Enrolment Session

Centre d’Accès Sécurisé aux Données

Secure Data HuB

CASD

Table of contents

- Presentation of the CASD
- Legal framework of data access
- IT
    - Infrastructure
    - Workflows
    - Demonstration
    - Terms and Conditions of Use
    - Information about access card issuing
    - Q&A
- Data
    - Anonymization techniques
    - Confidentiality rules
    - Input/Output
    - Data citation and return of publications
    - Q&A
- Support contacts
- Quiz

The CASD

The CASD

- **A public interest grouping** composed of five members: GENES, INSEE, CNRS, HEC Paris and École Polytechnique
- **A secure infrastructure to access confidential data** which benefited from “Equipement d’excellence” (EQUIPEX) funding of PIA (Programme d'investissements d'avenir) 1st edition

- **Our main mission:** organize and set up services of secure access to confidential data for users pursuing non-profitable research, study, evaluation or innovation purposes
  - Secondary mission: valorization of the technology in the private sector
- **Our goal :** provide a highly secure access for accredited data users in the best possible work conditions while minimizing access costs

The challenges of the secure data access

- **Security, a real challenge to allow data access:**
  - Ensuring a high level of security **in order for data producers to make more and more data available in confidence**
  - Strong authentication using biometrics
  - The enrolment session is mandatory and must be renewed every 4 years
  - A formal accreditation process before accessing the data
- **Uses:**
  - To meet the needs of users in terms of work environment (software, configuration…)
  - A shared work environment between project members
- **Fair treatment of all users**

➔Without this device, for example, it is certain that tax data access would not have been possible in 2013

➔The secure system put in place will allow access to other sources of confidential data

The CASD in figures

24 staff members organized in **3 poles**

Data Management

Data collection, output control, user support

IT

Infrastructure: IT infrastructure and user support

Development: CASD website and applications managed by CASD

Data science: implementation of Hadoop cluster, machine learning and GPCPU

Project Management

Transversal aspects, project reception, contracts and billing

Key figures: https://www.casd.eu/en/le-centre-dacces-securise-aux-donnees-casd/le-casd/

Certifications and CASD commitments

GDPR compliance by Bureau Veritas and CNIL authorization (n°2014-369)

ISO 27001 and ISO 27701 (GDPR) certifications

« Health data hosting » certification

Regular security audits

ISO 27001
Sécurité de l'information
FR055849

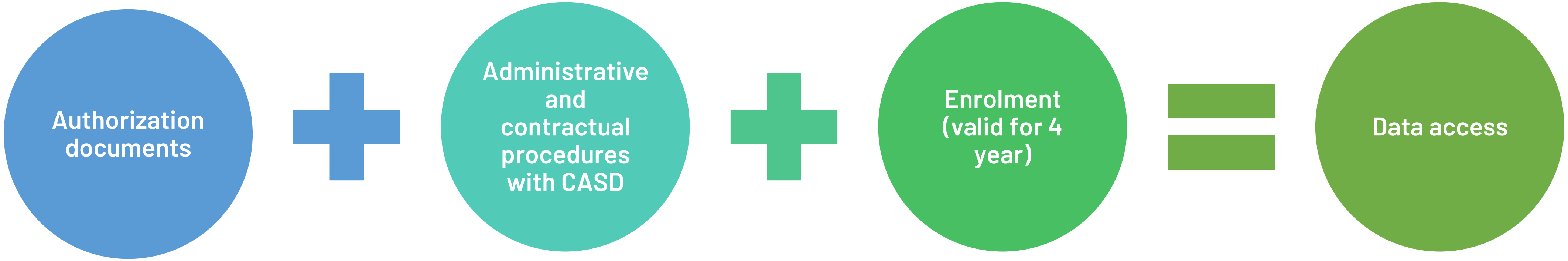
ISO 27701
Protection des données personnelles
RGPD / FR060159

HDS
Hébergeur de données de santé
FR055852

SNDS
Homologation au référentiel de sécurité des données de santé

CNIL.
Autorisation de traitement
2014-369

Steps towards Data access



Authorization documents + Administrative and contractual procedures with CASD + Enrolment (valid for 4 year) = Data access

Authorization documents:

- Statistical Secrecy Committee (CSS):
  - Documents signed by data producers and national archives
  - DGFIP authorization for tax data
- Data not covered by CSS: direct authorization by the relevant authority

Legal framework of data access

Statistical secrecy

- ➢ Access to data covered by the statistical secrecy is allowed under the conditions of **articles 6 and 7 bis of the 1951 law** (after consulting the Statistical Secrecy committee). These conditions were extended to tax data by the LPF (Livre des procédures fiscales – Tax procedures handbook) and to all administrative database by the CRPA (Code des relations entre le public et l'administration – Code of relations between the public and the administration)
- ➢ Data access is granted under cover of the statistical secrecy, the secrecy is shared

➔ **No dissemination, neither of individual data nor of results indirectly identifying persons or firms**

- ➢ Consequences in case of a secrecy breach (recalled by your commitment):
  - **You are personally liable (data access is strictly personal)**
  - **Criminal sanctions**:
    - articles 226-13 and 226-14 of the penal code (breach of professional secrecy):
      "the disclosure of secret information by a person who is in possession of it either because of his profession or his status, or because of his function or a temporary mission, is punishable by **one year imprisonment and a fine of 15 000 euros**";
    - articles 226-16 to 226-24 of the Penal code (violations of personal rights resulting from computer files or processing) in case of information related to individual firms
  - **Compensation in civil liability** for caused damages
  - **Not to mention damage to reputation…**

Processing of personal data

- **Data processing is subject to the obligations of the French Data Protection Act and the GDPR**

(specific provisions for organizations located on French territory).

- **Treatments prohibited by the GDPR:**

- A treatment with the final or intermediate purpose **of re-identifying one or more natural persons**
- A treatment with the final or intermediate purpose of taking **a decision against an identified natural person.**

- Consequences in case of breach:
  - According to the Data Protection Act: up to 5 years of imprisonment and a fine of 300 000 euros (section 5 of Chapter VI of Title II of Book II of the Penal Code)
  - According to the GDRP: administrative fines up to 20 000 000 euros or, in the case of companies, 4% of the total annual worldwide turnover of the previous fiscal year.

Reminder: personal data
Data referring to individuals, in other words all household sources, and individual firms in firm data sources

Processing of health data

- **The processing of health data is subject to the obligations of the Public Health Code (Article L4113-7).**

- **Treatments prohibited by the Public Health Code :**

- A treatment that would aim at **promoting health products** to health professionals or health institutions
- A treatment that would result in **the exclusion of benefits or the modification of insurance contributions or insurance premiums** for an individual or group of individuals

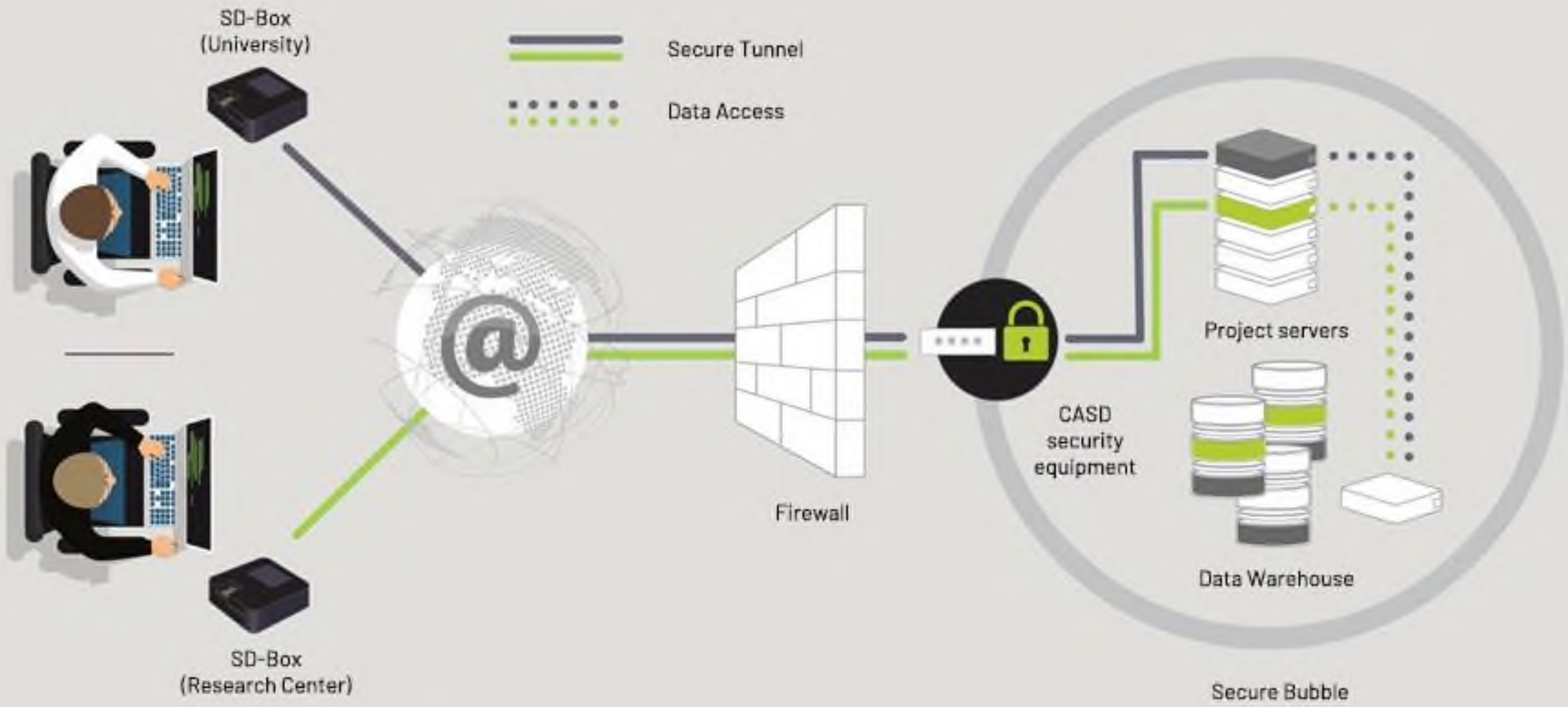
- **Any treatment that does not comply with the purposes declared to the CNIL is a prohibited treatment.**

Steps to take when processing personal data

- For the processing of personal data, here are the obligations to be complied with depending on the case:
  - registration of the project in the register of the processing operation
  - carrying out a privacy impact assessment (PIA) in the case of the use of so-called sensitive data (Art. 9 of the GDPR)
  - request of a CNIL processing authorization (in particular in the case of health data)

➔ *Contact your legal correspondent or your data protection officer when appropriate. More information on the website:* ***CNIL.fr***

IT Architecture


SD-Box
(University)
Secure Tunnel
Data Access
Firewall
CASD
security
equipment
Project servers
Data Warehouse
SD-Box
(Research Center)
Secure Bubble

[redacted]

Workflows

➢ **Remote access**

- With the SD-Box, you can work remotely on confidential data, while guaranteeing to the producer that no file can be retrieved by the user or inserted in the work environment (no copy/paste, no printing, no USB key, no internet connection…)

➢ **Inputs and outputs**

- There are specific workflows for importing or exporting files outside the secure environment, with verification by the Data Management and the IT departments
- Inter-project outputs ➔ transferring files from a project to another

➢ **Adapted and customizable work environment**

- Each project leader can contact us to implement updates – system or software – or to add software/packages

Demonstration

Demonstration

- Presentation of the SD-Box and the smartcard, logon procedure
- One dedicated virtual server per project
- The different spaces and file trees (DATA, IMPORTS, common space, etc.)
- The software installed by default in the environment (SAS 9.4 x64, Microsoft Office suite or LibreOffice, R, LaTeX - WinEdt, Stata MP, Scilab...). Software version management
- Installing an R package or a Stata ado file
- An example of results output: try to limit the number of inputs/outputs by working as much as possible within the secure environment
- Removing the smartcard closes only the local session, the session on the server remains open (the calculations continue to "run")
- Logout procedure

Terms and Conditions of Use (1)

- The SD-Box must be in a closed room and only the accredited user may see the screen
- The screen, the keyboard and the mouse must be provided and installed by the local IT manager. No other peripherals should be connected to the SD-Box
- Remove your smartcard from the SD-Box when you are away for even a short time (your calculations will not be interrupted)
- The smartcard is strictly personal
- You must not “lend” your session
- **No notes, no photo or video recording!**

Terms and conditions of use (2)

Security is an important challenge for confidential data access:

- This induces a certain number of constraints for the user that we have tried to make the least “unconformable” possible (dedicated environment with many software, dedicated equipment almost plug and play...) compared to other data access solutions in other countries.
- A large part of our system relies on the **trust placed in users** who can see the data (which is not the case in some other countries...)

Terms and Conditions of Use (3)

- Only the user is responsible in case of problem (output, disclosure…)
- The user contract specifies how CASD service can be interrupted (updates, maintenance…)
- Computer monitoring mechanisms are implemented in order to ensure compliance with security rules
- Concerning outputs, the user commits to export only non confidential data, it is the user that must do the checking if confidentiality rules are respected

Smartcard issuance

➢ Smartcard:
- Strictly personal
- Fingerprints saved only on the smartcard. Maximum two fingerprints (CNIL)
- Not a photograph, characteristic features. Impossible to recreate the fingerprints from the smartcard
- Data is stored only on the smartcard

➢ In case of suspected loss/theft, notify the CASD as soon as possible. We will take reversible measures to reduce the risk of fraudulent use. If the loss of the smartcard is confirmed, the card will be permanently cut off and you will have to return to the CASD to get a new one.

Anonymization techniques

Data anonymization

In order to respect the statistical secrecy of outputs, data must be **anonymized.**

« ***The anonymization****, according to the CNIL,* ***is a process using a set of techniques rendering impossible, in practice, any individual identification by any mean and in an irreversible way***»

→ Method recommended by the CNIL: **generalization**, which means transforming data to make them refer to a set of persons instead of a single person.

Three requirements

In order to be anonymized, data must respect three requirements:

1. **Non-individualization:** it must not be possible to isolate an individual from the dataset
2. **Non-correlation:** it must not be possible to link multiple datasets together concerning the same individual
3. **Non-inference:** it must not be possible to deduce near-certainly new information about an individual

Anonymization issues: thresholds

Two techniques:

- The aggregation: aggregate sufficiently the data in order not to have only X units per group

Age group	Number of individuals
20-24	2
25-29	7
30-34	6
35-39	1
40-44	3
45-49	9

Age group	Number of individuals
20-29	9
30-39	7
40-49	12

- Delete the information when it concerns less than X units

Enterprise category	Number of enterprises
Micro-enterprises	7
SME	2
Big enterprises	10

Enterprise category	Number of Enterprises
Micro-enterprises	7
SME	S
Big enterprises	10

Anonymization issues: thresholds

Two techniques:

- The aggregation: aggregate sufficiently the data in order not to have only X units per group

Age group	Number of individuals
20-24	2
25-29	7
30-34	6
35-39	1
40-44	3
45-49	9

Age group	Number of individuals
20-29	9
30-39	7
40-49	12

- Delete the information when it concerns less than X units

Enterprise category	Number of enterprises
Micro-enterprises	7
SME	2
Big enterprises	10
Total	19

Enterprise category	Number of Enterprises
Micro-enterprises	S
SME	S
Big enterprises	10
Total	19

- **Pay attention to the secondary secrecy !**

This technique presents a risk of re-identification because you have to pay attention to the fact that other available data may allow to recalculate the masked value. This is why, what we call the secondary secrecy must be applied: more than one value must be deleted. In the example that we have, if we find the total number of enterprises, we can easily recalculate the deleted number of SME.

Anonymization issues: the diversification

Achieve a distribution of group characteristics that is sufficiently diverse to reduce the risk of certain or near-certain deductions

Diagnoses taken care of during hospital stays for a given month and a given department						
Age group	**Number of patients**	**Hypertension**	**Diabetes**	**Asthma**	**Cancer**	**Respiratory deficiency**
20-29	13	0	4	13	0	0
30-39	36	6	10	9	5	7
40-49	52	15	9	11	16	8
50-59	49	14	11	6	10	8
60-69	53	12	9	8	11	23
70-79	58	8	31	12	6	56

Anonymization issues: high contributions

Avoid high contributions for amount variables

Business sector	Number of enterprises	Turnovers
Building construction	467	860 745
Civil engineering	389	1 696 872
Special construction work	804	973 610

Business sector	Maximum turnover	Maximum's percentage
Building construction	256 804	29,83%
Civil engineering	1 531 794	90,27%
Special construction work	41 947	4,30%

Confidentiality rules

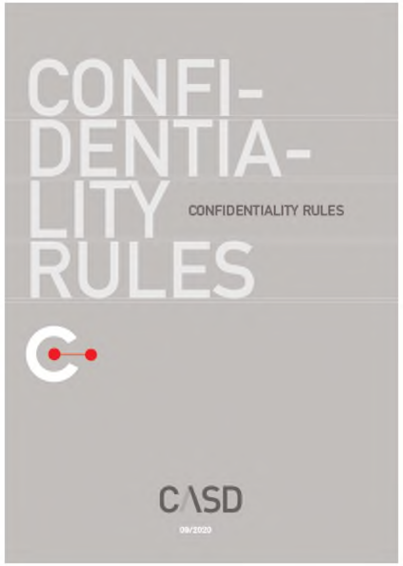
Confidentiality rules: general rules

- Household data
  - **The knowledge of an individual characteristic** cannot lead to **the knowledge of another one on the same individual**
  - The rules apply to natural persons (individual and individual enterprises)
- Firm data
  - No less than **3 units** per cell
  - A firm cannot account for more **than 85% of the total of an amount**
  - The rules apply to SIREN and not to SIRET
- Agriculture data:
  - No less than **3 units** per cell
  - A farm cannot account for more than **85% of the total of an amount**
  - The rules apply to farms and not to plots
- Tax data
  - Same rules apply for firms
  - For household data: no less than **11 individuals**
  - For tax-related data: a threshold of **11 units**, a unit cannot account for more than **85% of the total of a cell**

Confidentiality rules: specific rules

- In addition, specific rules per data sources exist, they are defined by the data providers

- Detailed in a document about confidentiality rules (sent by email after the enrolment session)

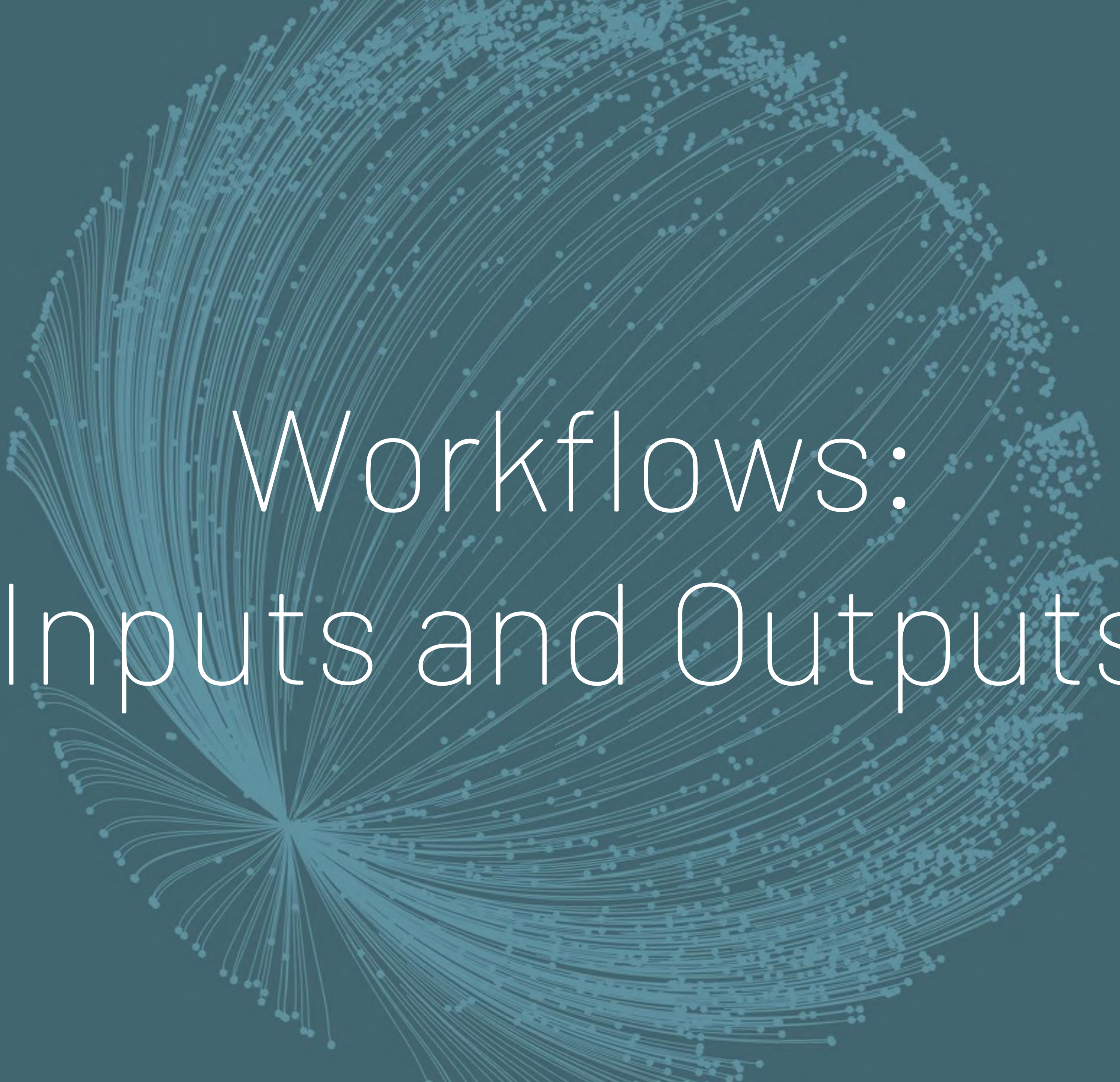


CONFI-
DENTIA-
LITY
RULES

CONFIDENTIALITY RULES

CASD

09/2020



Workflows: Inputs and Outputs

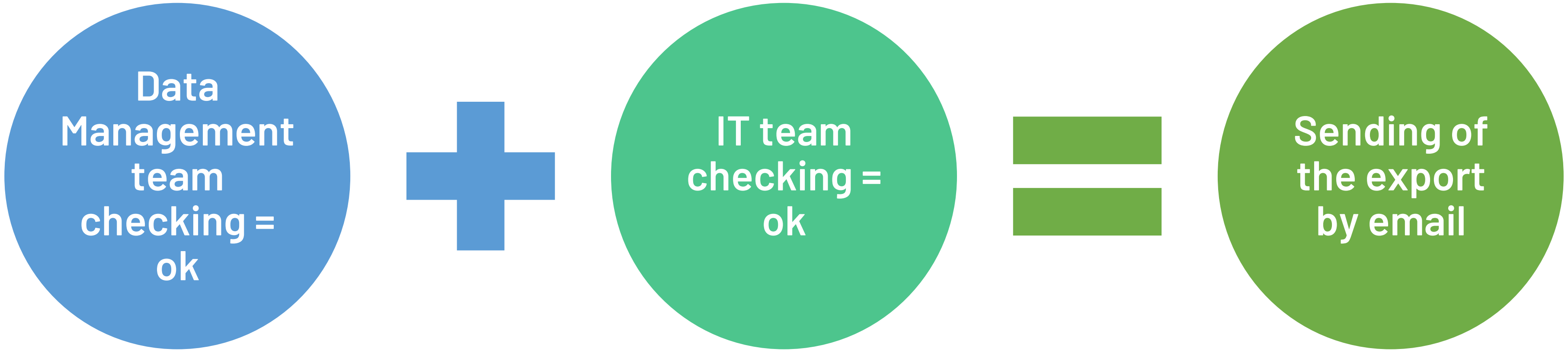
Output: definition

- Definition: output of non-confidential data outside of the secure environment

- Two types of outputs exist:
  - Manual output: needs the CASD checking of **every file**.
    - → The majority of projects with access to public statistical data.
  - Automatic output: files are directly sent without checking by the CASD.
    - → Mainly projects with access to health data, among others.

- **You are solely responsible for the respect of confidentiality rules in output files**

Manual output: two processing



Data Management team checking = ok

+

IT team checking = ok

=

Sending of the export by email

If the output respects confidentiality rules, and depending on its level of complexity, you should receive it within <u>48 hours</u>

Manual output: counting

- At the beginning of your project, you have a pack of **20 exports** (for all project members, not per one member)

- **One export** equals to **30 minutes** of processing time by the Data Management department to check one or multiple exports

- It means you initially have 600 minutes of checking time

  Example : 1st output : 5 min of checking time (do files without data)

  2nd output : 35 min of checking time (results tables)

  ➔ Counting : 40 min of checking time, being 1 export and 10 min use. You have 19 exports left

- If you reach 20 exports, you can order an additional pack (10 exports)

Manual output: how to proceed?

- Send an email to request an output in which you indicate **your project name** and the **reference of the output** (32 characters)
- Insert a **file describing** the output. It must indicates:
  - **the data used** so we can directly know which rule must be applied and the **signification of every used variable**
  - for regression, econometric models: the number of observations used
  - for maps and graphs: the population and the definition of the used variables
  - for aggregated results tables: variables description, each cell's number of observations and information on the highest contribution in each cell for results regarding **amount variables** (in a non-confidential control file)

Control file example

For results regarding amount variables, add columns indicating the **maximum** and **maximum's percentage** of amount variables.

➔**Create an additional file for the control, it will be deleted from your output before we send it to you**

Control file

Business sector	Number	Total turnover	Maximum turnover	Maximum's percentage
Buildings construction	43 467	860 745	256 804	29,83%
Civil engineering	22 389	1 696 872	1 531 794	90,27%
Specialized construction work	61 804	973 610	41 947	4,30%

Output file

Business sector	Number	Total turnover
Buildings construction	43 467	860 745
Civil engineering	**S**	**S**
Specialized construction work	61 804	973 610

Inter-projects export

- ➢ Transfer of an output from one project to another (whether you are a member or not)
  - ➔ Do an output request as indicated previously
  - ➔ In your email to service@casd.eu requesting the output, specify:
    - ➔ That the output is a transfer between two projects
    - ➔ The **name of the two projects concerned**
  - ➔ After checking, the export will be directly transferred from one project environment to the other
- ➢ **Warning!**

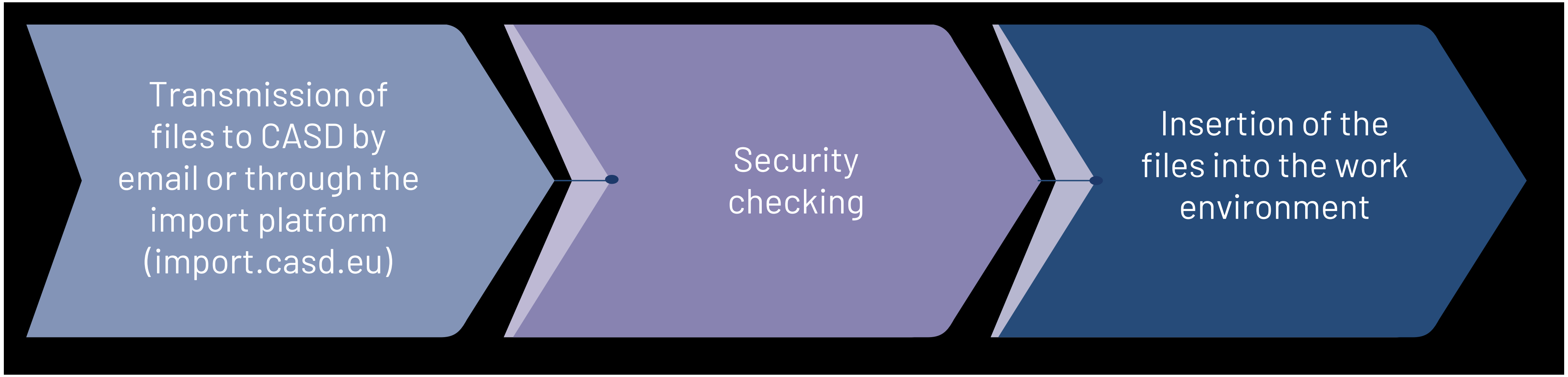
If the project of origin is accredited for different data sources than the project receiving the export, you will need to provide a detailed description of the files to be transferred in order for CASD to verify that the export does not contain data for which the receiving project is not accredited.

Automatic outputs

- An automatic output does not need CASD checking
- You will directly receive the files in your mailbox
- A form must be filled engaging you to some obligations (they differ depending on the data accessed)
- Thresholds of size and frequencies are implemented (they differ depending on the used data)
  - Example : 10 outputs per day per user with a maximum size of 10 Mo per output

Input: definition and process

- Definition : insertion by CASD into your work environment of files that you send us in a readable format


Transmission of files to CASD by email or through the import platform (import.casd.eu)
Security checking
Insertion of the files into the work environment

In the email of input request, specify your **project name** and add a **file description** (file type and its content)

Input: rules

- Input of **individual data** destined to be matched with other CASD data is only possible if it has been explicitly declared in your research project submitted for your authorization

- Data producer agreement is required to import any data unless it is public data

- Can only be imported « inactive » objects (non executable)

- It is only possible to import files in formats of the available software in the CASD environment (txt, csv, SAS, R, STATA...)

➢ **Pay attention** not to forget to encrypt confidential data

Open access data

Some open access data can be retrieved from the folder “Libre Acces” in the folder “Raccourcis”:

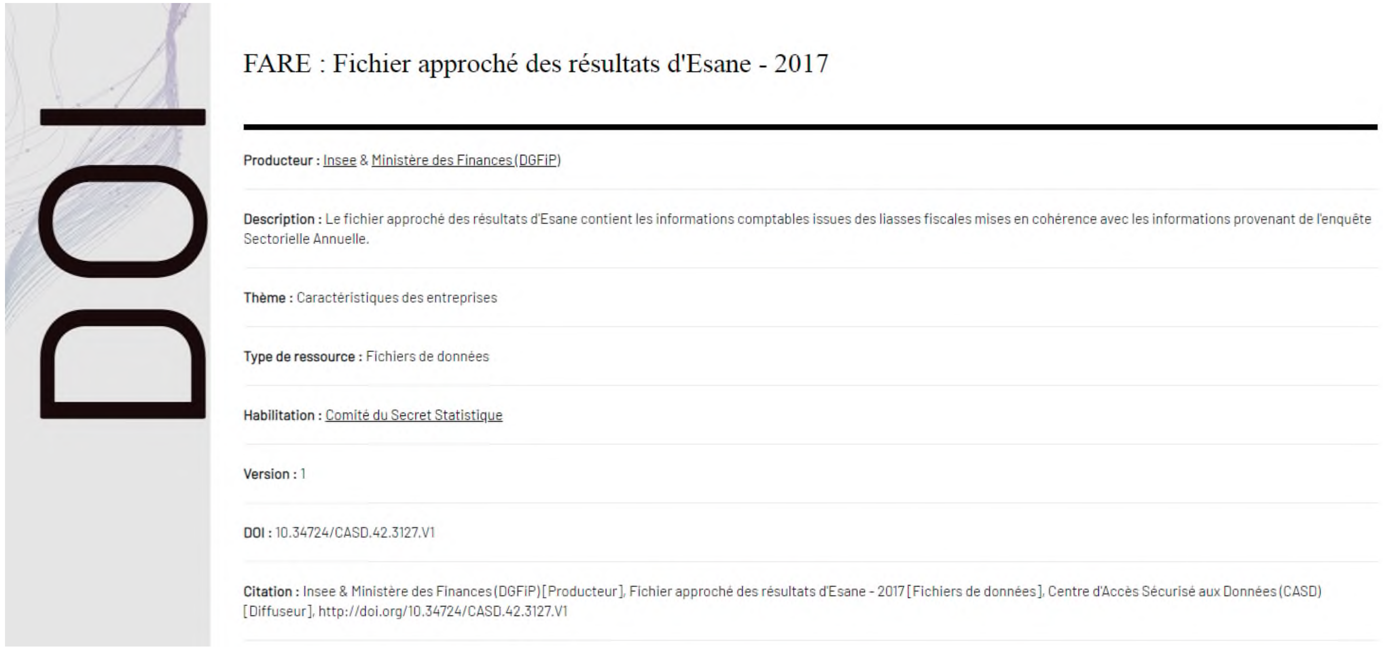
- GEOFLA, ADMIN-EXPRESS and Contour IRIS of the IGN database
- Sirene: register of enterprises and establishments – stocks on the 1st of January of each year
- Geographical Classifications and NAF, NES, PCS, COICOP, CPF, CJ
- Legal population
- ESANE documentation
- INSEE methodology sheets and SAS macro (CALMAR, CUBE, data analysis)
- National Register of Health and Social Establishments (FINESS)
- Drug and Tariff Information Base
- Other CASD documents:
  - CASD user guide
  - Confidentiality rules
  - Output good practices

Data citation and publications sharing

Data citation

- Cite the used data in your publication!
- A template for citing data and their DOI (Digital Object Identifier) is available on each source webpage in the section "Persistent identifiers"

Example for FARE 2017: Insee & Ministère des Finances (DGFiP) [Producer], Fichier approché des résultats d'Esane - 2017 [Data file], Centre d'Accès Sécurisé aux Données (CASD) [Diffusor], http://doi.org/10.34724/CASD.42.3127.V1



FARE : Fichier approché des résultats d'Esane - 2017

Producteur : Insee & Ministère des Finances (DGFiP)

Description : Le fichier approché des résultats d'Esane contient les informations comptables issues des liasses fiscales mises en cohérence avec les informations provenant de l'enquête Sectorielle Annuelle.

Thème : Caractéristiques des entreprises

Type de ressource : Fichiers de données

Habilitation : Comité du Secret Statistique

Version : 1

DOI : 10.34724/CASD.42.3127.V1

Citation : Insee & Ministère des Finances (DGFiP) [Producteur], Fichier approché des résultats d'Esane - 2017 [Fichiers de données], Centre d'Accès Sécurisé aux Données (CASD) [Diffuseur], http://doi.org/10.34724/CASD.42.3127.V1

Publications sharing

- If your project benefits from a subsidized rate, we ask you to mention the CASD in your publication, in the following terms:

« Access to some confidential data, on which is based this work, has been made possible within a secure environment offered by CASD - Centre d’accès sécurisé aux données (Ref. ANR-10-EQPX-17) » [English Version]

« L’accès à certaines données utilisées dans le cadre de ce travail a été réalisé au sein d’environnements sécurisés du Centre d’accès sécurisé aux données – CASD (Réf. ANR-10-EQPX-17) » [French Version]

- Do not forget to inform us about your publication by filling our online form: https://www.casd.eu/en/share-a-new-article/

Support

CASD Support

The Data

- Available data
- Data documentation
- Access procedure
- Enrolment
- Opening the right to access the data
- Importing files in your work space
- Results exports

Data Management Service

01 84 19 69 24

IT

- Your access tools: biometric card, SD-Box…
- Connection issues
- Technical issues on your project server
- Server configuration and modification: hardware, software

IT Service

01 84 19 11 37

Contracts and billing

- Your contracts
- Fees estimate
- Your billing
- Payments means

Project Management Service

01 70 26 69 32

Generic email address: service@casd.eu et acces.pmsi@casd.eu
CASD website: https://www.casd.eu

Quiz

- You just received an email with an access link
- 15 questions with only **one possible answer**
- The quiz is **mandatory**
- Goal: estimate your understanding and explain the parts which may have caused some understanding difficulties

We thank you for your attention

CASD