economics and Statistics (CREST), in collaboration with Inserm's Constances Joint Service Unit, ESSEC Business School and – regarding infrastructure and technical aspects - CASD, the project has begun its initial work. The project studies the development of relationships between patients and healthcare professionals, as well as the structure and evolution of these relationships. To do this, it exploits quasi-exhaustive SNDS data covering 11 years from 2008 to 2018, together with data from the healthcare professionals reference database (RPPS), representing several dozen Terabytes.

In 2024, CASD provided technical expertise which was essential to the Graph4Health project. Specifically, CASD provided:

- **An infrastructure and computing environment with data science software** (Python, R...) and specific hardware resources (double GPGPU NVIDIA H100 NVL)
- **This expertise consisted of implementing calculation "frameworks":**
    - SPARK for distributed calculation
    - SEDONA for work on geospatial data. Geospatial data are notably used to detect medical deserts. The CASD Data Science team has developed computing algorithms to identify the location of patients and doctors/hospitals (in coordinates or GPS location) and calculate the average distance between the two locations to detect the presence of medical deserts
- **The data catalogue in the Openmetadata tool** in which all the tables in the project are listed, along with their location, number of columns, and data lineage (i.e., which raw data tables were used to generate them). This catalogue of data specifically features SNDS data and geolocation data and comprises approximately 300 tables
- **A data matching operation between the postal code and INSEE hospital district code,** and correcting potential errors

SECURE ENVIRONMENTS FOR *CREATING NEW SOURCES*

02.

By their nature, CASD environments enable the processing of confidential data: large scale, sensitive information, etc. It is therefore technically possible, in these highly secure environments, to match various data sources, in particular based on direct identifiers (company registration (French SIRET) numbers, individual civil status, or even social security numbers). These data matching operations, most often performed on a large scale (several hundred thousand or even several million people), enable the creation of new data sources opening up a wealth of possibilities for research.

Public Statistics (with a capital S) have been conducting this type of data matching operations for several years now. This notably includes INSEE's RESIL system for matching individuals and housing, as well as other types of operations, such as surveys on autonomy matched with health data or the creation of inter-regime samples for monitoring contributors to the French pension system coordinated by the French social ministries.

To ensure a high-performance and secure working environment, some data producers use CASD through INFRASEC services, which provide highly customisable secure environments for data management and statistical analysis. Others contact CASD for data matching operations, in which CASD acts as a trusted third party responsible for data matching (see below). The same is true for certain projects led by researchers who, in partnership with one or more government agencies, may lead to the creation of a new source.

The ultimate goal is to generate enriched data sources that should ultimately be made available, once pseudonymised, to the research community.

// DATA PRODUCERS USING CASD SECURE ENVIRONMENT

The ESTRADD project, a collaboration between DREES and DARES (social ministries statistical services), aims to offer computing environments to statisticians / data scientists of the two MSS for their statistical production works: both organisations chose to partner with CASD for the parts that involve the most sensitive processing or require greater computing power combined with ad hoc tools and high stability.

Created in 2023, this project grew throughout 2024 and continues to expand. Each organisation has a secure bubble dedicated to data source administration and another shared bubble for easy information sharing. Other secure bubbles can be installed on request. By the end of 2024, DREES had 17 active bubbles; DARES had seven, including one providing access to the Spark calculating cluster dedicated to MIDAS (a study matching data on minimum social benefits, unemployment insurance rights, and employee career paths), which makes it possible to map the career paths of beneficiaries of unemployment insurance and minimum social benefit.

The statistical services of the Ministry of the Interior (SSMSI) and the Ministry of Justice (SSER) called on CASD to create two environments dedicated to the creation of a new data source reproducing the complete criminal record of individuals in their files. This processing requires a very high level of security, particularly in terms of securing the data from external tampering, the management of identifiers enabling data matching, and the traceability of access.

In the health sector, CASD operates computing infrastructures for IRDES (3 dedicated bubbles according to study and production topics) and for INSERM for the Constances cohort (5 bubbles dedicated to the various stages of cohort construction and exploitation). Constances also allows other research teams to use their data through a governance system specifically set up for this purpose. CASD is also involved in developing an application for processing SNDS health data received by the Constances team within a dedicated bubble. Development of the first module continued throughout 2024 and is expected to be completed in summer 2025.

estradd
DREES DARES
Secure work environment

estradd
Environnement sécurisé
de travail DREES DARES

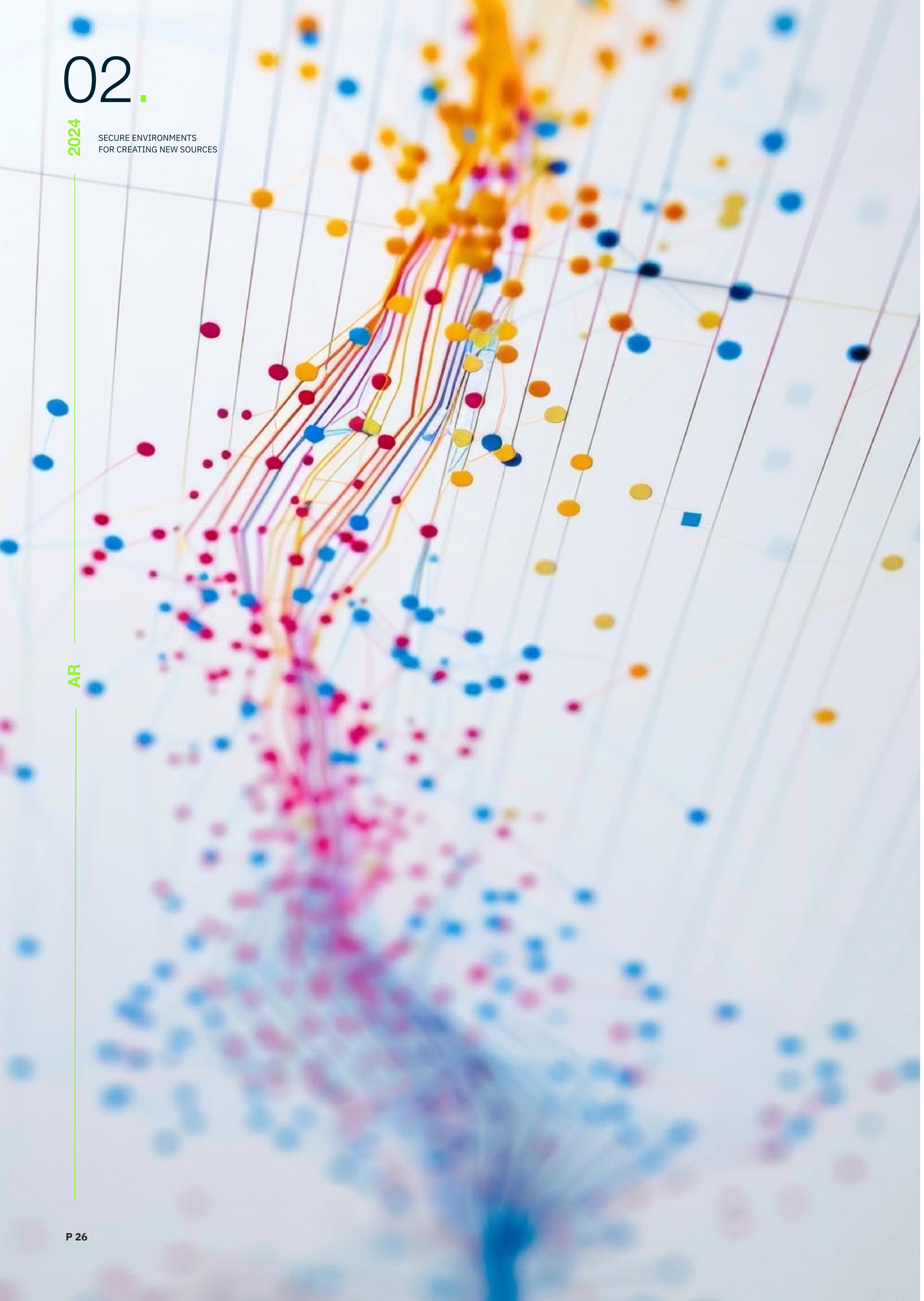
MINISTÈRE
DES SOLIDARITÉS
ET DE LA SANTÉ

Votre session est strictement personnelle,
ne la partagez pas

Prendre en photo ou filmer l'écran
est strictement interdit

Tout manquement à ces obligations
vous expose à des sanctions ou poursuites

DREES DARES – ESTRAD
Secure work environment



// CASD, *A KEY PARTNER* FOR TRUSTED THIRD-PARTY OPERATIONS

In addition to directly hosting production environments for several ministerial statistical services, CASD participates in the creation of new data sources by matching existing sources. These data matching operations, which are mostly based on large-scale administrative sources (several hundred thousand companies or several million people) and confidential data, require sound organisational and regulatory support, advanced data management skills, state-of-the-art security, and the trust of the organisations providing the data, as well as the trust of the data protection authorities. To this end, CASD acts as a trusted third party for matching data or ensuring that individuals are anonymised before data are made available to research teams.

In 2024, CASD continued its biannual contribution to DARES' FORCE and MIDAS data matching operations. FORCE matches data from France Travail (data on French unemployment) with labour movements (MMO data produced by DARES), while MIDAS also combines data from CNAF on recipients of minimum social benefits. These two data sources are made available to research teams upon request, always through the Statistical Confidentiality Committee accreditation process.

The CARE (Capacities and Resources of Senior Citizens) study conducted by DREES for its institutional facet (mainly residents of nursing homes), continues to monitor the mortality of respondents, for which CASD carries out operations in conjunction with INSEE, based on individual social security numbers.

For certain projects processing health data, CASD facilitates dialogue between researchers and CNAM departments (in charge of healthcare reimbursements) with a view to matching data with those of the National Health Data System (SNDS). CASD supports the research teams with data flow management, requiring several compartmentalised environments. During a joint presentation made in June 2024, CASD and CNIL detailed the data flows of varying complexity involved in such projects, aiming at guaranteeing the end-to-end security of these data matching processes.

This type of operation aims to ensure that none of the parties involved in a data matching operation (data producers and researchers) have the same identifiers, thereby strengthening data pseudonymisation measures.

// HEALTH AND *DATA MATCHING*

Data matching is a very valuable opportunity for many research projects to enrich health data.

See the replay

25th June 2024 **CNIL-CASD Webinar**: CNIL, in partnership with CASD, organised a webinar on the subject of data matching for the purpose of health research and studies. This webinar, which brought together nearly 300 participants, was an opportunity to present various use cases and also provided step-by-step illustrations of the data flow patterns to be anticipated when matching several data tables using social security numbers in the health sector. The replay of this webinar is available online.

CASD also participated in drafting practical guidelines with CNIL for data matching protocols with SNDS data.

VADEMECUM

CIRCULATION DU NIR
AUX FINS D'APPARIEMENT AVEC LE SNDS

Produit en collaboration avec le CASD

Version 1.0 – Juin 2024

CNIL.

VADEMECUM
USING SOCIAL SECURITY NUMBERS
FOR THE PURPOSE OF MATCHING WITH
SNDS DATA

03.

SECURITY *AND* *ADVANCED* DATA SCIENCE *TECHNOLOGY*

Data security is central to the CASD mission: in practical terms, more than 700 security measures (including 114 specified by the ISO 27002 standard) are implemented, being organisational, contractual, physical, or IT-related. The entire CASD “secure bubble” system is subject to regular technical audits by ANSSI-certified organisations. Any new functionality undergoes a security analysis right through from design to production. As CASD does not use any subcontractors who could have access to the data entrusted to us, the entire processing chain is controlled internally from start to finish.

Certification for the health data security standard and other certifications obtained by CASD guarantee that the data hosted and the associated processing operations remain optimally secure. CASD constantly monitors technology and European and national regulations in order to comply with these requirements. All procedures are documented, which also ensures the resilience of the teams. Certifications are subject to regular surveillance audits (twice-yearly), with the aim of continuously improving security and data protection management systems.

While security - particularly in terms of availability, integrity, confidentiality, and traceability - is absolutely essential for all stakeholders in the ecosystem, from data producers to end users, the secure bubble working environment must be adapted to users' analysis work, while meeting the highest standards for innovative and particularly resource-intensive calculation methods. In addition to the standard environment and tools dedicated to statistical work (R, Python, Stata, Julia, QGIS), data management (HeidiSQL, DuckDB, etc.) and results edition (Office and OpenOffice suite, LaTeX editor, etc.), CASD is committed to building architectures tailored to specific projects. In each case, the CASD Data Science team assists users in precisely defining their needs and using the tools made available to them.

As such, almost all R packages (CRAN) and Stata add-ons (REPEC) are available natively in the user environment, as are the most commonly used Python packages (and their add-ons). If other packages are provided by the user, CASD can add these to the secure bubble. The IT team can also, on request, install specific software required for certain data science projects. As always, this installation is carried out following a preliminary security analysis to ensure the integrity of the environment.

— THE SD-BOX®
AT THE HEART
OF THE SYSTEM

The SD-Box was designed in 2010 and has been improved over generations. It has become easier to use, more energy efficient and most of its components are now reusable and recyclable.

04

ÉVOLUTION OF THE SD-BOX

01

02

03

— *LAUNCH* OF THE BIO SSO CARD

A few years ago, each project had its own biometric smart card that was required to log into the corresponding virtual environment. Users who had access to multiple projects, and therefore multiple secure bubbles, had to have multiple cards and switch cards each time they wanted to switch between projects. The login time, which included numerous transactions for security checks (certificate validation, etc.), was rather lengthy.

CASD has developed a secure authentication interface that allows both:

- to compartmentalise authorisations into secure bubbles on the smart card, thus obtaining a multi-project card and streamlining the transition between secure bubbles, and
- to optimise the authentication chain validation process and the certificate validation process (by reducing transaction times, for example), thereby reducing connection times.

These developments were made to offer greater convenience for users while maintaining the high level of security required for processing confidential data.

1 172

The rollout has been progressive, and in 2024, 1,172 SSO BIO cards were implemented.

SECURITY, *A CENTRAL PRIORITY:* REGULAR AUDITS AND PREPARING FOR EUROPRIVACY CERTIFICATION

With security being central to all operations, certifications and audits are essential. CASD, which holds ISO 27001, ISO 27701, and HDS certifications, pays particular attention to these matters.

In 2024

- CASD initiated the process for the new RGDP - Europrivacy standard: official service certification as stipulated in Article 42 of the GDPR, which will enable us to become the first data host in Europe to be certified under this standard.
- CASD also launched an initiative to implement an ISO 9001 quality management system.
- CASD migrated to the new EBIOS RM risk analysis methodology (modified version of EBIOS).
- CASD conducts regular audits. The technical security audit carried out by Orange Cyberdefense in July 2024 confirmed the very high level of security of the CASD system.

RGPD
le référentiel Europrivacy
Europrivacy™
Applicant
CNIL.
Autorisation de traitement
2014-369
ISO
27001
Sécurité de l'information
FR078964
ISO
27701
Protection des données personnelles
Privacy / FR060212
BUREAU VERITAS
1828
HDS
Hébergeur de données de santé
FR060212
SNDS
Homologation au référentiel
de sécurité des données de santé
BUREAU VERITAS
1828

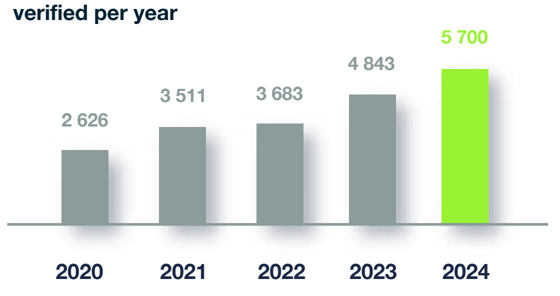
// RESPONDING TO TO GROWING *EXPORT REQUESTS*

he choice made by INSEE and Genes - the initiators of the CASD project - to use a remote access system enabling users to work within the secure project environment directly on the data until the desired results are obtained is a major advantage, and one which is still underdeveloped in many other countries. However, exports of final results must comply with the rules for data anonymisation set out by producers. Data producers choose the output checking method (most have opted for systematic manual checks conducted by CASD, while others allow automatic outputs).

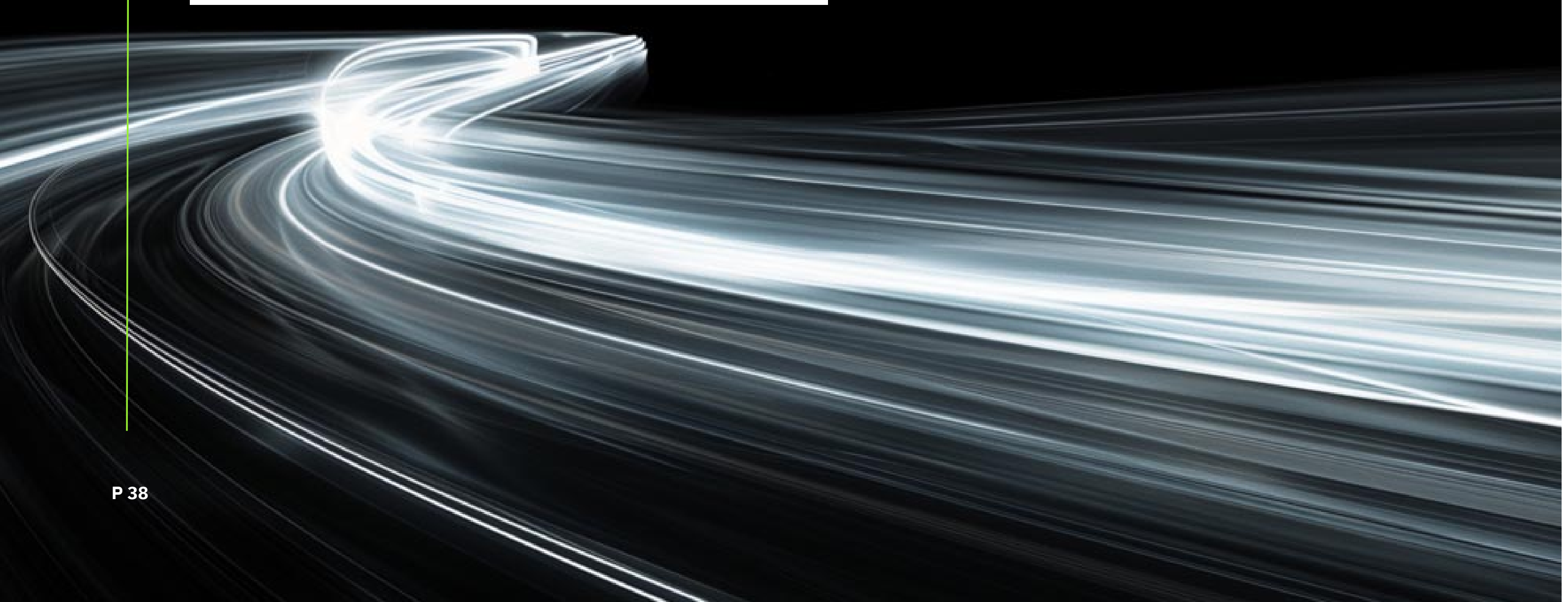
With 5,700 manual outputs verified in 2024, the average annual increase exceeded 20%, reflecting the growth in the number of projects. Processing time has nevertheless remained short.

5 732

Number of manual outputs verified per year



2024 saw the implementation of an internal tool to assist with output processing, saving CASD teams a significant amount of time: in practice, the flow of results exports is much faster and requires fewer steps for the teams. While the anonymity of outputs is verified manually, most of the other output processing steps are automatic, from the request to the delivery of the output to the user on the CDAP portal, which is expanding its functionality for CASD users. The level of security remains the same, as rigorous as ever.



// *ASTREAMLINING EXPORTS AND IMPORTS:* MAJOR PROGRESS

The exports and imports users may need to perform for their analysis must not compromise system security. On these two essential points, CASD strives to facilitate its users' work.

Developing machine learning algorithms and LLMs to verify the confidentiality of result outputs from the CASD secure bubble

At the end of their work on the CASD system, when researchers wish to retrieve the results of their analyses, they must submit their files to the CASD Data Management Service to verify that the files comply with confidentiality rules.

To assist this output checking process, CASD is developing a system for analysing result files and their content, based on generating characteristics from exported file groups and training a reinforcement model. The system draws on historical data from previous ‘manually’ performed reviews, where decisions (Accepted/Rejected) served as labels, and structured characteristics were extracted from the reviewed exports. The main objective was to identify situations that could pose a risk to confidential data and trigger alerts for review by an expert.

In 2024, CASD began to implement a new functionality to increase the reliability of the system. By leveraging large language models (LLMs), new predictive features were added that enrich the risk assessment model. More specifically, LLMs enable the detection of key attributes that were previously absent from the system, such as identifying columns containing headcounts and ensuring compliance with minimum value requirements.

Beyond digital validation, LLMs also improve the system's ability to analyse textual content, determining whether exported data consists of code, structured datasets, or free-text documents. By improving transparency and traceability, this integration, currently at the prototype stage, should enhance both the reliability and interpretability of compliance checks for data confidentiality.

Secure automatic import of scripts

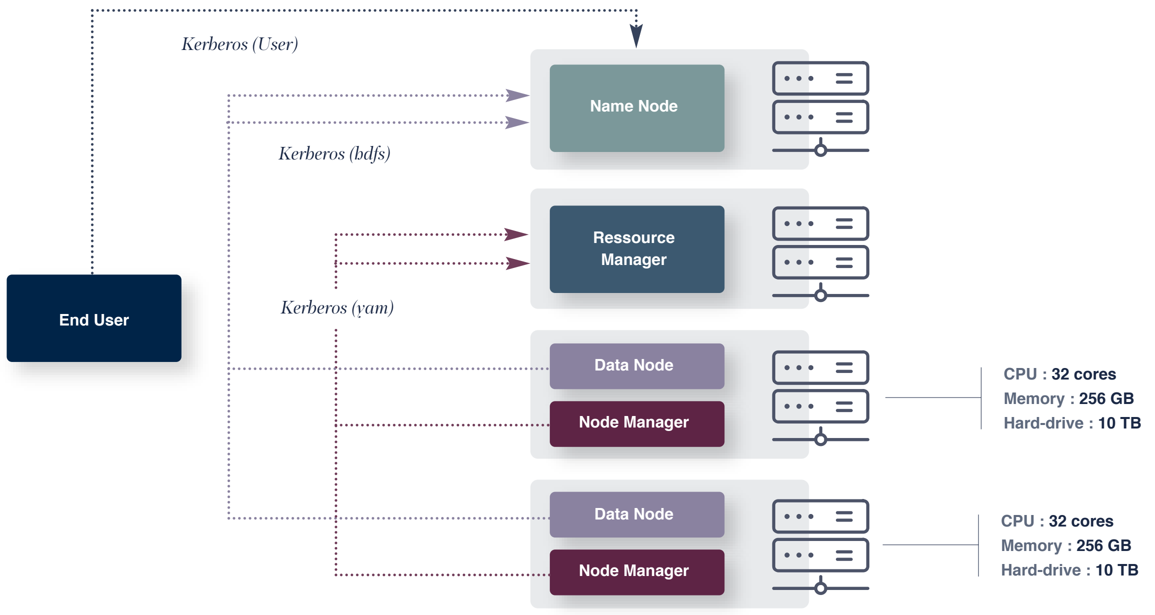
To facilitate imports, CASD has made it possible for the user to initiate a procedure to import text themselves within their secure bubble.

// *DEDICATED ENVIRONMENTS* FOR DATA SCIENCE PROJECTS

Upon request, the IT team can install specific work environments required for certain data science projects. As always, this installation is carried out following a preliminary security analysis to ensure the integrity of the environment.

As part of a DARES project processing the MIDAS data source, CASD deployed a SPARK/HDFS cluster, enabling calculations to be distributed as closely as possible to the data spread across 15 servers, combining:

- 150 vCPU,
- 2,8 To of RAM,
- 30 To of raw disk space



// A HIGHLY SPECIFIC ENVIRONMENT FOR *GENOMIC DATA*

Genetic data on CASD

In 2024, as part of Inserm's cross-disciplinary Genomic Variability program, also known as GOLD (Genomics variability in health & Disease), the GOLD-GENOPHENOMET project draws on genetic data from volunteers in the Constances cohort, generated as part of the France Médecine Génomique 2025 POPGEN pilot project, which is securely available on the CASD system.

The purpose of this project is to assess statistical and bioinformatic methods to be able to better understand the impact of genetic variations on health and thus, if possible, implement more effective measures in managing the population.

To do this, a specific working environment, allowing programming in a wide range of languages (C/C++, Java, Perl, Python, Bash, Awk) and with domain-specific software (Plink, LDPred2, snptest, bolt-lmm, etc.), has been installed in a dedicated secure bubble in Linux, accessible after authentication on an SD-Box.

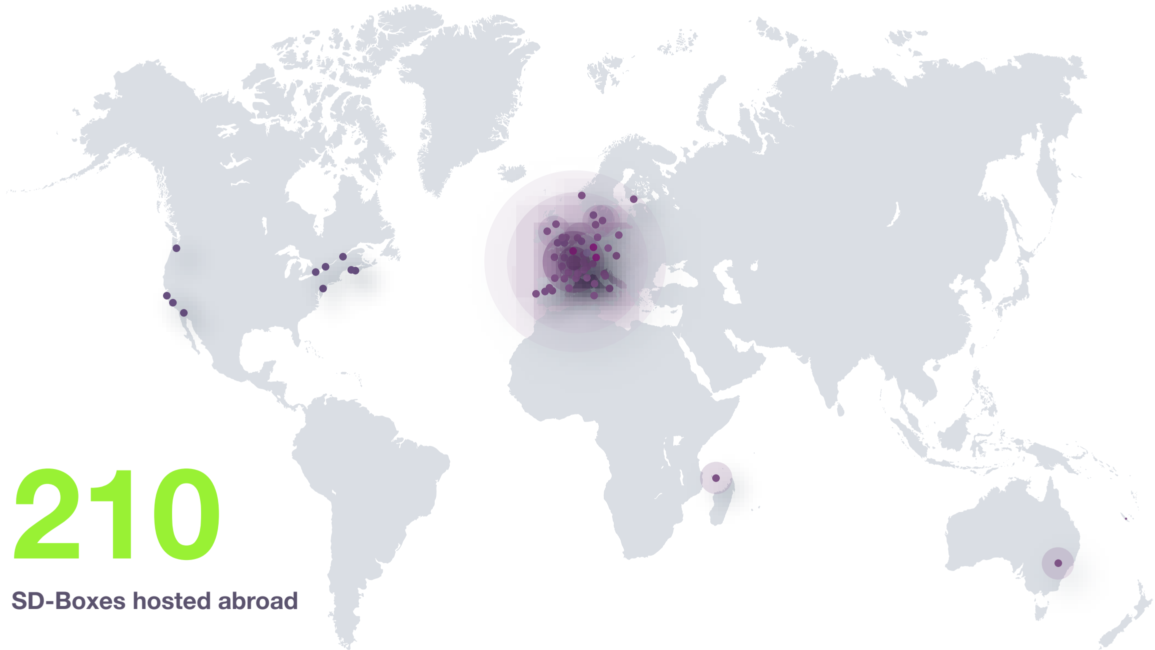
04. INTERNATIONAL

From its inception, and with the agreement of data producers, CASD has been committed to making confidential data available for international research. CASD is particularly proactive in this area.

Secure remote access is available from European Union countries and EU-associated countries under the same conditions as for researchers at universities and research centres in France. This is also true for countries that have received an adequacy decision from the European Commission for personal data processing and, for North American countries (the US and Canada), under certain conditions. This access from abroad is used both by French researchers working at universities abroad and by researchers from these countries, with many cooperative projects often involving several institutions from different countries working together in the shared research environment provided by CASD for each project.

CASD also participates in projects aimed at facilitating the use of confidential data across national borders. Through the IDAN (International Data Access Network) network, which CASD coordinates, we are working to build cooperation in this area with several secure centres in other countries.

// SD-BOXES *HOSTED ABROAD*



210

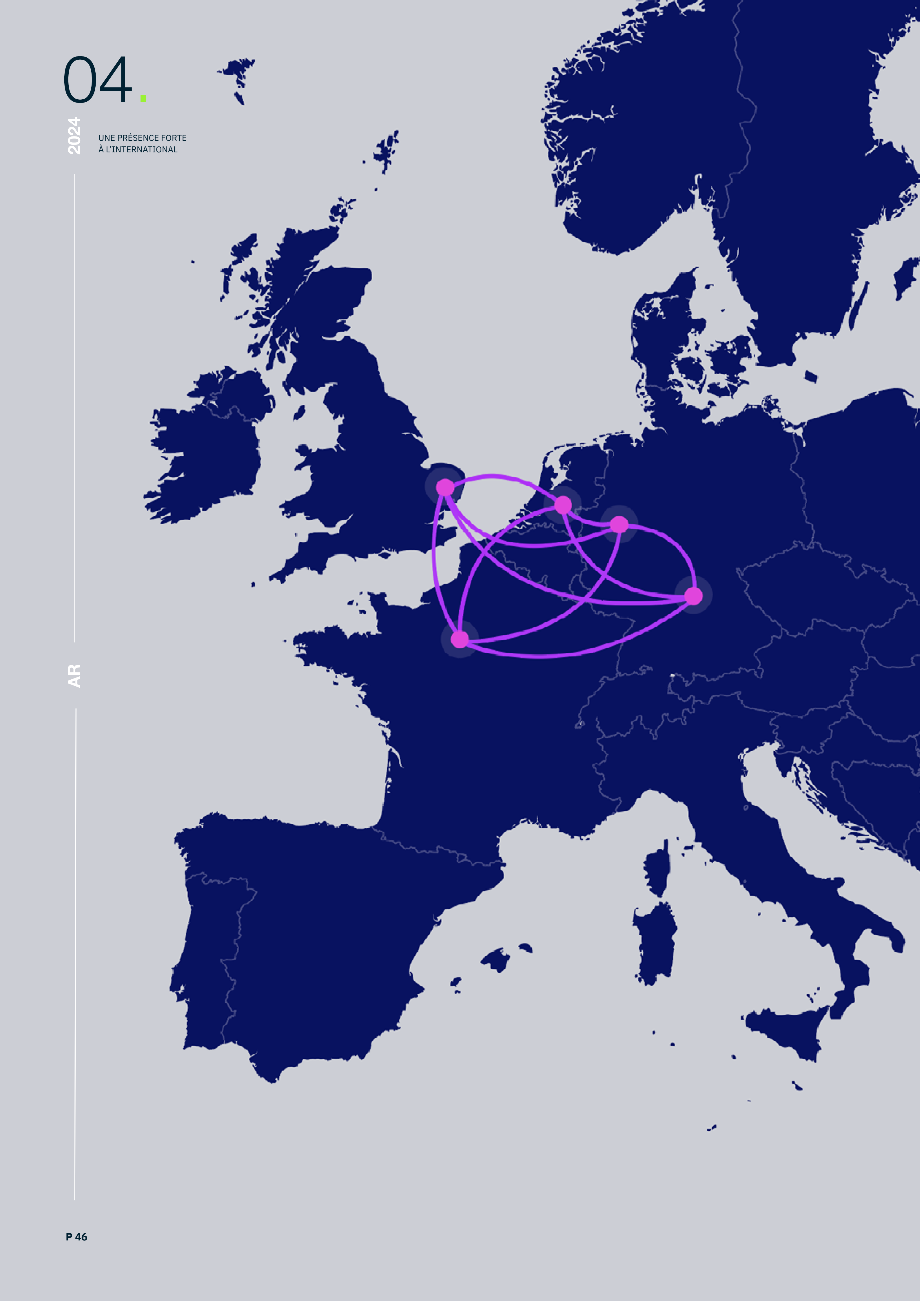
SD-Boxes hosted abroad

SD-Boxes abroad now represent approximately 20% of the total amount of SD-Boxes, a high and constant rate

In December 2024, 210 SD-Boxes were hosted abroad in 18 different countries, from which researchers work on French data.

The leading countries in 2024 in terms of the number of SD-Boxes hosted were:
The Netherlands, Italy, Belgium, the United Kingdom, Germany, the United States, Switzerland, Canada, Spain, Austria, Sweden, Luxembourg, Denmark, Norway, Ireland, Portugal, Finland and Australia.
Australia was also new to CASD in 2024, with the authorisation of a first project.

On-site enrolment at CASD is essential in order to be able to work on an SD-Box. While for many researchers the trip is also an opportunity to visit colleagues in France and discuss projects - some of which are joint projects involving French data - for others it can be a significant constraint, particularly if the journey is long and expensive. The pandemic has made this constraint all the more apparent. In response to this, the Statistical Confidentiality Committee gradually lifted the requirement for on-site presence for the accreditation procedure. After the operational validation of its remote enrolment technology, tested in 2023, CASD continued this work in 2024 to develop the protocol governing remote enrolment to reach same level of security as on-site enrolment.



// IDAN: *GERMAN AND BRITISH DATA* ACCESSIBLE THROUGH CASD

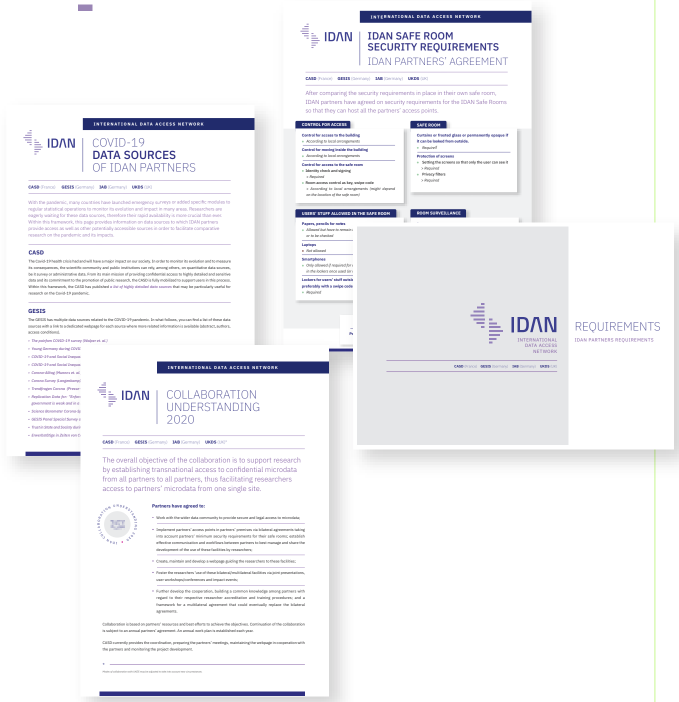
The aim of IDAN (International Data Access Network) https://idan.network/, the network of secure data centres coordinated by CASD, is to facilitate the secure use of data from multiple countries. In the first stage, agreements were made to allow remote access to all the partners' data from each partner's physical premises.

- As in previous years, several researchers have thus worked on IAB German data fdz.iab.de/en/data-access/ from the CASD IDAN safe room. Other users working on French data accessed these via GESIS in Mannheim www.gesis.org/en/aml/safe-room-mannheim
- In 2024, secure remote access to highly detailed or sensitive data available at UKDA (UK Data Archive www.ukdataservice.ac.uk), was established from CASD in France.
- CASD users can thus now work on German, British and French data.
- All information on catalogues, authorisation procedures, and reservation procedures once authorisations have been obtained can be found on the IDAN and CASD websites.

IDAN

IDAN
INTERNATIONAL
DATA ACCESS NETWORK

A collaboration between multiple Research Data Centres from European countries, open to enlargement, to facilitate research use of secured data, particularly the administrative ones between these countries



IDAN SAFE ROOM
SECURITY REQUIREMENTS
IDAN PARTNERS' AGREEMENT

COVID-19
DATA SOURCES
OF IDAN PARTNERS

COLLABORATION
UNDERSTANDING
2020

IDAN REQUIREMENTS

→

https://idan.network

// CASD INTERNATIONAL *RELATIONS*

CASD regularly participates in conferences and seminars.

CASD has been a partner in several international conferences

The COSMOS conference (11-12th April 2024) on the subject of metadata was organised by INSEE and gathered over one hundred participants from around twenty countries. CASD and INSEE together presented their work on metadata exchange, which aims to facilitate and improve the availability of metadata for users httD://cosmos-conference.org/2024/

COS MOS

CESS (Paris 15-16 October 2024) entitled "The Agenda beyond the GDP: Past, Present and Visions for the Future". This 2024 edition was organised conjointly, under the patronage of Esac and partnered with INSEE, Banque de France, Eurostat and the ECB, the Paris School of Economics (PSE) as well as CNIS and CASD. CASD was a partner in this conference. Kamel Gadouche, Director of CASD, and Olivier de Bandt from Banque de France presented the latest possibilities offered in terms of data which are the fruit of the partnership between the two organisations. Progress on access to administrative data within the IDAN network of secure centres coordinated by CASD was also presented. This presentation is available online at https://www.cess2024.insee.fr/programme/.

CESS 2024

CASD also presented at several other conferences in 2024.

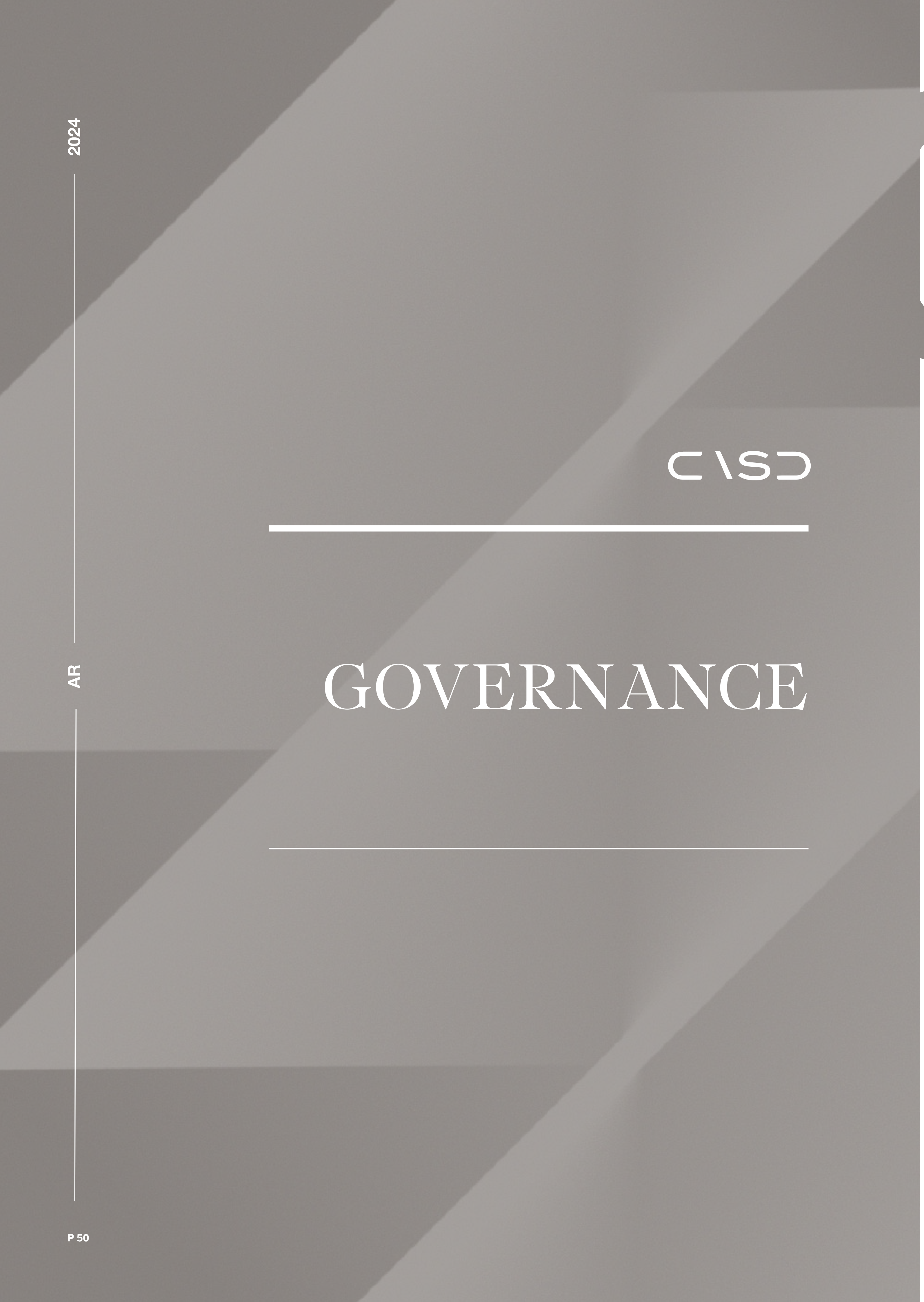
Qualidata : CASD participated in the 11th European Conference on Quality in Official Statistics, Q2024 https://www.q2024.pt/ organised by Statistics Portugal and Eurostat, held between 4th and 7th June 2024, during which Halima Bakia and Ifaliana Rakotoarisoa from the CASD Data Management team, together with Thomas Dubois from INSEE, presented the experiment CASD is currently conducting with INSEE on the exchange of documentation metadata using the DDI (Data Documentation Initiative) standard, which is enabling significant gains in terms of reliability of data documentation by speeding up the display of documentation on the CASD website.

CASD participated in the **Mapineq** seminar entitled “Improving Accessibility, Harmonisation and Data Linkage in Europe”. This seminar is available to view on YouTube.

CASD regularly welcomes visitors from foreign centres interested in its development, technology, and positioning. In 2024, CASD twice welcomed the UK Research and Innovation (UKRI) body responsible for developing research infrastructure, including UKRI director, 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 making research support interesting and rewarding CASD is showing us how to scale up research capacity in trusted research environment ”

Richard Welpton - Blog
Head of Data Services Infrastructure, UKRI



GOVERNANCE

05.

Created in 2010 by INSEE, and having received funding under the Investments for the Future programme (Equipment of Excellence program, or Equipex, from 2011 to 2019), CASD took the legal form of a Public Interest Group (GIP), a consortium created by an interministerial decree dated 29th December 2018. The consortium brings together the French State represented by INSEE, Genes, CNRS, École Polytechnique, HEC Paris, and Banque de France.

Created in 2010 by INSEE, and having received funding under the Investments for the Future programme (Equipment of Excellence program, or Equipex, from 2011 to 2019), CASD took the legal form of a Public Interest Group (GIP), a consortium created by an interministerial decree dated 29th December 2018. The consortium brings together the French State represented by INSEE, Genes, CNRS, École Polytechnique, HEC Paris, and Banque de France.

Its various bodies are:

The General Assembly

The Scientific Committee

The Committee of Data Producers

CASD, directed by Kamel Gadouche, is organised into several departments: PMS (Project Management Service); DMS (Data Management Service); IT-DS (IT-Data Science); and R&D (Research & Development)

// THE VARIOUS BODIES OF CASD *MEET REGULARLY* — *2024*

Catherine GAUDY
General Director of GENES

- The **General Assembly,** chaired by Catherine Gaudy, General Director of GENES, held meetings on 21st June 2024 and 2nd December 2024. It is composed of official representatives of the CASD consortium members (the French State represented by the General Director of INSEE, GENES, Banque de France, CNRS, HEC Paris and Ecole Polytechnique).

- The **Scientific Committee** welcomed two new members:

  - Romain Lesur, A senior administrator at INSEE and head of INSEE's Public Statistical Service Lab, he represents INSEE on the CASD Scientific Committee.
  - Paola Tubaro, Research Director at the French National Centre for Scientific Research (CNRS) and member of the French Centre for Research in Economics and Statistics (CREST). Her research focuses on the economics of digital platforms, global production networks in the artificial intelligence industry, the role of human labour in the development of automation, and digital inequalities.

  Two meetings were held on 31st January and 10th October 2024, chaired by Lars Vilhuber (Cornell University). During these sessions, presentations and discussions focused on:

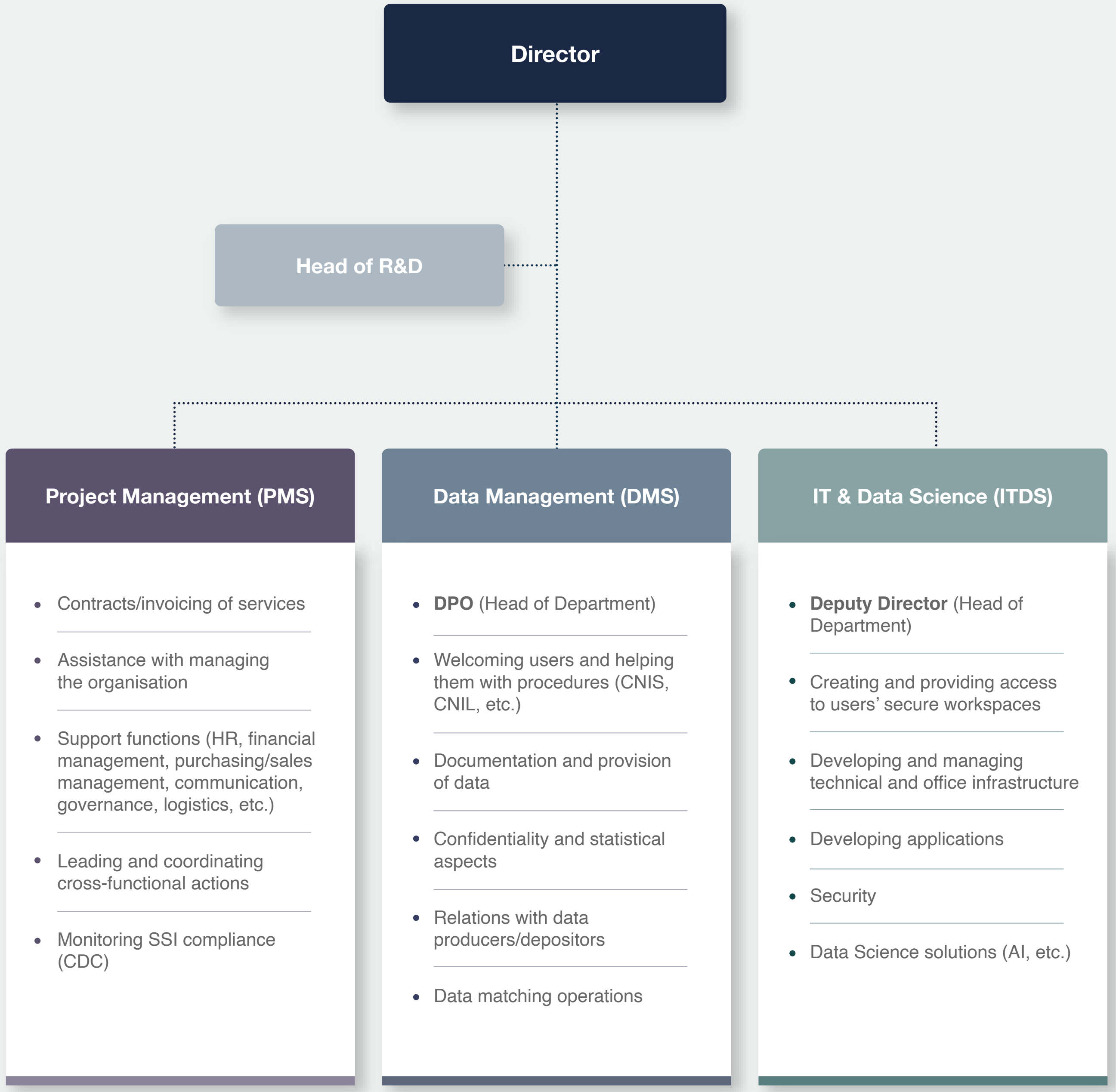
  - Progress on strategic CASD development priorities,
  - Anonymisation checks on result outputs and the Colysée project,
  - The latest developments in CASD (ergonomics, tools, interfaces, confidentiality...),
  - Future uses by researchers.

- **The Data Producers Committee,** chaired by Christel Colin, Director of Demographic Statistics at INSEE, met on 22nd November 2024. Discussions focused on:

  - Macro indicators from the CASD cost study,
  - Structuring a community of researchers using confidential data,
  - Data documentation,
  - The storage format of the data sources.

ANNEXES

ORGANISATIONAL CHART BY DEPARTMENT



Director

Head of R&D

Project Management (PMS)

- Contracts/invoicing of services
- Assistance with managing the organisation
- Support functions (HR, financial management, purchasing/sales management, communication, governance, logistics, etc.)
- Leading and coordinating cross-functional actions
- Monitoring SSI compliance (CDC)

Data Management (DMS)

- **DPO** (Head of Department)
- Welcoming users and helping them with procedures (CNIS, CNIL, etc.)
- Documentation and provision of data
- Confidentiality and statistical aspects
- Relations with data producers/depositors
- Data matching operations

IT & Data Science (ITDS)

- **Deputy Director** (Head of Department)
- Creating and providing access to users' secure workspaces
- Developing and managing technical and office infrastructure
- Developing applications
- Security
- Data Science solutions (AI, etc.)

— 2024 *KEY* INDICATORS

+ 19,5%	Increase in number of users compared to December 2023
+ 26%	Increase in number of SD-Boxes in use and monitored compared to December 2023
+ 20%	Increase in the number of co-contracting organisations compared to December 2023
+ 5%	Increase in the number of projects managed and invoiced compared to December 2023
+ 7 ETP	Evolution of the workforce managed at year-end compared to December 2023
+ 5%	Budget managed compared to 2023 budget (increase in turnover: + 29% compared to December 2023)

- To support the momentum of its business and cope with the increase in workload, CASD strengthened its teams by welcoming seven new employees in 2024

- CASD, a public interest group, has begun reviewing its internal organisation with a view to offering better support to its users by pooling certain tasks while safeguarding support functions:
  - With a view to improving the existing situation
  - To meet ever-increasing service quality and accountability requirements
  - In an effort to fulfil CASD commitments to our various stakeholders (CASD members, data producers, users, administrations, public authorities, etc.)

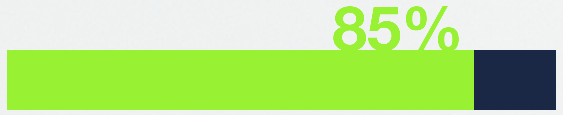
- Development priorities for 2022-2025: 2024 was a year of implementation and preparation for a new strategic cycle aimed at addressing new challenges in terms of scaling up and diversified usages (scaling up infrastructure, data science/AI, processing big data from large data platforms, quality assurance, structuring a community of data users, etc.).

— *PROGRESS* ON STRATEGIC PRIORITIES

PRIORITY 01

Develop and reinforce the technology offer

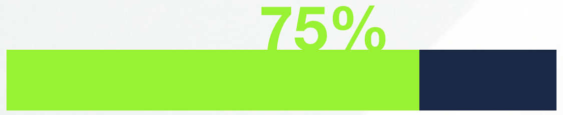
- **Develop private data science cloud technology (Spark, ML):** Large-scale SPARK / HDFS cluster development for the DARES MIDAS project
- **Develop automatic input of text/codes:** Completed (cofinanced by DREES/DARES)
- **Develop remote biometric enrolment:** Completed, tested, to be validated with CNIL and then put into production
- **Develop machine learning on file outputs (Colysée):** Completed, tested, in production
- **Analyse session recording (computer vision):** Proof of concept in progress
- **Improve time taken to connect and switch between Bubbles:** in production (reduced from 45 to 15 seconds)
- **Security:** successful ISO 27001, ISO 27701 and HDS certification



PRIORITY 02

Consolidate and expand the data offer

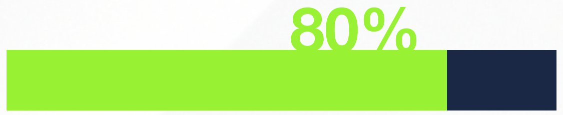
- **DDI data documentation / translation of documentation**
- **Develop a new interface for presenting documentation on the CASD website**
- **Data matching:** FORCE, MIDAS, BADS2 (PSE), SAMU
- **Develop collaborations with data producers** (Infrasec-DMA)
- **Training on corporate data,** Fideli, DGFiP....
- **VTL** (Validation et Transformation Language) integrated into the CASD infrastructure and DuckDB/Parquet integration



PRIORITY 03

Automation of internal operations

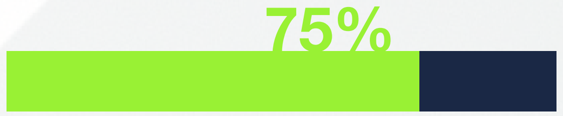
- **Develop the front office (CDAP) and back office (ROME)**
- **CDAP user interface** (monitoring projects/subscriptions), **Integration of BDF data**
- **Integrate the invoicing module into the internal Information System**



PRIORITY 04

International relations and relations with other HUBS

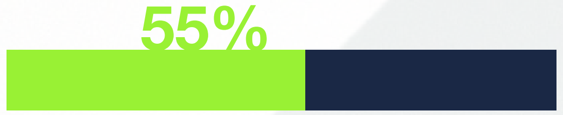
- **Develop collaborations with other centres (IDAN, etc.)**
- **Begin expanding IDAN** (International data access network) beyond its founding members



PRIORITY 05

Promote our technology

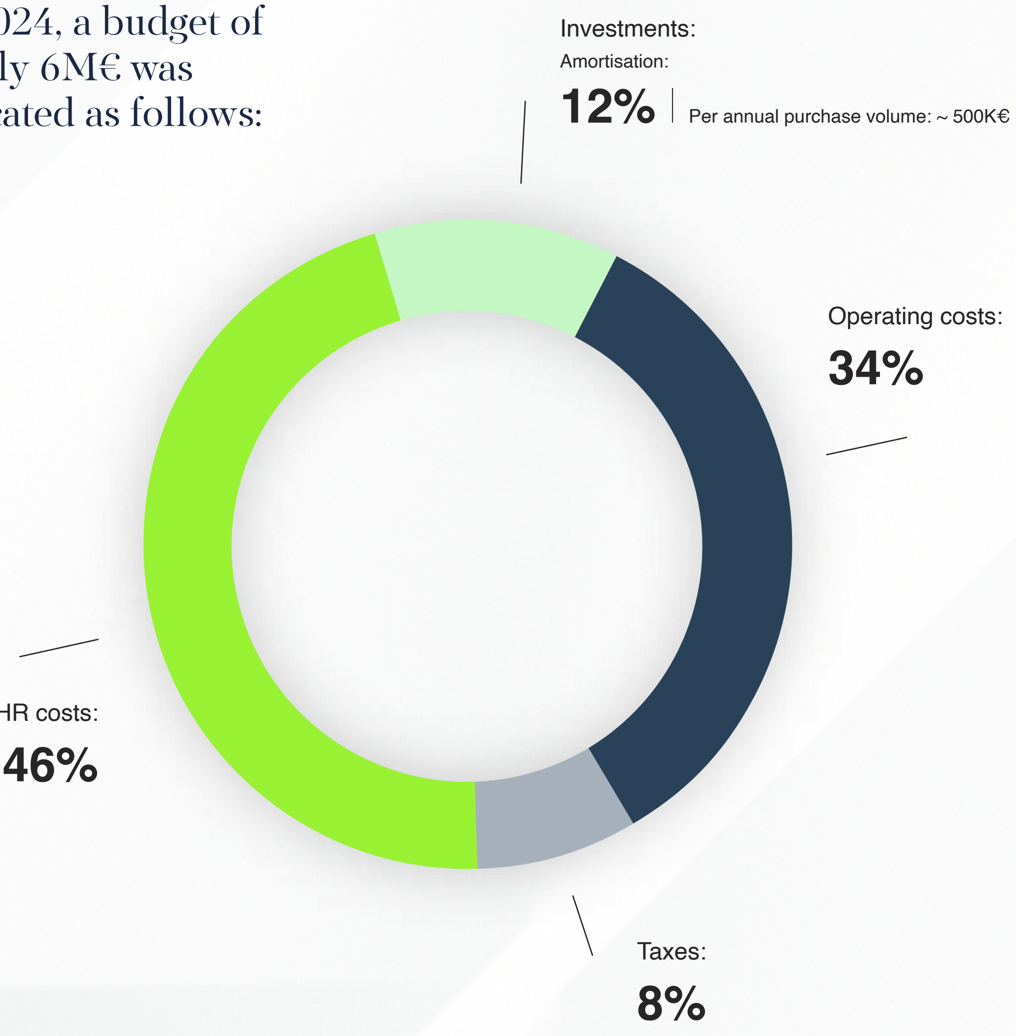
- **Commercial Development**
- Meta pilot project, TikTok for pilot projects, contact with Google, Microsoft and LinkedIn



— *FINANCIAL* MATTERS

6M€

In 2024, a budget of nearly 6M€ was allocated as follows:



C\SD

WHAT THE *DIFFERENT DEPARTMENTS* DO

The Department

PROJECT MANAGEMENT SERVICE (PMS)

Missions

- Sales administration: Relationship management and advice to users on contractual and financing aspects and on how to articulate their needs. Monitoring service implementation and invoicing
- Human Resources administration
- Budget management / accounts
- Facility management / logistics
- Governance support and secretariat to the management
- ISMS (Information Security Management System) and compliance with certifications
- Communication
- Support for the management on cross-functional projects within the organisation

Head of Department:
Tanguy Libes

PMS in 2024

- Recruitment: To accommodate the growing workforce, the PMS HR department has been strengthened with the addition of a second position.
- Start of work on automating key business processes (user project onboarding, discussions on automating service activation/deactivation, etc.) to concentrate efforts on higher value-added tasks and user support, in preparation for a web-based front office.
- Coordination of the rollout of new procedures for managing and onboarding specific projects, the challenges of which differ from those of traditional research and study projects.
  - Coordinating the terms and conditions for hosting and deploying the META pilot on Instagram data for research on well-being
  - Support for projects led by international user organisations, with legal guidance tailored to their specific needs (OECD, IMF, etc.)
  - Implementing new procedures for the management and administration of secure infrastructures provided to the social ministries’ statistical services (DREES / DARES) for producing and analysing statistics
- Contribution to the support of cross-functional projects (remote enrolment, biometric SSO, updating the contractual framework governing the deployment of services, management and improvement of the ISMS and associated documents, extension and security of the premises occupied, etc.)

The Department

RESEARCH & DÉVELOPMENT (R&D)

Missions

The R&D department of CASD was created in 2015 and has received increasing amounts of projects over time according to the needs of the organisation.

- **Conducting research, generally but not exclusively, in technological fields.**
- **Completing entire projects intended for production.**
- **Performing technology monitoring functions.**
- **Responsibility for the design, development, and monitoring of SD-Box manufacturing. As such, CASD is responsible for selecting and procuring electronic components, operating systems, and smart cards, as well as monitoring these components, including their associated software.**

R&D may be tasked with addressing an issue by management or the various CASD departments, generally relating to the SD-Box; this may concern hardware and software aspects, but also SD-Box interactions with the secure environment to which it connects and certain legal considerations such as air transport regulations, customs, and sometimes even issues beyond its usual sphere, such as the creation of free-cooling server rooms.

R&D can also undertake any subject, generally related to the CASD IT environment, in order to advise the management committee or the relevant departments.

The growth of CASD and the accumulation of generations of hardware and software components complicate the management and evolution of the existing system. For example, CASD currently supports nine variants of smart cards. Anticipating changes has thus become essential.

Head of Department:
Philippe Donnay

A few studies conducted by R&D in 2024

- Reviewing changes to certificates and their containers on smart cards.
- Assessment of microcode protections for SD-Box processors.
- System authentication protocols and their use by CASD.
- Remote renewal of smart card certificates.
- Hardware certification for SD-Boxes.
- Evolution of the system connecting SD-Boxes to the CASD infrastructure.
- Evolution of the SD-Box operating system.

The Department

DATA MANAGEMENT SERVICE *(DMS)*

Missions

The Data Management Service primarily performs a number of operational tasks: granting user access rights, receiving and making data available, liaising with producers and users, and checking data/result outputs. In addition to this main activity, the DMS also performs the following tasks:

- Regulatory monitoring of data protection, statistical confidentiality, and cybersecurity law: GDPR, Law 51-511, the Penal Code, CNIL guidelines, etc. The DMS is in regular contact with CNIL and participates in the Statistical Confidentiality Committee.
- Dialogue and monitoring agreements with data producers or project leaders
- Performing data matching as a trusted third party for operations requiring restricted access to directly identifiable data
- Documentation of the data sources made available, including variables, freely accessible on the CASD website
- Analysing CASD activity with regard to user projects: number of projects, number of outputs, etc., and usage profiles
- Implementation and operation of the CASD user satisfaction survey

Head of Department:
Rémy Marquier

In 2024, the DMS completed other important tasks:

- Data matching operations as part of CASD's role as a “trusted third party”: the FORCE and MIDAS projects run by DARES (Directorate for Research, Studies, and Statistics) continue to operate, and the DMS ongoingly supports users’ data matching projects.
- Regulatory and legal watch: dialogue with CNIL (French Data Protection Authority) particularly on standards relating to the processing of health data. The DMS also supports project leaders in their compliance procedures, particularly for implementing health data warehouses.
- Using the CASD internal project management database: statistically formatting this database, which covers all operational activities, will enable in-depth analyses to be carried out on how CASD is used (types of projects, users, etc.)
- Enriching producers’ data documentation and working on redesigning the documentation page of the casd.eu website. The DMS team also participated in two international conferences on international metadata standards and knowledge sharing in the field: COSMOS (Conference on Smart Metadata for Official Statistics) and QSTAT (European Conference on Quality in Official Statistics).
- Organising a presentation and training session on DGFiP (French Public Finances Directorate General) business data
- Specifications and tests for data-related internal tools: Colysée (semi-automatic verification of outputs from certain projects) and Vespa (improvement of the flow of user output results).

The Department

IT & DATASCIENCE *(IT&DS)*

Missions

The ITDS department comprises three complementary divisions, the main tasks of which are the design, deployment, maintenance, and optimisation of the infrastructure.

Three analyst-developers from the DEV-WEB team are responsible for the development, documentation, and maintenance of CASD web applications: the casd.eu website, all CDAP (confidential data access portal) modules, the internal ROME application supporting the operation of secure bubbles, and external websites (DOI documentation sheet, satisfaction survey, awareness quiz, etc.).

In 2024 their work mainly focused on the CASD interface of the CDAP, including the integration of new INFRASEC features (text import, changing IP addresses from home), automation of PMS service activation (triggering actions on a scheduled date or following the completion of contractual prerequisites: account activation, creation of project spaces, hardware configuration changes, invoicing, etc.) and accommodating Banque de France specific requirements for access authorisations to its data: smooth integration into the CSS module, management of additional multi-level contractualisation involving electronic signatures, and customised management of its data source repository.

The Data Science team comprises 3 data scientists who together form a pool of high-level expertise in a wide range of fields, intended for CASD users: Implementation of the Spark cluster, optimisation of Python code, assistance in choosing hardware configuration, performance comparisons between various software and hardware stacks and file formats. The Data Science team is also called upon to ensure the internal needs of CASD are met: use of models trained on historical data of accepted and rejected exports to prioritise controls (Colysée), a classification model for text-format entries, advanced detection of abnormal behaviour through the use of security logs, a program for converting very large file formats, documentation of the integration of data science/AI tools into CASD.

Their work in 2024 has strengthened the capabilities of Colysée (OCR on image files, continuous improvement of the underlying model, exploratory work on the use of LLM), tested tools based on large language models (LLM) within secure bubbles, and proposed both hardware (dedicated Hypervisor with 2TB of RAM, 2 NVIDIA H100 NVL) and software infrastructure (switch to Parquet format, high-performance management of distributed graphs using Spark)

Head of Department:
Éric Debonnel

2024 for the IT & Data Science Department

- Their main tasks are the design, deployment, maintenance, and optimisation of the CASD IT infrastructure (120 physical servers, 1500 virtual servers, 1300 SD-Boxes, 2GB backup capacity). They also provide user support and implement measures enabling CASD technical certifications, thereby strengthening our auditability.

- In 2024, efforts were focused on infrastructure availability (4 Internet connections, redundancy - regularly tested - of critical services, transformation of the secondary backup site into an equivalent of the main site, and cross-checking of backups), increasing automation of routine operating processes, offering new services (INFRASEC, VESPA, Linux environments), and the creation of a knowledge base to facilitate user support.

SECURE RESEARCH DATA CENTER

casd.eu

CASD IS A PUBLIC INTEREST GROUP (CONSORTIUM) BRINGING TOGETHER:

Insee

GENES

cnrs

ÉCOLE POLYTECHNIQUE

HEC PARIS

BANQUE DE FRANCE
EUROSYSTÈME