Conseil national
de l'information statistique

Paris, le 25 avril 2022 – n°61/H030

RENCONTRE « APPARIEMENTS DE DONNEES INDIVIDUELLES : ENTRE RICHESSE DE L'INFORMATION STATISTIQUE ET RESPECT DE LA VIE PRIVEE »

Rencontre du 28 janvier 2022

ACTES DE LA RENCONTRE
28 janvier 2022

RAPPEL DU PROGRAMME

INTRODUCTION 12
SESSION 1 – ETAT DES LIEUX DES PRATIQUES D'APPARIEMENTS 13
Les pratiques d'appariements de la statistique publique 15
Les appariements réalisés par les chercheurs 25
Echanges 29
SESSION 2 – QUELQUES EXEMPLES D'APPARIEMENTS DE LA STATISTIQUE PUBLIQUE 32
L'échantillon national interrégimes d'allocataires de compléments de revenus d'activité et de minima sociaux (ENIACRAMS) 33
Appariement entre l'enquête Emploi et le fichier historique de Pôle emploi pour comprendre les différences entre nombres de chômeurs et de demandeurs d'emploi 37
Mieux connaître l'insertion des jeunes : le système d'information InserJeunes 41
Echanges 45
SESSION 3 – LES PROJETS D'AVENIR 47
Le code statistique non signifiant (CSNS) 48
Le répertoire statistique des individus et des logements (RESIL) 52
TABLE RONDE – QUELS APPARIEMENTS POUR QUELS USAGES ? 61
Echanges 71
TABLE RONDE – QUELLE TRANSPARENCE, QUELLE INFORMATION DU PUBLIC ? 78
Echanges 91
CONCLUSION 99

Liste des participants

En présentiel :

ARTIGUELONG	Maryse	Ligue des droits de l'homme
AUBERT	Patrick	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
AVVISATI	Francesco	École d'économie de Paris
BARROT	Jean-Noël	Assemblée nationale
BAYLE	Jules	Assemblée nationale
BOZIO	Antoine	Institut des politiques publiques
CARON	Nathalie	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
CASES	Chantal	Société française de statistique (SFdS)
COLIN	Christel	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
D'ALESSANDRO	Cristina	Conseil national de l'information statistique (CNIS)
DUBOIS	Marie-Michèle	Conseil national de l'information statistique (CNIS)
DUPONT	Françoise	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
DURAN	Patrice	Ecole normale supérieure
ELBAUM	Mireille	Ministère des Solidarités et de la santé - Inspection générale des affaires sociales (IGAS)
ESPINASSE	Lionel	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
GADOUCHE	Kamel	Centre d'accès sécurisé distant aux données (CASD)
GUILLAUMAT-TAILLIET	François	Conseil national de l'information statistique (CNIS)
HUNYADI	Mark	Université Catholique de Louvain
LAGARDE	Sylvie	Institut National de la statistique et des études économiques (INSEE) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
LEFEBVRE	Olivier	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
MARTIN	John	IZA Institute of Labor Economics
MAUREL	Françoise	Conseil national de l'information statistique (CNIS)
MONTUS	Arnaud	Conseil national de l'information statistique (CNIS)
PAILHÈS	Bertrand	Commission nationale de l'informatique et des libertés (CNIL)
PASSERON	Vladimir	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
ROBERT	Philomé	France 24
TAGNANI	Stéphane	Conseil national de l'information statistique (CNIS)
TAVERNIER	Jean-Luc	Institut National de la statistique et des études économiques (INSEE) - Direction générale
TIMBEAU	Xavier	Observatoire français des conjonctures économiques (OFCE)

En visioconférence :

AUBERT	Magali	Institut national de la recherche agronomique (INRA)
ADAM	Lorraine	PROGEDO
AKIKI	Michel	Ministère de l'Agriculture et de l'alimentation - Service de la statistique et de la prospective (SSP)
ALAZARD	Antoine	Dijon métropole
ALKHOURY	Maria	Centre d'accès sécurisé distant aux données (CASD)
ALLAIN	Samuel	Direction régionale de l'alimentation, de l'agriculture et de la forêt - Aquitaine
AMARAL	Philippe	Ministère des Solidarités et de la santé - Direction générale de la cohésion sociale (DGCS)
AMBARD	Julien	Collectif les Morts de la Rue
ANDRE	Mathias	Institut national de la statistique et des études économiques (INSEE)
ANDREANI	Nicolas	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
ANGUIS	Marie	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
ANTUNEZ	Kim	Institut National de la statistique et des études économiques (INSEE) – Direction de la diffusion et de l'action régionale (DDAR)
ANXIONNAZ	Isabelle	Particulier
ARCHAMBAULT	Edith	Université Paris 1 Panthéon-Sorbonne
BACCAINI	Brigitte	Ministère de la Transition écologique - Conseil général de l'environnement et du développement durable
BAGEIN	Guillaume	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des

		statistiques (DREES)
BAILLY	Nathalie	Institut National de la statistique et des études économiques (INSEE) – Secrétariat général
BAKIA	Halima	Centre d'accès sécurisé distant aux données (CASD)
BALAVOINE	Angélique	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
BARRET	Emilie	Institut National de la statistique et des études économiques (INSEE) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
BAULNE	Jimmy	Institut de la statistique du Québec
BAYET	Alain	Institut National de la statistique et des études économiques (INSEE) – Direction de la diffusion et de l'action régionale (DDAR)
BÉGUIN	Jean-Marc	Particulier
BELLOC	Brigitte	Société française de statistique (SFdS)
BENABDALLAH	Said	Rectorat de Versailles
BENICHOU	Yves-Laurent	Institut National de la statistique et des études économiques (INSEE) – Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
BERSON	Clémence	Banque de France (BdF)
BERTHOLON	Raphaëlle	Confédération française de l'encadrement - Confédération générale des cadres (CFE-CGC)
BESSIERE	Sabine	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
BIANCO	Emma	Insee Auvergne - Rhône-Alpes
BLACHE	Guillaume	Pôle Emploi
BLANCARD	Patricia	Autorité de la statistique publique (ASP)
BLANDIN	Lola	Université Grenoble Alpes
BLAVIER	Pierre	CNRS - Observatoire sociologique du changement (UMR 7049)
BOIS	François-Xavier	Kernix
BONDON	Marine	Institut national des études démographiques (INED)
BONNANS	Dominique	Institut National de la statistique et des études économiques (INSEE) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
BONNET	Xavier	Institut National de la statistique et des études économiques (INSEE) – Inspection générale
BONNETÊTE	Félix	Ministère du Travail, de l'emploi et de l'insertion - Direction de l'animation de la recherche, des études et des statistiques (DARES)
BOREL	Marie	Ministère du Travail, de l'emploi et de l'insertion - Direction de l'animation de la recherche, des études et des statistiques (DARES)
BOUTIERE	Fabienne	Electricité de France (EdF)
BRIARD	Karine	Ministère du Travail, de l'emploi et de l'insertion - Direction de l'animation de la recherche, des études et des statistiques (DARES)
BRILHAULT	Gwennaëlle	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
BRINGE	Arnaud	Institut national des études démographiques (INED)
BRION	Philippe	Particulier
BRUNET	François	Institut National de la statistique et des études économiques (INSEE) - Inspection générale
BUNEL	Simon	Banque de France (BdF)
BURRICAND	Carine	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques
CAHOUR	Lisa	Santé Publique France
CARIOU	Sylvain	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (Inrae)
CARRASCO	Valérie	Ministère de l'Intérieur - Service statistique ministériel de la sécurité intérieure (SSMSI)
CARRERE	Amélie	École d'économie de Paris
CASTELLUCCIA	Claude	Commission nationale de l'informatique et des libertés (CNIL)
CAVIER	Bernard	
CAZALE	Linda	Institut de la statistique du Québec
CECCI-ANDREANI	Laury	Institut National de la statistique et des études économiques (INSEE) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
CHALEIX	Mylène	Institut National de la statistique et des études économiques (INSEE) - Direction du système d'information (DSI)
CHAMBAZ	Christine	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques d'entreprises (DSE)
CHAMKHI	Amine	Pôle Emploi
CHANTEUX	Alice	Conseil départemental de l'Isère
CHAPUT	Hélène	Institut National de la statistique et des études économiques (INSEE) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
CHARRANCE	Géraldine	Institut national des études démographiques (INED)
CHARRIER	Rodolphe	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
CHAUVEL	Brian	Université Paris Nanterre

CHAUVIN	Adrienne	Union sociale pour l'habitat
CHAUVIN	Pauline	Université Paris 5 - Faculté des Sciences Sociales
CHEJFEC	Thomas	Unedic
CHEVALIER	Pascal	Ministère de la Justice - Sous-direction de la statistique et des études
CHIN	Francis	Santé Publique France
CIESIELSKI	Henry	Ministère de la Transition écologique - Direction de l'habitat, de l'urbanisme et des paysages (DHUP)
CLERC	Marie	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
CLING	Jean-Pierre	Institut National de la statistique et des études économiques (INSEE) – Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
CLUSE	Margaux	Ministère des Solidarités et de la santé - Direction générale de la cohésion sociale (DGCS)
COCHET	Paul	Institut national des études démographiques (INED)
COLIN	Catherine	Conseil régional Occitanie
COLLIN	Christel	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de la jeunesse, de l'éducation populaire et de la vie associative (DJEPVA)
COMMANDEUR	Barbara	CARIF OREF Pays de la Loire
CORRE	Tifenn	Institut national de la recherche agronomique Toulouse
COSTENOBLE	Ophélie	Réseau des Carif-Oref
COSTER	Jean-Louis	Institut National de la statistique et des études économiques (INSEE) – Direction des statistiques d'entreprises (DSE)
COTREBIL	Philippe	Agence d'urbanisme de l'Oise la vallée
COTTET	Sophie	Institut des politiques publiques
COTTIN	Adeline	Insee Pays de Loire
COUDIN	Elise	Institut National de la statistique et des études économiques (INSEE) – Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
COUDRIN	Caroline	DEAL de La Réunion (Direction de l'environnement, de l'aménagement et du logement)
COULONDRE	Alexandre	Ecole des Ponts ParisTech
CROGUENNEC	Yannick	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
CRUAU	Natacha	Ministère du Travail, de l'emploi et de l'insertion - Délégation générale à l'emploi et à la formation professionnelle (DGEFP)
DAMPERON	Alexandre	Insee Ile-de-France
DE MIRAS	Christelle	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
DEFRESNE	Marion	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
DELAHAYE-ADAM	Elisa	Ministère de l'Intérieur
DELAME	Nathalie	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (Inrae)
DEMOLY	Elvire	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
DEMONSANT	Jean-Luc	Université de Toulouse
DEROSIER	Alice	Rectorat de Nantes
DEROYON	Thomas	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
DESPLANQUES	Guy	Particulier
DIAKHATE	Maryama	Ministère de la Justice - Sous-direction de la statistique et des études
DIARD	Karine	Institut National de la statistique et des études économiques (INSEE) – Direction des statistiques d'entreprises (DSE)
DIXTE	Christophe	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
DORNIER	Xavier	Institut français du cheval et de l'équitation (IFCE)
DOURAN	Manal	
DREUX	Cannelle	Conseil départemental 92
DUBOST	Claire-Lise	Ministère du Travail, de l'emploi et de l'insertion - Direction de l'animation de la recherche, des études et des statistiques (DARES)
DUC	Cindy	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques d'entreprises (DSE)
DUNNE	John	Office national de statistique d'Irlande (CSO)
DURR	Jean-Michel	CAOS-Consulting
DUSSUD	François-Xavier	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques d'entreprises (DSE)
DUVERNET	Laurent	Université Paris 10 Nanterre
EGHBAL	Sylvie	Institut national de la statistique et des études économiques (INSEE)
EIDELMAN	Alexis	Ministère du Travail, de l'emploi et de l'insertion - Direction de l'animation de la recherche, des études et des statistiques (DARES)

EL BOUHAIRI	Yacine	Centre d'accès sécurisé distant aux données (CASD)
FAIDHERBE	Thibault	Ministère des Outre-Mer - Direction générale des Outre-Mer (DGOM)
FAU	David	Dijon métropole
FAURET	Camille	Insee Ile-de-France
FAVARO	Antonin	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (Inrae)
FERON	Valérie	Observatoire régional de santé d'Île-de-France
FICHE	Dominique	Ministère de l'Agriculture et de l'alimentation - Service de la statistique et de la prospective (SSP)
FIRDION	Laëtitia	Institut Paris Région
FONTAINE	Roméo	Institut national des études démographiques (INED)
FRANCOZ	Dominique	Institut National de la statistique et des études économiques (INSEE) – Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
FREPPEL	Camille	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
FRESSARD	Lisa	Observatoire régional de la santé Provence-Alpes-Côte d'Azur
GALLIC	Gabrielle	Institut National de la statistique et des études économiques (INSEE) - Direction de la diffusion et de l'action régionale (DDAR)
GARBINTI	Bertrand	Centre de recherche en économie et statistique (CREST)
GARCIA	Cédric	Université Gustave Eiffel
GAUVIN	Charlotte	Ministère de l'Agriculture et de l'alimentation - Direction générale de l'enseignement et de la recherche
GÉLY	Alain	Confédération générale du travail (CGT)
GÉNIN	Gaëlle	Insee Nouvelle-Aquitaine
GEORGE	Estelle	Rectorat de Versailles
GESBERT BOULANGER	Florence	Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation - Direction générale de la recherche et de l'innovation (DGRI)
GIFFARD	Quentin	Biomasse Normandie
GILLES	Séverine	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
GIVOIS	Samuel	Ministère de l'Agriculture et de l'alimentation - Service de la statistique et de la prospective (SSP)
GODINOT	Alain	Particulier
GOLDBERG	Marcel	Institut national de la santé et de la recherche médicale (INSERM)
GOMOT	Eleonore	Ministère de l'Economie, des finances et de la relance - Direction générale des finances publiques (DGFIP)
GOSSIAUX	Sébastien	Institut National de la statistique et des études économiques (INSEE) – Direction de la diffusion et de l'action régionale (DDAR)
GOURDON	Olivier	Institut National de la statistique et des études économiques (INSEE) - Direction générale
GOURMELEN	Julie	Institut national de la santé et de la recherche médicale (INSERM)
GRISELLE	Patrick	Comité du label de la statistique publique
GUILLAUME	Stéphanie	Institut de recherche et documentation en économie de la santé (IRDES)
GUILLAUME	Thierry	Ministère de l'Agriculture et de l'alimentation - Service de la statistique et de la prospective (SSP)
GUILLEMOT	Danièle	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques d'entreprises (DSE)
GUILLOU	Sarah	Observatoire français des conjonctures économiques (OFCE)
GUIRCHOUN	Elodie	Conseil Régional d'Ile-de-France
HAAG	Olivier	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
HADDAK	Mohamed	Université Gustave Eiffel
HAGUET	Laurence	Cour des comptes
HARNOIS	Jérôme	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
HAYS	Olivier	Centre d'étude des supports de publicité (CESP)
HERVIANT	Julie	Insee Ile-de-France
HERZOG	Judith	PersonalData.IO
HUBERT	Jean-Paul	Université Gustave Eiffel
HURPEAU	Benoît	Institut National de la statistique et des études économiques (INSEE) – Direction de la diffusion et de l'action régionale (DDAR)
IDOHOU	Emmanuel	Particulier
ISNARD	Michel	Institut National de la statistique et des études économiques (INSEE) - Inspection générale
JALUZOT	Laurence	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
JARDIN	Marie	Observatoire régional de la santé Provence-Alpes-Côte d'Azur
JOUBERT-LECLERC	David	Institut de la statistique du Québec
JUDAS	Francis	Confédération générale du travail (CGT) - Fédération des Finances
KABLA-LANGLOIS	Isabelle	Insee Ile-de-France

KARKER	Chourouk	Union sociale pour l'habitat
KOLODZIEJ	Isabelle	Union des industries de la fertilisation
KOSSI	Dede	Institut de recherche et documentation en économie de la santé (IRDES)
KOUMARIANOS	Heidi	Institut National de la statistique et des études économiques (INSEE) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
KURKDJI	Patrick	Observatoire régional de la santé Provence-Alpes-Côte d'Azur
LABOSSE	Aline	Insee Auvergne - Rhône-Alpes
LACAILLE	Yves	Union nationale des professions libérales (UNAPL)
LAFARGUE	Loïc	Rectorat de Nantes
LAINÉ	Frédéric	Pôle Emploi
LAMARCHE	Pierre	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
LAMBREY	Serge	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
LAPINE	Malena	Institut national des études démographiques (INED)
LAVERGNE	Aurélien	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
LE CAIGNEC	Emilie	Ministère de la Justice - Sous-direction de la statistique et des études
LE ROLLAND	Lucie	Institut des politiques publiques
LEBRETON	Elodie	Santé Publique France
LEBUGLE	Amandine	Samu social de Paris
LECOCQ	Marie	FranceAgriMer
LECOUVEY	François	Centre d'études et de recherches économiques sur l'énergie (CEREN)
LEMERLE	Stéphanie	Ministère de l'Intérieur - Département des statistiques, des études et de la documentation (DSED)
LEON	Olivier	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
LEQUIEN	Matthieu	Institut national de la statistique et des études économiques (INSEE)
LEQUIEN	Laurent	Institut national de la statistique et des études économiques (INSEE)
LEROUX	Isabelle	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
LEROY	Claire	École nationale de la statistique et de l'administration économique (Ensae)
LEVI-VALENSIN	Michaël	Ministère de l'Agriculture et de l'alimentation - Service de la statistique et de la prospective (SSP)
LEZEC	Florian	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
LIOGIER	Valérie	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
LIXI	Clotilde	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques
LOMBRAIL	Pierre	Université Paris 13
LOONIS	Vincent	Institut National de la statistique et des études économiques (INSEE) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
LORRE	Geoffrey	Ministère du Travail, de l'emploi et de l'insertion - Délégation générale à l'emploi et à la formation professionnelle (DGEFP)
LOSTYS	Emilie	Ministère du Travail, de l'emploi et de l'insertion - Délégation générale à l'emploi et à la formation professionnelle (DGEFP)
LOUPIAS	Claire	Université d'Evry-Val-d'Essonne
LUNGARSKA	Anna	Institut national de la recherche agronomique Toulouse
MAHIET	Gilles	Orange Lab
MAKDESSI	Yara	Ministère de la Justice - Sous-direction de la statistique et des études
MALAGUTTI	Ornella	Ministère de l'Intérieur
MALHERBE	Lucas	Institut National de la statistique et des études économiques (INSEE) – Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
MALLÉJAC	Noémie	Particulier
MANDEREAU-BRUNO	Laurence	Santé Publique France
MARBACH	Léon	Sciences Po
MAREAU	Quentin	Conseil départemental de Meurthe-et-Moselle
MARQUIER	Rémy	Centre d'accès sécurisé distant aux données (CASD)
MARTIAL	Elodie	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
MATINET	Béryl	Ministère de l'Intérieur - Service statistique ministériel de la sécurité intérieure (SSMSI)
MEINZEL	Pauline	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
MEJEAN	Isabelle	Sciences Po
MERCIER	Alice	Samu social de Paris
MERLY-ALPA	Thomas	Institut national des études démographiques (INED)

MICHALLAND	Béatrice	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
MICHELOT	François	Institut Paris Région
MILCENT	Carine	Paris School of Economics - Université Paris 1
MISSEGUE	Nathalie	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
MONTAUT	Alexis	Insee Nouvelle-Aquitaine
MOREAU	Sylvain	Institut National de la statistique et des études économiques (INSEE) – Direction des statistiques d'entreprises (DSE)
MOREL	Claire	Particulier
MOTAMEDI	Kiarash	Ministère de la Transition écologique - Service de l'économie, de l'évaluation et de l'intégration du développement durable
M'PIAYI	Mélissa	Centre d'étude des supports de publicité (CESP)
NAUROY	Frédéric	Ministère de la Transition écologique
NAYMAN	Laurence	Centre d'études Prospectives et d'Informations Internationales (CEPII)
NGUYEN	Christine	Ministère du Travail, de l'emploi et de l'insertion - Délégation générale à l'emploi et à la formation professionnelle (DGEFP)
NGUYEN HUU CHIEU	Elise	Union nationale des professions libérales (UNAPL)
NIAY	Mathilde	Ministère de la Transition écologique - Service de l'économie, de l'évaluation et de l'intégration du développement durable
NICOLAU	Javier	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
NIJARI	Assia	Haut-Commissariat au Plan
OLIER	Lucile	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
OROZCO	Valérie	Institut national de la recherche agronomique Toulouse
ORZONI	Mathieu	Institut National de la statistique et des études économiques (INSEE) - Direction de la diffusion et de l'action régionale (DDAR)
PAILHÉ	Ariane	Institut national des études démographiques (INED)
PALAT	Blazej	Sciences Po
PAVARD	Clément	Agence nationale pour l'information sur le logement (ANIL)
PERREL	Céline	Institut National de la statistique et des études économiques (INSEE) – Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
PETORIN	Elodie	Centre national de la recherche scientifique (CNRS)
PICARD	Sébastien	Ministère de la Culture - Département des études, de la prospective, des statistiques et de la documentation (DEPS-Doc)
PIET	Laurent	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (Inrae)
PIGUET	Virginie	Institut national de la recherche agronomique (INRA) - Centre d'Économie et de Sociologie Appliquées à l'Agriculture et aux Espaces Ruraux (CESAER)
POLLET	Pascale	Autorité de la statistique publique (ASP)
POMÉON	Thomas	Institut national de la recherche agronomique Toulouse
PONS	Sébastien	Insee Bretagne
PORA	Pierre	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
POTREAU	Elisabeth	Insee Occitanie
POULHES	Mathilde	Ministère de l'Intérieur - Service statistique ministériel de la sécurité intérieure (SSMSI)
PRÉVOT	Pascal	Insee Nouvelle-Aquitaine
PROKOVAS	Nicolas	Confédération générale du travail (CGT)
PRUSKI	Cézane	Ministère des Solidarités et de la santé - Direction générale de la cohésion sociale (DGCS)
RAIN	Audrey	Institut des politiques publiques
RAKOTOARISOA	Ifaliana	Centre d'accès sécurisé distant aux données (CASD)
RAMAMONJY	V.	Centre d'études et de recherches économiques sur l'énergie (CEREN)
RAMBLIÈRE	Lison	Samu social de Paris
RANCOURT	Eric	Statistique Canada
RATEAU	Guillaume	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
RATHELOT	Roland	Centre de recherche en économie et statistique (CREST)
RATHLE	Jean-Philippe	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
RAVEL	Loïc	Ministère de l'Intérieur - Direction générale de la Police nationale (DGPN)
RAYNAUD	Philippe	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
REDOR	Patrick	Institut national de la statistique et des études économiques (INSEE)
REGOLO	Julie	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (Inrae)
REMILLON	Delphine	Institut national des études démographiques (INED)
RENUY	Adeline	Institut national de la santé et de la recherche médicale (INSERM)

REY	Grégoire	Institut national de la santé et de la recherche médicale (INSERM)
RIBON	Olivier	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
RICAU	Pascale	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
RICHARD	Mélanie	Agence nationale de l'habitat (ANAH)
RICHET	Jehanne	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
RICHET-MASTAIN	Lucile	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
RIEG	Christian	Ministère de la Transformation et de la fonction publiques - Direction générale de l'administration et de la fonction publique (DGAFP)
RIMBEAULT	Chloé	Initiative France
ROBERT-BOBÉE	Isabelle	Institut national de la statistique et des études économiques (INSEE)
ROBERTI	Vincent	Unedic
ROBIN	Yoan	Unedic
ROCHEREAU	Thierry	Institut de recherche et documentation en économie de la santé (IRDES)
RODRIGUES	Amandine	Insee Pays de Loire
ROSENWALD	Fabienne	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
ROTH	Nicole	Institut national de la statistique et des études économiques (INSEE)
ROUSSEAU	Sylvie	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
ROUX	Aliette	Maison des sciences de l'homme - Nantes
ROVERSI	Aurélia	Institut national des études démographiques (INED)
ROY	Delphine	Institut des politiques publiques
SABOT	Philippe	Ministère de l'Agriculture et de l'alimentation - Service de la statistique et de la prospective (SSP)
SALATHE	Manuelle	Ministère de l'Intérieur - Observatoire national interministériel de la sécurité routière
SAOUD	Ali	HCP Morocco
SAUVEUR	Jean	
SCHUHL	Pierrette	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques
SÉDILLOT	Béatrice	Ministère de la Transition écologique - Service des données et des études statistiques (Sdes)
SELZ	Marianne-Marion	Société française de statistique (SFdS)
SERIEYX	Yvon	Union nationale des associations familiales (UNAF)
SILBERMAN	Roxane	Centre national de la recherche scientifique (CNRS)
SIMON	Marion	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
SIQUEIRA	Juliana	Université Paris Nanterre
SOUAL	Hélène	Insee Auvergne - Rhône-Alpes
SOULLIER	Noémie	Santé Publique France
STEHLIN	Anne	Pôle Emploi
SUESSER	Jan Robert	Ligue des droits de l'homme
SUHARD	Véronique	Institut de recherche et documentation en économie de la santé (IRDES)
SULTAN	Joyce	Institut des politiques publiques
TACHFINT	Karim	Insee Centre
TARAYOUN	Tedjani	Ministère de la Justice - Sous-direction de la statistique et des études
TCHA	Stéphanie	Institut National de la statistique et des études économiques (INSEE) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
TERSEUR	Bruno	Direction régionale de l'environnement, de l'aménagement et du logement - Paca
TEYSSIER	Geoffrey	Institut national des études démographiques (INED)
THÉODOSE	Teddy	Université Paris 13
THOUMELIN	Claire	Ministère de la Culture - Département des études, de la prospective, des statistiques et de la documentation (DEPS-Doc)
TORELLI	Constance	Institut National de la statistique et des études économiques (INSEE) - Division de l'appui technique international
TORTOSA	Thomas	Insee Bretagne
TOULEMON	Léa	École d'économie de Paris
TOURNADRE	Emilie	Assemblée permanente des chambres d'agriculture (APCA)
TOUW	Alexandre	Université Paris Dauphine
TREYENS	Pierre-Eric	Insee Bretagne
VALLET	Louis-André	Centre national de la recherche scientifique (CNRS)
VANDERSCHELDEN	Mélanie	Institut National de la statistique et des études économiques (INSEE) – Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)

VAUTHIER	Mathilde	Assemblée permanente des chambres d'agriculture (APCA)
VESSILLIER	Delphine	Fédération française du bâtiment
VIDAL	Marie	Centre d'accès sécurisé distant aux données (CASD)
VIGLINO	Lionel	Institut National de la statistique et des études économiques (INSEE) – Direction de la diffusion et de l'action régionale (DDAR)
VILAIN	Annick	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
VINCENT	Roseline	Institut de recherche et documentation en économie de la santé (IRDES)
VIOLLIN	Guy	Comité du label de la statistique publique
VROYLANDT	Thomas	Unedic
WYCKAERT	Matthieu	Ministère des Armées - Observatoire économique de la défense (OED)
YOUSSEF	Youssr	Particulier
ZOLOTOUKHINE	Erik	PROGEDO

INTRODUCTION
Patrice DURAN, président du CNIS

Bonjour à tous. Par un hasard du calendrier, notre rencontre se déroule lors de la journée mondiale de la protection des données. Cette journée a été créée en 2007 pour informer le public sur la collecte et le traitement de leurs données à caractère personnel. Elle vise en particulier à le renseigner sur la motivation de ces collectes et sur leurs droits.

Je suis heureux de vous accueillir au sein du Centre de conférence Pierre-Mendès France, bien que nous aurions souhaité que cette rencontre se déroule dans de meilleures conditions. J'espère que le public pourra être présent lors de la prochaine rencontre qui se tiendra le 18 mai autour de la thématique des panels et des cohortes de la statistique publique, ici même, à Bercy.

Nous sommes réunis pour discuter autour de la question des appariements de données personnelles. Cette pratique soulève des questions d'ordre technique et méthodologique, mais aussi d'ordre juridique, éthique et sociétal.

Le premier objectif de cette rencontre consiste à favoriser la concertation entre les producteurs et les utilisateurs de la statistique publique. C'est du reste la finalité même de l'existence du CNIS qui s'inscrit autour de cette question.

Je salue les experts internationaux et les collègues de nos pays partenaires qui ont accepté de participer à nos débats et qui nous feront bénéficier de leurs expériences. Nous serions particulièrement heureux d'en apprendre plus sur leurs réussites comme sur les problèmes qu'ils rencontrent.

Il est notoire que nous enregistrons au CNIS une demande croissante d'accès aux données administratives et que la constitution de jeux de données appariant plusieurs sources se développe en complément des enquêtes. En effet, la montée en puissance des statistiques publiques construites à l'aide de données administratives, d'appariements de sources ou, plus généralement, de données massives, s'est très fortement accélérée au cours de l'année 2021.

Dans le même temps, les producteurs de statistiques se multiplient en dehors du service statistique public, questionnant de fait son périmétrage. En 2018, une rencontre du CNIS portant sur les enjeux des nouvelles sources de données invitait la statistique publique à relever le défi présenté par ces nouveaux producteurs en produisant des informations de qualité qui seraient plus riches, plus fraîches, moins coûteuses, tout en demeurant utiles au débat public et respectueuses de la vie privée.

De plus, l'avis général n°7 du moyen terme 2019-2023 du CNIS invite explicitement à « développer les appariements entre sources de données afin d'enrichir l'analyse des liens entre différents thèmes, en veillant au strict respect de la confidentialité lorsque les appariements reposent sur des informations « identifiantes ». La question de la transparence et du cadre juridique constitue donc une dimension importante de nos préoccupations.

Au-delà des questions méthodologiques et techniques et des questions juridiques, je tiens aussi à rappeler l'importance des appariements pour le pilotage des politiques publiques et plus largement pour la gestion de l'action publique. Comme le disait Aaron Wildavsky, un des grands maîtres de la *policy analysis*, les politiques publiques constituent toujours des

« mixtes de cogitation et d'interaction ». En effet, l'action publique est tout à la fois affaire de réflexion et de connaissance ainsi que d'action collective. Or les problématiques publiques actuelles en matière de gestion des problèmes publiques s'avèrent largement transversales aux nomenclatures administratives comme aux niveaux de gouvernement, ce qui ne peut que renforcer le besoin de coordination. De la sorte, les appariements, par les informations nouvelles qu'ils apportent, constituent une ressource non négligeable pour répondre à la difficile question de la coordination des politiques publiques. En effet, ils permettent tout particulièrement de briser la logique de silo qui existe au sein de l'administration publique.

Par exemple, la Direction de l'évaluation, de la prospective et de la performance (DEPP) a besoin de connaître des éléments concernant l'emploi, tandis que la Direction de l'animation de la recherche, des études et des statistiques (DARES) a besoin d'éléments touchant la formation.

De même, les apports du projet d'appariement des fichiers du Service statistique ministériel de la sécurité intérieure (SSMSI) et de la Sous-Direction de la statistique et des études (SDSE) du ministère de la Justice, sont considérables dans la mesure ils devraient permettre de documenter l'ensemble de la chaîne pénale et favoriser ainsi uen approche plus rationnelle et moins idéologique de questions trop souvent investies par les préjugés et les stéréotypes.

Aujourd'hui c'est tout un style et un système de gestion qu'il faut chambouler dès lors qu'il faut *avancer dans la voie de la désegmentation des interventions publiques et de leur mise en cohérence*. Ainsi, à l'évidence, les appariements correspondent bien au besoin actuel de réinvention d'une action publique nécessitant une meilleure coordination. Et cette période de pandémie que nous avons vécue est très largement emblématique d'une telle exigence. , Elle ne se limite pas en effet à une stricte affaire médicale, car c'est bien tous les champs de l'action publique qui se sont trouvés concernés.

Finalement, si les appariements présentent bien des enjeux méthodologiques et techniques, ils renvoient plus largement à un enjeu crucial de pilotage de l'action publique que nous ne pouvons ignorer et sur lequel nous aurons l'occasion de revenir aujourd'hui.

Je vous souhaite une excellente journée.

SESSION 1 – ETAT DES LIEUX DES PRATIQUES D'APPARIEMENTS

Présidente de la session : Mireille Elbaum, Présidente de l'Autorité de la statistique publique (ASP) ;

Sylvie Lagarde, directrice de la méthodologie et de la coordination statistique et internationale (INSEE) et Christel Colin, directrice des statistiques démographiques et sociales (INSEE), pour une présentation des pratiques d'appariements de la statistique publique ;

Kamel Gadouche, directeur du Centre d'accès sécurisé aux données (CASD), pour une présentation des appariements réalisés par les chercheurs.

Mireille ELBAUM

Je suis heureuse de vous retrouver pour présider cette première session.

Dans un premier temps, je m'étais étonnée de la formulation du titre de cette rencontre, centrée sur les appariements, car ce sujet n'était pas à mes yeux sans soulever des problèmes. Ayant commencé mes études entre la fin des années 1970 et le début des années 1980, nous étions alors en pleine discussion sur les menaces que pouvaient porter les croisements de fichiers, dans le cadre de questionnements autour des liens entre informatique et libertés.

De plus, j'ai toujours des interrogations concernant les sujets présentés sous un angle technique et non thématique, à partir des besoins des usagers, ainsi que sur le risque que ces outils techniques, qui fournissent des apports intéressants et des économies substantielles, nous enferment dans des données administratives pré-existantes, enfermement encouragé par nos contraintes financières. Par conséquent, bien que les appariements de données puissent concerner tant des données administratives que des enquêtes, je tiens à rappeler d'entrée qu'ils ne constituent pas une panacée et que s'ils aident à répondre aux questions posées sur les données que l'on souhaite apparier, ils ne permettent pas pour autant de répondre à toutes les questions.

Néanmoins, de l'eau a coulé sous les ponts depuis les années 1980. Le 22 septembre 2021, à la suite d'une mission de l'Inspection générale de l'Insee visant à faciliter les appariements de données individuelles au sein du service statistique public, l'Autorité de la statistique publique (ASP) a publié un délibéré visant à encourager et soutenir ce recours aux appariements. Ce délibéré souligne que la généralisation de cette pratique peut ouvrir la voie à des exploitations particulièrement novatrices et précieuses. Il précise que la statistique publique, contrainte ou non sur les moyens, ne peut pas manquer l'occasion de bénéficier de cet enrichissement des exploitations de données qui s'avère profitable pour les études statistiques et pour la recherche.

De ce fait, nous constatons que la possibilité d'apparier des données « métiers » et les données des politiques publiques avec d'autres types de fichiers, potentiellement avec des granularités très fines, offre des applications intéressantes pour le pilotage des politiques publiques, comme l'a indiqué Patrice Duran. Les appariements permettent ainsi de répondre à une demande croissante d'informations provenant des pouvoirs publics et d'enrichir les évaluations de leurs actions, y compris dans des dimensions qu'ils n'avaient pas envisagées au départ.

Par ailleurs, nous avons connu des progrès importants s'agissant du respect du secret statistique et du règlement général sur la protection des données (RGPD). L'Insee a notamment porté ces progrès par le biais d'un ensemble d'instruments et de structures dont nous parlerons aujourd'hui, tels que le code statistique non signifiant (CSNS) ou le Centre d'accès sécurisé aux données (CASD). Nous pouvons également compter sur des tiers de confiance, ou encore sur la « FOINisation » des données de santé, c'est-à-dire sur la fonction d'occultation des informations nominatives (FOIN), pour étudier de façon très fine les .

Dans un monde de « *data* », où nous peinons à distinguer le bon grain de l'ivraie, il nous faut mettre en avant le fait que les statisticiens publics disposent de compétences et d'atouts qui représentent un avantage comparatif spécifique par rapport à d'autres producteurs de données, dans la mesure où ils apportent, notamment par le biais d'appariements de données judicieux, des regards à la fois croisés et élargis à des fins d'information générale des acteurs sociaux et des citoyens. L'ASP mettra prochainement en ligne un délibéré sur ce point.

À cet égard, les appariements de données socio-fiscales et de données de santé, réalisées dans le cadre du CASD, constituent par exemple un outil essentiel pour étudier les inégalités sociales à l'aune de toutes les politiques publiques.

Cependant, il existe des *caveat* (garde-fous) autour de ces appariements, comme l'a relevé l'ASP dans son délibéré. Patrice Duran vient de mentionner l'aspect juridique de ces garde-fous, mais je souhaite souligner qu'il s'agit d'opérations lourdes et de procédures longues qui ralentissent la production d'informations, y compris au sein de la statistique publique. Ces procédures nécessitent un appui technique, juridique et institutionnel, qu'est notamment appelé à fournir l'Insee, comme l'expliqueront Christel Colin et Sylvie Lagarde.

Les chercheurs rencontrent également ces difficultés. J'ai pu le constater moi-même dans le cadre d'une mission réalisée pour l'Inspection générale des affaires sociales (IGAS) sur les cohortes épidémiologiques, où les problèmes technico-juridiques et les difficultés étaient d'autant plus grands lorsque ces cohortes étaient petites et centrées sur des données de santé très fines.

Enfin, je note que nous nous situons dans une étape intermédiaire en matière d'appariements. Certes, les statisticiens publics et les chercheurs franchissent des pas de géants, dans le cadre de procédures distinctes. Mais pour ce qui concerne les données de santé et en particulier du système national de données de santé (SNDS) dans le cadre du *Health Data Hub*, les avancées ne se font pas au même rythme, souffrant même de reculs qui peuvent poser des difficultés aux chercheurs. Ces derniers ont à cet égard diverses possibilités, pas toujours claires à leurs yeux, concernant le cadre et le statut de leurs travaux, selon que leur accès aux données s'effectue dans le cadre de l'Inserm, d'une convention avec le SSP ou sur la base d'une initiative propre.

Dans ce cadre, les chercheurs peuvent notamment être amenés à apparier des données de santé et des données socio-fiscales, ce qui peut s'avérer complexe et mobiliser de nombreux acteurs. Ainsi, j'ai constaté que Constances – la plus grande cohorte épidémiologique française, regroupant 200 000 personnes – intégrait des informations sur les revenus recueillies par interrogation directe des personnes en même temps que clles relatives à leur santé. Pourtant, des informations obtenues à partir d'un appariement avec des fichiers administratifs socio-fiscaux du CASD auraient fourni des résultats plus fiables, et c'est la voie dans laquelle ses responsables souhaitent finalement s'engager. Et donc, nous disposons en France de corpus juridiques ayant chacun leur légitimité, mais juxtaposés les uns à côté des autres, et pouvant générer *in fine* d'importants problèmes d'utilisation.

À présent, Christel Colin et Sylvie Lagarde, traceront un panorama des pratiques d'appariements réalisées dans la statistique publique, puis Kamel Gadouche mettra en lumière les apports du CASD en la matière.

Les pratiques d'appariements de la statistique publique

Christel COLIN

Bonjour à toutes et à tous. Nous allons présenter un panorama général des pratiques d'appariement dans le service statistique public (SSP), à savoir l'INSEE et les seize services statistiques ministériels (SSM).

<u>Pourquoi une rencontre ?</u>

Le SSP réalise des appariements de données sur les individus depuis les années 1960, pour répondre à des questions variées. Ils permettent de produire une information nouvelle avec un coût moindre que celui des collectes directes auprès des personnes.

Cette rencontre offre l'occasion de présenter cette longue pratique des appariements à travers différents exemples. Elle permet aussi de débattre autour des questions que peuvent poser ces appariements.

Nous nous interrogerons sur les méthodes d'appariements, sur leur cadre juridique, ou encore sur les informations à communiquer aux utilisateurs ou aux personnes concernées par les données.

Par ailleurs, le contexte général des appariements évolue, tant sur le plan des moyens techniques que sur celui des possibilités juridiques. Et les demandes adressées au service statistique public évoluent aussi. Il s'agit donc de voir quelles nouvelles questions se posent et d'identifier les garde-fous qui pourraient permettre d'y répondre.

Une longue histoire d'appariements

L'INSEE et les SSM disposent d'une longue expérience d'appariements de données sur les individus. Nous pouvons remonter au moins jusqu'en 1956, pour trouver une mobilisation de l'appariement dans le cadre de l'enquête « revenus fiscaux » (ERF) – renommée plus tard « revenus fiscaux et sociaux » (ERFS). Cette enquête rapprochait à l'origine des données du recensement de la population et des données fiscales pour un échantillon de personnes recensées, afin de calculer le niveau de vie en France et d'établir sa distribution. Cette enquête existe encore, et elle procède encore à des appariements. Elle est devenue annuelle depuis 1996, appariant désormais des données de l'enquête emploi et des données fiscales, auxquelles se sont ajoutées les données des prestations sociales depuis 2006.

Nous relevons un autre précurseur en matière d'appariement dans l'échantillon démographique permanent (EDP) mis en place en 1968. A l'origine, il s'agissait d'un appariement de données du recensement de la population et de données d'état civil, pour un échantillon de personnes sélectionnées sur leurs jours de naissance. Pour chacune de ces personnes, l'EDP s'enrichit chaque année de données provenant de différentes sources, dont le nombre s'accroît progressivement. L'EDP a aussi été récemment apparié par la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES) avec des données de santé pour construire l'EDP santé.

Au cours des années 1970, nous avons également assisté au développement de différents panels, qui permettent de suivre les personnes dans le temps. Les panels de la DEPP ont notamment été mis en place à partir de 1973 pour suivre les parcours scolaires. Ils ont été constitués à partir de différentes bases administratives provenant du système scolaire. Depuis les années 1990, ces panels intègrent les résultats aux évaluations nationales des acquis et des enquêtes auprès des élèves et des familles.

Dans la même période, le panel « déclarations annuelles de données sociales » (DADS) a été construit à partir de 1976, reliant des données sur les salaires et sur les périodes d'emploi déclarées par les employeurs, pour un échantillon de salariés. Ce panel, initialement limité au secteur privé, s'est peu à peu étendu au secteur public et inclut désormais tous les actifs, y compris les non-salariés.

En 1988, la DREES a constitué l'échantillon interrégimes de retraités (EIR), qui apparie, pour un échantillon de retraités, des données sur leurs montants de retraite dans les différents régimes afin de reconstituer leurs montants de retraite globale. En effet, la pension de retraite versée à un individu peut provenir de plusieurs régimes différents s'il a changé de régime en cours de carrière.

Un mode de collecte à part entière, soutenu par plusieurs décennies d'avis du CNIS

Ces exemples historiques montrent que depuis des décennies, les statisticiens publics ont cherché à répondre au mieux aux besoins et aux demandes des utilisateurs. Ils ont notamment tenté de satisfaire une demande croissante de données de plus en plus fines et pointues, pour cerner la complexité des situations et des parcours. Ces opérations ont pu être menées grâce à une utilisation intelligente et à un couplage de différentes sources d'informations, à savoir des enquêtes statistiques et des données administratives variées. Il s'agissait notamment de pondérer la qualité de ces sources et leurs apports respectifs, mais aussi d'intégrer des contraintes de coûts et de charges.

Les appariements permettent en effet de produire une information enrichie, à un coût raisonnable. C'est pourquoi il serait impossible de recueillir l'équivalent de cette information par une collecte directe. Cela tient à la fois à la richesse des données administratives qui sont souvent mobilisées, qu'à la puissance de l'appariement en tant que tel qui rapproche des sources différentes.

Finalement les appariements constituent un mode de collecte de l'information statistique à part entière. Ainsi, le CNIS soutient de longue date le développement de panels et d'appariements, tout en soulignant l'importance du respect de la confidentialité et de la vie privée, notamment dans le cadre de ses avis de moyen terme :

- Concertation pour le moyen terme 1999-2003 du CNIS : avis du CNIS sur l'insuffisance du suivi des trajectoires des personnes en matière sociale et d'emploi qui a débouché sur le développement de panels ;
- Concertation pour le moyen terme 1999-2003 du CNIS : « le Conseil demande à l'ensemble des producteurs de la statistique publique de développer les appariements entre sources de données afin d'enrichir l'analyse des liens entre différents thèmes, en veillant au strict respect de la confidentialité lorsque les appariements reposent sur des informations identifiantes ».

Les usages des appariements

Les appariements de données individuelles sont souvent mis en avant pour suivre des trajectoires ou pour évaluer ou piloter des politiques publiques. Ils permettent en fait une palette d'usages bien plus large que nous tenterons de décrire.

Tout d'abord, ces appariements permettent d'améliorer la qualité de l'information statistique. Nous le constatons notamment pour la mesure des revenus et des niveaux de vie. En effet, les revenus des ménages sont mieux appréhendés dans les données administratives – revenus déclarés à l'administration fiscale, prestations sociales versées, etc. – que sur la base d'enquêtes directes auprès des personnes. De ce fait, depuis les années 2000, nous assistons à un enrichissement généralisé des enquêtes auprès des ménages par des données administratives sur les différents types de revenus.

En outre, les appariements permettent de produire des informations à des niveaux géographiques fins, tandis que les échantillons d'enquêtes, qui sont souvent de taille trop limitée, couvrent au mieux une échelle régionale. Par exemple, le fichier localisé social et fiscal (FiLoSoFi) mesure les revenus et la pauvreté au niveau local – communes, quartiers, mailles d'un kilomètre carré – par le rapprochement de fichiers fiscaux et de prestations sociales exhaustifs, ainsi que par l'imputation de certains revenus.

Les appariements sont également précieux, voire incontournables, pour mesurer des phénomènes complexes et couvrir des périmètres complets. Par exemple, les bases « tous salariés » ou « tous actifs » offrent une vision complète des situations d'emploi des personnes sur une année : salarié du secteur public ou privé, de l'agriculture, de particulier employeur, non-salarié. Elles mesurent notamment la multiactivité. A l'inverse, les données administratives qui suivent ces populations sont souvent distinctes et n'en donnent donc chacune qu'une vision partielle. De même, les échantillons interrégimes de retraités et de cotisants, basés sur des appariements, donnent une vision complète des montants de retraites et des futurs droits à la retraite, tous régimes confondus. Il s'agit donc d'apports essentiels à la connaissance.

De surcroît, les appariements menés au sein de la statistique publique permettent d'améliorer la richesse de l'information statistique, notamment pour étudier des phénomènes situés à la croisée de plusieurs domaines et couverts par des sources différentes. Ils permettent par exemple d'étudier la mortalité ou la fécondité selon le niveau de diplôme, la catégorie sociale, le niveau de vie grâce à l'EDP ou d'étudier les inégalités sociales de santé grâce à l'EDP santé.

De plus, de nombreux appariements permettent de décrire des trajectoires individuelles, notamment les trajectoires d'insertion sur le marché du travail, les trajectoires professionnelles d'emploi et de salaire, ou encore celles de certaines populations telles que celle des bénéficiaires de minima sociaux. Ils peuvent permettre de décrire des trajectoires touchant différents domaines, par exemple l'évolution du niveau de vie ou la mobilité résidentielle lors du passage à la retraite en mobilisant l'EDP.

Aussi, il est bien connu que cette reconstitution des trajectoires individuelles permise par les appariements peut permettre d'évaluer des politiques publiques. Nous retrouvons de nombreux exemples de cet usage, notamment dans les SSM. L'appariement est par exemple employé dans le cadre d'évaluation de réformes ou de pratiques de l'Education nationale, avec la mobilisation des panels de la DEPP. Parallèlement, la DARES et la DEPP recourent également à cette pratique pour évaluer les effets des mesures d'aide à l'emploi sur l'insertion des jeunes passés par l'apprentissage ou la voie professionnelle à l'aide du dispositif InserJeunes qui sera présenté au cours de la journée.

Sur le plan méthodologique, les appariements permettent de mieux comprendre certains phénomènes, notamment en aidant à analyser les écarts entre des sources concernant les mêmes concepts ou des concepts voisins. Ainsi, l'appariement de l'enquête emploi et des déclarations des employeurs – déclarations annuelles des données sociales (DADS) et déclarations sociales nominatives (DSN) – vise à mieux comprendre le concept d'emploi et sa mesure. Le rapprochement de l'enquête Emploi, qui mesure le chômage au sens du Bureau international du travail (BIT), et du fichier historique des demandeurs d'emploi de Pôle emploi permet de mieux comprendre le concept du chômage et les écarts qui existent entre ses différentes définitions.

Comment apparie-t-on en pratique ?

Au-delà des exemples et des usages des appariements, intéressons-nous à présent à leur méthodologie. Apparier des données relatives aux individus consiste à rapprocher pour une même personne des données la concernant, à partir de différentes sources. Ces rapprochements peuvent concerner des enquêtes, à savoir des enquêtes différentes sur une même période, ou bien de mêmes enquêtes sur des périodes consécutives. Ils peuvent également mettre en jeu des enquêtes et des données administratives. Ce cas de figure permet soit d'enrichir des enquêtes à partir de variables issues de données administratives, soit de compléter des données administratives par des enquêtes menées sur des échantillons. Enfin, ils peuvent aussi relier uniquement différentes données administratives.

Rapprocher pour une même personne des données qui la concernent entre différentes sources paraît simple mais n'est pas toujours facile à mettre en œuvre. Il peut être difficile de vérifier que les individus en question soient bien les mêmes dans les différentes sources rapprochées.

Les rapprochements sont les plus simples lorsque les sources disposent d'un identifiant en commun qui permet d'identifier sans ambiguïté les personnes. Il peut s'agir d'un identifiant certifié, comme le numéro d'inscription au répertoire (NIR) géré par l'INSEE, autrement connu sous le nom de numéro de sécurité sociale, ou encore l'identifiant national des étudiants et des élèves (INE) géré par la DEPP.

Ces rapprochements peuvent s'avérer plus complexes dans d'autres cas, y compris lorsque nous disposons de données d'état civil complètes – nom, prénom, date et lieu de naissance, sexe. En effet, dans ce cas, l'appariement peut être plus ou moins aisé ou réussi, au sens où peu d'individus seraient non appariés ou mal appariés, selon la qualité de l'état civil. Typiquement, il est possible de retrouver des orthographes différentes de noms et de prénoms dans les différentes sources. L'appariement peut aussi se faire sur la base de données d'état civil incomplètes, par exemple en ne disposant pas du nom, complétées d'autres informations comme une adresse ou une commune de résidence. Il peut aussi utiliser d'autres caractéristiques individuelles permettant de rapprocher les sources (adresses, années de naissance, sexe, autre variable commune). Des exemples concrets présentés au cours de cette journée permettront d'illustrer cette diversité.

Au-delà de cette question d'identifiants et de clés d'appariements, diverses méthodes peuvent être utilisées pour apparier, que nous ne pourrons pas développer dans le cadre restreint de cet exposé.

Quels identifiants individuels pour apparier ?

Le recours aux identifiants individuels communs est donc le moyen le plus simple pour procéder aux appariements. Des identifiants à usage administratif tels que le NIR, utilisé largement dans la sphère administrative sociale, peuvent être mobilisés par la statistique publique, sous conditions, pour des traitements statistiques et de recherche, le temps des appariements. Des identifiants à usages statistiques ou de recherche peuvent aussi être mobilisés pour apparier des données. Pour diminuer leur sensibilité, le principe est de transformer le NIR, numéro d'identification signifiant, en un identifiant non signifiant n'offrant aucune information sur les personnes et ne permettant pas de remonter aux individus. Il s'agit de procédures de hachage et de cryptage, utilisées par exemple pour générer des pseudonymes via la procédure algorithmique FOIN qui concerne les données de santé. De la même

façon, le code statistique non signifiant (CSNS) est obtenu par opération cryptographique à partir du NIR.

Ce CSNS est une nouveauté introduite par la loi pour une République numérique de 2016. Il s'agit d'un identifiant spécifique à l'usage du service statistique public pour des finalités de production de statistiques publiques, sur lequel nous reviendrons au cours de la journée.

Comme va le développer Sylvie Lagarde, que le cadre juridique des appariements a en effet évolué et s'est assoupli au fil du temps.

Sylvie LAGARDE

Bonjour à tous. La longue histoire des appariements du service statistique public exposée par Christel Colin n'a pas été linéaire. Celle-ci a connu différentes périodes marquées par des évolutions des contextes législatif, technique et international. Je vais donc tenter de vous présenter ces évolutions, sans entrer dans le détail de toutes les périodes en question, en me contentant de mettre en lumière quelques-uns de leurs points saillants.

Un cadre juridique qui s'assouplit au fil du temps

Tout d'abord, le cadre juridique s'est assoupli au fil du temps. Cet assouplissement est sans doute lié à l'évolution du contexte. En effet, les évolutions du droit et de la société se nourrissent mutuellement.

Jusqu'en 2004, ce cadre juridique est essentiellement régi par la loi informatique et liberté de 1978. Tous les traitements de données relatives aux personnes qui étaient opérées par les administrations devaient être autorisés par la loi ou par des textes règlementaires, après un avis motivé de la Commission nationale de l'informatique et des libertés (CNIL). Ce cadre était donc très strict et impliquait des procédures assez lourdes. Ainsi, l'usage du NIR dans les appariements était particulièrement bien encadré, si bien qu'il nécessitait l'obtention d'un décret du Conseil d'Etat et un avis de la CNIL.

En 2004, nous avons assisté à une première évolution importante de ce cadre juridique, avec une modification de la loi informatique et liberté. Ce changement transpose l'évolution du cadre juridique européen, marqué par la directive européenne n°95/46/CE du 24 octobre 1995 relative à la protection des données. Cette directive a rendu compatible avec la finalité initiale des collectes de données leur traitement ultérieur à des fins statistiques ou à visée de recherche scientifique ou historique. De la sorte, cette modification de la loi informatique et liberté a permis de justifier par simple arrêté des traitements de données à finalité de statistique ou de recherche, avec le maintien d'une autorisation préalable de la CNIL. Néanmoins, ce cas de figure excluait les données sensibles et nécessitait le constat d'une absence d'interconnexion de fichiers dont l'intérêt public différait. De ce fait, l'appariement de fichiers sur la base du NIR demeurait très encadré, exigeant jusqu'en 2016 un décret en Conseil d'Etat et un avis de la CNIL.

Des étapes importantes ont ensuite été franchies, notamment en 2016, avec la loi pour une République numérique, qui a notamment introduit le CSNS, puis en 2018, avec l'entrée en application du règlement général sur la protection des données (RGPD).

Ainsi, les organismes chargés du traitement des données personnelles sont responsables au premier chef du respect du RGPD, sans saisine systématique de la CNIL en amont de

ces traitements. Cette responsabilité se fonde notamment sur des principes très stricts de minimisation des données et de durée de leur conservation. Ce RGPD encadre encore la rédaction des études d'impact, ou encore le rôle du délégué à la protection des données. Sa mise en place modifie donc profondément le contexte juridique du traitement des données.

De plus, le décret « cadre NIR » prévoit l'ensemble des utilisations possibles du NIR, y compris dans les traitements à finalités statistiques, ainsi que les conditions d'accès au répertoire national d'identification des personnes physiques (RNIPP). Néanmoins, le traitement contenant des données de santé fait alors l'objet d'une exception. Hors données sensibles, le service statistique public peut désormais utiliser sans décret en Conseil d'Etat un même identifiant pour chaque individu, le CSNS, pour rapprocher différents fichiers dans le cadre des traitements à finalité statistique. Il s'agit là d'une évolution importante.

Le cadre juridique spécifique des données de santé

Jusqu'en 2016, il existait plusieurs régimes distincts d'autorisation de la CNIL prévus par la loi informatique et liberté en matière de données de santé. Ces régimes étaient déclinés selon les finalités des traitements des données de santé. Ils pouvaient parfois nécessiter l'avis préalable d'un Conseil scientifique. Aussi, comme nous l'avons dit, un décret en conseil d'Etat restait nécessaire pour permettre l'accès au NIR.

La loi de modernisation de notre système de santé du 26 janvier 2016 a apporté d'importantes évolutions de ce cadre juridique. Elle a notamment institué le système national des données de santé (SNDS) – élargi en 2019 – qui prévoie un rapprochement des données du système national d'information interrégimes de l'Assurance maladie (SNIIRAM), des données hospitalières du Programme de médicalisation des systèmes d'information (PMSI), des données du Centre d'épidémiologie sur les causes médicales de décès (CépiDc), ainsi que de celles relatives au handicap avec un historique de vingt ans. Le SNDS est géré par la CNAM puis également par le *Health Data Hub*, créé par la loi n°2019-774 du 24 juillet 2019. Le NIR FOINisé est utilisé comme identifiant au sein du SNDS, sans possibilité de revenir au NIR.

En fin 2019, la création du *Health Data Hub* a offert un point d'accès unique aux données de santé pour des finalités de recherche, d'étude ou d'évaluation. Il assure également le secrétariat du Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (CESREES) qui émet un avis en vue de faciliter l'examen par la CNIL des demandes d'autorisation de traitement des données de santé à des fins de recherche, d'étude ou d'évaluation.

Le contexte récent induit un développement des appariements

Au-delà du contexte juridique, je souhaiterais aborder la question de l'évolution du contexte technique et international, qui favorise l'amplification actuelle des appariements.

Les capacités informatiques s'accroissent et nous exploitons de plus en plus des données administratives exhaustives. Or, autrefois, cette exhaustivité n'était pas permise techniquement. Par exemple, dans le cadre de l'exploitation des données de la DADS – remplacée actuellement par la DSN –, nous avons pu atteindre une exhaustivité à partir du début des années 1990, tandis que nous nous limitions auparavant à un échantillon au 1/25, puis au

1/12. De la même façon, nous avons été en mesure d'exploiter exhaustivement des données de revenus et de taxes d'habitation, puis celles des prestations sociales. En outre, la taille des échantillons de type EDP ou panels s'est accrue.

Nous assistons également à une prolifération des sources disponibles. En particulier, les données administratives sont plus nombreuses et elles deviennent plus accessibles. Les sources s'élargissent progressivement. Elles se rationalisent et se centralisent. Ainsi la DSN tend à élargir son champ à l'ensemble des salariés, en comprenant ceux du secteur public. La question du prélèvement à la source a également beaucoup fait évoluer de nombreuses données administratives, avec la mise en place du dispositif « passage des revenus autres » (PASRAU), du répertoire national commun de la protection sociale (RNCPS), ou encore du répertoire de gestion des carrières uniques (RGCU).

De cette façon, des bases de données importantes se mettent en place et offrent des opportunités d'accès pour les statisticiens. Cet accès se fonde sur un encadrement juridique porté notamment par la loi de 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques, la loi de 2016 pour une République numérique, ou encore le règlement européen 223/2009 relatif aux statistiques européennes.

Enfin, la culture de la donnée progresse chez les détenteurs des sources, à travers une méthodologie statistique plus efficace, ou encore à partir d'échanges entre pairs.

Les pratiques des autres instituts nationaux de statistique

Au sein des autres instituts nationaux de statistique, il s'établit également une stratégie explicite d'appariements et de mise en place de répertoires statistiques. D'autres pays recourent parfois à ces pratiques depuis plus longtemps que la France.

Les pays d'Europe du Nord – Danemark, Finlande, Norvège, Suède – disposent d'un système statistique fondé sur des registres depuis les années 1960-1980, cette date de démarrage variant selon les pays. Progressivement, cette pratique s'est déplacée vers les Pays-Bas qui ont installé depuis le début des années 2000 un système de registres et d'enquêtes interconnectés et normalisés, initialement dans un but de rationalisation des coûts. Plusieurs de ces Etats ne peuvent d'ailleurs pas légalement collecter d'informations dans une enquête si elles sont déjà disponibles dans des fichiers administratifs. Ces instituts ont donc été tenus de faire évoluer leurs pratiques, sur la base de ces obligations légales.

L'Irlande, dont le cas sera évoqué lors de la table ronde de cette rencontre, porte le projet PECADO qui regroupe différentes sources administratives pour produire des estimations de population.

De plus, Eurostat encourage le couplage de données et la mobilisation de données administratives pour les statistiques européennes, dans le cadre du projet *Administrative data sources* (ADMIN) de la Vision 2020. Cette direction de la Commission européenne finance ainsi certains travaux communs pour les statistiques européennes.

Parallèlement, des pays anglo-saxons et des pays d'Europe du Sud tels que l'Italie ou l'Espagne évoluent également rapidement sur ces questions. Cette pratique dépasse donc les différentes spécificités culturelles. De cette façon, les pays anglo-saxons situés hors de l'Europe et qui sont culturellement moins portés vers ces questions, tels que le Canada, l'Australie ou la Nouvelle-Zélande, réfléchissent aussi de concert sur les enjeux du couplage

de microdonnées, ainsi que sur l'intérêt de l'exploitation des données administratives et des fichiers appariés qui combineraient notamment des enquêtes et des données administratives :

- Statistique Canada s'appuie sur une directive de 2017 sur le couplage de microdonnées ;
- l'*Australian Bureau of Statistics* pilote le *Multi-Agency Data Integration Project* (MADIP) qui mobilise six agences et combine des données de santé, d'éducation et de démographie ;
- en Nouvelle-Zélande, il est mené un projet d'exploitation de données administratives avec une « colonne vertébrale » et des fichiers à apparier.

Il est intéressant de constater qu'il existe une réflexion très vive sur le plan international, portée en particulier par les pays anglo-saxons. Cette réflexion ne porte pas seulement sur les dimensions techniques ou méthodologiques des appariements. Il s'agit de s'interroger sur la concertation avec les utilisateurs, mais aussi sur des questions éthiques, ou des questions touchant le *social licence* (mandat social). Ce mandat social renvoie à la collecte des données et aux techniques d'appariements. Cette réflexion dépasse les cadres juridiques et techniques, englobant l'ensemble des champs de la société. Nous reviendrons sur ces questions importantes au cours des tables rondes de cette rencontre.

Quels garde-fous ?

Nous relevons donc l'existence de possibilités accrues permettant d'apparier des données plus largement et plus facilement. Or ces nouvelles possibilités nous invitent à réfléchir à la définition de garde-fous.

Tout d'abord, cette question appelle à réfléchir à l'élaboration d'un cadre juridique solide, qui constituerait un point très structurant et qu'il est important de maîtriser. Ce cadre se fonde notamment sur des finalités uniquement statistiques mettant en jeu le secret statistique. Il implique encore de ne pas permettre que des informations individuelles puissent être retransmises aux propriétaires des données administratives. Ces appariements ne doivent donc pas mener à des décisions portant sur des individus.

Il est également important de pouvoir discuter en amont des appariements avec des représentants de la société. A cet égard, le CNIS peut jouer un rôle très important, dans le cadre de ses différentes commissions. Il s'agit de mener des discussions systématiques sur les programmes de travail de la statistique publique, qui impliquent une présentation des travaux d'appariements avec des documents et des comptes rendus en libre accès. Cet enjeu a d'ailleurs été explicité dans le cadre de l'avis de moyen terme 2019-2023 du CNIS. Il s'agit encore de procéder à une mention systématique dans les programmes de travail présentés au CNIS de l'usage du CSNS dans les appariements du service statistique public, de manière à rendre cet usage transparent.

Par ailleurs, le principe de la collecte loyale de données et celui des traitements transparents renvoient à d'autres garde-fous. Pour répondre à ces principes, tous les traitements de données individuelles sont rendus publics sur les sites internet de l'INSEE et des SSM. Et lorsque des données d'enquêtes sont enrichies *ex post* avec, par exemple, des données administratives de revenus, l'enquêté en est informé dans la lettre d'information qu'il reçoit en amont de la collecte.

Aussi, le principe de minimisation des données constitue un garde-fou important. Il s'agit d'utiliser uniquement les données nécessaires. La question de la durée de conservation des données a également son importance. Ces éléments sont intégrés dans les études d'impact menées en amont des traitements statistiques.

Enfin, un travail mené sur la sécurité informatique constitue encore un garde-fou important. Il faut pouvoir restreindre et contrôler l'accès à un petit nombre de personnes responsables du traitement des données.

Conclusion

Pour conclure, le service statistique public s'appuie sur une longue pratique d'appariements individuels, qui était initialement plutôt cloisonnée par domaines thématiques, avec peu de partages entre pairs, aux échelles nationale et internationale.

Néanmoins, l'évolution du contexte juridique, technique et international, qui s'opère depuis quelques années, invite à changer d'échelle et à définir une nouvelle stratégie, afin de faciliter les appariements au sein du service statistique public. Il s'agit de s'appuyer de plus en plus sur le CSNS et, bientôt, sur l'offre de service portée par le programme du répertoire statistique des individus et des logements (RESIL), qui vous seront tous deux présentés. Lorsque ces appuis ne sont pas possibles, il reste notamment envisageable de recourir au NIR, en particulier dans le cadre d'appariements menés avec des organismes de protection sociale, pouvant notamment se trouver hors du service statistique public.

Enfin, il importe de davantage communiquer, informer et échanger avec la société sur les questions éthiques de respect de la vie personnelle suscitées par les appariements. Et il convient particulièrement de tenir compte de ces questions dans une société où le numérique prend une part croissante. La rencontre d'aujourd'hui nous donne ainsi l'occasion de débattre ensemble de ces enjeux.

Mireille ELBAUM

Merci à vous. Je souhaite apporter deux précisions le temps que le prochain intervenant s'installe. Dans son avis du 22 septembre 2021 concernant les appariements, l'ASP a bien indiqué que les programmes statistiques devaient indiquer de manière très claire « leurs objectifs, leurs contenus et leurs conditions de réalisation » de manière à pouvoir être pris en compte dans le cadre des travaux des commissions du CNIS.

J'apporte encore une précision concernant les données de santé. Une étape non négligeable a été franchie par un décret n°2021-848 du 29 juin 2021, permettant à l'INSEE d'avoir accès au SNDS. La situation antérieure explique notamment pourquoi l'EDP santé a été réalisé par la DREES et non l'INSEE. En fait, ce décret ouvre également de nouvelles possibilités, en particulier pour l'ensemble des équipes du Centre national de la recherche scientifique (CNRS), qui n'avait pas non plus accès au SNDS. Le décret permet aussi au directeur général de l'Institut national de la santé et de la recherche médicale (INSERM) de permettre l'accès au SNDS à l'ensemble de ses équipes.

Par conséquent ce décret donnera potentiellement lieu à des évolutions importantes. Néanmoins, la limite de conservation des données du SNDS demeurant à 19 ans, des problèmes se posent encore à des organismes tels que l'Institut national d'études démographiques

(INED) qui effectue des travaux historiques, ou plus largement, pour toute perspective de recherche historique.

Kamel Gadouche va à présent traiter la question des offres d'appariements dirigées vers les chercheurs.

Les appariements réalisés par les chercheurs

Kamel GADOUCHE

Merci à Mireille Elbaum pour ses propos introductifs, ainsi qu'à Christel Colin et à Sylvie Lagarde pour leur présentation des pratiques opérées au sein de la statistique publique. Je vais décrire des appariements réalisés par des chercheurs, dans une finalité de recherche. Entre statistique et recherche, la frontière est ténue et ces deux champs partagent des points communs. La recherche souffre d'un certain retard en matière de possibilités d'appariements, même si des avancées ont été permises ces dernières années.

Nous allons commencer par présenter quelques usages de chercheurs mobilisant des appariements, tout en nous intéressant aux difficultés posées par cette pratique. Puis, nous décrirons une nouvelle méthode d'appariement sécurisée basée sur des tiers de confiance et de sécurité, dans le cadre de bulles sécurisées. Enfin, nous mettrons en évidence quelques réalisations concrètes liées au CASD, qui ont été permises ces dernières années, notamment grâce à des avancées juridiques.

<u>Les usages des chercheurs et les difficultés posées par les appariements</u>

Depuis des années, les chercheurs ont accès à un nombre croissant de données via le CASD ou par d'autres intermédiaires. Ils ont davantage l'habitude de manipuler des données et ils foisonnent d'idées d'appariements, ce qui leur permet de répondre à des questions de plus en plus précises.

Ils mobilisent ainsi ces appariements pour étudier différents sujets, tels que : la mobilité intergénérationnelle des revenus entre les parents et leurs enfants ; l'évaluation de la réforme de la voie professionnelle ; les liens entre la carrière et la santé ; le suivi des épisodes de chômage et des trajectoires professionnelles ; ou encore l'efficacité des formations, qui captent des fonds publics importants.

Ainsi, si nous souhaitions étudier les liens entre la carrière et la santé, nous pourrions compter d'un côté sur des données de santé provenant du SNDS dotées d'identifiants propres basés sur la FOIN et de l'autre côté sur des données de carrière provenant de la DSN, disposant d'un autre type d'identifiants, lui aussi non signifiant.

Or lorsque nous souhaitons rapprocher ces deux types de données pour pouvoir suivre des individus, nous nous heurtons à une difficulté. En effet, ces deux identifiants ne se correspondent pas et il n'est pas possible de les combiner directement. Il est alors difficile de rapprocher des individus. Cet écueil provient d'une mesure volontaire de sécurité, chacun de ces fichiers disposant de leurs propres types d'identifiants.

Néanmoins, bien que cela puisse paraître a priori impossible, il s'avère possible de dépasser cette difficulté. A cette fin, il suffit de pouvoir remonter au NIR, un identifiant certifié. Ce NIR permet alors de repérer les individus et de procéder à un appariement.

Une nouvelle procédure d'appariement basée sur des bulles sécurisées

Remonter au NIR, qui est une donnée de santé très détaillée et donc une donnée sensible, implique nécessairement de disposer d'un environnement et d'une procédure sécurisés. Et cette question constitue un point fondamental de notre rencontre.

Actuellement, il existe de nombreuses sources de données auxquelles les chercheurs ont accès depuis ces dernières années, par le biais du CASD ou d'autres dispositifs. Ces sources sont riches et souvent exhaustives. Les chercheurs ont conscience que le service statistique français produit des données nombreuses et de bonne qualité. Il s'agit là d'un atout important pour la recherche et la possibilité d'apparier ces données ouvre la voie à de nouvelles possibilités.

La loi de 2016 pour une République numérique a permis de rendre juridiquement possible ce type d'appariements qui mobilisent le NIR. Toutefois, cette opération s'effectue dans un cadre précis. Pour se faire, il existe donc une procédure propre pour la recherche, qui partage des similitudes avec celle qui est attribuée à la statistique publique et que je ne décrirai pas.

Comme l'a précisé Mireille Elbaum dans son introduction, cette procédure est complexe et implique plusieurs acteurs. Cette procédure mobilise ainsi deux organismes externes qui font office de tiers de confiance et qui ne sont ni des producteurs, ni des utilisateurs de données.

Ainsi, un producteur A et un producteur B vont chacun créer leurs propres identifiants indexant ID1 et ID2 associés aux NIR. Dans un premier temps, ils envoient chacun les fichiers qu'ils souhaitent faire apparier, avec uniquement le NIR et cet identifiant indexant. En parallèle, ils transmettent chacun au second tiers de confiance les données informatives associées à cet ID1.

Ensuite, le premier tiers de confiance va hacher le NIR des deux fichiers provenant des deux producteurs. Ce hachage est une opération de cryptographie non réversible associée à une clé secrète. En appliquant une fonction de hachage, nous obtenons un code très long de ce type :

99740afc57d67b5879b664681d0f40789cae2109c74fe9d73a7a72c889ab01676aad8dc8a4731b8d39d36e9da36fa5dfd30c47d12d6547cf9d2033a5a67c6148

Le premier principe de cet algorithme implique que deux NIR identiques s'associent à un même code « H ». En second lieu, il faut empêcher de pouvoir remonter depuis ce code jusqu'au NIR, ce qui est rendu possible par la propriété de la fonction de hachage. Ainsi, comme deux NIR identiques provenant des deux fichiers différents fournissent un même code, il est possible d'apparier les fichiers.

Le premier tiers de confiance « identité » envoie alors au deuxième tiers les NIR hachés des fichiers transmis par les deux producteurs. Ce dernier rassemble alors méthodiquement les informations des fichiers avec les NIR hachés, pour intégrer in fine les fichiers appariés dans une bulle sécurisée, remplaçant au cours de cette étape ces NIR hachés par de nouveaux identifiants.

Il est important de noter que le premier tiers de confiance dispose du NIR et seulement du NIR, tandis que le second tiers dispose des données, mais pas du NIR. Le second tiers ne peut donc pas associer les informations des données à des codes identifiants. Aucun des acteurs de la procédure, qu'il s'agisse des deux tiers ou des deux producteurs, n'a accès à l'ensemble des informations.

Une fois que ce fichier apparié est créé, il se pose la question de son accès. Ainsi, cet accès se déroule dans le cadre d'un dispositif de bulle sécurisée, fournie par le CASD. Le Centre d'accès sécurisé aux données a pris en 2019 la forme d'un groupement d'intérêt public (GIP) à but non lucratif, regroupant l'INSEE, le GENES, le CNRS, l'Ecole polytechnique et HEC Paris. La mission principale de ce consortium consiste à établir des bulles sécurisées à des fins de recherche, d'étude, d'évaluation et d'innovation. Je vous propose de visionner une vidéo résumant les missions du CASD :

« CASD : Une bulle pour protéger les données. » (Le Monde)

« CASD, a single entry point to a large number of data producers. » (Science)

De nombreuses entreprises, administrations et producteurs de données veulent valoriser leurs données et peuvent pour cela avoir besoin de faire appel à une expertise externe. Parallèlement, de nombreux data scientists, chercheurs et consultants souhaitent obtenir l'accès à des données confidentielles pour leurs algorithmes, leurs études, ou pour leurs recherches. La transmission de copies de données directement à l'utilisateur n'offre pas de garanties de sécurité et de traçabilité. Pour leur sécurité, les données doivent rester confinées au sein d'une bulle sécurisée. Le CASD est un tiers de confiance mettant à disposition un équipement conçu pour permettre aux utilisateurs de travailler sur des données fournies par des producteurs dans des conditions de sécurité élevées.

« The center (CASD) holds information, including tax and medical data. » (The New York Times)

Après avoir signé un contrat avec le CASD, le producteur transmet de manière sécurisée une copie de ses données qui sont alors stockées dans l'infrastructure sécurisée du CASD en France. En mettant à disposition ses données, un producteur en reste propriétaire et définit les règles de leur utilisation. Le CASD offre aux détenteurs de données une opportunité de les valoriser tout en assurant leur sécurité et la conformité au RGPD. Le producteur détermine les personnes habilitées à accéder aux données dans le cadre de projets. Les données restent sécurisées et accessibles en permanence et deviennent facilement exploitables pour les utilisateurs. »

« CASD is an example of the type of infrastructure for sensitive health data. » (Nature)

Pour garantir la sécurité des données sensibles qui lui sont confiées, le CASD a créé la SD-Box, un boitier d'accès distant aux données intégrant une authentification biométrique, des communications chiffrées garantissant une sécurité de bout en bout du dispositif qui inclut plus de 300 mesures de sécurité. Une fois connecté à son serveur du CASD via sa SD-Box, un utilisateur peut accéder aux données sensibles qui sont autorisées pour son projet. Il peut ensuite les analyser grâce à une large palette de logiciels de traitement de données mise à sa disposition sur le serveur. Les serveurs accessibles en continu permettent aux membres habilités d'un projet de travailler conjointement sur les données et de partager leurs travaux. L'utilisateur peut également importer ou exporter des fichiers et scripts qui

doivent être conservés et vérifiés par les équipes du CASD, pour s'assurer qu'ils ne contiennent pas de données sensibles !

Ces bulles sécurisées sont portées par des serveurs hébergés sur un ancien site militaire situé en région parisienne, doté d'une ronde de garde. Au sein de ce bâtiment, il existe un dispositif de sécurité doté d'un contrôle d'accès biométrique à triple facteur – carte, main et code.

Quelques réalisations d'appariements permises par le CASD

Enfin, j'évoquerai succinctement quelques réalisations permises par le CASD. Tout d'abord, dans le cadre d'une étude sur l'innovation, il a été possible de croiser le fichier de l'impôt sur le revenu et le fichier des brevets. Par ailleurs, Antoine Bozio, qui est présent dans cette rencontre, a participé à un appariement entre le fichier d'actionnariat d'entreprise et celui de l'impôt sur le revenu. En outre, d'autres appariements ont été effectués à partir de la DSN avec des données de Pôle emploi et des formations professionnelles dans le cadre du dispositif Formation, chômage et emploi (FORCE). Aussi, le projet d'appariement MIDAS devrait permettre de relier les données de la DSN, de Pôle emploi et de la Caisse nationale des allocations familiales (CNAF).

Je me permets maintenant d'effectuer un focus sur l'appariement FORCE, dont les données sont actuellement utilisées par des équipes de recherche. Il s'agit d'un appariement de données provenant de la DARES et de Pôle emploi. L'appariement a fait intervenir DATASTORM, un premier tiers de confiance, ainsi que le CASD, qui a endossé le rôle du second tiers de confiance. Cet appariement a suivi la procédure que nous avons décrite dans cet exposé.

Conclusion

Finalement, des progrès significatifs ont été réalisés ces dernières années en matière d'appariements. Les projets de recherche menés sur la base d'appariements effectués dans le cadre du CASD et portant sur plusieurs domaines, sont répertoriés et consultables sur la page www.casd.eu/projets. Pour des besoins de transparence et pour la bonne information des citoyens, la description de ces projets est publiée tant sur le site du CASD que sur d'autres sites internet.

Je remercie les organisateurs de cette journée d'avoir permis la tenue de cette rencontre, centrée autour d'une pratique dont les enjeux resteront cruciaux durant les années à venir. En effet, les appariements constituent une source d'innovation et d'avancées de la connaissance. Ces avancées s'opèrent avec la recherche et pour la recherche. Elles bénéficient aussi à la société, à travers l'amélioration de l'évaluation *ex ante* et *ex post* des politiques publiques, le décloisonnement et la transversalité des études. Enfin, comme l'indiquait Patrice Duran, les appariements permettent de rompre les silos.

Le CASD est donc particulièrement enthousiaste à l'idée de participer, en coopération avec la statistique publique, au développement de ces sources de données qui s'avèrent porteuses d'innovations pour la recherche. Je souligne que la statistique publique a toujours été particulièrement bienveillante envers les chercheurs. Merci pour votre attention.

Echanges

Mireille ELBAUM

Parmi les adhérents du CASD, vous avez cité l'INSEE, mais pas le ministère de la Santé, ni l'INSERM. Et je relève qu'un dispositif de même nature devrait se mettre en place s'agissant du SNDS dans le cadre du *Health Data Hub*.

Or Patrice Duran expliquait que la crise du coronavirus n'était pas qu'une crise sanitaire. Et dans le cadre des missions du SSP, il est effectivement essentiel de croiser des données de santé avec d'autres sources. J'espère que le développement de ce type de rapprochements de données sera encouragé par davantage d'acteurs à l'avenir.

Par ailleurs, je remarque que lorsque des statisticiens réalisent une étude statistique ou une recherche, ils ne connaissent pas toujours à l'avance toutes les données dont ils auront besoin, les idées pouvant survenir au fur et à mesure des exploitations de données. Et donc, la conciliation de ces besoins avec le principe de minimisation des données et avec les lourdeurs des procédures n'est pas toujours évidente dans la pratique.

Yvon SERIEYX, UNAF

Les présentations de cette session ont montré la complexité des appariements touchant les données de santé. Franchir une nouvelle étape devrait demander du temps, mais aussi de l'argent, comme on l'observe avec les frais engendrés par le CASD.

Certes, nous devons veiller au respect de la vie privée et des règles de durées de conservation des données. Néanmoins, il semble difficile d'effectuer un nombre indéfini de procédures, au gré des nouveaux besoins qui peuvent surgir au cours des recherches.

Nous pensons donc que la question de la détermination *ex ante* des appariements nécessaires doit constituer un sujet à part entière sur lequel nous devons discuter dans le cadre des concertations et de l'ensemble de la comitologie préalable aux travaux d'enquête. Ainsi, un chapitre dédié à cette question devrait intégrer les dossiers présentés au CNIS pour des avis d'opportunité. Il s'agirait d'effectuer en amont un inventaire des données que l'on souhaite apparier et de toutes les connaissances que l'on veut faire progresser, après un examen de la littérature scientifique.

Mireille ELBAUM

Le délibéré de l'ASP du 22 septembre 2021 sur les appariements rappelle notamment l'importance de ce diagnostic des besoins qui incombe au CNIS et qui demande de connaître la réalité.

De plus, en matière sanitaire et sociale, la mission associant l'Inspection générale de l'INSEE et l'IGAS a donné lieu au diagnostic d'un certain nombre de besoins d'information liés aux appariements qui peuvent alimenter des discussions en amont des traitements de données.

Néanmoins, il ne convient pas de conditionner les appariements au même type d'avis d'opportunités ou autres avis auxquels sont soumises les enquêtes. En effet, ils donnent déjà lieu à une autre « comitologie » qui s'avère assez lourde.

Ainsi ce délibéré de l'ASP sur les appariements souligne simplement la nécessité de mentionner ces projets ou initiatives dans les programmes transmis au Cnis, de manière à ce que ses commissions thématiques puissent en tenir compte.

Christel COLIN

La question de Mireille Elbaum montre qu'il est utile de se poser les bonnes questions en amont des traitements de données, même s'il n'est pas possible de penser à tout avant de lancer les procédures d'appariements.

De plus, les nouvelles facilités techniques ou juridiques ne justifient pas l'emploi d'appariements tous azimuts. Nous devons envisager leur recours à partir de questions définies, tout en pensant en amont aux données et aux procédures nécessaires. C'est important pour la transparence des usages de ces appariements.

Cette importante phase *ex ante* intègre une réflexion méthodologique et statistique relative au choix des sources. Elle demande également de mener une concertation avec les utilisateurs. Il est désormais demandé que les appariements et leurs finalités soient précisément décrits dans le cadre des programmes de travail du Cnis. Dans cette optique, je précise que les sites de certains SSM décrivent bien les sources et les appariements, participant à cette démarche de transparence.

Il n'est pas facile de faire coïncider le principe de la transparence et le recours parfois nécessaire à des données supplémentaires en cours de recherche ou d'étude statistique. Néanmoins, les procédures se simplifient et peuvent faciliter le droit au remords.

Sylvie LAGARDE

J'ajouterais qu'une réflexion sur les besoins est menée en amont de la mise en œuvre de tous les traitements de données personnelles, dont les appariements, dans le cadre d'une analyse juridique. Cette réflexion est conduite en particulier durant l'élaboration du document de conformité à la protection des données (DCPOD).

Il s'agit d'intégrer aux opérations d'appariements cette réflexion, afin de définir des variables. Or cette démarche n'est pas si simple, notamment au regard de la prise en compte du principe de minimisation des données qui est contrôlée durant les échanges menés avec les utilisateurs et le délégué à la production des données (DPO).

Louis-André VALLET, CNRS

Comment est organisée la concertation entre la statistique publique et les établissements publics à caractère scientifique et technologique (EPST) tels que le CNRS, l'INRAE ou l'INSERM ?

Claude CASTELLUCCIA, CNIL

Le CSNS est-il persistant ou temporaire ? Est-il identique pour tous les appariements ?
Olivier HAYS, CESP

Pourquoi le GIP du CASD réunit-il l'Ecole polytechnique technique (X) et HEC, mais pas les autres écoles ou l'ensemble des écoles d'ingénieurs ?

Mireille ELBAUM

Au sujet de la concertation entre la statistique publique et les EPST, il faut signaler que les pratiques sont très variées. En effet, les chercheurs peuvent travailler sur des données à plusieurs titres. Sur le principe de l'indépendance des chercheurs, ceux-ci peuvent notamment procéder à leurs travaux, hors de tout lien de dépendance avec les organismes détenteurs des données, dans le cadre du CASD ou encore de l'INSERM.

Lorsque les chercheurs souhaitent apparier les données d'enquêtes qu'ils ont produites avec d'autres sources, ils peuvent mener des opérations conjointes avec le service statistique public. Ce cas de figure se rencontre par exemple dans le cadre de la récente enquête « épidémiologie et conditions de vie liées au covid-19 » (EpiCov) qui réunissait l'INSERM, l'INSEE et la DREES.

Enfin, les chercheurs peuvent travailler pour le compte d'organismes tels que les SSM sur la base de crédits d'études et de recherche. Ils peuvent alors exploiter des fichiers appariés ou des enquêtes, qui leur sont rendus accessibles sous la responsabilité de ces organismes. Ils peuvent encore suivre des procédures adjacentes pour pouvoir réaliser leurs études.

Finalement, il existe une palette de pratiques très large centrées autour de l'enjeu fondamental de la transparence de la statistique publique. A cet égard, nous tentons de promouvoir cette transparence auprès des SSM, en nous appuyant sur le niveau de transparence porté par certains organismes publics qui diffusent des statistiques tels que la CNAM ou la CNAF. Franchir de nouvelles étapes en cette matière devrait faciliter la recherche.

Christel COLIN

Pour répondre à la question touchant le CSNS, je précise que ce code n'est pas pérenne et qu'il dure dix ans, sauf si une faille de sécurité nécessite de le renouveler avant ce terme. Lionel Espinasse présentera le CSNS lors de la troisième session de cette rencontre.

Kamel GADOUCHE

En réponse à la question touchant la constitution du GIP, il faut savoir que les membres de ce consortium sont ceux qui contribuent au fonctionnement du CASD. Toutefois, les utilisateurs du CASD intègrent la communauté beaucoup plus large des chercheurs. Ainsi, presque tous les organismes de recherche français ont accès aux services du CASD.

Par ailleurs, le CASD a été créé dans le cadre du projet « équipement d'excellence » (EQUIPEX), intégré au Programme d'investissements d'avenir (PIA). Or l'appel à projets EQUIPEX visant à obtenir des financements se fondait sur une logique de localisation. Le choix des écoles du consortium s'est donc justifié dans le cadre d'une logique de campus.

En tout état de cause, tout organisme peut se présenter pour rejoindre ce consortium et des discussions se sont notamment tenues à cet égard avec l'INSERM et le ministère de la Santé. L'Assemblée générale de ce GIP décide de l'intégration de nouveaux membres.

SESSION 2 – QUELQUES EXEMPLES D'APPARIEMENTS DE LA STATISTIQUE PUBLIQUE

Président de la session : Antoine Bozio, Président de la Commission Services publics et services aux publics du CNIS, Institut des politiques publiques (IPP) ;

Patrick Aubert, sous-directeur de l'observation de la solidarité de la DREES, ministère des Solidarités et de la Santé, pour une présentation de l'échantillon national interrégimes d'allocataires de compléments de revenus d'activité et de minima sociaux ;

Vladimir Passeron, chef du département de l'emploi et des revenus d'activité, INSEE, pour une présentation de l'appariement entre l'enquête emploi et le fichier historique de Pôle emploi, pour comprendre les différences entre les nombres de chômeurs et de demandeurs d'emploi ;

Nathalie Caron, sous-directrice des synthèses au sein de la DEPP, MENJS, pour une présentation du système d'information InserJeunes visant à mieux connaître l'insertion des jeunes.

Antoine BOZIO

Bonjour à tous. Je suis maître de conférence à l'Ecole des hautes études en sciences sociales (EHESS), professeur d'économie à l'Ecole d'économie de Paris et directeur de l'Institut des politiques publiques, qui utilise beaucoup les données évoquées ce matin et qui contribue modestement à des appariements.

Cette session vise à mettre en évidence des exemples de réalisations d'appariements menés au sein de la statistique publique, sous l'angle de trois thématiques. Nous décrirons les différentes manières d'exploiter ces appariements, tout en faisant la promotion auprès d'un large public. Mon introduction ne sera pas plus longue, afin d'offrir plus de temps aux présentations et aux échanges.

Tout d'abord, Patrick Aubert, sous-directeur à la DREES, nous présentera l'intérêt de l'échantillon national interrégime d'allocataires de compléments de revenus et de minima sociaux (ENIACRAMS) qui permet d'étudier l'ensemble du parcours des allocataires des différents minima sociaux et prestations sociales.

Dans un second temps, Vladimir Passeron, Chef du département de l'emploi et des revenus d'activité de l'INSEE, montrera l'intérêt de l'appariement réalisé entre l'enquête emploi, conçue pour mesurer le chômage selon la définition du chômage du BIT, et les données de Pôle emploi. Cet appariement doit permettre d'analyser la relation entre la notion statistique du chômage et sa notion administrative.

Enfin, Nathalie Caron, Sous-Directrice à la DEPP, soulignera l'intérêt du suivi des parcours d'insertion professionnelle des jeunes à partir des données des appariements du système d'information InserJeunes.

L'échantillon national interrégimes d'allocataires de compléments de revenus d'activité et de minima sociaux (ENIACRAMS)

Patrick AUBERT

L'ENIACRAMS permet de relier cette rencontre avec celle qui se tiendra le 18 mai autour de la question des panels. Comme d'autres panels, l'ENIACRAMS est construit à partir d'appariements.

Le panel ENIACRAMS

L'ENIACRAMS existe depuis 2001. Ce panel vise à suivre la population d'intérêt constituée par les bénéficiaires des minima sociaux dits d'âge actifs, suivis jusqu'à l'âge de la retraite. Il s'agit de bénéficiaires de l'ancien revenu minimum d'insertion (RMI), de l'allocation de parents isolés (API), du revenu de solidarité active (RSA), de l'allocation de solidarité spécifique (ASS) et de l'allocation aux adultes handicapés (AAH). En somme, même si les prestations sociales ont évolué au cours de cette période, le panel couvre la sphère des minima sociaux.

Ce panel a augmenté son champ à partir de 2009 en intégrant des prestations relevant des compléments de revenus d'activité, à savoir la partie activité du RSA, puis la prime d'activité qui lui a succédé en 2016. Ainsi, le nom de ce panel avait alors évolué, passant d'ENIAMS à ENIACRAMS.

L'ENIACRAMS n'est pas exhaustif. En effet, le panel est né à une époque où les capacités informatiques permettaient difficilement l'exhaustivité, et le principe de minimisation des données a en outre conduit à opter pour un échantillon. Malgré tout, cet échantillon s'avère bien plus important que ceux étudiés dans les enquêtes, comprenant fin 2020 environ 170 000 allocataires ou conjoints d'allocataires du RSA, 90 000 bénéficiaires de l'AAH, 25 000 allocataires de l'ASS et 410 000 allocataires ou conjoints d'allocataires de la prime d'activité. Cet échantillon permet donc de mener une analyse très fine.

Il s'agit de suivre des individus de manière longitudinale, en nous intéressant à la fois à leurs parcours liés aux minimas sociaux et, le cas échéant, à la suite de leur parcours, après la fin de la jouissance de ces prestations. Ainsi, le panel peut suivre différentes trajectoires d'allocataires, qui peuvent sortir, puis revenir dans la sphère des minima sociaux. Ce panel présente donc l'intérêt particulier d'aider à mieux appréhender les trajectoires de sortie des minima sociaux.

Néanmoins, ce panel porte sur les situations en fin d'année, son pas étant annuel. Nous ne cherchons donc pas à entrer dans tout le détail des parcours, en nous limitant à suivre les allocataires d'une année sur l'autre.

Les informations contenues au sein de l'ENIACRAMS

L'ENIACRAMS est construit à partir d'appariements de données administratives provenant de deux grands types de sources. D'une part, ces données sont fournies par les Caisses verseuses de prestations sociales, à savoir la CNAF, la Mutualité sociale agricole (MSA) et Pôle emploi. D'autre part, l'INSEE apporte des compléments importants liés à la connaissance des individus, dont en particulier des éléments démographiques qui ont permis de définir le contour de cet échantillon. L'INSEE apporte encore des données permettant de

suivre la mortalité et enfin, il offre des données sur l'emploi provenant des panels « tous salariés » et « tous actifs ».

Étant donné que les minima sociaux renvoient à la sphère de la sécurité sociale, le NIR est présent dans toutes les sources administratives mises en jeu. Les appariements sont donc facilités.

Grâce à des procédures en double aveugle et à un numéro non identifiant, la DREES ne peut pas retrouver les NIR identifiants lors des appariements de données.

Les informations appariées sont très riches. Parmi elles, nous retrouvons des caractéristiques des bénéficiaires, ainsi que des précisions sur les prestations perçues.

En effet, les caisses verseuses des minima sociaux et des compléments de revenus d'activité versent de nombreuses autres prestations, apportant des données complémentaires sur les aides au logement, les prestations familiales ou encore les autres allocations chômage du régime assurantiel. En observant les parcours, nous nous intéressons tant aux minima sociaux en eux-mêmes, qu'aux autres prestations sociales venant aider les personnes. De plus, les données sur l'emploi permettent d'observer l'ensemble du parcours de l'emploi, pendant et après la sortie des minima sociaux, ainsi que de nombreuses caractéristiques sociodémographiques.

Par ailleurs, l'ENIACRAMS constitue un dispositif couplé qui, en plus de servir de base d'étude, peut aussi servir de base pour le tirage d'échantillons pour des enquêtes portant sur la population étudiée.

De cette manière, il est donc possible d'enrichir l'ENIACRAMS avec des données d'enquêtes qui offrent des informations particulières, que nous ne pourrions pas puiser dans les sources administratives. Nous pouvons donc recueillir des données plus personnelles, qui peuvent notamment nous informer sur des opinions ou des aspirations. C'est ce qui est fait dans le cadre des enquêtes auprès des bénéficiaires de minima sociaux (« enquête BMS »), menées toutes les 6-7 ans en moyenne (la dernière en 2018).

Ce point est à souligner, car nous avons plutôt tendance à effectuer l'opération inverse. En effet, habituellement, nous débutons par des enquêtes, avant de chercher à les enrichir en les appariant avec des données administratives. Or une telle procédure est souvent coûteuse, lourde et imparfaite, car nous ne retrouvons jamais 100 % de la population de l'enquête et que nous sommes contraints d'effectuer des imputations.

Ainsi, l'ENIACRAMS présente l'avantage de permettre de retrouver sans peine l'intégralité des correspondances, mais aussi de suivre la population étudiée après l'enquête. Nous pouvons par exemple mettre en relation toutes les caractéristiques observées dans l'enquête avec l'évolution des parcours au sein et hors de la sphère des minima sociaux, cinq ou dix ans plus tard, ou même annuellement. Ce système est intéressant et pourrait constituer un modèle pour d'autres champs et pour d'autres panels.

<u>L'ENIACRAMS, la source de référence des éléments de parcours des bénéficiaires de minima sociaux</u>

Concrètement, cette source est utilisée par la DREES pour produire les différentes statistiques annuelles portant sur les parcours liés aux minima sociaux, notamment dans le cadre

du panorama « minima sociaux et prestations sociales » publié chaque année à la rentrée scolaire. Ce panorama renferme différents indicateurs annuels clés, tels que le taux d'entrée et de sortie dans la sphère des bénéficiaires des minima sociaux, l'ancienneté des bénéficiaires, ou encore la récurrence du bénéfice de ces minima. Ces données sont également accessibles pour la recherche, dans le cadre du CASD.

À l'aide de l'ENIACRAMS, nous avons pu obtenir par exemple un graphique présentant entre autres le taux d'entrées dans les minima sociaux en fonction de l'âge des bénéficiaires. Ces taux sont calculés à partir du nombre de nouveaux entrants une année sur l'autre. Nous constatons que ce taux est le plus haut chez les jeunes, étant de l'ordre de 30 % pour les moins de 30 ans et diminuant progressivement, descendant aux alentours de 10 % pour les plus âgés.

Ce graphique présente également des informations sur les parcours de ces bénéficiaires, en distinguant parmi ces nouveaux entrants ceux qui avaient déjà bénéficié de minima sociaux au cours des dix dernières années, des « vrais » nouveaux entrants, qui n'en auraient pas bénéficié sur cette même période. Nous constatons alors que l'essentiel des personnes de moins de 30 ans est constitué par de « vrais » nouveaux entrants. À l'inverse, plus de la moitié des plus de 30 ans avaient déjà bénéficié de minima sociaux au cours des dix précédentes années.

Pour citer un autre exemple, l'ENIACRAMS nous a également permis de produire un tableau présentant la situation des anciens bénéficiaires du RSA ou de l'ASS, un an après leur sortie de ces régimes. Ces données sont obtenues grâce à l'appariement des informations liées au RSA et à l'ASS avec des données d'emploi et des allocations de chômage. Il est ainsi possible de suivre chaque année parmi ces sortants du RSA ou de l'ASS la proportion de personnes employées en CDI ou en CDD, en temps complet ou en temps partiel, ou encore la part des bénéficiaires de la prime d'activité. Nous incluons également dans ce tableau la part des personnes décédées. Ainsi, il est possible de créer des indicateurs particulièrement riches.

<u>Des possibilités d'appariements pour approfondir la connaissance des parcours</u>

Ce panel est déjà très riche, mais la question des trajectoires des bénéficiaires des minima sociaux est très vaste. Intéressons-nous donc aux appariements envisagés dans un avenir proche pour améliorer la connaissance de ces parcours.

Nous souhaitons éviter une logique de silo et décrire de façon globale la trajectoire et les conditions de vie des personnes. Aussi, nous souhaitons donc participer à cette petite révolution qui touche le monde de la statistique publique et qui rend les appariements beaucoup plus simples.

L'ENIACRAMS présente l'avantage de constituer un échantillon de grande taille qui laisse espérer l'obtention de correspondances avec de nombreuses sources. Ces correspondances sont notamment favorisées par le NIR auquel nous pouvons appliquer le CSNS. En outre, il faut préciser que les personnes sont sélectionnées dans l'ENIACRAMS selon leur jour de naissance, en incluant les jours de l'EDP. De ce fait, cet échantillon partage une caractéristique commune avec de nombreux panels, et permet donc de nombreux rapprochements. Nous espérons mettre en œuvre un certain nombre de nouveaux appariements dès cette année.

Les deux premiers projets d'appariements que nous allons présenter concernent le champ des parcours d'insertion des bénéficiaires des minima sociaux.

Tout d'abord, le dispositif Remontées individuelles sur l'insertion (RI-insertion), en cours d'élaboration, devrait permettre de faire remonter des données sur toutes les actions d'accompagnement et d'orientation réalisées par les organismes intervenant auprès de ces bénéficiaires – Conseils départementaux, CAF, Pôle emploi. En appariant ces données, nous serions en mesure de relier ces actions d'accompagnement avec les parcours. Il pourrait alors être possible d'identifier les disparités territoriales, ou encore les actions les plus efficaces pour aider à un retour à l'emploi.

Un autre appariement est envisagé avec les données du système d'information sur les mouvements de main-d'œuvre (SISMMO) créé par la DARES à partir de la DSN. Ce fichier est complémentaire aux données de l'INSEE. Il est axé sur les mouvements de main-d'œuvre, son pas temporel est très fin et ses données sont facilement accessibles. Le fichier du SISMMO permettrait d'enrichir les indicateurs relatifs aux sorties du bénéfice des minima sociaux grâce à l'emploi. Il aiderait encore à renforcer les indicateurs touchant les parcours d'emploi. Cet apport serait d'autant plus intéressant qu'un bénéficiaire de minima sociaux sur six est en situation d'emploi, dans le cadre d'emplois précaires à temps partiel ou à salaires trop faibles. Nous cherchons donc à couvrir l'ensemble du parcours de précarité des bénéficiaires des minima sociaux.

D'autres rapprochements sont imaginés et devraient sortir du champ traditionnel de l'ENIACRAMS. Ils s'insèrent plus particulièrement dans le cadre de cette révolution qui tend à sortir de logiques de silos, rompant ainsi avec le développement hermétique de travaux thématiques.

Une des premières grandes étapes qui pourrait être franchie dans ce sens pourrait concerner le rapprochement de différentes données touchant l'insertion et le monde de la retraite. Il serait effectivement possible d'apparier les données de ces parcours avec les données de l'échantillon interrégime de retraités (EIR) et de l'échantillon interrégime de cotisants (EIC).

Aujourd'hui, nous suivons très bien les parcours des bénéficiaires des minima sociaux, mais ceux-ci disparaissent de nos radars dès lors qu'ils jouissent de leur retraite. En effet, même si nous disposons d'éléments touchant une partie de leurs carrières qui permettraient d'estimer des droits à la retraite, nous manquons d'informations en la matière. Nous pourrons donc enfin vérifier la part des personnes qui passeraient directement du RSA au minimum vieillesse. Il devrait également être envisageable d'étudier l'influence des périodes sans validation de droits à la retraite sur les montants de retraite des bénéficiaires de minima sociaux. En tout état de cause, les liens entre la thématique de l'insertion et celle de la retraite sont très riches.

Par ailleurs, il serait aussi possible de mieux appréhender la question du handicap en appariant aux données de l'ENIACRAMS celles de l'enquête Vie quotidienne et santé (VQS) qui s'inscrit dans le dispositif d'enquêtes « Autonomie ».

Le lien entre le handicap et les minima sociaux renvoie à l'AAH. Cependant au regard des réponses des bénéficiaires lors des enquêtes, nous réalisons que le handicap concerne également des bénéficiaires d'autres minima sociaux tels que le RSA. Ces questions amènent donc à interroger l'action publique dans son ciblage des allocations. Un rapprochement

avec l'enquête VQS qui contient des informations sur les limitations fonctionnelles des personnes permettra de mieux connaître les caractéristiques des handicaps des personnes bénéficiaires de minima sociaux.

D'autres projets, plus lointains, visent également à dépasser les logiques de silos. Il est ainsi question de croiser des données de l'ENIACRAMS avec celles du panel Trajectoires des jeunes aux mesures actives du marché du travail (TRAJAM) de la DARES qui touche la garantie jeune. Des appariements seront aussi possibles avec des données sur l'hébergement social et sur les personnes sans domicile fixe provenant du logiciel SI-SIAO des services intégrés d'accueil et d'orientation (SIAO).

Il est encore envisagé d'effectuer un rapprochement avec les données de la protection de l'enfance du dispositif d'Observation longitudinale individuelle en protection de l'enfance (OLINPE). Cet appariement devrait aider à réfléchir sur les parcours des sortants d'aides sociales à l'enfance, notamment sur le plan de l'insertion professionnelle et des carrières.

Conclusion

Pour finir, au-delà de cette liste de projets, nous constatons qu'il reste difficile d'envisager un appariement avec des données de santé du SNDS. La réalisation de ce type d'appariements n'est pas impossible, mais elle est plus complexe et devrait prendre du temps. Je me permets donc de profiter de ma prise de parole pour relayer l'attachement de la DREES à faire rejoindre les univers du social et de la santé. Dans le cadre de l'ENIACRAMS, ce rapprochement constituerait un apport pour la compréhension de l'insertion et plus encore pour celle du handicap. L'appariement des données de santé constitue l'un des grands enjeux du service statistique public.

Appariement entre l'enquête Emploi et le fichier historique de Pôle emploi pour comprendre les différences entre nombres de chômeurs et de demandeurs d'emploi

Vladimir PASSERON

Bonjour à tous. L'appariement entre l'enquête Emploi, menée auprès de personnes tirées au sort, et le fichier historique de Pôle emploi s'est opéré il y a deux ans, dans des conditions plutôt artisanales par rapport aux appariements de l'ENIACRAMS.

Chômage et demandeurs d'emploi inscrits à Pôle emploi : deux mesures qui ont nettement divergé depuis 2009

Nous avons souhaité apparier l'enquête Emploi et le fichier historique de Pôle emploi en partant du constat que le nombre de chômeurs au sens du BIT, mesuré chaque année dans l'enquête Emploi, commençait à diverger nettement par rapport au nombre de demandeurs d'emploi inscrits à Pôle Emploi et classés en catégorie A (sans emploi).

Ainsi, nous pouvons constater que ces deux indicateurs se situaient à un niveau équivalent vers 2008 et qu'ils ont progressivement divergé jusqu'à aboutir à un écart de plus d'un million de personnes. Nous nous sommes donc concentrés sur la période 2013-2017, qui a vu cet écart se creuser, avec près de 200 000 personnes supplémentaires inscrites dans la catégorie A de Pôle emploi et environ 200 000 chômeurs en moins au sens du BIT.

Nous avons donc souhaité comprendre les phénomènes qui se déroulaient au niveau individuel, d'autant plus qu'au premier abord ces deux définitions sont très proches. A cette fin, des équipes de l'INSEE, de la DARES et de Pôle emploi ont donc œuvré ensemble.

Les chômeurs au sens du BIT et les demandeurs d'emploi inscrits à Pôle emploi

Je rappelle qu'un chômeur au sens du BIT est une personne sans emploi qui recherche activement un travail et qui est disponible dans les deux prochaines semaines. Cette définition du chômage permet de réaliser des comparaisons internationales. Ainsi, des questions très précises sont posées dans l'enquête Emploi pour distinguer les chômeurs en fonction de cette définition.

Inversement, le demandeur d'emploi en fin de mois inscrit à Pôle emploi en catégorie A est sans emploi et est tenu d'en rechercher un. Néanmoins, aucun critère ou question ne permet de mesurer le caractère actif de leur recherche d'emploi.

Ainsi, en comparant les populations selon ces deux définitions sur un diagramme de Venn, nous relevons qu'une partie des chômeurs au sens du BIT ne sont pas inscrits à Pôle emploi. Dans l'autre sens, nous constatons, qu'une partie des demandeurs d'emploi inscrits à Pôle emploi et en catégorie ne sont pas considérés comme des chômeurs dans la définition du BIT.

Cet appariement qui visait à enrichir l'enquête Emploi avec des informations provenant de fichiers historiques de Pôle emploi ne s'est pas opéré directement. En effet, nous n'avons pu compter que sur le prénom, la date de naissance et l'adresse, puisque dans l'enquête Emploi nous ne conservons pas les noms, dans la mesure où son échantillon est un échantillon de logements, il n'est pas nécessaire de les conserver pour la collecte. Il est en revanche important de parvenir à bien identifier les différents individus de chaque logement à partir de leurs prénoms et dates de naissance, afin de pouvoir les suivre une année sur l'autre, puisque l'enquête Emploi se réalise sur un panel à interrogation trimestrielle.

Cependant, bien que nous disposions, dans les deux bases de données à apparier et pour chaque individu, d'un prénom, d'une date de naissance et d'une adresse – par exemple Bruno, né le 1er janvier 1970 et résidant au 1 rue de Bercy – l'appariement entre ces deux bases n'a pas été simple.

D'une part, l'adresse postale dont dispose Pôle emploi pour pouvoir contacter les demandeurs d'emploi peut parfois être très différente de l'adresse d'habitation figurant dans les fichiers de la taxe d'habitation, qui est la base d'échantillonnage de l'enquête Emploi.

Des prénoms qui peuvent différer entre les deux sources

D'autre part, les prénoms officiels peuvent différer des prénoms recueillis par les enquêteurs, pouvant comporter notamment des variantes orthographiques – Mathieu et Matthieu, Lorène et Lorraine –, ce qui pose d'autant plus de difficultés avec les prénoms à consonance étrangère. En outre, nous acceptons également des surnoms tels que Cathy pour Catherine, le prénom visant surtout à identifier les personnes et à les suivre d'un trimestre sur l'autre.

Au passage, nous nous sommes demandé s'il n'aurait pas été intéressant de poser dès le départ des questions identifiantes dans le cadre de cette enquête Emploi. Nous aurions pu

récolter des noms, voire des NIR. Cela aurait permis de recueillir par appariement avec des sources administratives des informations que nous retrouvons aujourd'hui avec peine, notamment sur les revenus.

Quoi qu'il en soit, nous avons mis en œuvre différentes solutions pour dépasser cette difficulté liée aux prénoms. Notamment nous avons comparé les chaînes de caractère en acceptant de petits écarts d'orthographe, nous avons calculé la distance entre les chaînes de caractères et nous avons recouru à des prénoms phonétisés.

Des adresses qui peuvent différer entre les deux sources

De plus, nous devions encore tenir compte des logements dont les adresses pouvaient différer au sens fiscal et au sens de l'enquête Emploi. Il s'agit essentiellement de logements situés sur des immeubles donnant sur deux rues. Les maisons d'angles posaient davantage que les autres ce problème. Pour faire face à cette difficulté, nous avons donc géolocalisé les adresses des deux fichiers à apparier, avant de calculer les distances entre les deux adresses, et permettre ainsi un rapprochement entre individus d'un même logement mais dont les adresses diffèrent dans les deux sources.

Pour encore améliorer nos taux d'appariements, nous avons encore tenu compte d'une autre particularité. Nous avons relevé que certains groupes de prénoms partageaient des dates de naissance, mais étaient associés à des adresses différentes. Étant donné qu'ils avaient les mêmes dates de naissance dans les deux sources, il était très probable qu'il s'agisse des mêmes personnes.

Apparier pour observer le recoupement de deux populations

Au total, entre 2012 et 2017, environ 17 millions de personnes ont été inscrites dans la catégorie A de Pôle emploi. En parallèle, notre enquête Emploi portait sur 400 000 personnes. Nous sommes finalement parvenus à effectuer un appariement avec un taux de rapprochement de 85 %. Ce taux est à la fois faible et élevé. Mais nous avons validé la qualité de cet appariement à partir de différents tests de robustesse, comprenant des vérifications « manuelles ».

Finalement, cet appariement nous a permis de construire un diagramme de Venn croisant les catégories d'activité, d'une part au sens de l'enquête Emploi et d'autre part, au sens des différentes catégories statistiques de Pôle emploi.

Sans commenter l'ensemble des données du diagramme obtenu, intéressons-nous à l'écart entre les nombres de chômeurs inscrits en catégorie A à Pôle emploi et de chômeurs au sens du BIT.

Nous relevons au préalable que l'appariement retranscrivait bien les évolutions globales des années 2013-2017, ce qui en validait la robustesse et la qualité.

En 2017, 44 % des inscrits en catégorie A ne sont pas chômeurs au sens du BIT ...

Finalement, nous constatons qu'en 2017, sur 100 personnes inscrites en catégorie A à Pôle emploi, 44 ne sont pas chômeurs au sens du BIT. Ces chiffres sont plutôt structurels et ne varient pas sensiblement d'une année sur l'autre.

Ces 44 personnes sont en général plutôt des seniors proches de la retraite, qui, découragés n'effectuent pas de recherche active. Parmi eux, nous retrouvons aussi des personnes qui souffrent de problèmes de santé et qui ne peuvent pas faire de recherche active.

... et 33 % des chômeurs au sens du BIT ne sont pas inscrits en catégorie A

Sur 100 chômeurs au sens du BIT, 67 sont bien inscrits en catégorie A à Pôle Emploi, mais un tiers n'y est pas. Parmi ce tiers, nous retrouvons plutôt des jeunes, qui, ne pouvant pas prétendre à une indemnisation, ne prennent pas le temps de passer par Pôle emploi.

Les raisons de l'accroissement de l'écart entre les nombres de chômeurs au sens du BIT et des inscrits en catégorie A constaté entre 2013 et 2017

L'appariement a avant tout été motivé par la volonté d'observer des évolutions et de comprendre la divergence touchant les deux définitions des chômeurs. Je rappelle que cette divergence a été de l'ordre de 400 000 personnes entre 2013 et 2017.

Ainsi, nous constatons une augmentation de 300 000 personnes, au sein de la population inscrite en catégorie A à Pôle emploi et qui n'est pas composée de chômeurs au sens du BIT.

Nous retrouvons là aussi davantage de seniors. Nous avons donc cherché à expliquer cette évolution. Et cette explication se retrouve certainement du côté des mesures visant au recul de l'âge de départ à la retraite et de la fin des dispenses de recherche d'emploi à Pôle emploi. Cette augmentation touchant le nombre d'inscrits en catégorie A sans être chômeurs au sens du BIT touche ainsi ces seniors découragés qui s'orientent de moins en moins vers une recherche active d'emploi.

En outre, nous constatons que le nombre de chômeurs au sens du BIT qui ne sont toutefois pas inscrits à Pôle emploi a diminué à hauteur de 100 000 personnes. Cette baisse peut s'expliquer par un retournement du marché du travail, à compter de 2015, en faveur des jeunes. Néanmoins, comme les jeunes ne sont souvent pas inscrits à Pôle emploi, cette diminution ne se traduit pas par une baisse équivalente des inscrits en catégorie A.

Conclusion

Ainsi, cet appariement qui s'est avéré nécessaire et très utile nous a permis de répondre à de nombreuses questions sur les divergences que nous avions relevées entre ces deux mesures. Les données rapprochées permettent de relativiser la proximité entre les deux concepts du chômage qui semblaient très proches a priori. Nous mesurons en effet des écarts relativement consistants.

Nous avons dû dépasser de nombreuses contraintes méthodologiques pour cet appariement car ne disposant que de peu d'informations identifiantes. Nous avons cependant pu prouver sa robustesse (voir pour cela le document de travail en ligne sur le site de l'Insee). Il s'agit désormais de renouveler régulièrement ce rapprochement à intervalles réguliers, afin de pouvoir notamment observer les divergences postérieures à la crise du coronavirus.

Antoine BOZIO

Je vous remercie pour cette présentation qui rappelle que les appariements ne touchent pas que des rapprochements de données administratives. Relier des données d'enquêtes avec des données administratives apporte tout autant de richesse, tant pour renforcer une bonne compréhension des données d'enquêtes que pour mieux comprendre les informations des données administratives.

Mieux connaître l'insertion des jeunes : le système d'information InserJeunes

Nathalie CARON

Qu'est-ce qu'InserJeunes ?

Bonjour à tous. InserJeunes est un tout nouveau système d'information, sous co-maîtrise d'ouvrage de la DEPP et de la DARES, situé à la croisée des sphères de l'éducation et de l'emploi. Il s'agit d'un appariement de bases de données individuelles de la sphère de l'éducation avec une base individuelle de la sphère de l'emploi.

Les premiers résultats ont été diffusés en février 2021 et ont été suivis d'autres résultats publiés en décembre 2021. Il s'agit de plusieurs séries de production de données. Les projets d'appariements prennent toujours du temps et nous avons mis trois ans pour organiser celui-ci, avec l'aide d'un fonds de transformation de l'action publique (FTAP). Ce projet a été développé avec une équipe dédiée, dissoute à ce jour.

InserJeunes permet en premier lieu de mesurer le taux d'emploi des jeunes qui sortent de formations professionnelles de niveau CAP à BTS, par voie scolaire ou par voie d'apprentissage. Ce dispositif vise également à exploiter une très riche base de données qui permet de qualifier précisément les caractéristiques de cette insertion à travers différents indicateurs tels que le type de contrat, les secteurs dans lesquels les jeunes s'insèrent, l'adéquation entre la formation et l'emploi, ou encore le salaire.

La mesure du taux d'insertion, qui constituait la demande initiale, se base sur des indicateurs déclinés par formation au niveau national et régional, jusqu'au niveau des établissements. En effet, les appariements de données individuelles quasiment exhaustives permettent de descendre à des niveaux très fins, soit jusqu'aux lycées professionnels et aux Centres de formation d'apprentis (CFA).

A l'aide de ces résultats, nous pouvons produire des éléments de cadrage national et régional. Il est ainsi possible de mettre à disposition des familles et des jeunes ces données d'insertion, pour aider les élèves à s'orienter en fin de troisième ou en terminale.

Quatre points sur l'insertion sont effectués pour chaque cohorte de sortants. Ces points se situent à l'échéance de 6, 12, 18 et 24 mois après la sortie du système éducatif. Comme il s'agit de gérer simultanément deux cohortes, la production des données s'avère assez lourde à gérer.

La genèse d'InserJeunes

La création d'InserJeunes est à relier à une volonté ancienne de rapprocher les sphères de l'emploi et de l'éducation, pour obtenir des informations sur l'insertion professionnelle.

Jusque-là, nous disposions d'enquêtes, qui ne permettaient pas de disposer d'information à des niveaux très fins.

Or l'article 24 de la loi n°2018-771 du 5 septembre 2018 pour la liberté de choisir son avenir professionnel a posé de nouvelles exigences.

Cet article précise que « chaque année, pour chaque CFA et pour chaque lycée professionnel, sont rendus publics, quand les effectifs concernés sont suffisants : le taux d'obtention des diplômes ou des titres professionnels ; le taux de poursuite d'études ; le taux d'interruption en cours de formation ; le taux d'insertion professionnelle des sortants de l'établissement concerné, à la suite des formations dispensées ; la valeur ajoutée de l'établissement ». Ce dernier indicateur consiste à mesurer l'effet propre de l'établissement en tenant en particulier compte des caractéristiques scolaires et sociodémographiques des élèves.

Cet article de loi précise encore que « pour chaque CFA, est également rendu public chaque année le taux de rupture des contrats d'apprentissage conclus », cet indicateur étant important pour l'insertion par voie d'apprentissage.

La situation en 2018, lors de la publication de la loi pour la liberté de choisir son avenir professionnel

Lors de la mise en application de la loi n°2018-771 du 5 septembre 2018, la DEPP et la DARES ont été confrontées à la grande difficulté liée à l'obtention de taux d'insertion à un niveau aussi fin que celui des établissements.

En 2018, nous pouvions compter sur deux enquêtes menées par la DEPP, à savoir les enquêtes Insertion vie active (IVA) et Insertion professionnelle des apprentis (IPA). Ces enquêtes étaient exhaustives et concernaient donc l'ensemble des jeunes sortants.

Cependant, dans le cadre de ces enquêtes, nous ne disposions pas d'une base de données des jeunes sortants pour pouvoir les identifier et leur envoyer des questionnaires. Ces enquêtes étaient donc très coûteuses et très lourdes à gérer à la fois pour les chefs d'établissement, les équipes de la DEPP et les rectorats, qui intervenaient dans la collecte. Il fallait envoyer des questionnaires et effectuer des relances, qui étaient le plus souvent téléphoniques.

En outre, ces enquêtes n'étaient associées qu'à un seul point d'insertion, effectué sept mois après la sortie du système éducatif, soit en février de l'année suivant ces sorties.

Enfin, le taux de réponse était de l'ordre de 50 à 60 %, ce qui est assez bon, mais insuffisant pour obtenir des indicateurs au niveau des établissements ni même pour certaines spécialités au niveau national.

Les finalités d'InserJeunes

C'est pourquoi nous avons créé InserJeunes, dont la finalité initiale consistait à répondre aux exigences de cette loi de 2018. Il s'agissait de construire un système *ad hoc* rassemblant des données administratives exhaustives, qui étaient nécessaires pour descendre au niveau des établissements. Pour ce faire, nous avons rapproché des bases administratives scolaires, fondées sur les inscriptions des élèves et des apprentis, avec la base SISMMO de la DARES qui mesure les mouvements de main d'œuvre et qui repose sur la DSN.

Il a donc été possible de calculer le taux d'insertion professionnel des sortants de chaque établissement et la valeur ajoutée de ces établissements. En outre, nous avons constaté que ces rapprochements de bases permettaient aussi de calculer deux autres indicateurs prévus par cette loi, à savoir le taux de poursuite d'études et le taux d'interruption en cours de formation.

InserJeunes a également permis d'aller plus loin que ce que demandait cette loi, qui se limitait au niveau de l'établissement. En effet, nous avons pu construire des indicateurs à un niveau particulièrement fin, au croisement de l'établissement, de la formation suivie – CAP, baccalauréat, etc. – et de la spécialité, voire au niveau du diplôme de la voie professionnelle.

De plus, comme nous l'avons décrit, nous mesurons nos indicateurs à 6, 12, 18 et 24 mois après la sortie du système éducatif. Le délai de mise à disposition de ces taux est également beaucoup plus rapide que celui des enquêtes, puisqu'on obtient les informations du premier point à six mois concernant la situation en janvier de l'année n, dès le mois de décembre de l'année n.

Enfin, adossé à ce système d'information, il existe une base de données très riche qui rassemble les données de la DEPP sur les formations avec celles de SISMMO, et qui permet en particulier, comme je l'ai déjà précisé plus haut, de mesurer l'adéquation entre la formation et l'emploi ou encore le niveau d'emploi obtenu.

Les appariements sont au cœur d'InserJeunes

Le système d'information InserJeunes compte une vingtaine d'appariements. Certains sont simples et d'autres plus complexes. Les appariements les plus simples mobilisent les INE, identifiant national propre à chaque élève, étudiant ou apprenti qui a vocation à faciliter la gestion du système éducatif et à permettre le suivi statistique des élèves, étudiants et apprentis.

L'INE permet notamment de définir un fichier des sortants de l'éducation nationale, qu'on ne retrouve pas deux années de suite dans les fichiers d'inscription. Or pour apparier ce fichier de sortants avec les données du SISMMO, une difficulté apparaît. En effet, il s'agit d'apparier des fichiers dont les identifiants ne sont pas identiques, la source de la sphère de l'emploi se fondant sur le NIR. L'appariement est donc réalisé à partir des noms, prénoms, date de naissance, sexe et commune de naissance.

En somme, InserJeunes se construit autour d'une vingtaine d'appariements et dix d'entre eux ont été effectués de manière indirecte, dont celui réalisé avec SISMMO, qui était capital pour mesurer l'emploi. Ainsi, en 2018, nous nous sommes demandé quel outil choisir pour effectuer ces appariements.

L'outil utilisé pour les appariements

Nous avions besoin d'un outil robuste et assez rapide, tout en sachant qu'aucune reprise manuelle ne serait effectuée a posteriori. Nous avons donc essayé plusieurs méthodes et outils.

D'une part, nous disposions déjà d'outils pour donner ou vérifier l'INE à la DEPP. D'autre part, nous avons également essayé l'outil MatchID du ministère de l'Intérieur, ou encore des programmes SAS de plusieurs provenances.

Néanmoins, pour réaliser ces appariements indirects, aucun de ces outils ne nous donnait satisfaction. Nous avons donc construit un outil spécifique d'appariements en Python, en utilisant des librairies disponibles en open source. Cet outil peut a priori s'adapter à un certain nombre d'appariements, notamment car il a été construit sur des librairies disponibles en open source.

Les appariements réclament avant tout une bonne qualité des variables de départ et nous avons la chance de bénéficier de variables d'une qualité extrême. En effet, les noms, prénoms et autres traits d'identité sont de bonne qualité, tant avec l'INE que le NIR.

La diffusion des résultats par établissement : un site dédié et des données en open data

Ainsi, InserJeunes dont le projet a abouti en février 2021, a permis d'obtenir des premiers résultats, au niveau des établissements et au niveau national et régional. Nous projetons de les exploiter davantage. Les données sont mises en ligne en open data sur le site https://www.data.education.gouv.fr.

Parallèlement, un site spécifique a été construit pour permettre une large diffusion des résultats obtenus : https://www.inserjeunes.education.gouv.fr/diffusion/accueil. Ce site permet d'accéder facilement aux données des établissements, renseignant pour chaque spécialité, le taux de poursuite d'études, le taux de sortants du système éducatif et le taux d'insertion professionnelle des élèves ou des apprentis.

Quelques résultats au niveau national

Au niveau national, InserJeunes offre également des premiers résultats qui concernent pour le moment les cohortes de sortants de 2018, 2019 et 2020. Ces résultats n'incluent que les deux premiers points, réalisés à 6 et 12 mois, tandis que nous ne disposons que du premier point pour la cohorte de 2020. Les points suivants devraient être publiés dans les six prochains mois.

Quoi qu'il en soit, il est déjà possible d'observer l'effet de la crise du coronavirus sur l'insertion professionnelle de ces différentes générations de sortants.

Nous avons également publié des résultats sur le taux de poursuite d'études, ou encore le type de contrat de travail – CDI, CDD, intérim, contrat de professionnalisation – en distinguant les femmes et les hommes.

Conclusion

Pour conclure, nous avons la perspective d'intégrer à InserJeunes le champ des salariés du public. En effet, nous ne nous basons pour le moment que sur celui des salariés du privé, qui correspond au champ de la source de données sur l'emploi utilisée qui se fonde sur la DSN.

En outre, nous souhaitons intégrer les formations en lycées agricoles, puisque nous n'avons pour le moment pris en compte que les établissements relevant du ministère de l'Education nationale, de la Jeunesse et des Sports (MENJS) et l'ensemble des CFA.

Parallèlement, le projet « Trajectoires professionnelles dans l'ensemble l'enseignement supérieur » porté par le SSM Systèmes d'information et d'études statistiques (SIES) du ministère en charge de l'Enseignement supérieur en collaboration avec la DARES démarrera bientôt. Ce projet s'articule autour d'un objectif proche de celui d'InserJeunes, et concerne les étudiants de l'enseignement supérieur.

Finalement, il nous reste également à exploiter toute la richesse des données disponibles offertes par InserJeunes, pour mieux comprendre les conditions d'emploi des sortants du système éducatif.

Echanges

Nicolas PROKOVAS, CGT

Quelle est la source qui permet de saisir les caractéristiques de l'emploi ?

Thomas VROYLANDT, UNEDIC

L'ENIACRAMS contient des informations sur les demandeurs d'emploi. Contient-il aussi des informations sur les indemnisations du chômage ?

Claude CASTELLUCCIA, CNIL

Ces données sont-elles pseudonymisées ? Le cas échéant, comment faites-vous pour compléter ces données avec des enquêtes ?

Elisabeth POTREAU, INSEE

InserJeunes ne constitue-t-il pas un doublon avec l'enquête EVA ?

Laurence HAGUET, Cour des comptes

L'appariement entre des données déjà appariées et des données brutes a-t-il un sens ou une utilité dans l'évaluation des politiques publiques ? Le cas échéant, sur quels garde-fous ou préconisations pourrions-nous compter à cet égard ?

Stéphanie LEMERLE, DREES

Les résultats des taux d'insertion à un niveau très fin sont-ils publics ? Comment pouvons-nous y accéder ?

Patrick AUBERT

Les données relatives à l'emploi de l'ENIACRAMS sont issues du panel « tous actifs » de l'INSEE, qui utilise différentes sources provenant de la DADS, de la DSN et de sources touchant l'emploi public. L'ENIACRAMS souhaite donc également procéder à des appariements avec le fichier sur les mouvements de main-d'œuvre du SISMMO.

Au sujet des données de Pôle emploi, il existe effectivement des données d'indemnisation du chômage. Cependant, nous ne suivons les personnes qu'une fois qu'elles entrent dans la sphère des minima sociaux et pas en amont.

Enfin, à la DREES, les données sont pseudonymisées, donc nous ne disposons pas des identifiants. Comme je l'ai expliqué, l'appariement s'effectue en double aveugle, en faisant intervenir l'INSEE. Ainsi, lorsque nous avons besoin d'effectuer des enquêtes, nous remettons les numéros non signifiants à l'INSEE et aux Caisses qui nous permettent ensuite de retrouver les personnes à interroger.

Nathalie CARON

L'enquête EVA se fonde sur un échantillon, tandis qu'InserJeunes se base sur des sources exhaustives, ce qui permet d'obtenir des données disponibles à des niveaux plus fins. Vous trouverez le lien du site dédié à la diffusion de ces données, ainsi que celui renvoyant à l'open data à la fin de ma présentation d'InserJeunes.

Mireille ELBAUM

Au sujet de l'appariement de l'enquête Emploi et des données de Pôle emploi, nous remarquons qu'il s'agit d'une opération ambitieuse et difficile. Pourtant, cet appariement touche une question fondamentale, permanente et potentiellement évolutive, en fonction de l'évolution de la réglementation de l'indemnisation du chômage ou de celle de la retraite. Par conséquent, comment pourrions-nous réfléchir non pas à une pérennisation de cet appariement, mais à sa routinisation ? Serait-il possible de la rendre périodique, pour répondre à ce besoin permanent ?

De plus, Patrick Aubert a répondu que les données sur l'indemnisation du chômage ne sont pas suivies dans l'ENIACRAMS. Ainsi, bien qu'il s'agisse là aussi d'une question permanente relative à notre système de protection sociale, la division de nos ministères et des champs de compétences fait que les données sont produites de façon fragmentée, et sur la base de concepts différents, en matière de prestations sociales.

Par ailleurs, vous nous indiquez qu'InserJeunes ne suit pas les étudiants de l'enseignement supérieur. Cependant, je m'interroge sur les personnes qui ne se trouvent pas en situation d'emploi et qui ne sont pas des chômeurs inscrits, pouvant être inactifs, ou encore partis à l'étranger. Sur cette base, comment serait-il possible de compléter le panorama sur l'insertion des jeunes ?

Enfin, je relève une dernière difficulté touchant la DEPP. Les différentes données mises en jeu dans le cadre des différents appariements, sont-ils couverts par le secret statistique ? Une partie de ces données est-elle remise après son traitement à disposition des établissements et des acteurs administratifs du ministère concerné ? En effet, les traitements réalisés par la DEPP à partir de l'INE, y compris quand il s'agit de corriger des données existantes, ne sont pas toujours considérés comme des statistiques, à toutes les étapes des traitements.

Patrick AUBERT

Au sujet de l'indemnisation du chômage, je précise que le projet MIDAS est un projet d'appariement qui ne retiendra pas seulement les bénéficiaires de minima sociaux. De plus, le

projet d'appariement avec l'EIC qui contient des informations sur toute la carrière des individus, et donc également sur les indemnisations de chômage, permettra de mieux connaître les parcours des bénéficiaires de minima sociaux, y compris avant qu'ils ne bénéficient de ces prestations.

Vladimir PASSERON

Comme je l'ai précisé, nous comptons renouveler l'appariement portant sur l'enquête Emploi. Néanmoins, je ne suis pas sûr qu'il soit nécessaire de l'entreprendre tous les ans, sachant que les résultats sont plutôt d'ordre structurel. Cependant, l'écart entre les nombres totaux de chômeurs au sens du BIT et des inscrits en catégorie A à Pôle emploi est suivi plus régulièrement, chaque trimestre.

Quoi qu'il en soit, au niveau des données individuelles, j'ai évoqué tout à l'heure que nous nous interrogions sur l'intérêt d'ajouter des questions davantage identifiantes dans les enquêtes. Une telle opération permettrait de favoriser les futurs appariements et de faciliter les questionnements. En effet, il est actuellement complexe de collecter une information précise sur les salaires : pour être sûr de bien collecter le concept de salaire requis, il est notamment préférable que les enquêtés prennent le temps de chercher leurs fiches de paye. L'usage de questions identifiantes dans ces enquêtes, et des appariements ultérieurs pour enrichir les enquêtes, présenteraient donc des avantages. Néanmoins, le recours à cette pratique impliquerait un changement culturel : il pourrait apparaître contradictoire de nous rendre auprès des enquêtés pour présenter une enquête anonyme, tout en leur demandant leurs noms, prénoms et numéros de sécurité sociale, en plus des questions très précises propres à l'enquête. Une vraie réflexion est donc à amorcer en la matière.

Nathalie CARON

Pour répondre aux autres questions, il est vrai que la question des personnes qui ne sont pas en situation d'emploi dans le secteur privé reste ouverte. Nous ne savons pas s'ils sont au chômage, inactifs ou encore s'ils se trouvent à l'étranger.

Par ailleurs, nous avons réalisé qu'un accompagnement était nécessaire pour que les données que nous produisons puissent aider les familles et les jeunes dans leur orientation. Pour cela, il faut que le public puisse s'emparer correctement de ces données, de manière à pouvoir les interpréter. Nous travaillons donc sur cette question au sein du ministère avec la Direction générale de l'enseignement scolaire (DGESCO), afin de sensibiliser l'ensemble des acteurs concernés.

Enfin, les appariements menés dans le cadre d'InserJeunes s'opèrent avec des données hautement sensibles et la liste des destinataires des bases individuelles a donc été très clairement définie dès le départ, dans le respect du RGPD et de l'analyse d'impact relative à la protection des données (AIPD).

SESSION 3 – LES PROJETS D'AVENIR

Président de la session : Xavier Timbeau, Président de la Commission Environnement et développement durable du CNIS ;

Lionel Espinasse, Adjoint au Chef du département de la démographie, INSEE, pour une présentation du code statistique non signifiant (CSNS) ;

Olivier Lefebvre, Maître d'ouvrage du programme RESIL, INSEE, pour une présentation du répertoire statistique des individus et des logements (RESIL).

Xavier TIMBEAU

J'ai le plaisir de présider cette session sur les projets d'avenir. Je pense qu'il est clair dans tous les esprits que la question des méthodes des appariements est absolument cruciale pour la statistique nationale d'aujourd'hui et de demain.

Lionel Espinasse nous présentera le CSNS, introduit récemment dans le cadre de la loi pour la République numérique. Il nous fera part des différents enjeux de ce CSNS. Une longue histoire se trouve derrière les craintes qui ont poussé à générer cet identifiant non signifiant. Je pense notamment à l'épisode très agité de l'introduction par René Carmille du numéro d'identification. En effet, ce numéro qui visait à faciliter la mobilisation des classes d'âges successives et mieux combattre les nazis avait été accusé par la suite de servir les desseins du régime de Vichy.

Enfin Olivier Lefebvre nous présentera le répertoire statistique des individus et des logements (RESIL). Il mettra en lumière des problématiques importantes, en évoquant notamment les solutions envisagées pour réagir à la suppression de la taxe d'habitation. En effet, la suppression de cette taxe est parfois mal vécue dans l'univers statistique, puisque son fichier sert à échantillonner de nombreuses enquêtes.

Le code statistique non signifiant (CSNS)

Lionel ESPINASSE

Bonjour à tous. Le CSNS est une nouvelle offre de l'INSEE, dont l'usage est réservé au service statistique public. Sans revenir en détail sur ses objectifs qui ont été évoqués au cours de cette rencontre, je précise qu'il vise notamment à faciliter les appariements et à valoriser ainsi davantage les sources de données dont nous disposons.

<u>Qu'est-ce que le CSNS ?</u>

Le service CSNS se fonde sur une composante technique, avec la production d'une clé d'appariement (le code CSNS lui-même), ainsi que sur un aspect organisationnel.

Un code CSNS est affecté à chaque individu. Le calcul de cette clé doit produire la même valeur de CSNS, quelle que soit la source où l'on retrouve cet individu. Cette clé peut donc être utilisée pour apparier les informations de fichiers différents.

Le CSNS est non signifiant. Il ne doit donc comporter aucune information sur la personne en question.

Par ailleurs, le CSNS demeure pérenne sur dix ans. Si nous avions calculé un CSNS pour un individu dans une source en 2020, nous obtiendrons le même CSNS pour ce même individu dans une autre source en 2025. Au bout de ces dix années, une procédure de renouvellement peut être menée. En cas de faille de sécurité, nous pouvons également raccourcir cette période.

La composante organisationnelle du CSNS consiste à proposer un protocole d'usage qui comprend une phase de conventionnement entre les différentes parties. En outre, une application informatique dédiée permet aux utilisateurs de déposer des fichiers, pour lesquels ils souhaitent obtenir un CSNS. Ils trouveront en retour sur cette plateforme les résultats du calcul du CSNS générés par l'INSEE.

Un solide encadrement juridique

Sans revenir en détail sur la question de l'encadrement juridique, évoquée lors de la première session, je précise que le CSNS est issu de l'application de l'article 34 de la loi de 2016 pour une République numérique, qui le définit très précisément. Le CSNS est également porté par deux textes d'application de cette loi, à savoir un décret de 2016 et un arrêté de 2020.

Cet arrêté de 2020 précise les conditions techniques de sécurité du CSNS, notamment au sujet de la cryptographie, du chiffrement, des conditions de stockage et de son renouvellement.

Parallèlement, le CSNS est également encadré par l'article 30 de la loi informatique et libertés et par le RGPD, notamment dans son principe de minimisation des données qui guide assez fortement son protocole d'usage.

Une nouvelle offre de service

Cette offre de service attribue à chaque individu qui se trouve dans un fichier administratif ou d'enquête, quelle que soit la nature de l'opération statistique originelle, un code calculé pour chaque source et dont le résultat doit être unique pour chaque individu.

Ce calcul s'opère de deux façons. Si nous disposons d'une source statistique comportant un NIR, le calcul du CSNS est assez facile et il repose sur un hachage et sur du cryptage. Un service dédié à cette opération est déjà ouvert et fonctionne depuis octobre 2021. Il travaille déjà pour des utilisateurs, principalement pour la DREES. Quelques exemples d'appariements mobilisant ce CSNS ont été évoqués au cours de cette journée et les opérations fonctionnent bien.

La seconde façon de procéder permet d'obtenir un CSNS à partir de sources statistiques dépourvues de NIR. Il s'agit d'identifier les individus à partir de traits d'identité (état civil). Ces traits sont constitués par les noms, prénoms, sexe, et dates et lieux de naissance. Ces traits permettent alors de remonter au NIR, que l'on hache et crypte ensuite pour obtenir un CSNS. Un service dédié à l'application de cette seconde méthode devrait s'ouvrir au second trimestre de 2022.

Comme Kamel Gadouche, je vous montre à mon tour un exemple de CSNS. Ce code comporte 80 caractères et il ne permet pas d'identifier les personnes :

v1:3zUY-
VUpJFX9g7z3oi3zKSIKXB0yLIE4oGmxsQi1Z2atWI0n8lfnmVrxWUiJqeKHHTvcH+i0PvuO
2991M.

Utiliser le CSNS : les grandes étapes

En pratique, le protocole d'usage du CSNS comporte sept étapes. Les quatre premières étapes ne sont pas d'ordre technique, mais il s'agit de phases organisationnelles :

1. Les SSM doivent signer une convention de sous-traitance avec l'INSEE, tandis que les services internes de l'INSEE qui souhaiteraient aussi recourir au CSNS doivent signer une charte d'usage. L'INSEE, en tant que sous-traitant, n'est pas responsable du traitement. Il fournit uniquement la prestation du calcul du CSNS pour faciliter l'appariement.

2. Il faut ensuite établir une convention de collaboration entre les SSM et les services de l'INSEE qui souhaitent apparier entre eux des données. Cette convention mentionne toutes les motivations qui conduisent les deux organismes à apparier leurs fichiers. Elle détaille notamment les conditions de diffusion, ou encore les conditions d'étude. Le service CSNS n'intervient pas dans les négociations entre les deux parties, se contentant de veiller à ce que le protocole d'usage du CSNS soit bien respecté. Chaque partenaire dispose donc de la liberté de définir les règles de collaboration qui dépassent le cadre du protocole d'usage du CSNS.

3. Puis, le responsable du traitement doit déclarer le traitement, en citant le CSNS. Le service CSNS n'intervient pas dans cette étape.

4. Enfin, les SSM et les services de l'INSEE, doivent chacun inscrire explicitement leurs recours au CSNS dans leurs programmes de travail transmis au CNIS.

Viennent ensuite les étapes d'ordre technique :

5. Chaque partie dépose au service CSNS le fichier sur lequel il souhaite faire calculer un CSNS via une application informatique dédiée. Il s'agit essentiellement d'opérations de dépôts et de retraits de fichiers. Cette procédure a fait l'objet d'une homologation de sécurité obtenue en septembre 2021. Cette homologation sera à renouveler lors de l'ouverture du service dédié au calcul du CSNS passant par l'intermédiaire des traits d'identité.

6. Une fois que les fichiers ont été déposés, l'INSEE, en tant que sous-traitant, effectue le calcul du CSNS, soit à partir du NIR, soit à partir de traits d'identité – à compter du second trimestre de 2022 –, avant de le transmettre aux partenaires à travers la même application où les fichiers avaient été initialement déposés.

7. Les partenaires peuvent alors effectuer leurs appariements et produire des fichiers d'étude. Ils doivent alors respecter des règles liées à la durée et à la méthode de conservation des données, avec l'impératif de bien isoler les CSNS des variables d'intérêt des différents fichiers.

<u>Le calcul du CSNS</u>

Le calcul du CSNS s'effectue à partir du NIR, potentiellement obtenu après une étape d'identification statistique sur la base de traits d'identité. Une opération de hachage et de chiffrement du NIR est ensuite effectuée. Le hachage permet l'irréversibilité de l'opération, il n'est donc pas possible de retrouver un NIR une fois qu'il a été haché.

En revanche, le chiffrement est réversible. Il est donc possible de retrouver un NIR haché à partir du chiffrement qui lui est associé, en inversant la clé de chiffrement. L'étape du chiffrement rend alors possible la phase de renouvellement du CSNS prévue au bout de dix

ans. Pour renouveler un CSNS, nous n'avons donc pas à remonter jusqu'au NIR, mais simplement jusqu'à la signature hachée du NIR, ce qui constitue une garantie supplémentaire de sécurité.

Les modalités pratiques de l'utilisation du CSNS

Concrètement, deux partenaires A et B veulent apparier leurs fichiers. Le partenaire B est le responsable du traitement. Après les quatre premières étapes que nous avons décrites, ils procèdent alors au calcul du CSNS. Chaque partenaire demande alors séparément à l'INSEE de lui calculer des CSNS. Ces partenaires fournissent alors à l'INSEE soit leurs NIR, soit leurs traits d'identité, en vue d'obtenir en retour des CSNS. Différents cas de figure peuvent alors se présenter : seul un partenaire peut disposer du NIR ; seul un partenaire peut disposer de fichiers de très bonne qualité, etc.

Ensuite, le propriétaire du fichier A, qui n'est pas le responsable du traitement, va préparer pour le propriétaire B un fichier où il intégrera toutes ses variables d'intérêt en face des CSNS correspondants calculés par l'INSEE. Le propriétaire B pourra alors procéder à l'appariement sur la base du CSNS. L'appariement se déroulera alors plus ou moins parfaitement, avec parfois des CSNS d'un fichier qui ne trouveront pas toujours de correspondants dans l'autre fichier. Au final, le propriétaire B aura ses données, ainsi que celles du fichier A, bien réaffectées à chaque individu par le biais de numéros d'ordre. Les tables de passage contenant les CSNS sont à conserver de façon hermétique aux autres données.

Conservation du CSNS et sécurisation des données

Le service de l'INSEE délivrant le CSNS ne conserve aucune donnée, dès que le traitement est validé par le demandeur, d'autant plus si les fichiers contiennent des traits d'identité ou des NIR. Les propriétaires A et B doivent absolument conserver les CSNS hors des fichiers de variables d'intérêt avec des numéros d'ordre de manière à pouvoir les utiliser à nouveau. L'INSEE, en tant que sous-traitant, ne conserve alors plus rien, pas même les CSNS, dès lors que toutes les étapes techniques de l'appariement ont été validées.

Un focus sur l'identification statistique à partir de traits d'identité

Enfin, pour revenir au cas des NIR à retrouver à partir de traits d'identité, il se trouve que dans les fichiers administratifs ou d'enquêtes la qualité de remplissage des traits d'identité ne permet pas toujours de retrouver les bons NIR. Cette difficulté se retrouve notamment dans les cas où les personnes ont répondu sur un format papier soumis à une lecture optique. Ces difficultés peuvent également être accentuées par les différentes questions liées aux noms d'usage ou marital et aux autres éléments décrits dans la présentation de Vladimir Passeron.

Ainsi, lorsque l'on souhaite attribuer un CSNS à un trait d'identité et être certain que l'on puisse retrouver le même CSNS pour la même personne, quelle que soit l'opération statistique ultérieure, le processus méthodologique à suivre n'est pas si simple. Il faut trouver un équilibre entre la volonté d'identifier un maximum de personnes et celle de limiter le risque d'erreurs.

Il s'agit donc d'associer à chaque calcul individuel un indicateur de qualité. L'utilisateur devrait donc connaître *in fine* la qualité de l'appariement. Le projet est encore en cours et nous procédons actuellement à des tests avec quatre SSM, à savoir la DREES, la DARES, le

SDES et les SIES, afin d'ajuster au mieux cet équilibre. Différents principes techniques ont été retenus pour le moment, dont la recherche d'appariement exact avec le RNIPP. Cependant, si une personne a trois prénoms, qu'on ne retrouve que l'un de ces trois prénoms et qu'elle est la seule personne disposant de ce prénom qui soit née tel jour à tel endroit, nous pouvons légitimement être fondés à penser qu'il s'agit bien de la bonne personne. Nous réfléchissons donc au dosage du relâchement sur les correspondances de variables qui pourrait être toléré.

Le répertoire statistique des individus et des logements (RESIL)

Olivier LEFEBVRE

Bonjour à tous et à toutes. Je vais évoquer le projet du RESIL, un programme tout juste amorcé qui devrait aboutir en 2025. Il semblait important d'évoquer ce programme aujourd'hui, puisqu'il s'inscrit notamment autour de la question de la facilitation des appariements.

Tout d'abord, je vous présenterai le contexte et les objectifs de RESIL. Puis j'évoquerai plus concrètement le contenu des répertoires et les services que nous proposons de rendre, tels qu'ils sont envisagés aujourd'hui. Enfin, j'aborderai la manière dont nous tiendrons compte des enjeux juridiques et déontologiques relatifs à ces répertoires et aux pratiques d'appariements.

1. Contexte et objectifs

Le RESIL et les appariements

Parmi les finalités de RESIL, nous retrouvons la volonté de sécuriser, de fiabiliser et de faciliter les appariements, dans le respect des principes de la statistique publique et de la protection des données. Nous avons vu aujourd'hui à travers les différents exemples présentés que les appariements constituent une pratique extrêmement utile et que mutualiser certaines procédures pourrait bénéficier à tous.

De plus, RESIL est construit par l'appariement de plusieurs sources administratives, ce qui permet de disposer des données les plus complètes et précises possibles. Il faut également que RESIL soit suffisamment robuste pour résister à la disparition ou à la transformation de sources statistiques.

Le contexte

Nous devons faire face à la disparition programmée de la taxe d'habitation. Or celle-ci est très utile dans le processus de production statistique, puisqu'elle permet de discerner les contours des ménages, qui sont notamment essentiels pour les statistiques touchant les revenus ou les niveaux de vie. La taxe d'habitation permet également de fournir une base de sondage pour nos enquêtes. Enfin, elle sert d'intrant essentiel dans le cadre du recensement de la population.

Lorsque la disparition de cette taxe a été annoncée, nous avons dû donc trouver des solutions de court terme, opérantes dès 2023 et basées sur la mobilisation de données fiscales. Et nous avons également jugé utile de trouver des solutions de moyen terme en mobilisant différentes sources, en plus de ces sources fiscales.

Il existe une tendance de fond qui pousse l'INSEE, mais aussi l'ensemble de la statistique publique, en France et à l'international, à progresser dans l'utilisation des sources administratives et des appariements. Cette dynamique implique la mobilisation d'outillages techniques et méthodologiques.

Qu'est-ce qu'un répertoire ?

Un répertoire constitue une liste exhaustive d'objets – en l'occurrence RESIL concerne des individus et des logements – avec très peu de variables. Un répertoire est à la fois très haut, car il contient potentiellement toutes les observations d'un champ, mais aussi très étroit, car il ne renferme que peu de variables.

Les variables insérées dans un répertoire doivent permettre d'identifier sans ambiguïtés les unités qu'ils contiennent, notamment pour éviter des doublons et permettre de les relier avec d'autres éléments du système d'information. Le répertoire joue ainsi le rôle d'une colonne vertébrale du système d'information.

Les répertoires gérés par l'INSEE

L'INSEE s'inscrit dans une longue histoire et une longue expérience en matière de répertoires administratifs. L'INSEE gère ainsi le RNIPP depuis 1946. Plus récemment, en 2019, l'Institut a construit et pris en charge le répertoire électoral unique (REU) pour la gestion des listes électorales. Dans le domaine des entreprises, il a été confié en 1973 à l'INSEE la gestion du Système national d'identification et du répertoire d'identification des entreprises et de leurs établissements (SIRENE). En 2013, l'INSEE est intervenu dans la construction du répertoire *Legal Entity Identifier* (LEI), un répertoire international des entités qui interviennent dans les marchés financiers.

L'INSEE a également considéré qu'il fallait aller plus loin, en créant des répertoires statistiques situés en aval de ces répertoires administratifs. La statistique d'entreprise est d'ailleurs en avance par rapport à la statistique démographique et sociale. Nous avons en effet créé il y a quelques années le Système d'immatriculation au répertoire des unités statistiques (SIRUS), un répertoire statistique des entreprises et d'établissements, enrichi des informations collectées ou construites par la statistique publique – contours des groupes, niveau d'activité des entreprises, etc. Ainsi, nous souhaitons maintenant construire RESIL, un répertoire statistique propre aux domaines démographique et social.

Les spécificités du répertoire statistique

Passer d'un répertoire administratif à un répertoire statistique implique en premier lieu la définition de finalités différentes. En effet, ce type de répertoire vise uniquement à produire des informations statistiques et ses utilisateurs sont les membres du service statistique public. En outre, ce type de répertoires permet d'assouplir les règles de gestion. En effet, ces règles seront plus souples, car elles ne porteront pas de conséquences pour les entreprises ou les individus figurant dans ces répertoires, leur finalité n'étant pas administrative. En outre, il est possible d'ajouter dans ces répertoires quelques concepts statistiques tels que celui des ménages, mobilisé dans RESIL.

Les finalités de RESIL

RESIL doit permettre de répondre à quatre finalités :

- faciliter et fiabiliser les appariements mobilisant des sources administratives ou des enquêtes portant sur les individus, les ménages et les logements, afin de répondre à un réel besoin de l'ensemble des utilisateurs et des producteurs de données ;
- analyser la couverture des sources administratives utilisées au sein du service statistique public, afin de pouvoir caractériser les défauts de couverture et d'envisager la manière de traiter ces manques, dans le cadre d'un objectif d'exhaustivité ;
- Constituer des bases de sondages pour les enquêtes auprès des ménages ou auprès des individus ;
- Créer un point de référence commun pour la production de statistiques démographiques et sociales encore plus homogènes et comparables, le répertoire constituant le pivot ou la colonne vertébrale du système d'information.

2. Le contenu des répertoires et les services proposés

Concrètement, qu'est-ce que RESIL ?

Lors de sa mise en œuvre, RESIL sera constitué de deux répertoires statistiques. Le premier répertoriera les individus et l'autre les logements. Tous deux seront mis à jour en continu avec les naissances, les décès et différentes sources administratives, en ne retenant que des données d'identification. En effet, les autres données seront transmises aux systèmes d'information métiers qui les intégreront dans leurs processus de production.

Ces deux répertoires constitueront donc des entités vivantes et nous en prendrons des photographies une fois par an, ce qui est une fréquence raisonnable pour débuter. Nous constituerons donc notamment une liste des ménages, d'autant plus qu'il s'agit d'un intrant fondamental de certaines statistiques.

Mais la photographie du répertoire d'individus demande de pouvoir vérifier que les personnes répertoriées sont bien toujours résidentes sur le territoire. En effet, cette question se pose à tous les instituts de statistique qui fonctionnent à partir de répertoires et de données administratives. A cette fin, certains de ces instituts ont mis en place la méthode des « signes de présence ». Il s'agit de constater l'absence de trace de ces personnes dans les différentes sources administratives qui alimentent le répertoire, qui laisserait à penser qu'elles aient quitté le territoire.

La photographie de la liste des ménages devrait s'effectuer en fonction de deux concepts du ménage. Le premier de ces concepts, qui nous est le plus familier, renvoie aux personnes qui partagent un même logement. Le second concept, dont l'utilisation se développe, se centre plutôt sur le partage de ressources. Ce dernier permet notamment une plus grande pertinence des statistiques de revenus et l'analyse de certaines solidarités qui s'exercent au-delà du logement.

Parallèlement, le programme RESIL permettra de rendre deux services aux utilisateurs. Tout d'abord, il offrira un service d'accueil des données administratives qui permettra de transformer les données administratives brutes en une base de données statistique exploitable dans un processus de production statistique. Ce service permettra aussi de sélectionner les données d'identification qui alimenteront RESIL. Comme dans une gare de triage, ce service distinguera les variables qui intéressent les différents utilisateurs pour n'envoyer à chacun que ce dont il a besoin.

Par ailleurs, un service de production de fichiers enrichis par des appariements permettra de simplifier les appariements en fournissant les informations nécessaires à cette opération.

Un responsable de traitement qui souhaite recourir à ces services doit décrire les finalités de l'appariement qu'il souhaite mener, effectuer une étude d'impact, remplir les conditions de transparence et les autres obligations du RGPD. Il précisera alors les variables dont il a besoin sur un champ d'intérêt donné, dont les unités statistiques se rapporteraient aux individus, aux logements, ou aux ménages.

Les sources retenues pour alimenter RESIL

Pour alimenter RESIL, nous avons retenu différentes sources. Nous utiliserons le RNIPP, mais aussi des sources fiscales décrivant le foncier ou les foyers fiscaux. Nous espérerons aussi faire appel à la future source issue du processus « Gérer mes biens immobiliers » – déclarations par les propriétaires de la destination de leurs biens immobiliers et identité de l'occupant principal. Cette source pourra notamment aider à reconstituer le lien entre les logements et leurs occupants. En outre, nous souhaitons recourir à des sources sociales touchant les bénéficiaires de prestations sociales, telles que des sources touchant les prestations de la Caisse Nationale d'allocations familiales (CNAF) ou de la MSA. Nous voulons encore utiliser le RNCPS qui compile la liste des bénéficiaires ou des ayants droit de différents régimes de protection sociale. Enfin, nous devrions faire appel à la DSN et au dispositif PASRAU. A chaque fois, il s'agira de retenir uniquement les éléments d'identification ou d'adresse qui figurent dans ces fichiers. Ces sources couvrent une portée assez large.

Nous envisageons encore de recourir à d'autres sources pour renforcer la couverture de RESIL, en nous intéressant à certaines catégories de personnes. Parmi ces sources possibles, nous pourrions penser aux fichiers d'inscriptions dans l'enseignement supérieur ou scolaire, ou encore aux fichiers relatifs aux titres de séjour, pour avoir une vision des entrées sur le territoire. Là encore, il s'agirait d'utiliser uniquement des informations d'identification.

Les informations contenues dans RESIL

RESIL contiendra des identifiants et des clés d'identification qui permettront d'assurer le rôle de référentiel et d'éviter ainsi les doublons. Ces identifiants et ces clés devraient aussi aider à réaliser des appariements. Parmi ces clés, des identifiants internes qui ne sortiront jamais de RESIL seront associés aux individus et aideront à gérer la base de données du répertoire. RESIL contiendra également des CSNS pour faciliter les appariements, ainsi que des identifiants techniques des différentes sources utilisées. Pour les logements, RESIL associera encore une clé d'identification interne, qui sera différente de l'adresse. Il recueillera enfin des identifiants d'adresses issus du référentiel d'adresses que l'INSEE devrait construire.

RESIL devrait aussi contenir d'autres variables, telles que des données d'état civil, conservées dans une base distincte, ou encore des données permettant de relier les individus, les logements et les ménages. En outre, RESIL devrait intégrer un indicateur de résidence sur le territoire français reposant sur la méthode des signes de présence que j'ai décrite. Enfin, il devrait contenir différents éléments assez classiques dans la gestion de répertoires ou de bases de données, qui permettent d'assurer la traçabilité des traitements proposés : dates de mises à jour ; dates des changements de valeur pour certaines variables de RESIL, etc.

Le calendrier du projet RESIL

Nous espérons que le projet RESIL, qui se trouve encore dans sa phase initiale, puisse aboutir en 2025. Nous avons conduit une première phase exploratoire en 2020-2021 qui nous a permis de cibler des objectifs. Nous entrons à présent dans une phase de projet. Cette phase débutera par un travail important d'élaboration des textes juridiques de traitement que nous souhaitons voir publier en début 2023. Le travail d'ingénierie statistique et technique est prévu entre 2022 et 2024. Nous pourrons alors initialiser le répertoire en 2024 et déboucher enfin en 2025 sur une mise en service de RESIL et de ses offres de services.

3. Enjeux juridiques et déontologiques et principes d'action

L'encadrement juridique du projet RESIL

Le cadrage juridique de RESIL est encore en cours d'élaboration. Nous considérons que les traitements doivent être justifiés par un texte de niveau assez élevé et nous espérons donc obtenir un décret en Conseil d'Etat, soit le même niveau que celui des textes qui encadrent le NIR ou le CSNS. En outre, il est important que ce décret puisse imposer un avis de la CNIL. Dans cette optique, nous échangeons actuellement avec la CNIL dans le cadre d'une demande de conseil. Nous procédons également à une analyse d'impact relative à la protection des données, puisqu'il s'agit d'un répertoire qui se veut exhaustif sur les individus et qui permet d'effectuer différents appariements.

Concertation et information du public

En parallèle du mandat juridique conféré au programme RESIL par les textes de loi, il est important que ce programme puisse aussi bénéficier d'un « mandat social ». Les services de RESIL pourraient alors être mandatés et légitimés pour procéder à des traitements, en se fondant sur les attentes des utilisateurs et sur le sens de leurs missions, tout en veillant à maintenir la confiance qu'ils inspirent. Ce mandat social se construit, se traduit et se maintient, à l'aide d'une communication et d'une transparence sur les missions du programme. Cette transparence est particulièrement importante en matière d'appariements.

Des garanties pour la protection des données

Il est essentiel de souligner que RESIL est associé à une garantie de confidentialité, étant donné qu'il s'agit d'un répertoire à vocation statistique, couvert par le secret statistique. En outre, les traitements de RESIL et ceux qu'il permet de pratiquer reposent sur le principe de la transparence. A cet égard, nous pourrions nous inspirer des pratiques liées au CSNS. RESIL se fondera également sur le respect du principe de nécessité et de minimisation des données. En effet, il contient des données très fortement cloisonnées. Seules les informations permettant d'identifier sans ambiguïté les individus et les logements sont conservées, les données métier étant segmentées dans d'autres bulles.

Enfin, RESIL devra offrir des garanties de sécurité des données. Nous nous conformerons donc à l'état de l'art en cette matière durant les opérations de construction et lors de la mise en œuvre de ce système de répertoire. Il s'agira de cloisonner les données, de limiter leur accès, d'effectuer des classements sous haute protection, mais encore de disposer d'outils de sécurisation auxquels nous recourons largement à l'INSEE.

Echanges

Marcel GOLDBERG, INSERM

Le CSNS peut-il être utilisé par un chercheur d'EPST ou par un universitaire hors des SSM ?

Antonin FAVARO, INRAE

Le service du CSNS peut-il être mobilisé dans le cadre d'un projet de recherche, pour effectuer un appariement, par exemple avec le recensement agricole, le recensement de la population, ou encore avec d'autres enquêtes, telles que l'enquête Emploi, ou l'enquête Patrimoine ?

Thomas MERLY-ALPA, INED

Le CSNS peut-il être employé pour apparier l'enquête d'un organisme extérieur aux SSM avec d'autres données, si cette enquête dispose de l'avis d'opportunité du Comité du label du CNIS ?

Claude CASTELLUCCIA, CNIL

Dans votre exemple, les fichiers A et B disposent des correspondances entre CSNS et NIR. De ce fait, quel est l'avantage d'utiliser le CSNS au lieu du NIR ?

Laurent Piet, INRAE

Même s'il n'est pas signifiant comme le NIR, existe-t-il ou est-il prévu un équivalent du CSNS pour les numéros SIREN et SIRET ? En effet, en prenant l'exemple des enquêtes sur les entreprises agricoles, nous constatons qu'elles incluent souvent des données sur les agriculteurs eux-mêmes. Par conséquent, le SIRET de l'exploitation permet alors d'accéder à d'autres informations sur les individus, dont certaines sont très facilement accessibles sur internet.

Lionel ESPINASSE

Je précise que le CSNS est réservé aux SSM. Les chercheurs peuvent donc soit bénéficier des services offerts par le CASD, soit travailler pour un SSM. Ces différentes modalités de travail ont été explicitées par Mireille Elbaum lors de sa précédente prise de parole.

Par ailleurs, le CSNS peut être mobilisé dans le cadre d'un projet de recherche pour toute enquête, sans limitation dans la liste des sources, à partir du moment où ces sources sont d'une qualité suffisante. Cependant, elles ne peuvent être traitées que par un utilisateur habilité. Par exemple, le SSM du ministère de l'Agriculture est tout à fait autorisé à utiliser le service du CSNS pour travailler sur l'enquête du recensement agricole.

Le CSNS peut être utilisé dans le cadre du traitement d'enquêtes qui n'ont pas été produites par le service statistique public, dès lors que le responsable de son traitement est un SSM.

Il n'est pas d'actualité pour le moment de concevoir un type de service équivalent à celui du CSNS pour les numéros SIRET.

Enfin, l'avantage du CSNS par rapport au NIR réside dans le fait que le CSNS simplifie la procédure. Il permet ainsi d'éviter la demande d'un décret en conseil d'Etat pour procéder à un appariement.

Xavier TIMBEAU

Ainsi, il est actuellement difficile d'apparier deux sources sur la base du NIR.

Lionel ESPINASSE

Effectivement, cet appariement est difficile à réaliser sans passer par le CSNS.

Xavier TIMBEAU

Qu'en est-il dans le cas où je souhaiterais apparier deux bases de données, dans deux services statistiques, avec la présence de NIR dans chacune de ces deux bases ?

Lionel ESPINASSE

Pour les apparier, il faut y être éligible au sens du décret de 2019 encadrant l'utilisation du NIR. Autrement, il me semble que cette procédure serait très compliquée à réaliser.

Xavier TIMBEAU

J'en déduis que le CSNS permettrait de trouver une solution dans ce cas de figure. Néanmoins, le CSNS empêche-t-il de réaliser le travail présenté par Vladimir Passeron, où il était question de mener une identification des individus à partir de traits d'identité de manière astucieuse, avec, par exemple, des prénoms qui n'étaient pas associés à la même adresse, mais à la même date de naissance ? Le CSNS empêche-t-il également de mener le travail de la DEPP qui nous a été présenté et qui se basait sur des traits d'identité de données d'état civil de bonne qualité provenant de sources administratives ?

Lionel ESPINASSE

La réponse à cette question est ambivalente. Dans le sens où le CSNS propose une procédure robuste pour retrouver des traits d'identité et leur affecter un NIR, puis un CSNS, alors le CSNS peut constituer une bonne solution. Mais dans l'étude présentée par Vladimir Passeron, il existait notamment des variables basées sur l'adresse ; or le CSNS ne traite pas cette variable. En somme, lorsque l'on procède à des appariements sur le champ de l'état civil, la méthode du CSNS est la plus robuste. Autrement, s'il s'agit de réaliser des appariements à partir d'autres informations, telles que des adresses, le CSNS n'apporte pas d'aide.

Olivier LEFEBVRE

Et c'est dans ce dernier cas de figure que RESIL pourra apporter un service supplémentaire. En effet, RESIL permettra de mobiliser d'autres variables telles que des adresses, afin d'effectuer un pas supplémentaire dans les appariements. RESIL offrira donc des services tout à fait complémentaires par rapport à ceux fournis par le CSNS.

Xavier TIMBEAU

Pour revenir à RESIL, est-il possible d'effectuer un lien entre l'identifiant interne du RESIL et le numéro de point de livraison utilisé par les distributeurs d'énergie ?

Olivier LEFEBVRE

Cette question renvoie à notre travail mené sur les sources du répertoire des logements, qui se trouve encore tout au début de sa phase exploratoire. A ce stade, nous considérons que le socle de notre répertoire se basera plutôt sur le répertoire des locaux de la DGFIP. Il se pose effectivement la question de la mobilisation des données des fournisseurs d'énergie et nous devons vérifier si la mobilisation de cette source pourra renforcer la couverture ou la robustesse de notre système. Cette question reste ouverte.

Xavier TIMBEAU

Je précise que mon interrogation portait davantage sur des questions d'appariements que sur celles liées à la couverture.

Mireille ELBAUM

Des parlementaires pourront-ils demander une exploitation de RESIL pour trouver le taux de bénéficiaires de prestations sociales qui ne résideraient pas à l'adresse indiquée et qui ne vivraient donc pas dans le cadre d'un ménage qui leur ouvrirait les droits à ces prestations ?

Xavier TIMBEAU

Vous relevez cette obsession de la fraude que peuvent avoir certains parlementaires.

Olivier LEFEBVRE

Il est effectivement possible que des parlementaires aient l'idée de recourir à RESIL dans ce sens. Néanmoins, RESIL étant à vocation statistique, nous ne pourrons en extraire aucune caractérisation individuelle se rapportant à tel ménage ou à tel fraudeur présumé. Par ailleurs, RESIL se fondera sur des considérations statistiques relatives à la construction des populations résidentes et des ménages pour définir les adresses principales. Ainsi, le fait qu'une adresse statistique diffère de celle d'un fichier donné ne renvoie pas nécessairement à une fraude. Il faut encore tenir compte de considérations techniques, mais la réponse à votre question reste difficile.

Xavier TIMBEAU

Si je comprends bien, l'identifiant de RESIL demeure interne au service et personne ne connaît son numéro dans RESIL.

Olivier LEFEBVRE

Effectivement, personne ne connaîtra son numéro dans RESIL.

Xavier TIMBEAU

Et a priori, le dispositif légal garantit que ce numéro demeure au sein de RESIL.

Olivier LEFEBVRE

Le numéro RESIL n'est utilisé que pour la gestion de notre base de données et ne sortira jamais de RESIL. Sur le plan de la segmentation des droits d'accès au répertoire, je précise que les personnes qui auront accès à la table dans laquelle figureront les identifiants seront très peu nombreuses.

Xavier TIMBEAU

Pour répondre à la question de Mireille Elbaum, il sera tout de même possible de demander une étude statistique sur la fraude, en mobilisant un appariement de fichiers avec les informations de RESIL. Le service de RESIL pourra alors recevoir des fichiers à apparier et les renvoyer ensuite, mais sans les numéros d'identification. Cette étude statistique pourra alors mettre en évidence des estimations de fraude basées sur des différences d'adresses de résidence effective et d'adresses déclarées aux organismes de protection sociale.

Olivier LEFEBVRE

S'il s'agit d'une finalité d'étude statistique, que cette finalité est légitime et que le responsable du traitement déclare son traitement, ses finalités et les sources mobilisées, le service de RESIL pourra répondre à une sollicitation, dans le respect du RGPD.

Yvon SERIEYX, UNAF

Des concertations avec la société civile et avec le monde universitaire sont-elles prévues à une étape ou à différentes étapes de l'élaboration de RESIL ?

Un invité à distance

Quel lien est-il fait entre le programme RESIL et les travaux sur l'identifiant unique des logements réalisé par l'INSEE en lien avec la DGFIP ?

Olivier LEFEBVRE

Pour répondre à la question tournant autour de la concertation, d'une part, nous avons jugé essentiel de bien communiquer sur la construction de notre système de répertoires. Il s'agit de présenter les différents services possibles que pourra offrir le programme RESIL. D'autre part, notre mandat social qui s'articule autour des besoins des utilisateurs implique de mener une concertation continue. Ainsi, durant les quatre ans qui viennent, nous espérons que le CNIS nous invitera régulièrement, durant les différentes étapes importantes du projet, pour mieux répondre aux attentes des utilisateurs. Quant à savoir s'il serait nécessaire de mener une concertation spécifique avec le monde universitaire, il reste à déterminer à quel moment cette concertation serait utile et quel objectif elle permettrait d'atteindre.

Pour ce qui est du lien entre RESIL et les projets d'identifiants des locaux, en première intention, le répertoire de logement RESIL devrait mobiliser le fichier des locaux d'habitation géré par la DGFIP. S'il émergeait, au niveau inter-administratif, un projet d'identifiant dédié aux logements, nous devrions saisir cette opportunité pour alimenter notre répertoire des logements. Nous sommes donc particulièrement attentifs aux évolutions des projets en cours, sur lesquels nous pourrions nous appuyer, le programme RESIL étant en construction.

Xavier TIMBEAU

Ainsi, cet identifiant unique des logements se retrouverait a minima dans le répertoire des logements de RESIL.

Olivier LEFEBVRE

Si un tel identifiant était créé au niveau inter-administratif, il serait important de pouvoir le retrouver dans RESIL, étant donné qu'il faciliterait les appariements de sources relatives aux logements.

François GUILLAUMAT-TAILLIET, Adjoint au Secrétaire général du CNIS

En cette fin de matinée, je remercie les orateurs, le public, ainsi que les interprètes et les techniciens. Je relève que les invités témoignent de l'excellence de la traduction qui leur est proposée.

TABLE RONDE – QUELS APPARIEMENTS POUR QUELS USAGES ?

Président de la table ronde : Philomé Robert, Journaliste et Présentateur, France 24 ;

Francesco Avvisati, directeur du programme Innovation, données et expérimentations en éducation (IDEE), J-PAL Europe/PSE – Ecole d'économie de Paris ;

John Dunne, directeur du Centre des données administratives, Office national de statistique d'Irlande (CSO) ;

John Martin, président du Comité consultatif du gouvernement irlandais sur le marché du travail, IZA *Research Fellow*, ancien directeur de l'emploi, du travail et des affaires sociales de l'Organisation de coopération et de développement économiques (OCDE) ;

Jean-Noël Barrot, député des Yvelines, Vice-Président de la Commission des Finances.

Philomé ROBERT

Bonsoir à tous et merci de continuer à suivre cette rencontre. Tout d'abord, je souhaite remercier Cristina D'Alessandro et Françoise Dupont de m'avoir fait l'honneur de me proposer d'animer cette table ronde. Je dois vous dire qu'avant qu'elles m'aient sollicité, je ne connaissais pas l'existence des appariements. J'ai donc pu me renseigner sur ce sujet que je ne maîtrise pas. Je pourrais néanmoins poser quelques questions aux invités, que j'écouterai religieusement.

Francesco Avvisati, vous êtes directeur du programme Innovation, données et expérimentation en éducation (IDEE) au bureau J-PAL Europe basé à l'Ecole d'économie de Paris, que vous avez rejoint en 2021. Vous avez travaillé durant onze ans à l'OCDE, principalement autour du Programme international pour le suivi des acquis des élèves (PISA). Vous êtes docteur en économie, ancien élève de l'Ecole normale supérieure (ENS).

John Dunne, vous êtes statisticien au bureau central irlandais des statistiques (CSO). Vous y avez travaillé dès l'obtention de votre diplôme universitaire de premier cycle. Depuis début 2020, vous avez pris la tête d'une direction qui explore deux frontières de la donnée sur

lesquelles nous reviendrons. Vous avez créé et dirigé le Centre pour les données administratives (ADC). Vous avez obtenu un doctorat à l'université de Southampton. Par ailleurs, vous avez notamment développé des méthodes d'estimation de la population à partir de données administratives.

John Martin, vous êtes le président du Comité consultatif du gouvernement irlandais sur le marché du travail et vous êtes membre du Conseil national des statistiques de votre pays. Vous avez été directeur de l'Emploi, du travail et des affaires sociales à l'OCDE. Vous avez publié dans des revues spécialisées de nombreux articles et vous avez écrit plusieurs livres dans votre domaine de travail.

Vous parlerez tous deux en anglais, mais vous pourrez être compris et entendu par tous grâce à un service de traduction.

Jean-Noël Barrot, vous êtes député de la seconde circonscription des Yvelines, vice-président de la Commission des finances, secrétaire général du Modem, mais aussi enseignant. Vous avez travaillé au *Massachusetts Institut of Technology* (MIT) et vous vous intéressez fortement aux questions liées à l'évaluation des politiques publiques.

Je passe d'emblée la parole à Francesco Avvisati dans le cadre de notre table ronde.

Quels sont les apports des appariements pour les politiques publiques de l'éducation ?

Francesco AVVISATI

Le monde de l'éducation a déjà été évoqué plusieurs fois ce matin et il est clair que les apports des appariements y sont indéniables. Ainsi, InserJeunes donne lieu à la construction d'indicateurs qui permettraient de mieux comprendre la phase de transition entre les univers scolaires et professionnels.

Dès lors que nous nous situons sur une telle problématique de transition et que nous quittons le strict champ de l'éducation, nous avons besoin d'effectuer des appariements pour pouvoir mettre en évidence des informations pertinentes, qui permettent de piloter une politique.

Au cours de cette journée, nous avons également cité des panels et en particulier des panels sectoriels. L'utilisation de ces panels permet d'évaluer les politiques publiques sur un long terme, ce qui est particulièrement important en matière d'éducation. Il faut pouvoir disposer d'un grand recul et suivre des trajectoires sur un très long terme. En effet, un grand nombre de politiques éducatives visent des résultats sur de très longues échéances.

De plus, les objectifs des politiques éducatives dépassent largement la sphère du système scolaire. Effectivement, l'école est un levier particulièrement important des politiques redistributives et elle constitue un vecteur de mobilité sociale. Il est possible de voir dans l'école un remède contre les inégalités, celle-ci pouvant détruire les déterminismes sociaux.

Pour prendre des exemples concrets, dans l'objectif de favoriser l'égalité des chances et d'améliorer les perspectives des personnes issues de milieux défavorisés, nous pouvons souhaiter vérifier la pertinence d'une réduction de la taille des classes au sein de l'ensemble de l'école primaire ou plus spécifiquement au sein d'établissements situés dans des zones

défavorisées. Nous pouvons encore nous demander s'il est intéressant de mener des politiques tardives et ciblées, telles que l'attribution de bourses individuelles pour l'accès aux études supérieures, ou s'il est préférable d'intervenir plus précocement en investissant dans la scolarité à deux ans. Or il est difficile de comparer les effets de ces différentes politiques à partir de données immédiates.

Dans le monde de l'éducation, il existe un consensus sur l'intérêt d'une intervention précoce, au cours des premières années de l'école primaire. Cette intervention serait la plus efficace et pourrait avoir des effets à très long terme, aidant à éviter que les inégalités sociales se perpétuent entre les générations.

Un article de référence, écrit par des chercheurs suédois et hollandais, intitulé « Effet de long terme de la taille des classes » a été publié en 2013 dans le *Quarterly Journal of Economics*. Cet article a suivi des élèves suédois nés avant 1982, ceux-ci ayant entre 27 et 42 ans lors de l'étude. Il était possible de comparer les salaires de ces anciens élèves.

Les chercheurs se sont basés sur des districts de la carte scolaire des écoles publiques qui étaient plutôt similaires. La taille des classes était le fruit du hasard, étant donné qu'il existait une règle qui fixait l'ouverture d'une seconde classe lorsqu'une école regroupait 30 élèves, puis d'une troisième classe, lorsque l'effectif atteignait 60 élèves. Les chercheurs ont donc identifié des écoles de 29 ou 59 élèves ou de 31 ou 61 élèves, de manière à distinguer des classes de tailles très différentes.

Il a ainsi été possible de mesurer l'effet propre de la taille des classes, en comparant les résultats scolaires. Et surtout, les chercheurs ont enrichi les données avec des informations sur les salaires de ces anciens élèves qui avaient atteint l'âge adulte.

Cette étude, comme de nombreuses autres études, a pu démontrer que l'exposition à des petites classes renforçait les résultats scolaires, cet effet ayant cependant tendance à s'estomper au cours de la scolarité.

Mais surtout, les auteurs de cet article ont pu montrer que les effets de la taille des classes sur les résultats scolaires persistaient sur le marché du travail. En effet, les anciens élèves qui ont été exposés aux classes les plus petites sont ceux qui bénéficient des salaires les plus élevés et qui ont le moins recours aux prestations sociales.

Donc, en appariant des données d'enquêtes avec des registres de l'éducation, puis en appariant ces derniers avec des registres d'impôts, il a été possible de démontrer que ces politiques touchant la taille des classes pouvaient être efficaces dans un contexte donné. Cependant, pour tirer des conséquences plus générales, il reste encore à vérifier si la situation de la Suède d'avant 1982 correspond bien à la situation à laquelle on souhaite appliquer une réduction de la taille des classes.

Finalement, très peu d'études ont pu être menées pour prouver le consensus qui tournait autour de la taille des classes. Ainsi, il faut souligner que les appariements menés sur des données récoltées sur un long terme constituent un apport essentiel.

Cette temporalité qui renvoie à des périodes d'une vingtaine d'années rend difficile le lien entre ce type d'appariements et le pilotage des politiques publiques. En revanche, ces appariements permettent de forger des consensus très influents qui aident à imaginer les politiques futures.

Philomé ROBERT

Quelles sont les limites que vous rencontrez pour réaliser ces appariements ?

Francesco AVVISATI

En France, des études ont déjà montré les effets de la taille des classes sur les résultats scolaires, en se basant sur des seuils d'ouverture de classes et sur des panels. Néanmoins, le système d'information permettant de suivre les trajectoires scolaires des enfants n'a été créé que tardivement en France. L'INE n'existe qu'à partir de 2002 et il touche dans un premier temps uniquement le secondaire et le supérieur. L'extension de l'INE à la sphère de l'école primaire n'a eu lieu qu'en 2017. Nous ne pouvons donc pas mobiliser de données administratives pour étudier des trajectoires allant du primaire à la vie active.

De surcroît, ce type d'appariements basés sur des fichiers administratifs engendre des difficultés techniques d'appariements. Comme nous l'avons vu lors de la seconde session de cette rencontre avec InserJeunes, l'appariement entre des données des sphères de l'éducation et de l'emploi demande le recours d'identifiants différents qui ne peuvent être appariés que par le croisement de traits d'identité. Or cette difficulté s'avérerait d'autant plus importante pour apparier des données qui ont vingt ans d'écart, notamment à cause des changements de noms, ou de situations familiales. Il faudrait donc s'attendre à un appariement moins précis que celui réalisé en Suède.

Philomé ROBERT

Je vous remercie. Je rappelle qu'un dispositif est mis en place pour pouvoir récolter vos questions, qui pourront nourrir un temps d'échange. En attendant, rendons-nous du côté de l'Irlande, afin de retrouver John Dunne.

Quelles sont les raisons qui ont amené l'Irlande à se doter d'un répertoire ?

John DUNNE

Pour répondre à la question, l'Irlande n'a pas de registres de population tels qu'on les imagine en Europe. Ces registres sont généralement utilisés par les autorités locales pour gérer les services publics et les gens s'enregistrent là où ils vivent. En revanche, l'Irlande dispose d'un numéro d'enregistrement officiel qui est utilisé dans le cadre de l'authentification et de l'identification dans ses systèmes d'administration publique. Ce numéro est attribué à un enfant à la naissance, lorsque le parent commence à bénéficier du versement des allocations familiales universelles. Le numéro est ensuite utilisé par cette personne dans ses engagements ou démarches auprès de l'État tout au long de sa vie. Ces transactions ont lieu du berceau à la tombe, car la personne est confrontée à l'éducation, les impôts, la sécurité sociale, la santé, la retraite et enfin le décès. En réalité, il est très difficile de vivre dans un État démocratique moderne sans s'engager auprès des autorités publiques.

Depuis longtemps, le CSO a reconnu le potentiel statistique inexploité contenu dans les systèmes d'administration publique. La loi irlandaise sur les statistiques, promulguée en 1993, prévoit l'utilisation de ces données administratives à des fins statistiques. Le CSO a accédé à ces données administratives depuis 1993 et probablement même avant. En 2009, le CSO a centralisé son infrastructure pour collecter, trier et traiter ces données afin de les

rendre sûres et disponibles à des fins statistiques. Le Centre de données administratives ou CDA joue le rôle de centre d'échange, en veillant à ce que des procédures appropriées de gouvernance des données soient en place. Par exemple, l'analyse statistique irlandaise n'est effectuée que sur des données pseudonymisées, tous les identifiants étant remplacés par des clés d'identification protectrices, ou PIK. Une clé d'identification protectrice permet d'établir des liens sûrs sur cette clé à tout moment entre les sources de données, garantissant ainsi la protection de l'identité originale.

Un élément clé de la réforme du secteur public en Irlande est l'utilisation ou le déploiement de numéros d'identification officiels pour les propriétés, les entreprises et les achats dans les systèmes de l'administration publique. L'accent mis sur des données plus intelligentes a également permis de renforcer les capacités d'analyse des données, non seulement pour le CSO, mais aussi au sein des organismes du secteur public eux-mêmes. La finalité première de l'utilisation des données administratives est de fournir des statistiques de qualité pour éclairer la prise de décision à un coût moindre. Bien entendu, l'un des principaux objectifs est de pouvoir sensibiliser la population chaque année à faible coût, à l'instar de ce qui se fait aux Pays-Bas et dans les pays nordiques depuis de nombreuses années. Des objectifs similaires sont poursuivis dans l'ensemble de l'UE, car les États membres devront à l'avenir produire des estimations annuelles de type recensement. C'est là où l'Irlande en est actuellement avec l'utilisation de données administratives.

Philomé ROBERT

A quel stade se trouve le projet que vous décrivez et quel bénéfice en avez-vous retiré pour le moment ?

John DUNNE

En dehors des capacités analytiques, depuis sa création, le CDA a également eu une réaction rapide, en particulier en temps de crise. En 2009, il y a eu la crise financière à laquelle il a pu répondre. Plus récemment, c'est la pandémie qui a permis d'informer la réponse du gouvernement. L'utilisation de données administratives permet d'obtenir des résultats statistiques très détaillés et permettent une analyse longitudinale des cohortes de la population. Par exemple, il est possible de réaliser une analyse statistique sur les lieux d'emploi des diplômés, x années après l'obtention de leur diplôme, sans avoir à réaliser l'enquête. Ce type de projets se déroule dans notre île.

L'Irlande dispose désormais de ces capacités, et elle travaille actuellement au déploiement d'un nouveau code postal, qui a été lancé en 2015 sur notre système d'administration publique. Ce système de code postal est assez nouveau et différent des autres codes postaux, car il est associé à la boîte aux lettres plutôt qu'à un groupe de maisons dans une zone spécifique. Par exemple, si on donne son code postal à quelqu'un, celui-ci peut l'entrer dans Google Maps et obtenir l'itinéraire directement vers sa maison. Le déploiement de ce code postal a considérablement amélioré les capacités de l'Irlande en matière de statistiques géographiques détaillées. En général, les habitants des codes postaux plus anciens ont ce code postal universellement utilisé par les services de livraison, quel que soit l'objet de la livraison.

En ce qui concerne la combinaison d'estimations démographiques de type recensement, le CSO a entrepris des recherches sur la manière dont il peut composer de telles estimations de manière robuste. Les résultats obtenus à ce jour suggèrent que le CSO peut être

optimiste quant à la mise en œuvre d'un nouveau système robuste d'estimations annuelles de la population à l'avenir.

Pour finir, nous trouvons continuellement de nouvelles opportunités pour fournir des informations statistiques améliorées dans divers domaines.

Philomé ROBERT

Je vous remercie, veuillez rester avec nous, le public aura des questions pour vous. En attendant, nous donnons la parole à John Martin.

Quels sont les apports des appariements pour les politiques publiques de l'emploi et quelles pratiques d'appariements avez-vous relevé au niveau international ?

John MARTIN

En ce qui concerne la valeur ajoutée du couplage statistique des données administratives pour le marché du travail et les politiques sociales, il a eu un impact énorme dans le domaine de l'évaluation des impacts dans ce domaine. Cela a été amplement démontré ces dernières années par les méta-analyses à grande échelle de l'impact de ces politiques réalisées par le dernier prix Nobel d'économie, David Card, et ses co-auteurs Jochen Kluve et Andrea Weber. Bon nombre des évaluations incluses dans leurs articles couvrent un large échantillon de pays et s'appuient sur le croisement de données administratives comme élément de base essentiel.

Il illustre cela en montrant comment la capacité de relier des données administratives provenant de différentes sources a eu un impact énorme sur l'évaluation du marché du travail et des politiques sociales dans son propre pays, l'Irlande. Jusqu'à la création de ce que l'on appelle la base de données longitudinale des demandeurs d'emploi (Jobseekers Longitudinal Database, JLD) par le ministère irlandais de la protection sociale en 2013-2014, il y avait essentiellement peu d'antécédents de mise en relation de sources de données administratives en Irlande et d'exploitation de celles-ci pour évaluer les politiques publiques. La base de données longitudinale des demandeurs d'emploi a été développée par des analystes du département de la protection sociale, en étroite collaboration avec des statisticiens détachés du Central Statistics Office d'Irlande, le bureau central des statistiques. Elle suit les demandes de prestations, l'emploi, la formation et la participation aux politiques de l'emploi de tous les demandeurs d'emploi et parents uniques qui ont fait une demande de prestations depuis l'année 2004. Le principal identifiant de ces personnes est ce que l'on appelle en Irlande le numéro personnel de service public, qui, comme l'a souligné John DUNNE dans sa présentation, est nécessaire à tous les individus pour accéder aux services publics en Irlande. Pour protéger la vie privée, toutes les données individuelles sont pseudonymisées et cryptées et l'accès à la base de données longitudinale des demandeurs d'emploi qui en résulte est strictement contrôlé par le ministère de la protection sociale.

La base de données longitudinale des demandeurs d'emploi rassemble des données provenant de registres tenus par le ministère de la protection sociale, par le fisc irlandais (Revenue) et par SOLAS, l'agence irlandaise pour l'éducation et la formation. La base de données qui en résulte contient plusieurs millions d'épisodes individuels de demandes de prestations, de statut professionnel, d'éducation et de formation, de participation à des programmes du marché du travail, ainsi que des données sur les revenus et les paiements

d'impôts. L'une des raisons pour lesquelles les autorités irlandaises ont investi dans le développement de cette base de données longitudinale sur les demandeurs d'emploi était de l'utiliser comme un élément clé pour évaluer rigoureusement l'ensemble des programmes du marché du travail, d'éducation et de formation qui existent en Irlande. Au cours des cinq dernières années, le ministère de la protection sociale a passé des contrats avec des consultants externes, des instituts de recherche et des universitaires, afin de produire des évaluations d'impact de nombreux programmes de ce type. Le personnel du ministère de la protection sociale entreprend également des évaluations en utilisant la base de données longitudinale des demandeurs d'emploi. Les évaluations qui en résultent sont toutes publiées après un examen rigoureux par des pairs.

Les évaluations d'impact réalisées à l'aide de méthodes économétriques de pointe ont mis en évidence les programmes qui ont atteint les objectifs fixés et ceux qui n'ont pas fonctionné. Les preuves produites par les évaluations utilisant la base de données longitudinale des demandeurs d'emploi ont considérablement enrichi la base de connaissances sur la conception et la mise en œuvre des politiques de l'emploi en Irlande et ont entraîné des changements politiques significatifs en faveur de ces programmes. Le prochain programme à être évalué cette année est le programme d'emploi communautaire, qui est de loin la plus grande politique active du marché du travail en Irlande, tant par le nombre de participants annuels que par le montant des dépenses publiques. Cette évaluation est entreprise par ses anciens collègues de l'OCDE et du Centre commun de recherche de la Commission européenne.

Les données longitudinales sur les demandeurs d'emploi ont fait leurs preuves, mais elles présentent certaines limites qu'il convient de mentionner. Tout d'abord, l'horizon temporel pour l'évaluation des résultats post-programme est généralement assez court, de six à douze mois. Idéalement, le chercheur souhaiterait disposer d'une fenêtre temporelle allant jusqu'à cinq ans ou plus après le programme afin d'évaluer les impacts à long terme du programme. Une caractéristique irlandaise spécifique rend plus difficile le suivi des résultats à long terme, à savoir la longue tradition d'émigration de l'Irlande, en particulier parmi les cohortes d'âge plus jeunes. Deuxièmement, les données longitudinales sur les demandeurs d'emploi manquent actuellement de données sur le niveau d'éducation et les compétences, qui constituent l'une des principales variables observables et qui, selon toutes les recherches, sont fortement corrélées aux résultats sur le marché du travail. Au lieu de cela, les chercheurs doivent se contenter des professions antérieures du demandeur d'emploi, qui constituent un substitut plutôt médiocre du niveau d'éducation et des compétences. Néanmoins, le ministère de la Protection sociale est actuellement engagé dans l'extension de la base de données pour inclure un plus large éventail de prestations sociales et de transitions sur le marché du travail en vue de rendre la base de données plus utile. Comme l'a souligné John DUNNE, l'Irlande investit également dans la liaison d'autres ensembles de données administratives, preuve qu'il s'agit d'une priorité élevée pour les statistiques officielles à l'avenir.

La deuxième question posée était d'examiner la possibilité de relier les données administratives pour évaluer les impacts des politiques de l'emploi dans une optique comparative. Bien entendu, les pays nordiques, en particulier le Danemark, la Norvège, moins la Suède, et la Finlande, ont été des pionniers dans ce domaine. Leur rôle prépondérant s'explique par leurs registres de population dotés d'identifiants personnels uniques, qui permettent aux chercheurs de suivre les mêmes individus pendant de longues périodes, alors qu'ils transitent entre différents états du marché du travail et interagissent avec un large éventail d'organismes publics. Les chercheurs de ces pays ont ensuite

appliqué diverses méthodes économétriques, expérimentales et quasi-expérimentales, pour évaluer ex post l'impact de la participation à divers programmes du marché du travail et/ou du recours à différentes prestations sociales, sur un large éventail de résultats. Il ne s'agit pas seulement des résultats sur le marché du travail, mais aussi des revenus et des perspectives de carrière, de la pauvreté, de la santé, de la formation des familles et de la retraite. Les chercheurs d'Amérique du Nord, aux États-Unis et au Canada, utilisent depuis longtemps des données administratives croisées pour analyser l'impact des systèmes d'assurance chômage et d'aide sociale sur le marché du travail. Il convient d'ajouter que les chercheurs de ces deux pays ont également utilisé les flux bruts sous-jacents à leurs enquêtes régulières auprès des ménages, pour aborder ces questions et d'autres relatives au marché du travail.

Si l'on laisse de côté les pays nordiques, dépourvus de registres de population et gênés par des préoccupations liées à une sécurité jusqu'alors menacée, les autres pays européens ont été plus lents à percevoir le potentiel de la liaison des données administratives à des fins de recherche. Toutefois, la situation a considérablement évolué ces dernières années, des pays comme l'Allemagne, les Pays-Bas et la Suisse ayant fait de grands progrès dans ce domaine. Dans le cas de l'Allemagne, les réformes Hartz de 2003 à 2005 ont imposé aux autorités d'évaluer le marché du travail et les programmes sociaux introduits par les réformes. Pour ce faire, elle a donné une impulsion majeure pour relier les sources de données administratives afin de faciliter ce processus d'évaluation. Elle permet également certaines innovations, comme le fait que l'Allemagne et la Suisse soient en mesure d'établir un lien entre les demandeurs d'assurance-emploi actuels et les travailleurs sociaux qui traitent leurs dossiers. Cela permet de déterminer si les travailleurs sociaux et les différentes approches qu'ils adoptent à l'égard de leurs clients font réellement une différence dans les résultats obtenus après la participation. Les nations d'Europe du Sud ont peut-être été un peu plus lentes à reconnaître le potentiel de la mise en relation des données administratives à des fins de recherche sur le marché du travail, peut-être en raison d'un passé et d'une préoccupation concernant les questions de confidentialité des données. Cependant, ils sont maintenant beaucoup plus actifs dans ce domaine, comme le montre clairement ce matin le cas de la France. En revanche, les pays d'Europe centrale et orientale et la Grèce sont à la traîne et passent clairement à côté des possibilités offertes par la mise en relation de données administratives provenant de sources différentes.

Pour conclure, la mise en relation des données administratives afin d'évaluer les résultats des politiques sociales et du marché du travail a clairement démontré leur valeur et de plus en plus de pays l'ont reconnu.

Philomé ROBERT

Je vous remercie pour cet éclairage. Vous nous avez décrit l’apport des appariements pour les politiques publiques de l’emploi en Irlande avant de nous faire voyager au Canada, aux Etats-Unis, ou encore en Scandinavie. Néanmoins, vous semblez faire état de certains retards français en la matière. Je me tourne donc du côté de Jean-Noël Barrot, qui est à la fois un élu et un économiste.

Quelle est la valeur ajoutée des appariements de données pour la préparation et l’évaluation des politiques publiques ?

Jean-Noël BARROT

Il est notoire que depuis quelques années et en particulier depuis la crise du coronavirus, la France a franchi un pas en avant important, grâce à la mobilisation de la statistique publique, vers un meilleur abord des politiques publiques, en amont et en aval de leurs conceptions. En effet, nous avons pu compter sur des données chiffrées permettant d'examiner les sujets que nous traitons de manière très sérieuse et rigoureuse. Cette avancée a été rendue possible par la mise à disposition croissante de données.

En ce qui concerne les sujets qui m'intéressent plus particulièrement et notamment ceux qui ont trait à l'entreprise, je pourrais citer la mise à disposition par la banque publique d'investissement (BPI) d'un certain nombre de ses données d'intervention, qui s'est opérée avant la crise sanitaire. Avant cela, la France souffrait d'un certain retard, puisque la *Small Business Administration* – agence indépendante du gouvernement américain dédiée aux petites entreprises – mettait à disposition des chercheurs ses données d'intervention depuis bien plus longtemps.

A la suite des premières lois de finances du quinquennat actuel, qui ont profondément modifié la fiscalité du capital, nous avons assisté à l'ouverture aux chercheurs d'un certain nombre de panels, tels que le fichier des déclarations de revenus fiscaux (POTE). Cette ouverture devait permettre une évaluation de ces réformes fiscales, et en particulier la mise en place du dispositif de prélèvement forfaitaire unique, ainsi que la suppression de l'impôt sur la fortune (ISF) remplacé par l'impôt sur la fortune immobilière (IFI).

Cette ouverture a bénéficié à l'évaluation de la politique publique en question, mais aussi à la communauté scientifique, puisque les chercheurs peuvent accéder à ces bases de données, tout en respectant les procédures du Comité du secret statistique.

De plus, lors de la crise du coronavirus, grâce à l'INSEE, la France a probablement su avant les autres pays comment mobiliser des données pour comprendre les phénomènes en cours. Elle s'est appuyée sur du *nowcasting* (prévision immédiate), en mobilisant des données qui n'étaient pas habituellement employées par les instituts statistiques nationaux. En outre, la France a mis en place en un temps record, dès le mois d'avril 2020, l'enquête « flash » Activité et conditions d'emploi de la main-d'œuvre (ACEMO) qui a permis de suivre mensuellement les effets de cette crise sur les entreprises.

De la sorte, un effort important de mise à disposition de données a été mené pour aider à évaluer les politiques publiques.

Philomé ROBERT

En tant qu'économiste, quels besoins identifiez-vous pour la recherche, afin qu'elle puisse offrir les informations nécessaires au pilotage de l'action publique ?

Jean-Noël BARROT

Pour vous répondre, je vais devoir faire appel au chercheur qui est en moi, de manière à voir comment il pourrait aider le député que je suis. Tout d'abord, j'estime que les chercheurs sont particulièrement bien traités en France. En effet, ayant effectué des recherches aux Etats-Unis, j'ai pu constater qu'il était particulièrement difficile d'accéder aux données administratives américaines et de les apparier.

En effet, pour effectuer mes recherches outre-Atlantique, je devais me rendre à Cambridge (Massachusetts), dans les locaux du *National Bureau of Economic Research* (NBER) du *Boston Research Data Center* (BRDC). Il fallait que je badge à plusieurs reprises avec ma carte, avant de pouvoir me rendre dans une salle sans fenêtres, qui bénéficiait d'un minimum de ventilation. Dans cette salle, je ne pouvais pas utiliser mon propre ordinateur. En outre, cet accès était associé à un coût important. Certains chercheurs habitaient littéralement dans ce petit local du NBER, mais je ne pouvais pas y passer toutes mes journées. De ce fait, la recherche semblait faire des progrès endogènes, portée par une catégorie de chercheurs spécialistes qui restaient enfermés sur des données particulières, tandis que d'autres ne travaillent pas du tout sur ces données, les recherches transversales n'étant pas favorisées.

En France, le CASD a permis d'apparier des données à distance, dans un cadre sécurisé et agréable. Le CASD a tout autant pensé à la sécurisation des données, qu'à l'expérience de l'utilisateur. Le CASD a ainsi permis aux chercheurs de demeurer dans leurs bureaux et de basculer au sein de leurs mêmes écrans sur des données publiques ou privées. Ils peuvent travailler sur des données administratives à partir de leurs propres outils de travail. A ce niveau, nous sommes particulièrement gâtés en France.

Je précise que nous disposons d'un socle de données administratives que le monde entier peut nous envier, tout du moins en ce qui concerne les entreprises. Il s'agit là de la contrepartie de notre fort encadrement administratif en matière fiscale. D'ailleurs, la qualité des données concernant les entreprises est si bonne, que de nombreuses études de référence concernant les entreprises sont construites à partir de données françaises.

Ces dernières années, nous avons eu une mise à disposition très importante de données nouvelles pour les chercheurs, qui permettent de répondre à des questions sur lesquelles nous ne pouvions pas répondre auparavant. Et l'appariement des bases de données administratives ou associatives est permis par le CASD, un outil que je trouve plus performant que celui dont disposent par exemple les universités américaines, étant donné qu'il est possible de travailler dans un cadre sécurisé depuis son bureau. Cet outil formidable permet à la France de bénéficier, pour le pilotage de ses politiques publiques, d'un éclairage de la recherche qui devient de plus en plus précis, tant sur la nature que sur la temporalité des données.

Ainsi, une nouvelle habitude a pu s'établir grâce à l'ouverture aux chercheurs d'un certain nombre de bases de données. Lors des discussions du budget à l'Assemblée nationale, l'Ecole d'économie de Paris organise chaque année une analyse très détaillée de l'effet sur les entreprises et les ménages du projet de budget du gouvernement. Cet exercice est mené en quelques jours, voire en une à deux semaines, le projet de budget n'étant pas connu à l'avance. Ces simulations sont permises grâce aux facilités permises par la statistique française, à savoir la disponibilité des données et la rapidité des appariements.

Cette analyse donne alors lieu à un débat toujours inconfortable pour les députés de la majorité. Tout juste après que nous sommes sortis de la Commission des finances, l'Ecole d'économie de Paris nous communique le véritable impact du budget avec un niveau de détail stupéfiant. C'est alors que les députés de l'opposition s'appuient sur cette analyse pour renforcer leurs arguments et critiquer notre budget, alors que nous n'avions pas pu anticiper certains de ses effets.

De plus, l'année dernière, certains résultats de l'Ecole d'économie de Paris différaient des simulations opérées par le gouvernement dans le cadre de l'élaboration du budget. Par conséquent, les économistes de cette école ont pu ensuite prendre le temps d'expliquer d'où pouvaient provenir ces différences d'analyses.

A partir de là, j'estime que nous entrons dans un cercle vertueux qui s'inscrit autour de la présentation du budget. Cette dynamique est centrée autour d'un enrichissement mutuel ou d'un contrôle, rendu possible par l'ouverture des données aux chercheurs. Je pense que cette démarche va dans le bon sens et que nous sommes plutôt bien lotis en France.

Echanges

Philomé ROBERT

Merci, monsieur le député et merci à vous quatre. Passons à présent à la séquence des échanges, en recueillant les questions en provenance de la salle et du tchat.

Yvon SERIEYX, UNAF

Dans le cadre de l'enquête PISA, est-il possible d'utiliser dans certains pays des appariements pour remplacer le questionnaire "parents" – qui est, je crois, facultatif – afin de recueillir les informations sociodémographiques des familles des élèves ?

Francesco AVVISATI

Les enquêtes internationales ne sont pas si simples que cela, étant donné qu'il s'agit d'obtenir des informations de même qualité et de même type, dans tous les pays. Pour l'instant, il n'est donc pas question d'imposer le recueil de données administratives pour remplacer des questionnaires, tout simplement parce que tous les pays qui participent à ces enquêtes ne sont pas en mesure d'effectuer cette tâche.

En revanche, une enquête internationale se base sur des données d'enquêtes nationales, composées d'un volet national et international. Les données de ce second volet sont transmises à l'organisme chargé de superviser l'enquête internationale. Or le tirage de ces enquêtes s'opère dans de nombreux pays à partir d'identifiants d'élèves ou d'établissements. Il est alors possible d'enrichir les enquêtes internationales avec de nouvelles données. Au Chili, dans le cadre d'études, des chercheurs ont mobilisé des enquêtes PISA et des bases de données administratives grâce aux identifiants d'élèves, opérant ainsi des suivis longitudinaux.

Finalement, au niveau international, il semble beaucoup plus réaliste de remplacer par des données administratives non pas les questionnaires touchant les informations sociodémographiques des familles, mais plutôt ceux qui concernent les établissements.

Roxane SILBERMAN, CNRS

Le CSO réalise-t-il des appariements pour des projets de chercheurs ?

John DUNNE

Pour l'OSC d'Irlande, le principal moteur de la production de ces données croisées était le potentiel des informations statistiques pour améliorer la prise de décision. On a également

reconnu le potentiel de recherche de ces données et l'intention de permettre aux chercheurs d'y accéder à des fins de recherche, ce qui a été fait dans une certaine mesure. Cependant, ils ont adopté une approche prudente au fur et à mesure qu'ils développaient leur propre compréhension de ce qu'était une recherche sûre et appropriée. Ils voulaient s'assurer que leur approche était sûre et, dans cette optique, ils ont utilisé ce que l'on appelle un cadre en cinq étapes pour réfléchir à chaque étape du développement de l'infrastructure à venir. Le cadre en cinq étapes était un cadre ICE qui configure des projets sûrs. Il prévoit un cadre sûr pour la recherche ; des données sûres, en veillant à ce que les données soient correctement anonymisées/pseudonymisées pour le but recherché ; des personnes sûres, des chercheurs crédibles qui entreprennent le travail ; et que le résultat de la recherche soit également sûr et non divulgué. Les données administratives recèlent un potentiel important, et il suffit d'adopter une approche prudente pour s'assurer que ce qu'ils ont et font est sûr et approprié pour l'avenir.

John MARTIN

Je souhaite préciser que *the Jobseekers Longitudinal Database*, la base de données longitudinale que j'ai présentée, a été construite par les statisticiens du département ministériel de la Protection sociale. Ceux-ci ont travaillé en étroite collaboration avec des statisticiens du CSO. Et cette collaboration se poursuit pour enrichir et améliorer les bases de données longitudinales qui concernent les demandeurs d'emploi irlandais.

Kamel GADOUCHE

Est-il envisagé de formaliser, législativement ou par une autre voie, les contributions et les échanges indépendants effectués avec les chercheurs dans le cadre de la construction du budget, sur le fondement de l'*evidence-based policy* (politique publique fondée sur des éléments de preuve) envisagée par le Congrès américain ?

Jean-Noël BARROT

Cette idée est intéressante, néanmoins, il ne me semble pas qu'une telle décision soit prévue à l'Assemblée nationale. Sans réellement y parvenir, nous avons tenté d'inscrire dans la loi le fait que le Parlement puisse disposer de davantage de moyens d'expertise, pour mieux analyser en amont des lois et en particulier des textes budgétaires.

Néanmoins, dans cette optique, grâce à la bienveillance de l'INSEE, une cellule de spécialistes des données créée à l'Assemblée nationale et composée de trois ou quatre personnes, a pu concevoir une interface permettant aux députés de simuler les effets d'un amendement déposé sur certaines lois : https://leximpact.an.fr/. Cette interface aide à réfléchir sur les projets de loi de finances, ou encore sur des lois touchant l'impôt sur le revenu, certaines prestations sociales et les dotations de l'Etat aux collectivités. Cet outil a été conçu pour les députés et il s'appuie sur certaines bases de données de chercheurs. Il permet déjà de répondre à certaines questions et il est envisagé de l'étendre à des textes de loi touchant d'autres impôts. Il s'agit d'un petit pas en avant, en direction de l'*evidence-based policy* que vous évoquez.

Par ailleurs, l'Assemblée nationale a récemment instauré le Printemps de l'évaluation, en inscrivant son principe dans la loi organique. Cette pratique se déroule sur trois semaines, entre la fin du mois de mai et le début du mois de juin. Dans ce cadre, les ministres responsables de leurs administrations viennent rendre des comptes sur l'exécution des budgets

qui leur ont été confiés l'année précédente. Ils rendent encore des comptes sur l'efficacité des politiques publiques qu'ils ont été chargés d'exécuter. Ces ministres sont interrogés par les députés, qui disposent désormais de ressources internes à l'Assemblée nationale, pouvant s'aider en particulier d'administrateurs de l'Assemblée, pour rédiger un rapport d'évaluation. Tous les députés ne mobilisent pas des travaux fondés sur des appariements de données ou sur des données administratives, mais je souhaite qu'ils puissent y recourir davantage au cours de cette évaluation.

Jimmy BAULNE, Institut de la statistique du Québec

L'accès aux données à distance opéré par le biais du CASD se réalise-t-il sur des fichiers qui auraient fait l'objet d'un certain masquage qui permettrait de réduire le risque de divulgation ?

Kamel GADOUCHE

Les données mises à disposition par le CASD font l'objet d'une pseudonymisation. Les identifiants directs ont été remplacés. En revanche, l'identifiant direct a été conservé pour les données des entreprises. Néanmoins, le CASD ne recourt pas à des perturbations ou à d'autres techniques d'anonymisation sur les données qui lui sont confiées. Par conséquent, une traçabilité élevée est mise en place. Tout ce qui est réalisé par les utilisateurs est filmé et enregistré. Ce matin, j'ai pu vous présenter les mesures de sécurité qui étaient mises en œuvre. Néanmoins, l'objectif du CASD est bien de mettre à disposition des chercheurs les données les plus détaillées possibles, de manière à permettre des études très précises.

Philomé ROBERT

Je me tourne vers John Martin. Vous avez mis en lumière la mobilisation des appariements dans le cadre des politiques publiques de l'emploi. Nous avons évoqué tout à l'heure quelques éléments de comparaison avec des pays tels que le Canada, l'Irlande, ou encore la Suisse. Je me demandais donc si vous pourriez détailler davantage d'éléments de comparaison avec la France. Par exemple quelles sont les spécificités propres à la France ou au Canada en matière d'appariements ?

John MARTIN

Cette question est large. Au Canada, par exemple, de nombreuses études sont focalisées sur le système d'indemnisation du chômage. En effet, il existe dans ce pays un régime assez particulier qui varie dans ses différentes provinces. De nombreuses études qui ont mobilisé des appariements de données ont ainsi montré l'existence d'un travail saisonnier très important qui est subventionné par ce système d'indemnisation canadien.

Par exemple, dans les provinces maritimes (Nouveau-Brunswick, Nouvelle-Ecosse, Ile-du-Prince-Edouard), des études démontrent que le taux de chômage est plus élevé à cause de travailleurs saisonniers qui bénéficieraient de cette indemnisation. De plus, une étude très intéressante, qui a employé des données longitudinales administratives, a montré que les individus apprenaient rapidement les règles du système d'indemnisation du chômage. Ceux-ci parvenaient à bénéficier de ces indemnités plusieurs fois dans leurs parcours, jusqu'à faire monter les taux de chômage, en particulier dans les provinces maritimes. Ce sujet est très sensible au Canada, étant donné que les règles du chômage relèvent des provinces.

Roxane SILBERMAN, CNRS

D'autres pays que la France ont-ils entrepris d'établir des outils semblables au CSNS de l'INSEE ?

Sylvie LAGARDE

John Dunne, j'ai cru comprendre que le *Personnal Identification Code* (PIC) irlandais était un équivalent du CSNS ?

John DUNNE

L'un d'eux était purement basé sur un générateur de nombres aléatoires qui était conçu pour contenir l'identifiant original. L'autre était un système appelé "salt and hash", dans lequel l'identifiant était préfixé d'un texte secret que seules deux ou trois personnes connaissaient. Ensuite, le flux de texte concaténé était haché pour obtenir l'identifiant. Cette opération était effectuée de manière systématique pour chacun des identifiants, avec une routine de salage et de hachage différente pour chaque identifiant.

Philomé ROBERT

Je reviens vers vous, monsieur le député. Vous organisez des journées sur l'évaluation des politiques publiques, quelles leçons en tirez-vous ? Quelles seraient les marges de progression en la matière ?

Jean-Noël BARROT

J'ai le sentiment qu'un nombre croissant de travaux universitaires mobilisent des données administratives. Ces travaux s'intègrent dans le débat public, voire créent des débats dans l'opinion. Je trouve cette démarche vertueuse, étant donné qu'elle offre un apport à la conception des politiques publiques.

Le Printemps de l'évaluation a été organisé à deux reprises, à deux années d'intervalle, mais pas cette année. En effet, nous avons estimé que nous situant à quelques semaines d'une élection présidentielle, le climat n'était plus propice pour permettre une évaluation des politiques publiques.

Il nous paraissait cependant important que le monde de la décision publique et celui de la recherche puissent se rencontrer au sein de l'Assemblée nationale, qui est mandatée pour évaluer les politiques publiques d'après l'article 24 de la Constitution. Il était utile que ces deux univers puissent se rencontrer, apprendre à se connaître et sachent élaborer un langage commun. Il est effectivement important que le monde de la décision publique sache mieux utiliser et mobiliser les résultats des travaux universitaires et inversement, que le monde universitaire puisse mieux cerner les attentes des décideurs, de manière à leur fournir le matériel qui leur serait utile. Je pense que cette démarche a participé au mouvement général vertueux qui tendait à intégrer davantage les travaux des chercheurs dans le débat public.

Ainsi, dans l'actualité de ces derniers jours, le sujet de la fiscalité des donations et des successions s'est inséré dans le débat de la campagne présidentielle, essentiellement à la suite d'une note publiée par le Conseil d'analyse économique (CAE). Ce débat n'est pas

associé à un poids budgétaire conséquent, mais il a pris une ampleur importante au regard de la symbolique forte liée à cette fiscalité. Le CAE cherchait à répondre à certaines questions à partir de données administratives et il a pu apporter des éléments nouveaux. Il s'est notamment intéressé au coût de certaines exonérations fiscales. Or il se trouve que l'Assemblée nationale avait déjà interrogé le gouvernement sur cette matière, demandant un rapport sur le coût aux finances publiques de différentes exonérations, dont l'avantage successoral porté sur l'assurance vie. Mais le gouvernement n'avait pas été capable de répondre à cette question, car les données sur les transmissions n'étaient pas suffisamment connues, en France comme dans beaucoup de pays. Finalement, cette note du CAE a permis de résoudre cet écueil en parvenant à chiffrer ces coûts par diverses manières. Ces vérités nouvelles ont été intégrées au débat public par le CAE, passant du monde universitaire au monde politique.

Patrice DURAN

La problématique de l'efficacité des politiques publiques, qui est actuellement centrale dans la légitimation des gouvernements, suppose l'intelligibilité du monde social. Par conséquent, la question des appariements devient un élément essentiel, puisqu'elle permet une meilleure réflexivité de l'action publique et par là même un approfondissement des connaissances.

Pour en revenir à la question du rôle du Parlement dans l'évaluation des politiques publiques pilotées par le gouvernement, je relève l'existence d'une grande faiblesse de l'Assemblée nationale en la matière. En effet, pour avoir connu le General Accounting Office, maintenant appelé *Government Accountability Office*, organisme du Congrès américain chargé du contrôle des comptes publics du budget fédéral, mais aussi très investi dans l'évaluation des politiques publiques, on ne peut que s'interroger sur la différence de moyens entre le Congrès américain et le parlement français quand on voit que le GAO disposent de moyens à peu près équivalents à ceux de l'INSEE ! Or, dans le cas français, ce n'est pas par la mise à disposition d'une partie de la Cour des comptes au Parlement pour l'assister dans l'évaluation des politiques publiques qu'on fera un grand pas en avant, d'autant que cette institution n'est certainement pas la plus spécialisée en matière d'évaluation des politiques publiques.

Certes, comme vous l'indiquez et comme je l'ai également signalé, il est vrai que tous les rapports soulignent l'importance de revoir le lien existant entre la recherche, l'enseignement et le monde de l'action publique, tant du côté du Parlement que des ministères. Mais, comme mon voisin Jean-Luc Tavernier, je m'interroge sur la place historique de la statistique publique tant dans l'enseignement qu'au sein même de la gestion publique. Au CNIS, nous avons pu constater de ce point de vue à quel point la pandémie a permis d'avancer considérablement dans la prise de conscience de tous les acteurs publics du caractère indispensable et décisif d'une statistique publique pertinente et de qualité. En effet, la statistique publique a été longtemps mal perçue en France à cause, bien souvent, de l'insuffisance de connaissance sur sa nature et de fait sur son utilité. Comment peut-on agir pour offrir à la statistique publique toute sa place à l'avenir ? C'est là un enjeu central pour nous comme pour tous les acteurs publics.

Jean-Noël BARROT

Je pense que vous avez raison. Non seulement nous nous apercevons que la statistique publique est très utile pour pouvoir comprendre les phénomènes en cours, mais il est aussi

urgent que le gouvernement français et les autres gouvernements se dotent des compétences nécessaires pour pouvoir traiter les données. Sans quoi, les gouvernements subiront les dictats que leur imposeront ceux qui détiennent la maîtrise de ces données.

Il existe une réponse parmi d'autres, que vous avez peut-être évoquée implicitement dans votre question et qui consiste à intégrer plus naturellement et de manière plus fluide des économètres statisticiens et des docteurs dans la fonction publique de l'Etat. Cette réponse pourrait être apportée de manière progressive et permettrait à la fonction publique de s'acclimater aux enjeux de la statistique. Il existe encore des ministères dits thématiques, tels que celui du Logement, qui sont encore loin d'intégrer cette approche. S'ils se désinvestissent de la statistique, d'autres acteurs tels que Google risquent de prendre le relais.

Philomé ROBERT

Comment pouvons-nous combler ce fossé ?

Jean-Noël BARROT

Pour répondre à l'inefficacité de certaines politiques publiques liée à l'absence de la pratique des traitements de données, d'autres acteurs risquent d'intervenir en offrant des services prétendument gratuits, mais rémunérés par de la publicité.

Je rejoins l'avis de Patrice Duran sur la faiblesse du Parlement en matière d'évaluation des politiques publiques, sans vouloir paraître cassant ou déprimé.

L'OCDE se fonde sur une espèce de taxonomie des institutions financières indépendantes (IFI) qui challengeraient les pouvoirs exécutifs dans leurs choix budgétaires. Parmi ces institutions, d'une part, elle distingue celles qui sont logées à l'enceinte des parlements tels que le *Congressional Budget Office* (CBO) ou encore *l'Ufficio parlamentare di bilancio* (UPB). D'autre part, elle distingue d'autres institutions logées hors des parlements, dont le meilleur exemple est l'*Office for Budget Responsability* (OBR) britannique, qui met au défi les hypothèses et les choix du Trésor. Je ne sais pas ce qu'il en est en Irlande.

Lorsque j'ai commencé à lancer des recommandations sur le sujet, j'ai préconisé de loger une IFI au Parlement, en étant peut-être capturé par ma corporation. Cette proposition pose certaines questions, puisqu'il s'agit de refaire l'exercice du gouvernement de projection des dépenses.

Ensuite, le rapport de la Commission sur l'avenir des finances publiques présidée par Jean Arthuis, dans le cadre duquel certains d'entre vous avaient été auditionnés, a réinterrogé l'efficacité de la dépense publique. Or ce rapport a plutôt préconisé l'établissement d'une IFI externe au Parlement.

L'OCDE identifie actuellement le Haut-Conseil aux finances publiques (HCFP) comme étant l'IFI française. Et elle la situe parmi les dernières IFI dans ses classements, quelques soient les indicateurs retenus, qu'il s'agisse du nombre de ses membres, de son budget, ou encore de ses missions. Je salue d'ailleurs ses membres qui sont présents dans cette rencontre.

Or le rapport de Jean Arthuis proposait d'installer cette IFI au HCFP. Si cette institution s'avérait la plus adaptée pour remplir cette mission d'évaluation, il faut donc l'alimenter au

plus vite. Il serait regrettable de demeurer dans la situation actuelle où rien n'existe à l'Assemblée nationale à part la petite équipe de trois ou quatre personnes que j'ai évoquée, et où le HCFP se contente de vérifier des hypothèses à l'échelle macro.

Finalement, heureusement que nous disposons d'un service statistique public et de laboratoires de qualité, qui remplissent ce rôle d'évaluateurs, et qui permettent de rendre notre système plus cohérent. A cet égard, j'avais présenté l'exemple de l'évaluation du budget opérée par l'Ecole d'économie de Paris. Toutefois, il est vrai qu'il faut encore résoudre cette question sur le plan institutionnel.

Un invité à distance

Les différents organismes comparables au CASD au sein des différents pays pourraient-ils s'interconnecter de manière à permettre aux chercheurs de bénéficier d'un accès à des données provenant de différents pays ?

Kamel GADOUCHE

Le CASD, notamment par le biais de Roxane Silberman, coordonne le réseau de centres d'accès sécurisés *International Data Acces Network* avec l'Allemagne, le Royaume-Uni et les Pays-Bas, pour monter un *core network* (noyau) visant à permettre des travaux de recherche transnationaux.

Je profite donc de la présence de John Martin et de John Dunne pour leur demander s'ils trouvent du sens dans cette possibilité d'échanger des accès de données et de permettre ces interconnexions, en vue d'enrichir les connaissances mutuelles de chacun de ces pays.

John MARTIN

Ces interconnexions se retrouvent au niveau européen, notamment à travers le réseau EUROMOD, dont font partie l'Irlande et un certain nombre de pays. Il s'agit d'un modèle de micro-simulation permettant d'évaluer et de comparer les effets des politiques socio-fiscales. Je citerais encore le *Luxembourg Income Study* (LIS), qui produit une base transnationale de données microéconomiques sur les revenus. Il s'agit de réseaux internationaux qui tendent à partager des modèles et des méthodes. Néanmoins, ils n'échangent pas de données individuelles. A ma connaissance, de telles pratiques n'existent pas véritablement au niveau européen.

John DUNNE

Dans le premier cas, il s'agissait d'une approche standardisée de la déclaration statistique à travers l'UE. Dans un deuxième temps, c'est principalement là qu'ils ont effectué une grande partie de leur travail de normalisation et de comparaison en termes d'informations statistiques. L'étape suivante consistait à créer des fichiers de microdonnées sûrs, ce qui était fait sur des enquêtes sociales distinctes, éventuellement. Il était possible de relier ces enquêtes dans des centres de recherche spécialisés. Le partage de microdonnées entre pays soulève de nombreuses questions sensibles, notamment en ce qui concerne le partage de données personnelles en dehors du territoire irlandais, où la législation est incertaine. La stratégie européenne en matière de données et la loi sur la gouvernance des données, qui est sur le point d'être finalisée, ont suscité beaucoup d'intérêt au niveau euro-

péen. Le paysage des données semble être en pleine mutation, mais les agences statistiques nationales s'engageront probablement avec prudence dans ce domaine. Il reste à voir ce qui va se passer à l'avenir.

Philomé ROBERT

N'y ayant pas d'autres questions, nous clôturons cette table ronde. Merci beaucoup à vous trois.

TABLE RONDE – QUELLE TRANSPARENCE, QUELLE INFORMATION DU PUBLIC ?

Présidente de la table ronde : Chantal Cases, Société française de statistique ;

Eric Rancourt, directeur des méthodes statistiques et de la science des données, Statistique Canada ;

Bertrand Pailhès, directeur des technologies et de l'innovation, CNIL ;

Mark Hunyadi, philosophe, professeur de philosophie morale et politique à l'Université catholique de Louvain ;

Maryse Artiguelong, vice-présidente de la Ligue des droits de l'homme.

Chantal CASES

Bonjour à toutes et à tous. Cette table ronde rassemble deux participants présents dans la salle et deux autres qui se trouvent à distance. Je les remercie tous les quatre. Notre table ronde s'intéresse à la transparence et à l'information du public en matière d'appariements. Au cours de cette journée, nous avons montré en détail toute la richesse des projets de répertoires et d'appariements, en soulignant essentiellement leurs apports en termes de connaissances.

Mais finalement, nous avons peu réfléchi sur la question de mandat social qui a été évoquée ce matin et je crois qu'il est temps de l'approfondir. La transparence et la garantie du respect de la confidentialité des données apparaissent essentielles et déterminantes. Elles permettent de conférer un véritable sens à ces projets.

De surcroît, je pense que nous devons communiquer les objectifs de ces projets d'une manière très large, y compris auprès du grand public qui nous fournit des données, tant aux spécialistes qu'à ceux qui connaissent moins leurs aspects techniques dont l'abord est difficile. Un certain nombre de participants de cette réunion a ainsi pu se rendre compte que les appariements sont techniquement complexes et qu'ils ne s'avèrent pas aussi évidents qu'on aurait pu se l'imaginer.

Nous nous demanderons donc comment assurer cette transparence, la bonne information du public, mais encore une bonne concertation sur ces projets de la statistique publique. Nous tenterons de répondre à cette question en interrogeant quatre grands témoins.

Tout d'abord, nous donnerons la parole à Eric Rancourt, directeur général des méthodes statistiques modernes et de la science des données à Statistique Canada. Il est riche d'une

expérience extérieure qui pourrait nous éclairer sur le mandat social associé aux projets statistiques.

Je vais l'interroger dans un premier temps sur ses pratiques en matière d'appariements de données. Je sais que Statistique Canada, tout comme l'INSEE et quantité d'autres organismes statistiques, utilise depuis longtemps des données administratives pour élaborer des statistiques. Ainsi, avant d'entrer dans le vif du sujet touchant à l'information du public, j'aimerais connaître la position actuelle de Statistique Canada concernant la pratique des appariements de fichiers. J'aimerais encore que vous nous présentiez les occasions qui vous ont conduites à approfondir vos réflexions sur vos relations avec le public en matière d'appariements.

Dans un second temps, je souhaiterais que vous nous expliquiez plus concrètement la manière dont vous organisez l'information du public. Sylvie Lagarde a évoqué ce matin le couplage de microdonnées réalisé par Statistique Canada. J'aimerais connaître les enjeux de ce couplage et avoir une idée des données mises en jeu. Enfin, j'aimerais que vous nous présentiez les sujets particulièrement sensibles pour le public canadien.

Eric RANCOURT

Bonjour à tous, je suis très heureux de participer à cette table ronde organisée par le CNIS. Je remercie Chantal Cases et Françoise Dupont de m'avoir invité. Je vais évoquer la question du couplage d'enregistrements (appariement) au Canada, en me concentrant sur les pratiques de Statistique Canada, bien que ce sujet puisse déborder au niveau des provinces. En revanche, je n'aborderai pas la question des centres de données de recherche, mais je précise tout de même que nous commençons à mettre en œuvre une possibilité d'accéder à distance aux données.

Tout d'abord, préalablement à la présentation des événements qui se sont produits il y a quatre ou cinq ans et qui ont beaucoup changé la situation en matière de mandat social, voici un historique succinct du recours aux données administratives du Canada.

Avant les années 1940, nous recourrions uniquement à des pseudo-recensements, ainsi qu'à des fichiers administratifs, dans une plus grande mesure. Cette situation a perduré jusqu'à l'article célèbre de Jerzy Neyman de 1934 qui a introduit l'échantillonnage probabiliste. A la suite des travaux américains de Morris Hansen, nous avons mis en place en 1945, au Canada, une enquête sur la population active, grâce au travail de Nathan Keyfitz. Les enquêtes ont alors commencé à proliférer dans une très large mesure, jusqu'aux années 1990 et aux années 2000. Nous pouvions alors parler de la cristallisation d'une logique de production de l'information centrée sur les enquêtes.

Néanmoins, nous avons continué dans le même temps à recourir à des données administratives, surtout dans le cadre de bases de sondage liées à la sphère sociale. Nous les avons également mobilisées en l'état dans bien des cas, pour produire des statistiques économiques.

Les données administratives ont été utilisées de manière croissante dans les années 1970, si bien qu'une division dédiée à ce type de données a été créée autour de 1979 et 1980. Ces données sont donc inscrites dans l'ADN de Statistique Canada, d'autant plus que nous sommes régis par une loi qui vise à prévenir le double emploi de la collecte statistique dans

les différents ministères et à favoriser la production de statistiques intégrées. Cette loi confère seulement un mandat global et ne mentionne pas de couplage ou d'appariements de fichiers. Nous déduisons entre autres le recours à cette pratique dans cette prévention du double emploi de la collecte. En outre, une clause spécifique indique que Statistique Canada a accès à tous les fichiers, produits au niveau fédéral, provincial, municipal, voire même dans le domaine privé.

Jusqu'il y a quelques années, le recours aux appariements avait beaucoup progressé, essentiellement dans un but d'amélioration ou de vérification de la qualité et pour réduire les frais des enquêtes. Graduellement et de manière croissante, ils ont été mobilisés pour remplacer des données d'enquêtes. Nous avons alors créé des environnements d'appariements sécurisés et nous avons établi un centre d'expertise dédié à cette pratique au sein de la Direction de la méthodologie. Nous avons établi des processus d'accès aux fichiers assez contraignants et complexes. Nous avons également créé un comité devant lequel devaient passer tous les projets d'appariements jusqu'en 2017 et qui devait évaluer en profondeur des indicateurs relatifs à la protection de la vie privée. Ensuite, une fois les appariements réalisés, les informations pouvaient éventuellement figurer sur le site de Statistique Canada.

De cette façon, depuis plusieurs années, nous avons traité des milliers de fichiers et nous avons réalisé toutes sortes d'appariements. En fait, dans les années 2010, tant au Canada que dans d'autres pays, nous avons assisté à une prolifération de production de données.

Dans le même temps, la population a eu une plus grande connaissance de l'existence de ces données, faisant croître des demandes d'informations très détaillées. Mais dans le même temps, le taux de réponse aux enquêtes a diminué de façon constante et importante.

Face à ce phénomène, Statistique Canada a mis en œuvre un programme de modernisation, dans l'optique de renverser le paradigme de production de l'information. Il s'agissait de partir des données administratives pour les compléter par des enquêtes et d'ancrer ce processus dans un système caractérisé par une chaîne d'inférences valide. En procédant ainsi, nous avons utilisé quelques nouvelles sources de données administratives en plus de celles déjà existantes. Cependant, nous les avons utilisées d'une manière différente.

En 2017, la loi sur la statistique a été modifiée pour des raisons particulières qui ne sont pas reliées avec l'emploi des données administratives. Cette loi a été additionnée de différentes clauses, touchant notamment la nomination du statisticien en chef, ou encore la dépénalisation de certaines non-réponses. Mais comme nous utilisons des données administratives et que nous entreprenions la modernisation de nos approches statistiques, certaines personnes ont commencé à associer la modification de cette loi avec le développement de l'utilisation des données administratives. A partir de là, certaines personnes ont commencé à se montrer très frileuses sur l'utilisation de ces données.

De la sorte, en 2018, un tollé a été généré après que nous ayons discuté de la possibilité d'utiliser les données des transactions bancaires spécifiques aux individus en obtenant des informations auprès des banques. Ce projet n'a pas vu le jour et la polémique est allée jusqu'au Premier ministre, qui a été interpellé au Parlement à ce propos. Cet épisode a alors modifié de manière durable l'image du mandat social des institutions publiques en général et en particulier celle de Statistique Canada.

Dans nos travaux de développement en cours, , nous avons alors tenu compte de cet épisode. Ainsi, nous avons augmenté considérablement la transparence. Nous partageons des informations sur notre site internet, à la fois avant, pendant et après les appariements. Ces informations sont donc communiquées au public et au Ministre au moins trente jours avant que nous ne procédions à l'appariement. L'opération est donc plus complexe, mais elle est très transparente.

Nous avons ensuite développé un processus ou un cadre touchant les acquisitions de fichiers administratifs et par extension pour les appariements qui est basé sur la nécessité et la proportionnalité. Ce cadre a été rendu beaucoup plus robuste et il a inclus des considérations éthiques tenant compte de la vie privée, de la transparence, de l'équité, de la confiance, ou encore de précautions de non-malfaisance. Ce cadre est centré sur les bénéfices de l'utilisation des données administratives et des appariements, non pas seulement pour le système statistique, pour les décideurs ou pour les commanditaires des statistiques, mais d'abord et avant tout pour le public. Il s'agit de se demander en quoi la récolte de données administratives et les appariements peuvent correspondre à l'intérêt général en ensuite mettre en œuvre une approche proportionnelle à ce besoin.

Revenons en particulier sur la directive de 2017 sur les appariements. Celle-ci prévoit tous les types de couplages de données et présuppose l'acquisition des données. Comme je l'ai mentionné, l'acquisition de tout fichier est prévue par la loi sur la statistique. Le mandat légal est donc bien établi. Néanmoins, depuis 2018 ou 2019, l'acquisition du mandat social a été entravée. Par conséquent, le cadre de nécessité et de proportionnalité a pu aider à obtenir ce mandat social. Il s'agissait de justifier de manière très détaillée pourquoi nous utilisons des données, à quelle fin, et en quoi leur utilisation répondrait au principe de proportionnalité, même si la loi sur la statistique nous permettait pleinement cette utilisation.

Ainsi, étant donné que nous nous situons sur une logique de données administratives et pas seulement sur une logique d'enquêtes, au lieu d'effectuer une évaluation des facteurs liés à la vie privée à chaque demande de couplage – en sachant que nous en opérons des milliers chaque année –, nous avons créé une autorisation et une évaluation générale. Il s'agit d'éviter la procédure visant à demander des appariements à toutes les activités de routine, telles que l'utilisation de données administratives dans la construction de bases de sondage, l'évaluation de la qualité, ou encore l'évaluation du codage.

En revanche, l'ajout de nouvelles variables à un ancien ou à un nouvel appariement nécessite d'effectuer une demande, qui passe par un formulaire, par une évaluation des facteurs liés à la vie privée, et qui nécessite dans certains cas de passer devant un comité. La conformité de ces demandes avec le mandat de Statistique Canada est alors examinée, devant correspondre à l'intérêt public et au respect de la confidentialité. Le comité vérifie encore les effets potentiels du traitement de données sur les valeurs canadiennes telles que l'équité ou la non-malfaisance, tout en veillant au respect de la nécessité de proportionnalité et en vérifiant l'intérêt scientifique et méthodologique de l'appariement. La collecte se déroule ensuite en mobilisant une anonymisation et une clé de couplage, dans un environnement sécuritaire.

Parallèlement, l'information est publiée en amont sur internet. Et à la fin de l'année, un rapport est présenté au Parlement avec l'ensemble des appariements réalisés. Dans tous les cas où des facteurs liés à la protection de la vie privée sont évalués, les appariements sont présentés au commissaire à la protection de la vie privée, un haut fonctionnaire qui relève directement du Parlement.

Avant de procéder à un appariement, il faut procéder à l'acquisition des fichiers. L'acquisition de tout nouveau fichier est soumise à une évaluation sommaire de la sensibilité potentielle des données. Cette évaluation est ensuite suivie de l'évaluation de la nécessité, de l'efficacité, de la proportionnalité et des éventuelles alternatives. Cette procédure est transférée au Comité d'éthique dès lors que l'évaluation met en évidence des considérations éthiques. Ensuite, le conseiller principal en éthique et de l'intégrité scientifique détermine si le projet peut démarrer et adresse une recommandation en ce sens au gestionnaire de programme. Le cadre de nécessité et proportionnalité s'applique à l'ensemble des acquisitions de fichiers et des appariements.

Par ailleurs, nous travaillons actuellement sur l'établissement de centres de recherche virtuels permettant aux chercheurs d'accéder aux données depuis chez eux, sur la base des cinq niveaux de sécurité décrits par John Dunne.

Nous travaillons également à l'établissement d'une échelle de sensibilité basée sur l'analyse des non-réponses dans les enquêtes, mais également sur des groupes de discussion et des rencontres visant spécifiquement à recueillir des informations sur la sensibilité des Canadiens sur différents sujets. Ainsi, les données d'enquête ou les données d'utilité publique – abonnements d'électricité, de câbles, etc. – sont beaucoup moins sensibles que les données financières transactionnelles par exemple. Je ne peux pas vous présenter aujourd'hui cette échelle, puisque nous en sommes encore à nos travaux préliminaires. Nous souhaitons élaborer une échelle qui puisse permettre d'aider dès le départ les gestionnaires des appariements à déterminer le niveau de sécurité et de transparence nécessaire, les permettant de bien doser les efforts liés à la documentation et à la justification de leurs programmes.

Enfin, nous développons un système central de registres. En effet, nous disposons au Canada d'un registre des entreprises, ainsi qu'un registre des adresses qui a été converti en registre des immeubles. Nous travaillons donc actuellement sur un registre des individus, à l'aide d'une nouvelle infrastructure sécurisée dédiée à l'intégration des données. Ce projet s'inscrit dans un agenda similaire à celui du programme RESIL. Nous devrions ainsi être dotés d'un système complet permettant de faciliter l'ensemble des appariements à l'horizon de 2025.

J'ai donc présenté globalement les travaux actuels de Statistique Canada, tout en répondant à votre question. Je vous remercie de votre attention.

Chantal CASES

Cette perspective d'outre-Atlantique que vous nous présentez est intéressante. Je vous en remercie, ainsi que de vous être levé avant votre aurore pour assister à l'ensemble de la rencontre, ce qui est très courageux de votre part. Je donne la parole à Bertrand Pailhès, directeur des technologies et de l'innovation au sein de la CNIL. Dans un passé récent, il a été directeur de cabinet d'Axelle Lemaire, secrétaire d'Etat au numérique. Il a conduit à ce titre l'adoption de la loi de 2016 pour une République numérique, qui a notamment facilité les appariements pour la statistique publique et la recherche.

Pour faire le point sur le respect de la confidentialité et sur l'information du public, auprès de qui sont collectées les données personnelles, quels sont les garde-fous élaborés par la CNIL pour réaliser des appariements de fichiers ou des répertoires tout en préservant la vie

privée ? Pourriez-vous encore nous rappeler comment les risques et les bénéfices potentiels de ces opérations sont évalués au sein de la CNIL ? Nous avons constaté qu'il n'était pas toujours simple de comprendre en détail tous ces projets.

Dans un second temps, je vous demanderai de répondre à une question plus polémique. Les garde-fous et le respect de la confidentialité sont indispensables. Néanmoins, ces garde-fous peuvent devenir des carcans. De ce fait, comment construire des garde-fous qui n'empêchent pas la production d'appariements et d'analyses utiles au bien collectif ? Peut-être que Mark Hunyadi nous en dira davantage sur cette question, mais cette question m'est chère et je pense qu'elle l'est également pour la présidente de l'ASP. Ainsi, l'appariement de données de santé et de données sociales est encore aujourd'hui extrêmement complexe à réaliser, ce qui nous empêche de connaître de nombreuses informations sur les inégalités sociales de santé et d'accès aux soins. Or ces informations seraient particulièrement utiles pour remédier à ces inégalités.

Bertrand PAILHÈS

Bonjour à tous. Je m'excuse de ne pas pouvoir être parmi vous, mais le contexte sanitaire m'oblige à rester à distance. Pour répondre à votre première question sur les garde-fous relatifs aux appariements, je pense que les différents pays évoqués aujourd'hui ont des cultures marquées par la protection des données personnelles. La France est aussi historiquement associée à une protection importante de ces données, qui se manifeste notamment autour du fort encadrement du NIR.

Il est d'abord important de préciser que les appariements sont soumis au cadre du RGPD, le cadre général qui demande à chaque étude ou enquête, chaque « traitement » au sens du RGPD, d'évaluer le principe de proportionnalité et la finalité qui permettent de justifier le traitement de données personnelles. Ce principe général s'applique en particulier pour les appariements.

Il faut également noter que le RGPD, qui est un texte européen sur la protection des données personnelles, a tout de même introduit un régime spécifique pour les traitements statistiques, dans l'objectif de lever certaines obligations d'information ou encore certains droits d'opposition ou d'accès que pourraient avoir les personnes vis-à-vis des traitements statistiques qui les concernent. Il est d'ailleurs paradoxal d'évoquer cet élément dans une table ronde associée à la question de la transparence. Finalement, l'intention du législateur européen vise à permettre de manière assez large l'exploitation statistique des données, y compris dans des finalités différentes de celles qui avaient animé leurs collectes. Cette intention s'inscrit dans l'idée d'une utilité collective de ces traitements de données.

En outre, dans le cadre de notre évaluation des risques et des bénéfices potentiels, le RGPD prévoit le recours à une analyse d'impact sur la protection des données (AIPD). Il s'agit d'une sorte de document qui vise à matérialiser les risques pour les personnes et les mesures à prendre pour limiter ces risques. L'AIPD n'est pas requise pour tous les traitements, mais en pratique, les traitements opérés sur une grande échelle et qui touchent des données sensibles nécessitent a priori l'usage de cet outil, ce qui correspond à de nombreux cas de projets d'appariements.

Au sein de la CNIL, nous avons produit à cette fin, avec nos collègues européens, une méthode qui vise à fournir un maximum d'éléments pour tenter d'évaluer l'impact de ces traitements sur les droits et les libertés des personnes. Cette démarche rejoint sans doute

les travaux menés par Statistique Canada autour de l'échelle de sensibilité décrite par Eric Rancourt. Cette méthode s'inscrit dans une approche qui est sans doute en partie héritée d'une approche centrée sur la sécurité des systèmes d'information. Mais cette méthode se veut surtout multidimensionnelle, s'adaptant au type de données, à l'acteur qui les traite, au type de menaces qui peut se porter sur ces données. Dans certains cas, des données qui peuvent sembler peu sensibles dans un traitement peuvent l'être davantage dans un autre contexte, risquant par exemple d'être divulguées et générer ainsi un effet direct sur les personnes.

Cet exercice d'analyse des risques invite en particulier à réfléchir en termes de droits et de libertés des personnes et à se demander en quoi un traitement de données peut affecter ces droits fondamentaux.

Il existe encore d'autres garde-fous. Ainsi pour la CNIL, le fait d'apparier n'est pas une finalité en soi. Par conséquent, tout appariement de données répond à des finalités données. Des données peuvent par exemple être appariées pour étudier la réussite professionnelle en fonction d'un parcours scolaire, comme l'a décrit Francesco Avvisati. En soi, nous aborderons toujours ces appariements à l'aune de leurs finalités.

Je précise que des actes réglementaires permettent néanmoins d'apparier des données de santé, et la réflexion sur leur finalité est d'ailleurs particulièrement approfondie. Pour ce faire, il existe le *Health Data Hub*, la plateforme des données de santé, qui vise spécialement à réaliser des appariements.

De plus, un autre garde-fou est porté par la gouvernance. Eric Rancourt a évoqué différents comités qui contrôlent les traitements de données et il en existe également en France. Ainsi, la CNIL participe au Comité du label de la statistique publique pour les projets d'enquêtes de la statistique publique, ou encore au Comité du secret statistique pour les recherches. Une partie mineure mais constante des activités de la CNIL vise donc depuis très longtemps à contribuer à l'évaluation de l'opportunité des traitements de données dans le cadre d'études et d'enquêtes. La CNIL cherche encore à identifier les risques qui n'auraient pas été identifiés par les communautés de la recherche et de la statistique, exerçant un regard supplémentaire.

Dans ce contexte, la CNIL porte une attention particulière aux données sensibles. Il existe un débat régulier en France à cet égard. Ce débat se rapporte en particulier aux données d'appartenance ethnique, qui constituent des données sensibles au sens de la loi et qui doivent faire l'objet d'une attention particulière. Le contrôle de proportionnalité y est particulièrement important.

Les avis de ces comités peuvent par exemple inciter à rendre facultative la collecte de certaines données sensibles, de manière à laisser la possibilité aux personnes enquêtées de ne pas répondre, pour protéger leurs libertés. En pratique, cette possibilité ne semble cependant pas applicable à toutes les enquêtes.

Cette gouvernance agit également dans le cadre des recherches de santé, mais dans un cadre juridique beaucoup plus contraignant qu'ailleurs. Ce cadre est caractérisé par des régimes d'autorisations, qui peuvent notamment être délivrées par la CNIL, ou encore par le Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (CESREES), qui sont interrogés sur l'intérêt public des traitements de données.

Je réponds à présent à votre deuxième bloc de questions, qui me semble particulièrement intéressant. Il existe effectivement en France une tradition axée sur la protection des données personnelles. En 1974, le système automatisé pour les fichiers administratifs et le répertoire des individus (SAFARI), qui devait créer de grands fichiers et interconnecter des données, avait été dénoncé. Les conditions d'accès ont alors été particulièrement durcies, sur le motif de deux craintes principales. La première renvoyait au fait que le NIR est identifiant, puisqu'il correspond à un numéro d'ordre de naissance dans une commune donnée et à une date donnée qui permet d'identifier directement des personnes. La seconde était associée à l'inquiétude que pouvaient générer de grands volumes de données identifiantes.

Afin d'éviter les carcans relatifs à ces garde-fous, il existe deux approches principales. La première solution est d'ordre juridique. Il s'agit de changer la loi, en considérant que les préventions opérées dans les années 1970 ne sont plus nécessaires et qu'il convient plutôt de se rapprocher du système des pays dits « à registres » qui disposent de répertoires détaillés de populations et de logements et d'assouplir ainsi l'utilisation de certains numéros identifiants.

Le cas irlandais qui a été présenté est d'ailleurs intéressant à cet égard. Je relève que dans d'autres pays tels que l'Estonie qui a connu une dictature communiste, le fait de disposer d'une grande base de données avec des numéros identifiants ne semble pas si grave en comparaison de la période soviétique. Or la France, qui n'a pas connu la même histoire, est encore très attachée à ce type de protections. Cette solution relève d'un choix politique qui doit passer par le Parlement et la CNIL ne fait qu'appliquer la législation en vigueur.

La seconde solution, qui me semble la plus intéressante étant donné que je suis à la tête de la Direction des technologies et de l'innovation à la CNIL, fait appel à la technique. Ainsi, le CSNS renvoie véritablement à cette démarche. Lorsque j'avais occupé une première fonction à la CNIL dans les années 2010, j'avais constaté que la CNIL était souvent interrogée sur des blocages. Pour effectuer un appariement de données basé sur le NIR, il fallait obtenir un décret en Conseil d'Etat, ce qui était complexe pour des chercheurs. La CNIL a alors lancé une réflexion tournant autour d'un code non signifiant dérivé du NIR.

Lorsque j'ai travaillé au secrétariat d'Etat chargé du Numérique, il a été question d'intégrer cette idée dans la loi pour la République numérique, de manière à simplifier l'accès des chercheurs aux données. L'appariement de données est effectivement crucial pour orienter la conduite des politiques publiques et favoriser leur efficacité. Il a fallu trouver un entre-deux qui puisse nous prémunir contre les risques liés à la protection des données personnelles. Nous avons alors intégré un article à cette loi, de manière à proposer la solution du CSNS, qui s'avère plus lourde qu'un emploi direct du NIR dans les appariements, mais qui permet de couvrir ces risques. Le numéro n'est plus identifiant et si une base de recherche s'évapore dans la nature, il ne sera pas possible de les rapprocher aisément des identités réelles, grâce aux mécanismes cryptographiques mis en place.

Parmi ces solutions techniques, il est également question de construire de grands hubs de données sécurisées et d'assumer l'existence d'un accès plus large de données, qui reste cependant soumis à un cadre exigeant de sécurité. Et c'est dans cette démarche que s'inscrit le CASD, dont le développement récent a été très bien accueilli par la recherche. Le CASD a notamment permis d'améliorer l'ergonomie du travail des chercheurs, ce qui constitue également un très bon point. Toujours dans cette démarche, nous retrouvons le projet du gouvernement du *Health Data Hub*. En matière de données de recherche, le législateur

souhaitait vraiment rassembler des données dans un lieu très sécurisé, comme en témoignent la loi de 2016 et celle de 2019, évoquées au cours de cette rencontre. Pour la CNIL, qui joue un rôle de régulateur, il est finalement plutôt rassurant de disposer d'un système de ce type, géré par des professionnels, plutôt que de disposer de bases de données qui se déplacent dans des laboratoires de recherche, sans que la sécurité soit toujours assurée.

Pour ce qui est du croisement entre les données de santé et les données sociales, je n'ai malheureusement pas de réponse pour le moment. Peut-être que la solution proviendra d'infrastructures telles que le CASD, qui a récupéré une grosse partie des données du SNDS l'an passé avec l'accord de la CNIL. Je pense qu'il existe aujourd'hui la possibilité de réaliser des rapprochements de données qui semblaient relativement impossibles dix ans auparavant. J'espère donc que nous parviendrons à conserver notre rythme dans le cadre de la compétition internationale, tout en maintenant un haut niveau de protection.

Chantal CASES

Nous reviendrons sur vos propos, lors du temps d'échange. Je vais maintenant donner la parole à Mark Hunyadi, professeur de philosophie morale et politique à l'université catholique de Louvain et membre du comité d'éthique d'Orange nouvellement créé, mais aussi du comité d'éthique commun de l'Institut national de la recherche agronomique (INRAE), du Centre de coopération internationale en recherche agronomique pour le développement (CIRAD), de l'Institut français de recherche pour l'exploitation de la mer (IFREMER) et de l'Institut français de recherche pour le développement (IRD). Il a également publié différents ouvrages, dont *Au début est la confiance,* paru en 2020, qui sera probablement évoqué dans nos discussions.

Tout d'abord, différents projets de répertoires et d'appariements ont été présentés ce matin. Il existe encore d'autres projets réalisés à l'international et qui n'ont pas été présentés. Il se trouve que, fort heureusement, ces projets fournissent une occasion pour les statisticiens d'approfondir leur réflexion sur leur éthique professionnelle. J'aimerais donc recevoir votre point de vue sur l'éthique des acteurs de la statistique publique.

De plus, puisque vous avez beaucoup travaillé sur la notion de confiance, je souhaiterais que vous nous offriez des conseils et que vous partagiez votre point de vue sur la manière dont la statistique publique peut maintenir la confiance existante du public, en particulier en matière d'appariements de données individuelles.

Mark HUNYADI

Bonjour à tous. Je remercie Chantal Cases et Françoise Dupont de m'avoir invité à cette table ronde. Je n'interviens pas en tant que statisticien, mais en tant que philosophe éthicien. J'ai donc le sentiment de pénétrer par effraction dans un club dont je ne partage pas tous les codes et les présupposés.

En matière d'appariements de données, les acteurs de la statistique tels que l'INSEE se trouvent devant un contexte évolutif, caractérisé par l'émergence de nouvelles sources de données et de nouveaux acteurs. Ce contexte entraîne une remise en question des pratiques et de l'identité des organismes de la statistique publique.

Or cette remise en question de l'identité constitue déjà en soi une question éthique. Elle ne renvoie pas à des problèmes d'ordre technique, mais bien d'ordre éthique. Il s'agit d'une

remise en question de votre rôle, de vos fonctions, de votre utilité, mais aussi des principes et des valeurs qui gouvernent ces nouvelles pratiques et qui guident vos actions.

Je pense que vous ressentez tous l'existence de cette question d'identité. Tout du moins, j'ai ressenti l'existence de cette préoccupation en interagissant avec vous. Or cette question d'identité renvoie à une éthique de l'identité. Je précise qu'il ne s'agit pas d'une éthique identitaire, l'identité renvoyant à une redéfinition de soi, étant entendu qu'une identité universaliste constitue également une identité.

Pour répondre plus précisément à votre question portant sur l'éthique, il est toujours intéressant de distinguer deux types de questions éthiques, de natures extraordinairement différentes.

Le premier niveau de question renvoie à ce que j'appelle la « petite éthique » et je vous prie de m'excuser d'emblée de cette gradation qui peut paraître péjorative. Cette petite éthique est centrée sur l'individu et elle est en jeu dans les questions que nous nous posons tous sur la sécurité des données, ou encore la protection de la vie privée. La CNIL s'intéresse exactement à cette catégorie d'éthique.

Dans nos sociétés, cette éthique est concrétisée par l'éthique des droits de l'homme, l'élément suprême de l'architecture normative de nos sociétés occidentales avancées. Je nomme donc « petite éthique » cet ensemble de problèmes qui relèvent de la défense et de la protection des droits et des libertés individuelles. Bertrand Pailhès a illustré à merveille le fait qu'il soit possible de résoudre en principe ce type de problèmes grâce au droit et à la technique, par exemple, par le biais de l'établissement de garde-fous.

Dans le cadre de cette petite éthique et de notre réflexion sur l'appariement des données, j'ai été assez inquiété ce matin par le contenu des différentes présentations. Je m'interroge vraiment sur notre capacité à concilier le principe de minimisation de données avec celui de la maximisation des appariements. Ainsi, ce que j'ai entendu me fait froid dans le dos, et en particulier pour ce qui concerne RESIL, quels que soient les garde-fous que l'on pourrait imaginer. J'ai bien compris le principe des différents garde-fous juridiques et techniques que vous avez présentés. Néanmoins, j'attire votre attention sur le fait que ces garde-fous peuvent aussi être retirés d'un trait de plume, à partir de modifications de lois, de contexte normatif, ou de gouvernement. Je ne peux donc que m'inquiéter de cette éventualité.

Toutefois, je considère que cette éthique reste une petite éthique, centrée sur l'individu, et qu'il ne s'agit pas d'une éthique globale. Le deuxième niveau d'éthique que je nomme « éthique globale » ou « grande éthique » surgit lorsque nous nous interrogeons sur la finalité des appariements, sur leur rôle et sur leur utilisation globale dans le cadre d'un projet de société dessiné par l'utilisation des statistiques. En somme, cette éthique apparaît lorsqu'il s'agit de réfléchir au sens de cet immense projet que constitue le projet statistique.

D'une manière générale, nous nous posons une question de « grande éthique » dès lors que nous nous interrogeons sur le monde où nous mène cette volonté de quantifier intégralement le monde. Cette volonté instaure un rapport au monde médiatisé par les chiffres.

Et nous connaissons toute l'importance que cette médiation par les chiffres occupe dans la sphère politique. L'importance de cette médiation est d'ailleurs tout à fait normale et compréhensible, puisque nos sociétés modernes et complexes requièrent tout à fait logiquement une gestion par les statistiques. C'est pourquoi la statistique est née au XIX^{e} siècle,

non pas à partir des mathématiques, mais à partir d'une branche de la politique sociale, sur la base d'une demande provenant de la gouvernance et de la société.

La statistique est régie par une valeur suprême qui est la production d'un certain type de connaissance, à savoir ce qui est mesurable et objectivable. Or ces intérêts de connaissance sont en permanence menacés par une instrumentalisation du pouvoir. En effet, s'il existe un intérêt de connaissance pour la recherche, ces intérêts suscitent également l'émergence d'un projet de gouvernementalité. Nous parlons beaucoup actuellement du concept de la gouvernementalité algorithmique portée par les GAFAM et je crois qu'il est également possible de parler de gouvernementalité statistique.

Ce projet de gouvernementalité statistique mobilise des appariements et renforce les problèmes de sécurité liés à la petite éthique. Mais ce projet s'inscrit également dans la grande éthique, étant donné qu'il s'agit d'un projet de société qui fait appel à la statistique pour contrôler le pilotage des politiques publiques et aider à la programmation de ces politiques.

Sans vouloir donner une vision purement négative de ce projet, je pense que dans nos sociétés complexes, il est inéluctable. Et ce caractère inéluctable peut s'avérer angoissant. Or, dans le cadre de ce pilotage statistique, on s'adresse aux citoyens non pas comme à des sujets libres, dotés d'une inventivité, d'une imagination et d'une volonté politique, mais plutôt comme à des êtres que l'on peut piloter et programmer. Je force un peu le trait, mais dans le cadre de cette gouvernementalité statistique de la politique, la statistique sert alors à effectuer cette espèce de programmation.

Je suis convaincu qu'il existe un intérêt légitime de la statistique. Cependant, je constate un phénomène très ambivalent. Ainsi, je veux montrer qu'il existe à la fois la nécessité requise par la complexité de notre société de répondre à cet intérêt de connaissance statistique ; mais que dans le même temps, cet intérêt nous place dans une proximité très problématique de l'apprivoisement et surtout de la domestication du réel et des personnes que la statistique aide à gouverner.

Une grande responsabilité est donc associée aux statistiques, qui façonnent les représentations sociales des gens, qui cherchent la preuve par les chiffres. Ce matin, Bercy était en grand sourire, car l'INSEE venait de montrer que la croissance atteignait 7 % en 2021. En France, nous sommes accrocs aux chiffres qui façonnent les représentations sociales et la manière dont on se représente les individus qui font l'objet des statistiques. Ces individus sont alors considérés plutôt comme des êtres à piloter, plutôt que comme des volontés libres et autonomes. Or ces considérations ne sont pas sans poser de problèmes dans un contexte démocratique.

Ces questions ne renvoient pas à la petite éthique. En parlant de ce projet de société qui transpire à travers ce projet de gouvernementalité statistique, je renvoie plutôt à une conception générale de la société. Il me semble que ces questions renvoient aux vrais problèmes éthiques, étant donné que ceux posés au niveau de la petite éthique peuvent en principe être résolus plus ou moins aisément d'une manière juridique et technique. En revanche, en matière de « grande éthique », nous faisons face à un problème d'une toute autre nature, pour lequel aucun comité technique ou comité d'éthique n'est prévu. En effet, nous ne retrouvons dans l'agenda d'aucun comité d'éthique ces questionnements liés au projet de société qui se dessine autour du numérique et de la statistique. Ils ne s'intéressent pas à ces grands mouvements qui nous emportent actuellement à la manière d'une lame de fond et sur lesquels nous n'avons pas d'emprise. Aucun comité d'éthique n'est là pour

réguler ces projets, ce qui pose un véritable problème. Toute notre éthique est centrée autour de la petite éthique et nous tentons de les régler, sans comprendre que nous sommes également emportés par les problèmes de la grande éthique.

Pour répondre à votre question portant sur la confiance, je commencerais par préciser qu'il convient déjà de savoir ce qu'est la confiance. Comme vous l'avez rappelé, mon dernier livre s'intéresse à ce concept. Tout le monde entend ce mot, sans forcément savoir à quoi il se rapporte.

En quelques mots, la confiance est toujours reliée à des attentes que l'on peut nourrir à l'égard de quelque chose. Nous sommes assis sur des chaises et nous nous attendons qu'elles nous soutiennent. Avoir confiance, c'est parier sur le fait que ces chaises pourront nous soutenir, celles-ci pouvant tout à fait s'effondrer.

Il en est de même avec les personnes, pour qui nous nourrissons des attentes liées à des comportements. Faire confiance à quelqu'un consiste à s'attendre à ce qu'il se comporte d'une certaine manière. Et il en est encore de même pour les institutions. Ainsi, nous entendons souvent l'expression « nous avons confiance dans la justice de notre pays », qui signifie que nous avons confiance dans le fait que la justice et ceux qui la représentent agiront conformément aux idéaux de la justice, tels que l'impartialité ou encore la multiplication des points de vue.

Ainsi, avoir confiance dans une institution signifie que l'on s'attend à ce que l'institution en question soit à la hauteur des attentes qu'elle a elle-même produit. En prenant l'exemple de l'INSEE, nous nous attendons donc à ce qu'elle produise une connaissance saine. Pour vérifier que cette connaissance soit saine, il faut vérifier qu'elle soit conforme à des standards mis au point par des statisticiens. Avoir confiance en une institution, consiste donc à parier sur une attente normative qui soit conforme à ce qu'on attendait. Et c'est à vous de savoir ce qu'on attend de l'INSEE.

Chantal CASES

Je vous remercie de nous avoir un peu bousculés. Je pense que votre présentation est très utile dans la réflexion que nous menons sur les appariements.

Je laisse maintenant le soin à Maryse Artiguelong d'effectuer la dernière présentation de cette table ronde. Elle vient du monde de l'informatique et elle est invitée en tant que vice-présidente de la Ligue des droits de l'homme et de la Fédération internationale des droits de l'homme. Elle anime le groupe de travail de la Ligue des droits de l'homme « liberté et technologie de l'information et de la communication ». Par ailleurs, elle est membre de l'observatoire des libertés et du numérique. Nous sommes donc particulièrement intéressés par son point de vue.

Tout d'abord, quelles réflexions vous inspirent les projets d'appariements de données et de répertoires présentés aujourd'hui et qui concernent la statistique publique ? Quels avantages et quels risques percevez-vous pour les personnes concernées par ces données ? Comment vous positionnez-vous sur les points qui relèvent des questions de la « grande éthique » ?

Enfin, étant donné que nous tentons d'organiser une information transparente et une concertation sur ces projets, j'aimerais que vous nous fassiez des suggestions à cet égard,

notamment sur les procédures qu'il serait nécessaire de mener en amont et en aval des projets. De plus, le CNIS est un lieu typique de cette concertation et il en existe d'autres. Je souhaiterais donc savoir si vous estimez que ce cadre est suffisant et si vous avez des idées à proposer pour l'améliorer.

Maryse ARTIGUELONG

Bonjour à tous. Je remercie l'INSEE d'avoir invité la Ligue des droits de l'homme. Il est vrai que nous avons une habitude ancienne d'opérer avec vous des coopérations ponctuelles sur certaines interrogations. Je signale que la Fédération internationale des droits de l'homme commémore cette année ses cent ans. Elle avait été créée en 1922 par les Ligues françaises et allemandes qui sentaient venir des jours difficiles sans parvenir à empêcher leur arrivée.

Il est vrai que la place de la Ligue des droits de l'homme dans la lutte contre les discriminations et dans la protection des données personnelles et de la vie privée n'est pas aussi importante que nous l'aurions souhaité, mais il s'agit là d'une partie de nos combats quotidiens. J'ai appris énormément de choses aujourd'hui et je remercie mon collègue qui m'a poussé à accepter cette invitation. Ainsi, j'aurais appris que je travaille au quotidien autour de la « petite éthique ».

Plus sérieusement, étant donné que je travaille depuis longtemps sur la question de la protection de la vie privée et des données personnelles, pour moi les appariements renvoyaient plutôt aux interconnexions de fichiers de police et de justice. Nous avions d'ailleurs mené une action récemment contre l'appariement entre le fichier de traitement des signalements pour la prévention de la radicalisation à caractère terroriste (FSPRT) et le fichier des personnes hospitalisées sans consentement en psychiatrie (HOPSYWEB). Ce projet d'appariement nous paraissait assez abominable, car il touche des questions de santé et génère une suspicion lourde et scandaleuse sur des personnes internées d'office.

Néanmoins, j'ai été impressionnée par l'éthique, la rigueur et le suivi de la doctrine, qui régissent tous les projets qui ont été présentés. Pourtant, la masse de données traitées me paraît vertigineuse. En fait, je ne sais pas si tout le monde réalise que chaque personne se trouve dans ces énormes fichiers administratifs.

Ainsi, même si de nombreux verrous ont été présentés aujourd'hui, je peux m'inquiéter de les voir sauter, d'autant plus qu'il a été constaté qu'ajouter une question dans une enquête ne s'avère pas nécessairement neutre. Même si je sais que l'Union européenne surveille les instituts de statistique publique de chaque pays de l'union, je m'inquiète de l'éventualité qui conduirait un gouvernement à opérer un changement dans le cadre éthique de ces appariements de données.

Par ailleurs, je rejoins les propos du député Jean-Noël Barrot qui s'inquiétait d'un éventuel remplacement des instituts de statistique par les GAFAM qui obtiennent rapidement des données plus ou moins fiables à partir de leurs systèmes de big data. Si ce cas se présentait et que ces instituts cessaient d'être financés, nous risquerions de perdre la fiabilité de l'information et ne plus pouvoir répondre à nos questions à partir de vérités.

En outre, je relève que la présentation de RESIL a suscité de nombreuses questions qui rejoignent certaines de mes inquiétudes.

Les appariements et les répertoires que vous avez décrits présentent des avantages. Il est vrai qu'il est utile pour la population de bénéficier de politiques publiques basées sur des statistiques publiques aussi construites et fiables que celles que vous avez présentées. A cet égard, je suis rassurée. Néanmoins, nous pouvons nous demander si la statistique doit être le seul fondement des politiques publiques.

Effectivement, les données des statistiques publiques fournissent des résultats relatifs à un instant donné. Néanmoins, elles ne peuvent pas prévoir l'imprévisible. Qui aurait dit que nous serions confinés durant deux ans et condamnés pour beaucoup au télétravail ?

Je réponds à présent à votre seconde question, portant sur les concertations et les échanges qui pourraient avoir lieu en amont des projets d'appariements. Finalement, nous avons appris beaucoup de choses aujourd'hui, grâce aux partages qui ont pu s'opérer dans le cadre de cette rencontre, et nous sortons enrichis de cette journée. Je pense qu'il faudrait organiser des réunions plus courtes et plus restreintes, mais avec des juristes et des techniciens qui ne proviendraient pas des instituts de statistiques, mais aussi les ouvrir à la société civile, qu'elle soit organisée ou non.

A cet égard, je note que la Commission nationale consultative des droits de l'homme (CNCDH) travaille actuellement sur un avis relatif à l'intelligence artificielle et aux droits de l'homme. Dans ce cadre, de nombreux experts juridiques sont entendus, mais aussi d'autres acteurs. Et il se trouve que parmi eux, nous avons entendu ATD Quart monde, qui a relaté l'organisation d'ateliers avec des personnes en difficulté dont le niveau de connaissance dans le numérique n'était pas très élevé. Il s'est avéré très intéressant d'entendre cette voix et j'estime que nous avons tout à gagner à entendre tout le monde. Cette opération est peut-être compliquée, mais elle est forcément positive.

Enfin, je pense qu'il faut oser communiquer au grand public. Certes, celui-ci sait que nous sommes recensés de temps en temps, mais il ignore le détail des différentes réalisations du service statistique public. Tout le monde entend parler des chiffres de l'INSEE, mais personne ne sait d'où ils proviennent.

Echanges

Chantal CASES

Je précise que ATD Quart monde et d'autres acteurs de la société civile participent aux réunions des commissions du CNIS, dans le cadre des concertations réalisées en amont des opérations statistiques.

Il est temps de recueillir les questions en provenance du tchat et de la salle.

Des invités à distance

Pourriez-vous apporter des précisions sur le gradient de sensibilité des traitements statistiques présenté par Eric Rancourt ? En particulier, je souhaiterais comprendre le lien entre cette échelle et la non-réponse aux enquêtes, ou encore son lien avec l'opposition aux appariements.

Prévoit-on de réaliser une enquête ou une méta-analyse sur les données disponibles, de manière à savoir ce qu'en pensent les premiers intéressés, à savoir les personnes interro-

gées ? Que sait-on aujourd'hui sur leur confiance ou sur leur défiance vis-à-vis des garanties sur la vie privée qui sont affichées par les acteurs de la statistique publique ? Enfin, a-t-on analysé les motivations du refus de réponse ? Que sait-on sur ce qui les inquiète ou ce qui les rassure particulièrement ?

Pour établir la confiance, quelques expériences ont été menées, notamment dans les pays nordiques ou anglo-saxons. Celles-ci mettent en jeu des comités de citoyens reliés à de grandes enquêtes qui intègrent des appariements ou des données sensibles, ou qui recourent à de grands centres de données. Pouvons-nous y voir une piste pour favoriser la confiance du public ? Où placer ces comités entre la petite et la grande éthique ?

Eric RANCOURT

Je vais répondre à la première question, tout en fournissant des éléments de réponse aux deux autres questions.

J'ai fait allusion au lien entre les non-réponses aux enquêtes et le projet d'échelle de sensibilité qu'est en train d'élaborer Statistique Canada. Cette échelle devrait permettre de situer différents sujets, de manière à guider nos efforts en matière de transparence, d'éthique, d'équité, de sécurité, ou encore de protection de la vie privée. Il ne s'agit pas pour autant d'opérer des traitements non sécuritaires ou non éthiques. Néanmoins, nous manquons de balises pour nous assurer du respect de ces principes. En fonction de la sensibilité mesurée, nous pourrions par exemple être amenés dans certains cas à préférer recueillir des données agrégées au lieu de données de niveau micro. Nous pourrions aussi être conduits à réduire la taille d'un fichier en nous basant sur un petit échantillon plutôt que sur un fichier administratif complet. Il serait encore possible de réduire le nombre de variables.

Ce projet d'échelle de sensibilité fait suite entre autre aux discussions qui avaient entouré notre projet d'étudier les transactions bancaires et les données des crédits pour tenter d'améliorer la compréhension des ménages en situation précaire. Ce projet revêtait une finalité sociale très importante, mais la société n'a pas été disposée à conférer un mandat social pour permettre la récolte de microdonnées exhaustives. La prise en compte de cette sensibilité repositionne donc la méthode ou la nature d'un projet.

Je précise que pour réaliser notre échelle de sensibilité, nous chercherons à comprendre pourquoi les non répondants ne répondent pas aux enquêtes. Nous nous intéressons aussi au point de vue des enquêteurs qui ont interrogé ces personnes. Parallèlement, nous organisons des groupes de discussion comprenant des non répondants, de manière à obtenir des éléments qualitatifs. Des chercheurs en psychologie spécialisés dans la participation sociale nous aident à comprendre ces non-réponses.

Enfin, notre ouverture sur la société dépasse le simple fait de publier des informations sur notre site internet. En effet, nous tentons de contacter les citoyens de différentes façons. Nous avons notamment mis en place un comité d'éthique externe. Nous essayons encore de développer des enquêtes de satisfaction ou de confiance, en interrogeant le public sur les facteurs qui semblent influer le plus sur sa confiance envers Statistique Canada, mais aussi sur les moyens qui pourraient permettre de l'améliorer.

Chantal CASES

Peut-être que d'autres participants de cette table ronde souhaitent répondre à ces questions. Je céderais probablement la parole à certains membres de l'INSEE pour évoquer les questions d'enquête de satisfaction ou encore celles touchant à l'analyse des non-réponses. Mais comme le terme « confiance » a été prononcé, je pense que Mark Hunyadi souhaiterait réagir.

Mark HUNYADI

Il est certain que les comités citoyens augmentent la pertinence politique des statistiques. Je ne connais pas ces expériences des pays nordiques que vous citez et il faudrait notamment vérifier dans quelle mesure ces citoyens prennent part aux décisions.

Par ailleurs, dans ce nouveau contexte de multiplication des sources de données, provenant notamment d'acteurs privés tels que les GAFAM et où les instituts de statistique publique sont concurrencés en permanence, je tiens à souligner l'existence d'une carte à jouer importante pour la statistique publique.

Les GAFAM ne font qu'enregistrer des comportements qui ont effectivement eu lieu et ils ne le font pas dans un intérêt cognitif, mais dans une optique prédictive. En effet, ils souhaitent prédire des comportements futurs, essentiellement à des fins commerciales. Or ils ne font que tirer des conclusions *ex post* de comportements enregistrés, car l'outil numérique employé fonctionne de cette manière.

De la sorte, face aux GAFAM, les organismes de statistique publique disposent d'un atout extraordinaire, étant donné qu'ils ne se limitent pas à enregistrer des faits établis, tels que des faits administratifs. En effet, ils effectuent également des enquêtes. C'est pourquoi, il serait intéressant de revaloriser l'enquête, puisque celle-ci permet non seulement de savoir ce que font les individus, mais aussi ce à quoi ils aspirent et ce qu'ils aimeraient faire. Or aucun dispositif numérique des GAFAM n'est en mesure d'obtenir ces informations.

De la sorte, un paradigme méthodologique davantage orienté sur l'enquête pourrait augmenter la pertinence démocratique des statistiques. D'une certaine manière, nous rapprocherions véritablement le point de vue du citoyen du point de vue surplombant de la personne qui ne fait que mesurer des phénomènes. Il s'agit de demander aux personnes ce à quoi elles aspirent, dans différents domaines, tels que la mobilité professionnelle. Et seule l'enquête permet d'appréhender ces aspirations. Il me semble que nous pouvons voir ici une clé importante pour offrir toute sa place à la statistique dans ce contexte dominé par l'enregistrement numérique des données.

Chantal CASES

Je vous remercie de réhabiliter l'enquête et ses questionnements sur les comportements et les aspirations. Je rappelle que nous avons souligné au cours de cette rencontre et dès son introduction la complémentarité entre les enquêtes et les fichiers administratifs.

Maryse ARTIGUELONG

Je précise que les GAFAM étudient le passé pour influencer le comportement des personnes. Comme le disait Etienne Klein, les algorithmes servent à prédire l'avenir à condition qu'ils ressemblent beaucoup au passé.

Chantal CASES

Je peux peut-être apporter une précision sur la manière dont on mesure la confiance ou la satisfaction des populations, bien que Jean-Luc Tavernier devrait probablement aborder ce point dans sa conclusion. A cet égard, je relève l'existence de quelques enquêtes de satisfaction. En outre, il me semble que les non-réponses sont généralement analysées en amont des enquêtes et en aval, lors de la préparation des analyses.

Christel COLIN

En guise de complément, je souhaite rebondir sur un élément indiqué par Eric Rancourt. Statistique Canada a été contraint de réagir à la forte chute du taux de réponse dans les enquêtes menées auprès des ménages. Or, clairement, nous n'observons pas ce phénomène dans les enquêtes de l'INSEE. Il peut exister une érosion du taux de réponse, mais celui-ci ne chute pas. Je suppose que cela tient aux efforts réalisés pour convaincre les ménages de répondre, notamment en proposant des questionnaires adaptés et dont la longueur reste limitée.

En outre, nous disposons de retours d'ordre qualitatifs touchant les réactions des enquêtés sur les questions qui leur sont adressées. Ils montrent que les personnes qui ne souhaitent pas répondre ne répondent pas le plus souvent pour une question de temps et non pour des motifs ayant trait à un manque de confiance ou à une défiance. En revanche, certains enquêtés invités à répondre à une enquête par internet demandent comment leur adresse électronique a été trouvée, ce qui témoigne de l'existence d'une certaine sensibilité sur cet aspect.

Enfin, vous venez de citer l'existence d'enquêtes de satisfaction touchant des indicateurs produits par l'INSEE et sur l'INSEE en général. Néanmoins, je ne suis pas la mieux placée pour fournir des précisions sur ces enquêtes. Quoi qu'il en soit, ces enquêtes de satisfaction, qui font notamment appel à des internautes, sont menées chaque année et permettent de suivre la confiance du public envers les indicateurs de l'INSEE.

Jean-Luc TAVERNIER, directeur général de l'INSEE

Je précise que je n'ai pas prévu de parler de la question de la confiance dans ma conclusion. Il existe effectivement des enquêtes de satisfaction visant notamment à mesurer la confiance en notre institution. D'un côté, nous interrogeons des internautes qui se rendent sur le site de l'INSEE, tout en sachant qu'il s'agit d'un public acquis. De l'autre côté, nous interrogeons la population générale.

Et dans ce second cas de figure, nous ne nous attendons pas à des résultats formidables. En tant qu'expert officiel du domaine public, nous sommes frappés de deux infamies, à savoir celle liée à l'expertise et celle associée au fait d'appartenir à la sphère publique et officielle.

Pour autant, la confiance que nous mesurons de manière plus générale, indicateur par indicateur, n'est pas associée à ces types d'infamies. En effet, lorsque nous posons la question « Croyez-vous aux indicateurs statistiques publics officiels ? », les réponses ne sont pas très positives. En revanche, lorsque nous contextualisons notre question en précisant que l'INSEE produit chaque mois tel ou tel indicateur et en en interrogeant le public sur sa confiance en ces indicateurs, les réponses s'avèrent bien plus favorables.

Ainsi, dans le cadre de ma fonction, je considère qu'il existe une véritable difficulté à conserver la confiance dès lors qu'il existe, partout dans le monde et notamment dans ce pays, un mouvement général où l'on tend à adopter une posture de défiance devant tout ce qui ressemble à une expertise ou à une parole officielle.

Un invité à distance

Dans le cadre de ce projet global de "gouvernementalité statistique", comment interpréter l'absence d'exemples d'appariements concernant le champ de la transition énergétique, qui permettraient de mieux suivre la rénovation des logements ou du parc automobile, ou encore la consommation énergétique des ménages ou des entreprises, ainsi que les aides qui y sont associées ? Les réalisations d'appariements semblent révéler une hiérarchie dans les priorités politiques, construites de longue date et en relation avec les résistances des acteurs.

Bertrand PAILHÈS

Je vais présenter un point de vue qui n'est pas celui de la CNIL, mais qui provient de mon expérience, ainsi que des discussions entreprises en 2016 dans le cadre de l'élaboration de la loi pour une République numérique. Dans ce cadre, nous avions été notamment guidés par l'idée consistant à permettre un accès à des bases de données administratives à des chercheurs, sans nous limiter à des « chercheurs officiels ». Il me semble que le CASD a obtenu récemment un accord avec la Banque de France. Or, il y a cinq ou six ans, pour effectuer des recherches sur la Banque de France, il était nécessaire d'obtenir l'agrément de cette institution.

Je pense que la confiance se fonde également sur une certaine pluralité de la recherche et des regards, dans le cadre du ciblage des questions et des appariements. En effet, il se peut que des structures historiques ou des pouvoirs puissent orienter de nombreuses études autour de la croissance et très peu sur la rénovation des logements. C'est pourquoi il faut sans doute parvenir à construire un cadre qui permette justement à d'autres acteurs d'investir le terrain, d'accéder aux données et d'effectuer des recherches sur d'autres sujets. De cette manière, nous pourrions lutter contre les deux infamies évoquées par Jean-Luc Tavernier, liées à l'incarnation des pouvoirs publics.

Enfin, je pense aussi que la solution se trouve aussi dans le fait de pouvoir challenger le secteur public par une recherche indépendante qui serait issue d'universités publiques ou d'autres provenances et qui permettrait de faire progresser une vision plurielle du monde.

Chantal CASES

Selon vous, l'open data généralisée, ouverte dans de bonnes conditions, favoriserait-elle la confiance ?

Bertrand PAILHÈS

Je ne parle pas au nom de la CNIL, mais je peux dire qu'il s'agit là effectivement de l'une des idées portées par la loi pour une République numérique, ou encore par la plateforme data.gouv.fr lancée antérieurement par François Fillon. Mais je pense aussi à l'accès aux données par d'autres voies que l'open data. A cet égard, je pense notamment à des travaux sur la fiscalité, réalisés par certains chercheurs. Ainsi, je me demande comment les chercheurs peuvent avoir accès aux données, à partir du CASD ou par d'autres voies, de manière à ce qu'ils puissent effectuer des analyses qui ne sont pas réalisées ailleurs.

Chantal CASES

Les statisticiens publics sont déjà habitués à se faire challenger par la recherche et ils se trouvent tout à fait en accord avec ce principe. La question de l'ouverture aux données au-delà de la sphère de la recherche mérite effectivement d'être posée, en distinguant ses dangers et ses avantages.

Bertrand PAILHÈS

Tout à fait, à la CNIL, nous sommes plutôt prudents en matière d'open data. Je pense qu'il existe effectivement des dangers et des avantages à cette ouverture. Cette ouverture permettrait d'aboutir à une forme de contextualisation. Il serait possible de faire comprendre comment les chiffres ont été obtenus et redonner ainsi de la confiance. De plus, la mise en open data des données signale au public qu'il peut toujours reprendre les chiffres et produire sa propre analyse, dans le cas où il ne croirait pas en la fiabilité de l'analyse produite par des chercheurs ou des statisticiens. Néanmoins, cette démarche n'est pas possible pour de nombreuses données et en particulier lorsque des données personnelles sont mobilisées.

Chantal CASES

Je pense aussi que la pandémie a bien mis en évidence l'importance de cette question de la confiance.

Patrice DURAN

Je souhaite vous faire part d'un ensemble de réactions. Les appariements complexifient les questions liées à la statistique. De la sorte, il est important, pour ne pas dire vital, que l'INSEE et le CNIS jouent un rôle pédagogique. Nous faisons ce que nous pouvons en la matière, mais nous devons aller plus loin dans ce sens.

En effet, trop d'erreurs existent qui sont parfois publiées sous des plumes pourtant savantes. J'en veux pour preuve le livre *La gouvernance par les nombres* d'Alain Supiot, membre du collège de France. Ce brillant juriste de droit du travail développe une attaque de la statistique publique sans en avoir une connaissance précise, d'où bien des erreurs viennent encombrer sa prose que Jacky Fayolle, ancien de l'Insee, a, à juste titre, sévèrement critiqué par la complaisance que cette thèse entretient avec des thèses idéologiques sur le rôle des « chiffres » dans le gouvernement des hommes ! Or, le manque de connaissance de beaucoup sur ce qu'est réellement la statistique publique est aujourd'hui un vrai problème. Les gens confondent trop souvent données, faits et statistiques, et malheureusement une telle méconnaissance est encore très, trop forte dans l'administration publique

dans son ensemble, qu'il s'agisse de l'État comme des collectivités territoriales. Il faut donc être prudent lorsque nous parlons de gouvernance par les nombres, de plus il ne faut pas confondre les statistiques et les usages qui peuvent en être réalisés. Il existe ainsi un véritable problème de formation sur le sujet qui interroge autant l'éducation nationale que la formation des fonctionnaires.

Dans le cadre de cet effort de pédagogie de l'INSEE et du CNIS, nous avons invité différents acteurs tels qu'ATD Quart monde. Dans ce cadre, nous organisons également des groupes de travail, dont le prochain, lié à une demande du Défenseur des droits, devrait porter sur les discriminations.

De plus, il est clair que la réalité de ce qu'est statistique publique est trop méconnue, même des acteurs dont on pourrait penser qu'ils sont mieux informés que d'autres. Ainsi, nous avons été interrogés par des membres de la Haute Autorité de Santé (HAS) au sujet de la santé de l'enfant. Or, même s'ils étaient conscients que ce sujet n'était pas qu'un sujet strictement médical, ils ont tout de même été étonnés d'apprendre que cette question était beaucoup plus documentée qu'ils ne le pensaient, lorsque nous leur avons exposé les différentes données émanant des différents SSM. Finalement, c'est bien l'ignorance de ce qu'est la statistique publique et de ce qu'elle fait qui constitue aujourd'hui qui n'est plus supportable et qu'il faut absolument prendre en compte.

Si l'on veut combattre la défiance que beaucoup manifestent à l'égard de l'action publique, nous devons savoir expliquer et informer sur ce que nous faisons en matière de statistique publique. Aujourd'hui la confiance est d'autant plus nécessaire que le monde est plus complexe. Le juriste et sociologue allemand Niklas Luhman a publié un livre *Vertrauen : ein mechanismus der reduktion sozialer komplexität* (« La confiance, un mécanisme de réduction de la complexité sociale ») dans lequel il expliquait à juste titre que la confiance constituait un moyen de gérer des problèmes complexes. Lorsqu'on ne peut pas maîtriser une question tout simplement parce que nous n'avons pas la formation suffiante pour comprendre la politique mise en œuvre pour la gérer, ne pas avoir confiance devient problématique. On a ainsi pu voir avec le mouvement des gilets jaunes en quoi la défiance à l'égard du politique pouvait avoir des conséquences graves sur le fonctionnement social. C'est aussi la confiance dans les institutions qui se joue ici.

Or cette question renvoie à la « petite éthique » que vous avez présentée. Il se trouve que les problèmes posés par cette petite éthique se sont présentés dans le cadre de la pandémie. En effet, les refus de porter des masques ou de se faire vacciner s'appuyaient sur l'invocation de la liberté individuelle. C'est bien toute la question du rapport entre les droits de l'homme et les droits du citoyen qui se joue ici, car il révèle des enjeux dont la portée est loin d'être la même.

Il est d'ailleurs intéressant de voir comment le droit administratif français a rencontré ce problème de la petite éthique et intégré à sa manière la défense de l'individu propre au libéralisme politique et la préservation de l'intérêt général. Ainsi, le droit administratif distingue la responsabilité « sans faute » de celle « pour faute ». La responsabilité pour faute correspond assez classiquement à la logique de production des organisations publiques où des usagers peuvent demander réparation d'un dommage causé par un mauvais fonctionnement de nature organisationnelle. Il est clair bien évidemment que l'accroissement de l'intervention de l'État et donc de ses administrations multiplient les opportunités de litiges que ce soit suite à des erreurs de gestion ou à l'inapplication des lois et décrets. L'administration a une obligation sinon de résultat tout au moins d'effectivité.

Intéressante et complexe est la responsabilité sans faute dont la logique pourrait être une préfiguration d'une conception moderne de la responsabilité du fait de l'action publique.

Et la responsabilité sans faute apparaît en quelque sorte comme une manière de combiner la puissance de l'État avec la défense des droits de l'homme. L'enjeu ici n'est pas celui de la réparation, mais celui du dédommagement. Une action publique d'intérêt général peut avoir des conséquences dommageables pour des individus. À ce titre, ils méritent d'être dédommagés sans pour autant que soit rejetée la politique mise en œuvre par les pouvoirs publics. Ce fut ainsi tout l'enjeu de l'arrêt du Conseil d'État dit « Ville Nouvelle-Est » dans lequel le Conseil d'État défendu la théorie du bilan dans une situation où une ville avait procédé à des expropriations pour la création d'un complexe universitaire.
La question de l'action publique commande donc à ce que soient conjugués l'individuel et le collectif dans une réflexion tant sur les finalités de l'action que sur leurs conséquences. On peut probablement regretter que le droit de la responsabilité administrative ne soit pas porteur, ou tout au moins représentatif, d'une doctrine moderne de l'action publique. C'est d'ailleurs tout l'enjeu aujourd'hui des évaluations d'impact ex ante, même si le Conseil d'État n'a pas été satisfait par la qualité de ces dernières !

J'ai été président du Groupement d'intérêt scientifique (GIS) Démocratie et participation, où j'ai pu également constater l'extrême complexité du rapport entre ces deux termes. La re-composition des registres de justification du pouvoir politique crée de nouvelles contraintes dans l'exercice même du pouvoir. La revendication de légitimité de nos gouvernants ne peut plus se satisfaire de la seule légalité de leurs actes indépendamment de leur portée. Quand l'efficacité et la performance sont devenues dans la plus large part des États modernes les mots d'ordre d'une doctrine d'action publique, la mise en évidence des résultats de l'action publique s'impose et la statistique publique en est une des voies importantes. Comme le dit Pierre Rosanvallon, c'est bien la « démocratie d'exercice » qu'il faut inventer. Malheureusement le thème de la participation tel qu'il trouve à s'incarner dans la notion de pluralisme n'est pas exempt de lourdes ambiguïtés. Mais ce n'est le problème qui nous occupe ici.

Par conséquent, la question de la petite éthique est d'une complexité terrible. Pourtant, s'il est clair que nous avons besoin de connaissance pour agir efficacement en matière d'action publique, des considérations relevant de cette petite éthique peuvent potentiellement freiner les connaissances les plus pertinentes. C'est bien là tout le débat sur la question des statistiques ethniques. Certes, nous devons nous mobiliser contre des usages malfaisants de ces statistiques, mais nous devons également connaître les personnes qui peuplent notre environnement. Comme l'a évoqué, Jean-Noël Barrot, si nous souhaitons bénéficier de politiques efficaces, il faut savoir de quoi est fait le monde.

Chantal CASES

Je ne peux pas clore cette table ronde sans laisser un petit temps de parole à Mark Hunyadi, s'il souhaite réagir, étant donné que Patrice Duran vient d'évoquer les questions de la grande et de la petite éthique.

Mark HUNYADI

J'approuve tout ce qu'a dit Patrice Duran, à l'exception de ses commentaires sur Niklas Luhman et Alain Supio, mais il s'agit là vraiment de points de détail.

Chantal CASES

Vous pourrez discuter de ces détails après la clôture du séminaire. Je constate qu'un invité souhaite encore poser une question.

Un invité à distance

La petite éthique comprend des éléments qui donnent des garanties à la société qui sont internes à l'INSEE, à savoir l'ensemble des décisions internes qui s'y prennent. Elle comprend aussi des garanties provenant de processus externes et qui se tiennent dans le cadre du CNIS, ou encore du Conseil d'Etat. Or, lorsque l'on passe du NIR au CSNS, on bascule entre ces garanties internes et externes. De ce fait, comment est pensée aujourd'hui cette articulation, et comment tient-elle compte des fragilités et des forces de ces deux types de processus ?

Chantal CASES

Je ne suis pas sûre que cette question s'adresse spécifiquement aux grands témoins de cette table ronde.

Mark HUNYADI

Je ne saisis pas les enjeux techniques de cette question qui me semblent néanmoins importants. En tout cas, je précise que lorsque je parle de « petite éthique », je ne veux pas dire qu'elle n'est pas importante, mais simplement qu'elle est petite.

Sylvie LAGARDE

Le CSNS interne aux statistiques publiques est encadré juridiquement et permet de simplifier les appariements. Nous pourrions travailler quasiment en interne juridiquement, tout en veillant au respect du RGPD. Mais je considère qu'il est important de porter la question du CSNS hors de l'INSEE, d'où l'intérêt d'en parler au CNIS, de le rendre visible et d'échanger au sujet de son usage, au-delà de la communauté des statisticiens, avec l'ensemble des utilisateurs. Il s'agit d'un point très important de notre travail.

Chantal CASES

Je pense aussi que ce point est essentiel. Au cours de cette table ronde, nous avons entendu parler de différents comités, d'éthique, d'une communication à destination du grand public à renforcer, ou encore d'une ouverture plus grande à la société civile. Nous avons donc beaucoup de travail à réaliser. Je vous remercie beaucoup d'avoir participé à cette table ronde, pour discuter de sujets qui ne vous étaient pas toujours familiers, tout du moins pour deux d'entre vous. Je remercie également tous les participants.

CONCLUSION
Jean-Luc TAVERNIER, directeur général de l'INSEE

Bonsoir à tous. Il me revient la tâche de conclure cette rencontre en cette fin de journée. Je déclinerai mes observations en quatre points.

<u>L'intérêt des appariements pour la connaissance des faits sociaux</u>

Historiquement, la loi de 1951 définissait deux piliers dans la statistique publique, à savoir les données administratives et les enquêtes. Récemment, nous avons beaucoup parlé des données privées et nous évoquons aujourd'hui les appariements de données administratives, une pratique adossée aux données administratives et qui s'avère particulièrement importante.

L'INSEE a débuté la pratique des appariements, dans le cadre de l'enquête « Revenus fiscaux », de manière plutôt timide, près de dix ans après sa création, avant d'y recourir plus largement. Parallèlement, la DEPP a été le premier SSM à recourir aux appariements à partir de 1973, autour d'un panel d'élèves. Le service statistique public mobilise ainsi depuis longtemps cette pratique.

Pour autant, une approche plus systématique des appariements n'a été adoptée que récemment et je voudrais en rappeler l'intérêt. Comme l'a indiqué Jean-Noël Barrot, nous ne manquons pas de données fiscales, notre pays réservant un domaine étendu à la fiscalité. Nous ne manquons donc pas de données administratives.

Néanmoins, ces données ne représentent que des photographies. En effet, si ces données concernent des personnes, elles les photographient à tel moment et sous différents statuts – élève, étudiant, demandeur d'emploi, salarié, stagiaire, etc. De facto, il n'est pas étonnant de constater que les phénomènes qui s'inscrivent dans une dynamique quelconque nécessitent soit des enquêtes, soit des observations menées à différents instants. Or cette nécessité demande absolument l'emploi d'appariements, la donnée statistique en elle-même ne suffisant pas.

John Martin a expliqué que cette pratique s'était beaucoup développée pour documenter des *labour market policies*, ou encore pour mieux comprendre l'incidence des politiques éducatives, les réformes de l'assurance chômage, le *workfare*, ou encore la mobilité sociale. Un participant a évoqué la possibilité intéressante d'user de l'appariement pour mieux comprendre les questions concernant la rénovation thermique. A cet égard, il faudrait pouvoir comparer des données récoltées avant et après ces rénovations.

En théorie, il est possible de produire des informations sur ces faits sociaux à partir d'enquêtes, mais leur recours pose certains problèmes. Avant de décrire ces difficultés, je précise qu'il faut continuer à effectuer des enquêtes et qu'elles font l'objet d'une demande sociale très forte. A cet égard, si Chantal Cases a indiqué qu'il fallait réhabiliter l'usage de l'enquête, je pense que sa langue a fourché, car les enquêtes sont et demeurent très largement mobilisées. Mais il se trouve que le recours aux enquêtes pose un problème de coût. De plus, les enquêtes font appel à la mémoire des déclarants. Or cette mémoire s'avère assez courte et il est difficile d'obtenir des précisions sur des épisodes de vie lointains. Enfin, les enquêtes présentent des limites de granularité et ne peuvent pas offrir des données sur des petits territoires, alors que l'élaboration de nombreuses politiques publiques demande d'avoir accès à des informations territoriales.

Par conséquent, nous pouvons voir toute la place que peuvent occuper les appariements. Ceux-ci sont particulièrement utiles pour obtenir des informations sur des faits sociaux tels que la mobilité intergénérationnelle, qui s'inscrit dans une profondeur temporelle importante. Ainsi, les travaux d'Emmanuel Saez ont pu mettre en évidence des mécanismes de cette mobilité sociale aux Etats-Unis. Ses travaux ont été permis par l'existence d'un identifiant associé aux enfants dans la déclaration fiscale de leurs parents, qu'ils conservent ensuite. Il a donc pu alors apparier les déclarations fiscales des enfants et des parents.

A cet égard, en me fondant sur l'effet réverbère – parabole de l'ivrogne qui cherche ses clés la nuit dans une vaste rue et qui limite ses recherches sous l'espace restreint éclairé par un réverbère – je ne suis pas loin de penser que le fait que nous attachions en France autant d'importance à la redistribution statique des richesses s'explique par cette difficulté à mesurer la mobilité sociale au cours de la vie active ou dans la temporalité intergénérationnelle. A cause de ce manque de données, les recherches se focalisent sur la réduction des inégalités par un système de prélèvement et de transfert dans une modalité statique. Ainsi, l'effet du manque de données sur notre manière d'appréhender des phénomènes rend vitale la nécessité de développer les appariements, dans une profondeur temporelle maximale.

Cependant, par rapport aux enquêtes, les appariements présentent la limite de rendre plus difficiles les comparaisons internationales. En effet, ils se fondent sur des données administratives dont la nature varie dans les différents pays. En outre, John Martin a signalé que l'Irlande était un petit pays très ouvert aux migrations et que les mouvements de personnes donnent lieu à des zones d'ombre dans une partie des parcours individuels. En effet, les systèmes d'observation basés sur des données administratives demeurent éminemment nationaux.

Le développement de la connaissance par la multiplication des bases de données appariées et le respect du droit des personnes

J'ai été surpris de constater que nous avons progressé tant au niveau des appariements que du respect de la « petite éthique » pour reprendre la terminologie présentée plus tôt. Si nous pouvions nous situer très loin de la frontière technologique il y a quelques décennies, nous tendons à présent à nous en rapprocher. Cette rencontre montre que davantage de données sont appariées et ouvertes à la recherche et qu'il s'agit d'un point positif pour la connaissance. Bertrand Pailhès, que je salue et que je remercie pour notre collaboration lors de l'élaboration de la loi de 2016 pour une République numérique, nous indique que le CASD a énormément avancé dans la sécurisation et la protection des données. Je remercie d'ailleurs Kamel Gadouche pour sa contribution dans la construction du CASD. Nous avons ainsi sans doute procédé au mieux dans cet arbitrage entre la multiplication des bases de données et le respect de la petite éthique.

Par ailleurs, j'ai été frappé par la crainte relative à la protection des données reliées à un éventuel passage vers un régime non démocratique. Je pense en effet que cette crainte est infondée. Si nous considérons l'infrastructure du CSNS, il faut savoir que nous pouvons la détruire en quelques minutes. Il est tout à fait possible de brûler nos vaisseaux en cas de changement de régime. Et inversement, un régime non démocratique pourrait tout à fait établir un système totalitaire très rapidement.

Néanmoins, nous constatons que réaliser des appariements qui permettent d'améliorer les connaissances tout en respectant les libertés individuelles représente beaucoup de travail, notamment en France. Mais si toute l'Union européenne baigne dans la matrice de la Global Privacy Assembly (GPA) et du RGPD européen, tous ses membres ne se fondent pas sur les mêmes préoccupations dans le cadre de cet arbitrage. J'ai cité l'exemple de l'Estonie qui a un passé démocratique récent et qui se préoccupe moins de la protection des données. Mais il est étonnant de constater que l'Espagne aussi mobilise très largement dans ses appariements l'usage de l'identifiant du Padrón et de celui de la carte nationale d'identité. Aussi, les pays nordiques qui ont un passé démocratique utilisent massivement des identifiants et facilitent beaucoup leurs appariements.

Il ne m'appartient pas de juger ces différences, elles relèvent plutôt d'arbitrages politiques. Néanmoins, je souligne que nos appariements nécessitent énormément de travail, du fait de la multiplicité des identifiants et de la forte limitation de l'usage du NIR, dont nous connaissons bien l'histoire. Kamel Gadouche nous a montré que les appariements menés dans le cadre du CASD impliquaient la participation de deux tiers et étaient assez complexes. Lionel Espinasse a également montré une complexité comparable avec le CSNS. Ces appariements nécessitent des efforts et ils peuvent même représenter un défi. Ainsi, Vladimir Passeron a pu montrer qu'il était possible de procéder à des appariements à partir de traits d'identité différents, malgré quelques pertes.

Le lien des appariements avec le monde de la recherche

Il existe une filière statistique et une filière de recherche. La loi pour une République numérique décrit d'ailleurs ces deux filières. Dans ce cadre, je me demande si nous devons imaginer des appariements déjà réalisés par la statistique publique et ouverts spécialement aux chercheurs, comme nous commençons à le faire, ou bien si nous devons, en sens inverse, recycler les travaux des chercheurs en tant que facilitées essentielles pour la statistique publique. En fait, je me demande s'il faut que les appariements soient réalisés une fois pour toutes pour ces deux filières. Cette question pose notamment un problème de confidentialité, dans la mesure où l'appariement est alors plus durable à la portée de davantage d'utilisateurs.

J'ai le sentiment que jusqu'à présent en France, la tendance générale consistait à considérer que l'appariement constituait une part d'un projet de recherche d'un laboratoire économique et qu'il revenait au chercheur de décrire des critères d'arrêt, des critères de qualité et de définir le nombre d'étapes d'appariements pour obtenir un rapprochement maximal. Mais une réflexion qui doit encore mûrir pourrait nous orienter vers l'élaboration d'un portefeuille d'appariements clés en main, à mettre à disposition des chercheurs. Faut-il se diriger dans cette direction ? Est-il licite et justifié de demander aux chercheurs qui ont procédé à des appariements de pouvoir éventuellement les mettre à disposition d'autres chercheurs ou de la statistique publique ?

La question de la petite éthique

Enfin, sans revenir sur la « grande éthique », qui pourrait faire l'objet de riches discussions, ou sur la confiance, évoquées au cours de la table ronde précédente, je reviens sur la question de la « petite éthique ». Je pense que pour le moment, le public, qui fournit des données administratives, n'est pas conscient que ces données peuvent donner lieu à des appariements. Il n'est pas conscient que dans ce cadre, leurs données peuvent potentiellement devenir réidentifiantes. Il ne connaît pas non plus les enjeux éthiques des appariements. Nous devons donc réaliser un effort particulier pour informer le public.

La statistique publique ne fait pas l'objet de polémiques à cet égard. Néanmoins, il convient de les anticiper. Si les appariements se développent, nous devons réfléchir au cadre de transparence que nous devons aux répondants. Sur ce plan, nous avons déjà décidé que le CNIS devait connaître tous les programmes de travail des appariements projetés. De plus, nous veillons au respect de la minimisation et de la proportionnalité des traitements de données.

Il est vrai que nous devrions sans doute réfléchir à d'éventuelles validations externes du respect de ces principes, qui semblent manquer. Je pense que nous devons penser à ces

questions, sans attendre de faire face à des polémiques. Il ne s'agit pas simplement de reposer sur nous-mêmes, mais bien de disposer d'une validation externe qui vérifierait l'effectivité des efforts que nous déployons dans le cadre d'un processus déterminé pour faire respecter ces principes de minimisation et de proportionnalité ce que fait le comité du label pour les enquêtes.

Enfin, je remercie tous ceux qui ont eu la ténacité de rester avec nous jusqu'au bout de cette journée. Plus de deux cents auditeurs étaient connectés ce matin et peut être que plus de cent le sont encore. Je pense que la thématique de cette rencontre ne passionne pas que les producteurs de la statistique publique. Je remercie tous les intervenants et j'adresse une pensée particulière à Eric Rancourt, qui a bien mérité de prendre son petit déjeuner à l'heure qu'il est depuis le Québec. Je remercie tous ceux qui ont accepté de participer à cette journée. Je souligne notamment la qualité de leurs diaporamas. Je remercie enfin tout particulièrement Françoise Dupont et ses collègues, qui ont constitué la cheville ouvrière de cette rencontre. Merci à tous et bon week-end.
