Le Centre d'Accés Sécurité aux Données

la sécurisation des données pour la recherche en science sociale, Semaine DATA-SHS

Ahmed Tritah
Université de Poitier, Laboratoire d'Economie de Poitiers, Institut des Migrations, FR CNRS TEPP, Mines Paris PSL
ahmed.tritah@univ-poitiers.fr

Poitiers, 12 décembre 2023

Sécurisation des données

- Production importantes de données administratives dans le cadre des activités des entreprises et des individus,en société : données fiscales..
- Production de données dans le cadre des parcours de soins, scolaires, judiciaire –> suivi longitudinal
- Production de données par des organsimes publics très variés (éducation, santé, travail, fiscalité, justice, agriculture, environnement, ...)
  1. Depuis la mise en application du RGPD, les garanties de sécurité sont devenues des exigences juridiques fortes.
  2. Mais également demande sociétale plus forte pour la transparence, l'évaluation (Réforme constitutionel de 2008)...et pour affiner le microscope du chercheur en SHS (**observer, mesurer pour comprendre**)
- Comment concilier les deux ? : solution technologique

Solution apportée par le CASD : un écosystème vertueux

1. La **SD-Box**, boîtier informatique sécurisé d'accès, permet d'accéder à distance à une infrastructure sécurisée où les données confidentielles sont sanctuarisées.
   –> lieu de **stockage** et de **traitement** des données est appelé « **bulle sécurisée** ».
   La SD-Box,est très simple à installer, facile à remplacer avec une mise à jour à distance régulière.
2. Couplée à l'infrastructure informatique centrale « étanche »,

- (1) et (2) forme un ensemble cohérent de services maîtrisés de bout en bout garantissant un très haut niveau de sécurité que le CASD fournie aux producteurs de données –> **gagner la confiance**

Quelles données ?

- **INSEE** : ministères de la Justice, de l'Éducation nationale, de l'Agriculture et de l'alimentation, de l'Economie et des Finances (données fiscales)...
  –> un décret d'application précise que l'accès doit s'effectuer par le biais du CASD.
- En **santé** : données sur les séjours hospitaliers publics et privés en France (données PMSI de l'ATIH), données de cohortes de santé (Constance, Gazelle, ...)
- Dans le **privé**, partenariat avec le CASD : offrir une sécurité des données pour permettre un accès externe dans le cadre de collaboration (chercheurs, start-up, consultants, etc.)
  $\Longrightarrow$ démarche actuelle d'«open innovation» mêlant à la fois savoir-faire métier, recherche, et capacités d'innovation des start-up.

Quelques chiffres

- 510 sources de données mises à disposition de façon sécurisée : Données Diponible
- Un total de 1447 projets gérés et hébergés depuis son lancement : SELECTION DE PROJET
- 982 institutions utilisatrices pour 4977 utilisateurs de données sécurisées depuis son lancement
- Plus de 400 publications et communications (articles, chapitres d'ouvrages, ouvrages, thèses, rapports, conference papers, etc.), référencées et réalisées par les utilisateurs des données mises à disposition par le CASD

L'accés aux données : procédures administratives

Statistique publique ou de données fiscales

1. Soumettre un projet de recherche au **comité du secret statistique (CSS)**
   Objectif : lever le secret statistique ou fiscal (protecteur) pour les membres (uniquement eux) du projet

2. Le CSS vérifie que le projet peut être qualifié de projet de recherche scientifique et que les porteurs du projet sont des chercheurs.
   Comité composé des producteurs des données et des représentants des chercheurs
3. Après cette instruction, le comité émet un avis, suivi d'une décision de l'administration des Archives nationales, ou du ministre du Budget (données fiscales). **La Cnil intervient s'il s'agit de données personnelles**.

L'accés aux données concrétement : procédures administratives

4. **Séance d'enrollement** (Palaiseau) nécessaire avant tout accès

   au CASD pour les chercheurs habilités
   - Formation et sensibilisation aux lois de protection et confidentialité aux respect des règles du secret statistique
   - Règles de sécurité informatique (conditions d'utilisation signées par chaque utilisateur) : accès strictement personnel, obligation de retirer sa carte d'authentification lorsque l'on s'absente du poste SD-Box, etc.
   - conditions d'hébergement du boîtier (contrat entre le CASD et l'établissement où sera installée la SD-Box).
     la SD-Box doit être installée dans un local fermé à clé, l'écran ne doit être visible que par son utilisateur, etc.
5. À l'issue de la séance, on obtient sa carte d'accès et on l'on appose ses empreintes digitales

Synthèse phase accés aux données

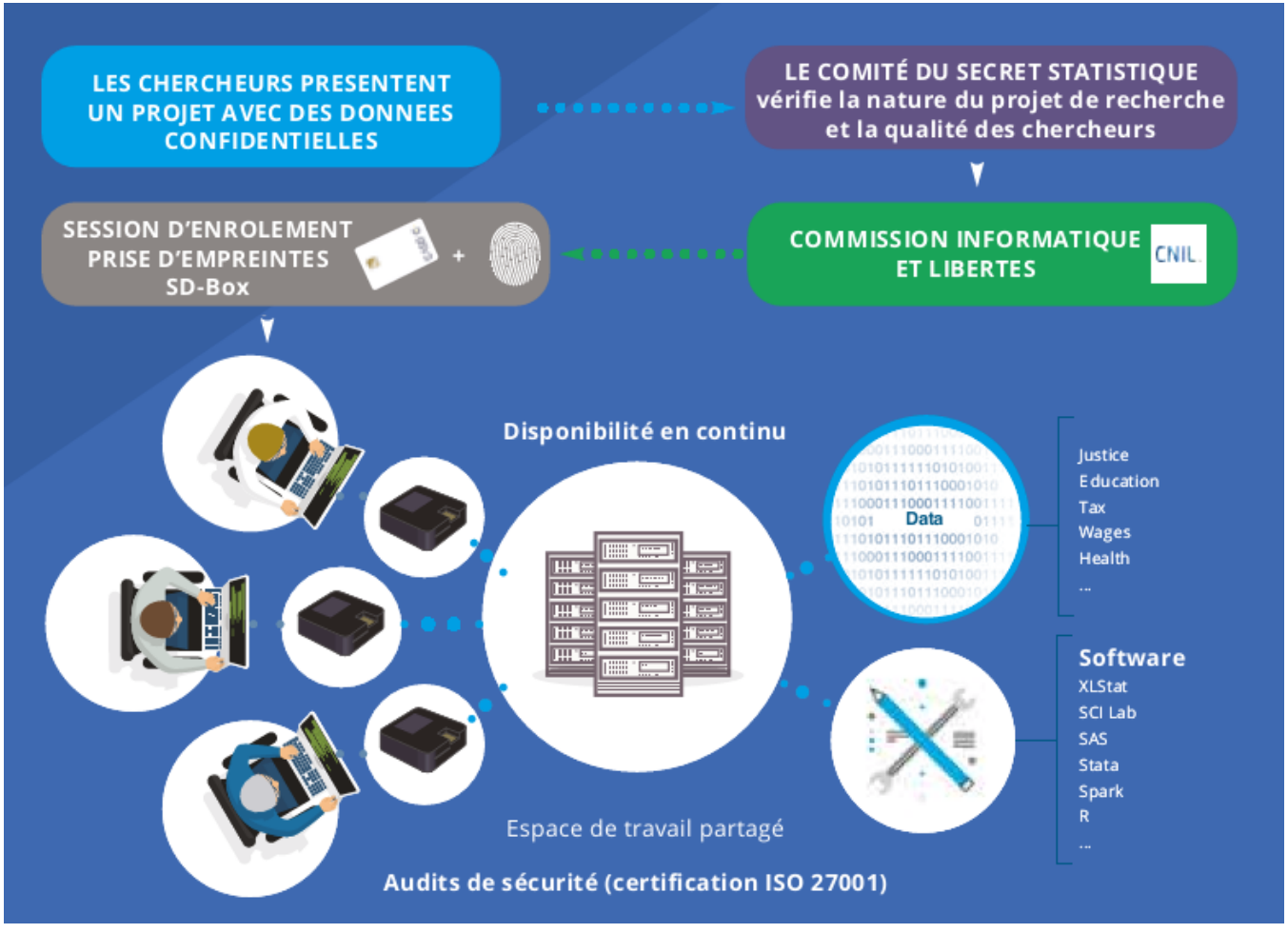

FIG.: Source : Economie et Statistique (2019)

Coûts

- Relativement couteux au regard des moyens dont disposent les labo en SHS : (location du boitier par mois, coût par utilisateur dégressif avec le nombre d'utilisateur, et progressif avec la configuration choisie, tarif spécifique pour les doctorants)
- Nécessité de prévision budgetaire : cycle de production de la recherche
- Possibilité de mettre le projet en "hybernation" (suspendre l'abonnement pendant la phase de soumission, etc.)

Le travail sur le CASD au quotidien une autonomie avec contrôle

- Après l'enrôlement, envoie et installation de la SD-Box dans l'établissement de ratachement du chercheur : brancher un écran et un clavier et se connecter au réseau.
- Travail en tout "autonomie" mais en respectant les contraintes de sécurisation :impossible de récupérer aucun fichier à partir du boitier (directement).
- Fichiers récupérés aprés requêtes qui comprend l'envoie des codes de production de la requête, une description de la requête et des variables de sorties
- Les résultats sont dépôsés dans un espace du serveur réservé à cet usage

Attention : Nombre de requête illimité mais payantes (10 export par défaut, au-dela achat de pack d'export...)

Le travail sur le CASD au quotidien une autonomie avec contrôle

Le contrôle de confidentialité

- Impression de fichiers, transfert de fichiers, opérations de copier-coller, etc. ; impossibles.
- Contrôle a priori effectué par les gestionnaires du CASD : vérifient que l'utilisateur a bien réalisé le nécessaire pour garantir que les fichiers de résultats satisfont aux règles de confidentialité définies par le producteur des données.
- **Cas des données de santé** : procédure automatique, sans vérifications manuelles.
  - on renseigne un formulaire électronique d'engagement a respecté les règles de confidentialité. Fichier transmis automatiquement par messagerie avec lien de téléchargement sécurisé. Copie des fichiers conservée cinq and par le CASD pour des contrôles a posteriori.

Avantage pour les producteurs de données

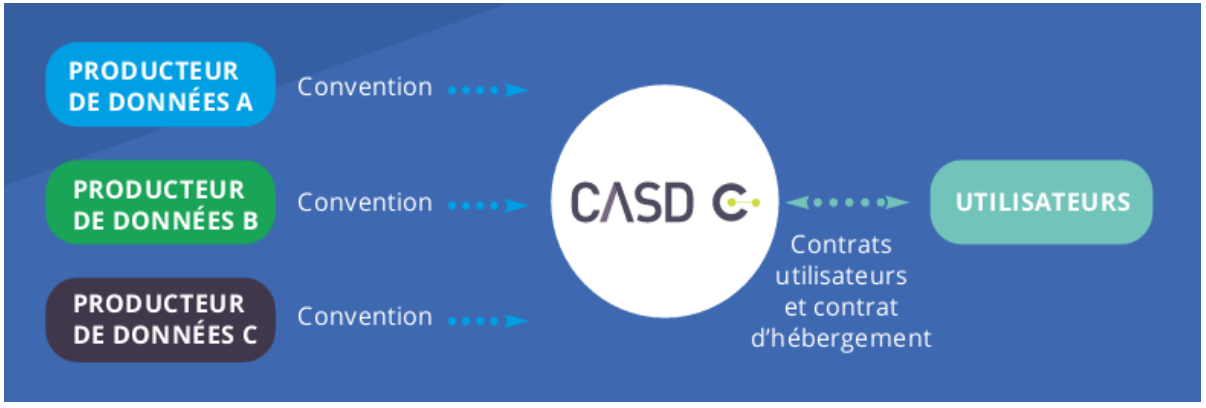

FIG.: L'intermédiation du CASD allège la charge du producteur et facilite la mutualisation des sources
Source : Economie et Statistique (2019)

Intérêt pour les chercheurs

- Possibilité de mobilisées des données provenant de plusieurs producteurs par utilisation conjointe ou par appariement au sein d'un même environnement de travail.
  Exemple : données de l'Insee couplées/appariées avec les données DGFiP
- Avantage très important par rapport à d'autres pays (Ex. Royaume Unie)
- Un modèle de spécialisation et de mutualisation qui présente un avantage pour l'utilisation des données françaises.
  –> offre probablement unique au monde
- De plus en plus de chercheurs européens et aux Etats-Unis demandent désormais l'accès aux données françaises.

Importance des appariements pour saisir les aspects multidimensionels des problématiques

- Démultiplier la puissance d'information et d'explication
- Beaucoup d'études et d'évaluation sont possibles ou enrichis par les possibilités d'apparier
  - Liens entre revenus salariaux et les revenus de remplacement
  - trajectoires scolaires et trajectoires professionnelles
  - trajectoires santé et trajectoires d'emploi, etc.
- Concevoir, mettre en place et évaluer des politiques publiques
- Par rapport aux enquêtes : moins couteux, et exhaustivité
- Appariement (fait par le CASD parfois) :
  - effectué sur la base du NIR (Numéro d'Inscription au Répertoire national d'identification des personnes physiques) depuis 2016 (version criptée "NIR hâché")
  - SIREN/SIRET pour les entreprises

Les enjeux de la reproductibilité

- Reproductibilité : nécessité pour la garantie de la scientificité de la recherche
- Les revues demandent de déposer données et code pour que les résultats publiés puissent être reproduits
  –> difficultés s'agissant de données confidentielles
- Solution côté CASD : partenariat avec une agence de certification (Cascad)
  - L'agence vérifie la conformité des résultats, en amont des soumissions à une revue scientifique.
  - Certification attribuée à l'issue d'un processus d'évaluation à partir des données sources présentes sur le CASD et de l'ensemble des codes informatiques mis à disposition par le chercheur.
- Objectif : augmenter les chances de publication les revues académiques.

Ouvertures internationales

- Possibilité de mutualiser des projets autour autour de plusieurs centres de depôt sécurisés en Europe : études comparatives par exemple
- France, Allemagne, UK, Pays-Bas

Conclusion

- Le CASD : écosystème d'avantgarde par sa technologie et le nombre de données déposées.
  –> auditionnés par la congrés américain (American Congress, 2017).
- Confiance créée du côté des producteurs <–> nombre croissant de données disponibles
- Perspective importante en santé : articulation avec les sciences économiques et sociales de plus en plus importants.
- Côté chercheurs
  - aversion aux contraintes à passer par un système sécurisé
  - Coûts : nécessité de financemet externe (AAP)
- Limites compensées par le nombre et la richesse des nouvelles données
- Barrières du financement : enjeux de transparance, équité et diversité scientifique des approches en SHS

Stimuler un cercle virtueux d'offre de demande

Qualité de l'information pour une recherche de qualité