

Des données confidentielles sécurisées pour les datascientists

Kamel Gadouche, Directeur du CASD.

Le CASD est un équipement permettant aux chercheurs de travailler à distance, de manière hautement sécurisée, sur des données individuelles très détaillées. On peut qualifier ces données de confidentielles car elles sont le plus souvent couvertes par un secret : secret professionnel, secret des affaires, secret statistique, secret fiscal, secret médical etc.

Les données présentes sur le CASD sont donc toutes d'une grande précision, identifiantes ou indirectement identifiantes, et contiennent une grande richesse d'information. La mise à disposition de ces données ne peut se faire que dans des conditions de sécurité très élevée garantissant leur confidentialité ainsi que leur traçabilité.

C'est pour répondre à ce besoin de sécurité que le CASD a conçu en 2009 un dispositif spécifique reposant sur un petit boîtier d'accès dédié (appelé SD-Box) totalement sécurisé et autonome qui permet d'accéder à distance à une infrastructure sécurisée où les données confidentielles sont sanctuarisées. Cet endroit de stockage et de traitement des données est appelé « bulle » ou parfois « enceinte ». Le principe est qu'aucune donnée ne peut sortir de cette bulle sans contrôle et ceci afin de prévenir tout risque d'évasion de fichiers de données. Le contrôle d'accès de l'utilisateur est réalisé à l'aide d'une authentification forte

s'appuyant sur une carte à puce contenant un certificat de sécurité et un lecteur biométrique d'empreintes digitales. Conformément à la loi, ce traitement a fait l'objet d'une demande d'autorisation à la commission informatique et liberté qui a été accordée (CNIL - délibération n°2014-369).

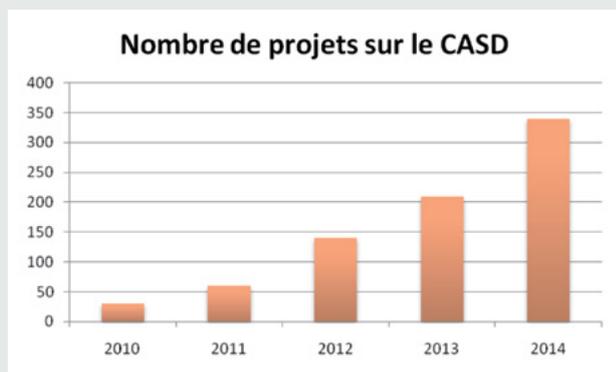
La technologie du CASD a été initialement conçue et développée en 2009 au sein du Genes qui était alors une direction de l'Insee. L'objectif pour l'Insee était de pouvoir mettre à disposition des chercheurs ses données sans prendre le risque de voir ces fichiers de données circuler de main en main, sous tout type de supports : portable, clé usb... Ce risque est réel et potentiellement très préjudiciable pour l'institut en cas d'incident : perte du support, support qui se retrouve entre de « mauvaises » mains...

La mise à disposition de données confidentielles à des tiers se confronte de manière systématique au risque de dissémination qu'elle pourrait induire. Cette problématique n'est pas uniquement vraie pour les données de l'Insee. De nombreux autres organismes y sont confrontés que ce soit dans le public ou dans le privé.

Dans le public, le CASD met aujourd'hui à disposition des données des ministères de la justice, de l'éducation, de l'agriculture, des finances pour

LE CASD EN QUELQUES CHIFFRES

Aujourd’hui, plus de 120 sources de données sont disponibles sur le CASD pour près de 350 projets, soit près de 1000 utilisateurs en France et en Europe.



Bien que ce ne soit pas exhaustif, il a été possible de recenser une centaine de publications dans des revues scientifiques s’appuyant sur des travaux réalisés sur le CASD.

Le CASD est une entité du Genes qui compte vingt personnes et qui est composée d’un secrétariat général, d’un service statistique, d’un service infrastructure et développement informatique et d’une cellule datascience.

Le budget du CASD est d’environ 2M€ par an provenant d’une part des subventions obtenues à la suite d’appels à projets français (investissements d’avenir) ou européens et des contributions des partenaires du projet (Insee, CNRS, ENS Cachan, Ecole Polytechnique et HEC), d’autre part des recettes provenant de la facturation du service : soit auprès de l’utilisateur final, soit auprès des sociétés privées qui font appel aux services du CASD. Cette part provenant des recettes de la facturation est en forte croissance et est amenée dans les prochaines années à couvrir l’ensemble des coûts du service.

les données fiscales... Pour ces dernières, il a été nécessaire de modifier la loi (loi ESR de 2013) et qu’un décret soit publié en 2014 pour qu’elles puissent être mises à disposition des chercheurs. Le décret précise explicitement que l’accès ne peut s’effectuer qu’au moyen du centre d’accès sécurisé aux données (CASD) du Genes.

Dans le privé, plusieurs sociétés se sont adressées au CASD parce qu’elles désiraient accroître la sécurité de leurs données en accès externe dans le cadre de collaborations avec des chercheurs, des start up ou des consultants. C’est le cas par exemple de la banque postale, de Generali, de la banque publique d’investissement (BPI), de la société MAPP, d’Erdf, de RTE... Dans ce dernier cas, il s’agissait de mettre en place un data-lab sécurisé pour travailler sur des technologies big data dans un environnement distant et sécurisé. Pour ces sociétés, confier leurs données au CASD a été un moyen de répondre à leur besoin de sécurité et de permettre ainsi la réalisation de projets collaboratifs innovants impliquant de nombreux acteurs externes. Ces exemples s’inscrivent parfaitement dans la démarche actuelle d’« *open innovation* » mêlant à la fois le savoir-

faire métier, la recherche, et les capacités d’innovation des start up.

Les données de santé : un enjeu important pour le CASD

Le CASD est aujourd’hui confronté à une forte demande venant du domaine de la santé.

La loi de santé de Marisol Touraine qui vient d’être examinée par l’Assemblée nationale, prévoit dans son article 47 un accès aux données médico-administratives facilité pour les chercheurs, à condition que cet accès soit suffisamment sécurisé pour garantir la confidentialité et la traçabilité des données.

C’est dans ce contexte que des tests sont actuellement menés au CASD pour l’accès aux données de cohortes et aux données médico-administratives. Le CASD démarre en ce moment une expérimentation de mise à disposition pour les chercheurs des données de la cohorte Memento (suivi de 4000 patients atteints de la maladie d’Alzheimer). Des tests vont aussi démarrer au CASD pour l’accès aux données de la cohorte

Constances, la plus importante cohorte épidémiologique « généraliste » de France constituée d'un échantillon représentatif de 200 000 adultes âgés de 18 à 69 ans.

Dans le domaine de la santé, le besoin de sécurité est au moins le même que dans les autres domaines, mais la nature et le volume des données changent ainsi que les usages associés. Ceux-ci sont beaucoup plus diversifiés. On observe en ce moment un grand nombre de nouveaux profils d'utilisateurs du CASD : des médecins, des épidémiologistes, des bio-statisticiens, des data-analystes de la santé etc.

Dans ce domaine, les données peuvent rapidement devenir volumineuses, en particulier lorsque celles-ci contiennent des informations génomiques ou de l'imagerie. Les nouvelles possibilités de croisement de données augmentent aussi mécaniquement le volume des données à traiter. Le CASD a d'ores et déjà commencé à intégrer dans son architecture sécurisée des technologies issues du monde du big data, ouvrant ainsi de nouvelles possibilités d'exploitation des données de gros volume dans le domaine de la santé.

Le CASD est fortement impliqué dans deux projets européens

En 2013, le CASD a mis en place une infrastructure d'accès distant dans le cadre d'un pilote pour le projet collaboratif européen DARA, avec l'Allemagne, la Grande-Bretagne, la Hongrie et le Portugal pour le compte d'Eurostat. En 2014, le CASD a poursuivi le maintien de ce pilote en vue de sa présentation aux différents acteurs

européens impliqués dans la potentielle généralisation de cette solution.

Le CASD participe à un grand projet européen DwB (*Data without Boundaries*) impliquant 21 pays ayant pour objectif de favoriser l'accès aux micro-données par les chercheurs. Un des sous-projets concerne la réalisation d'un réseau de centres d'accès sécurisés. Le CASD a réalisé en 2014-2015 une architecture technique pour ce réseau EURAN (*European remote access network*) s'appuyant sur sa technologie. Dans le cadre de ce projet, le CASD a également participé à d'autres actions, comme celle de permettre à des chercheurs européens non-français de travailler sur des données françaises individuelles et très détaillées par le biais du CASD. Par ailleurs, le CASD suscite également l'intérêt d'instituts nationaux statistiques étrangers qui souhaiteraient pouvoir mettre en œuvre un centre d'accès sécurisé.

Les défis futurs du CASD

Après la mise à disposition des données statistiques lors de la création du CASD, un grand pas a été franchi en 2014 avec la mise à disposition des données fiscales. Aujourd'hui, le CASD dans sa phase de développement doit faire face à au moins quatre grands défis : la mise à disposition des données de santé qui devrait monter en puissance à la suite de l'adoption de la loi Santé, l'intégration de plus en plus importante des technologies du big data dans son offre, son développement dans le cadre européen, et enfin la valorisation de plus en plus importante de la technologie dans le secteur privé. Pour ce dernier défi, le Genes a déjà largement bénéficié du soutien de la communauté des Alumni et compte beaucoup encore sur elle... ■



La SD-Box.