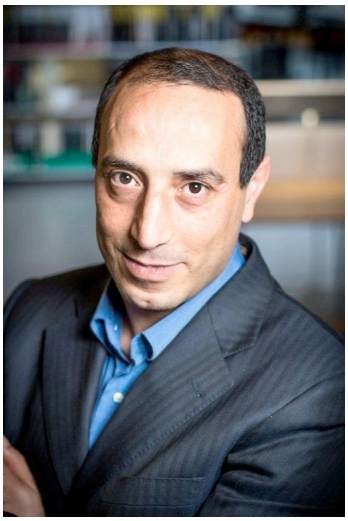
PROGEDO
DATA INFRASTRUCTURE

actu
en 13 968 caractères

n°2 - avril 2015



EDITO

Kamel Gadouche,
Directeur du Centre d'Accès Sécurisé à Distance (GENES)

La future loi de santé proposée par Marisol Touraine, faisant grand bruit dans les médias et dans la communauté de recherche, pourrait permettre un bond en avant pour la recherche scientifique dans le domaine de la santé.

Actuellement en examen à l'Assemblée, l'article 47 de cette loi permet en effet d'ouvrir plus largement l'accès aux données de santé et facilite de ce fait les travaux de nombreux chercheurs.

L'enrichissement des données disponibles proposé par le texte, permettrait aux scientifiques de trouver directement les données nécessaires à leurs travaux centralisées dans le nouveau Système National des Données de Santé (SNDS), au lieu de s'adresser à différents instituts. Cette base de données, dont la création est prévue par l'article 47, serait une mine d'informations regroupant les données de l'assurance maladie (SNIIRAM), des séjours hospitaliers (PMSI), les causes de décès, les données sur le handicap, et un échantillon de données de l'assurance maladie complémentaire. Cette base constituerait en elle-même une petite révolution car elle serait pratiquement unique au monde.

L'article 47 prévoit aussi la création d'un Institut national des données de santé qui permettrait d'unifier la procédure d'accréditation pour l'accès aux données, actuellement réalisée auprès de différents organismes que les chercheurs peinent bien souvent à contacter en raison de leur dispersion. Cet article prévoit également un allégement des formalités pour pouvoir réaliser des traitements impliquant le numéro de sécurité social (NIR). Aujourd'hui, un chercheur souhaitant employer le NIR doit faire publier un décret en Conseil d'Etat pour en obtenir l'autorisation ; autant dire que cette procédure est inaccessible dans la plupart des cas. A l'avenir, l'article 47 prévoit une procédure moins lourde et plus rapide concernant les projets de recherche ou d'études en santé pour lesquels les chercheurs pourraient se contenter d'obtenir une demande d'autorisation de la CNIL.

Toutes ces avancées impliquent la mise en œuvre d'un niveau de sécurité beaucoup plus élevé concernant l'accès aux données, niveau qui en assure la confidentialité, l'intégrité et la traçabilité.

C'est tout le travail auquel nous nous consacrons au CASD depuis plusieurs années.

DONNEES DE SANTE

L'accès aux données de santé : un enjeu majeur pour la recherche

La future loi de santé (voir édito) prévoit 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 des séjours hospitaliers: le Programme Médicalisé des Systèmes d'Information (PMSI).

Ces derniers mois, une collaboration étroite entre le CASD et l'Institut thématique multi-organismes de Santé publique (ITMO) Santé publique et l'alliance AVIESAN s'est établie. De nombreuses rencontres entre le CASD et des propriétaires de données de cohortes de santé ont eu lieu récemment. C'est ainsi que 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). Plusieurs discussions sont également en cours dans le cadre d'accès aux données de la cohorte constances (200 000 personnes âgées de 18 à 69 ans) sur le CASD.

Dans le domaine de la santé, les données peuvent rapidement devenir volumineuses, en particulier lorsque celles-ci contiennent des informations génomiques ou de l'imagerie. De plus, les nouvelles possibilités de croisement de données augmenteront mécaniquement le volume des données à traiter. C'est pour cette raison que le CASD a d'ores et déjà commencé à intégrer dans son architecture des technologies issues du monde du Bigdata, ouvrant ainsi de nouvelles possibilités d'exploitation des données de santé de gros volume.

Un mouvement de fond s'est mis en place ces dernières années en France depuis le rapport de R. Silberman en 1999, et d'une manière encore plus marquée en Europe notamment grâce à la mise en place du CESSDA, pour répondre à un besoin d'accès aux données de plus en plus important pour la communauté de la recherche. C'est ainsi que la mise à disposition des données a été repensée de manière à accroître la sécurité des données, simplifier les procédures d'accès et en réduire les coûts de mise à disposition.

Le CASD, en pratique

Le CASD appartient au GENES (Groupe des Écoles Nationales d'Économie et Statistique) et fait partie de la TGIR PROGEDO. Il dispose d'un équipement permettant aux chercheurs de travailler à distance, de manière hautement sécurisée, sur des données individuelles très détaillées le plus souvent couvertes par le secret statistique, fiscal, médical etc. La richesse de ces gisements de données, pour la plupart longtemps demeurés inaccessibles aux chercheurs, offre aujourd'hui de nouvelles perspectives pour la recherche scientifique.

Ces données présentes sur le CASD étant toutes d'une grande précision, et même pour la plupart quasi-identifiantes, elles ne peuvent être mises à disposition d'utilisateurs que dans des conditions de sécurité très élevées. C'est pour répondre à ce besoin de sécurité que le CASD a conçu en 2009 un dispositif d'accès spécifique qui repose sur un petit boitier dédié (appelé SD-Box) totalement sécurisé et autonome permettant au chercheur, après s'être authentifié de manière biométrique, de travailler à distance sur les données et ce, sans qu'à aucun moment, il ne puisse extraire des données détaillées via par exemple un copier/coller, une clé USB ou une imprimante.

La technologie du CASD a été initialement conçue dans le but de fournir un accès sécurisé aux données individuelles très détaillées de l'Insee. Très rapidement, il est apparu que ce besoin n'était pas spécifique à l'Insee, bien au contraire : en peu de temps de nombreux autres organismes détenteurs de données ont demandé au CASD, parfois sous l'impulsion des chercheurs, d'héberger et de mettre à disposition leurs données. C'est par exemple le cas des ministères de la justice, de l'éducation, de l'agriculture, de la banque publique d'investissement, et du ministère des finances pour les données fiscales. Aujourd'hui, plus de 80 sources de données sont disponibles sur le CASD pour environ 1000 chercheurs qui les utilisent pour leurs travaux.

DONNEES

Nouveautés 1er trimestre 2015

SHARE

les biomarqueurs, une grande innovation pour la sixième vague

SHARE, la grande enquête européenne sur la santé et le vieillissement, innove en vague 6 avec la collecte de biomarqueurs par auto-prélèvement de gouttes de sang. L'inclusion d'un tel protocole dans une enquête en sciences humaines et sociales est une nouveauté qui a constituée un véritable défi méthodologique pour l'équipe SHARE France. Ces données biologiques compléteront les mesures de santé objective déjà effectuées dans les précédentes vagues (test de souffle, force de préhension, tests cognitifs). En France, un quart de l'échantillon longitudinal est concerné par l'introduction des biomarqueurs soit plus d'un millier de personnes. Certains des biomarqueurs pris en compte (cholesterol, C-reactive protein et interleukin-6) sont classiquement utilisés pour diagnostiquer les personnes fragiles. D'autres ont pour finalité de détecter des maladies chroniques, par exemple l'hémoblogine glyquée dans le cas du diabète. Ces nouvelles données permettront de franchir un pas important dans l'étude des relations entre les situations économiques et sociales, l'état de santé objectif et la consommation de soins chez les personnes de plus de 50 ans.

Panel Elipss

première enquête en ligne sur le portail Quetelet

Le panel internet Elipss (Étude Longitudinale par Internet Pour les Sciences Sociales), de l'équipement d'excellence DIME-SHS, vient de rendre disponible en ligne une première enquête.

Ce panel internet, représentatif de la population française, est constitué de 6000 personnes qui sont invitées régulièrement à participer à des recherches d'intérêt général dans de nombreux domaines (santé, environnement, politique, sport et loisirs). Elaborées par des chercheurs de l'université et du CNRS qui ne disposent pas de leurs propres moyens d'enquête par questionnaire, ces recherches ont une finalité exclusivement scientifique et permettent d'étudier l'évolution des comportements, des situations et des opinions dans la société française.

La première vague de l'enquête annuelle ELIPSS, réalisée en 2013, a été ajoutée au catalogue du portail Quetelet et ouvre la voie à la diffusion des données recueillies par ce dispositif d'enquête.

ESS

terrain de la 7ème édition achevé

Les données de la 7ème édition de l'enquête ESS ont été collectées de la mi-novembre 2014 à la mi-février 2015 : cette phase de terrain a permis d'interroger plus de 1900 personnes en face à face à leur domicile. Le questionnaire, d'une durée moyenne légèrement supérieur à une heure, a permis de recueillir un nombre très important de variables. Cette 7éme édition questionne particulièrement les «attitudes envers l'immigration» (module déjà posé lors de la 1ère édition de l'enquête) et les «inégalités sociales de santé». Le reste du questionnaire porte sur les modules qui sont répétés depuis la première vague réalisée en 2002: «confiances interpersonnelle et institutionnelle», «intérêt et participation politiques», «valeurs morales politiques et sociales», «intégration et exclusion sociale», «identités religieuses, ethniques et nationales» et «santé et bien-être».

La phase de vérification et de validation par l'équipe de coordination nationale est actuellement en cours. Les données seront ensuite transmises au centre de données norvégien (Norwegian Social Science Data Services - NSD), qui se chargera de la publication sur le site ESS Data au cours du second semestre 2015.

CASD

accès aux données fiscales

Suite à une modification législative et la publication d'un décret d'application, les chercheurs français accèdent désormais aux données fiscales sur le centre d'accès sécurisé aux données (CASD) du Groupe des Ecoles Nationales d'Economie et Statistique (Genes).

Ces données, collectées par l'administration fiscale, sont une source d'information incontournable pour l'analyse et l'évaluation des politiques publiques.

La possibilité d'accéder à ces données riches, précises et variées constituait un enjeu majeur pour la recherche scientifique en sciences économiques et sociales.

Fortement encadré par la loi, l'accès à ces données ne pourra se faire qu'après avoir obtenu un avis favorable du Comité du Secret Statistique et uniquement dans l'environnement sécurisé du CASD. Désormais, des publications d'études scientifiques innovantes s'appuyant sur ces données feront progresser la connaissance de notre société, et éclaireront le législateur et les pouvoirs publics

EVENEMENTS

DNSS2015

journée d'études sur la collecte, structuration, analyse et diffusion des Données Numériques en Sciences Sociales

Co-organisée par la Plate-forme Universitaire de Données de Lille, l'équipe du futur master Réseaux sociaux et numériques de l'Université de Lille et la MESHS, la journée d'étude DNSS2015 du 10 avril 2015 s'intéressera aux différents aspects du lien entre les données numériques et les sciences sociales.

Les moyens de récupérer ces données, de les transformer, de les exploiter et de les diffuser seront au cœur des échanges et alimenteront la réflexion autour des questions juridiques et éthiques qu'elles soulèvent.

Toutes les informations sur cette journée d'études sont sur http://dnss.meshs.fr/

OpenDataCamp Élections

le Hackathon des données électorales

A l'occasion de la mise à disposition les résultats électoraux du CDSP sur le site data.gouv.fr, le CDSP, le ministère de l'Intérieur et Etalab, la mission du secrétariat général pour le modernisation de l'action publique ont organisé conjointement à Sciences Po un hackathon le 23 février 2015.

80 étudiants, chercheurs, datascientists, et acteurs socio-politiques ont répondu à l'invitation à venir travailler ensemble sur les données électorales.

Des sujets aussi variés que la caractérisation du vote radical, l'analyse des noms des listes aux élections régionales 2010, les relations entre données socio-démographiques et comportement électoral, ou encore l'analyse des méta-données politiques des journaux télévisés ont émergé de ces collaborations.

Retrouver le compte-rendu de la journée sur le site d'Etalab (https://www.etalab.gouv.fr/retour-sur-lopen-data-camp-elections).

ACTU DES DEPARTEMENTS

3 médailles de cristal du CNRS au CDSP

Anne Cornilleau et Anne-Sophie Cousteaux - coordinatrices de l'équipe quantitative du CDSP - et Geneviève Michaud - coordinatrice de l'équipe informatique du CDSP - ont été récompensées de la médaille de cristal 2015 du CNRS pour leur engagement et l'excellence de leur travail au CDSP depuis 2005 et dans le cadre élargi de l'equipex DIME-SHS depuis 2011.

Attribuées par le Président du CNRS et le Directeur de l'Institut des sciences humaines et sociales, ces médailles sont une marque de reconnaissance de l'établissement pour le travail accompli et témoigne plus généralement de la très grande importance du rôle des ingénieurs et techniciens dans la recherche.

CIMES, le nouveau catalogue de la statistique publique européenne

Dans le cadre du projet européen DwB, le CASD et ses autres partenaires du Réseau Quetelet ont développé une base de données documentant des enquêtes de la statistique publique au niveau européen.

CIMES (Centralising and Integrating Metadata from European Statistics) présente les données, les types de fichiers disponibles, et les procédures d'accès de plus de 31 pays européens.

Directeur de la publication : Pascal Buléon
Responsable éditoriale : Amélie Vairelles

UMS PROGEDO Réseau Quetelet
190 avenue de France, 75013 PARIS

www.progedo.fr