Conseil national
de l'information statistique

Paris, 25th of april 2022 – No 60/H030

MEETING ENTITLED “PERSONAL DATA MATCHING: BETWEEN A WEALTH OF STATISTICAL INFORMATION AND RESPECT FOR PRIVACY”

Meeting of 28 January 2022

MINUTES OF THE MEETING
28 January 2022

PROGRAMME REMINDER

List of participants

In attendance:

ARTIGUELONG	Maryse	Human Rights League (*Ligue des droits de l'homme*)
AUBERT	Patrick	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (*Direction de la recherche, des études, de l'évaluation et des statistiques* – DREES)
AVVISATI	Francesco	Paris School of Economics (*École d'économie de Paris*)
BARROT	Jean-Noël	French National Assembly
BAYLE	Jules	French National Assembly
BOZIO	Antoine	Public Policy Institute (*Institut des politiques publiques*)
CARON	Nathalie	Ministry of National Education, Youth and Sport – Directorate of Evaluation, Forecasting and Performance Monitoring (*Direction de l'évaluation, de la prospective et de la performance* – DEPP)
CASES	Chantal	French Statistical Society (*Société française de statistique* – SFdS)
COLIN	Christel	National Institute for Statistics and Economic Studies (*Institut national de la statistique et des études économiques* – INSEE) – Directorate of Demographic and Social Statistics (*Direction des statistiques démographiques et sociales* – DSDS)
D'ALESSANDRO	Cristina	National Council for Statistical Information (*Conseil national de l'information statistique* – CNIS)
DUBOIS	Marie-Michèle	National Council for Statistical Information (*Conseil national de l'information statistique* – CNIS)
DUPONT	Françoise	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
DURAN	Patrice	*Ecole normale supérieure*
ELBAUM	Mireille	Ministry of Solidarity and Health – Social Affairs Audit Unit (*Inspection générale des affaires sociales* – IGAS)
ESPINASSE	Lionel	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
GADOUCHE	Kamel	Remote Secure Access Data Centre (*Centre d'accès sécurisé distant aux données* – CASD)
GUILLAUMAT-TAILLIET	François	National Council for Statistical Information (*Conseil national de l'information statistique* – CNIS)
HUNYADI	Mark	Catholic University of Leuven (*Université Catholique de Louvain*)
LAGARDE	Sylvie	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (*Direction de la méthodologie et de la coordination statistique et internationale* – DMCSI)
LEFEBVRE	Olivier	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
MARTIN	John	IZA Institute of Labour Economics
MAUREL	Françoise	National Council for Statistical Information (*Conseil national de l'information statistique* – CNIS)
MONTUS	Arnaud	National Council for Statistical Information (*Conseil national de l'information statistique* – CNIS)
PAILHÈS	Bertrand	National Commission on Informatics and Liberty (*Commission Nationale de l'Informatique et des Libertés* – CNIL)
PASSERON	Vladimir	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
ROBERT	Philomé	France24
TAGNANI	Stéphane	National Council for Statistical Information (*Conseil national de l'information statistique* – CNIS)
TAVERNIER	Jean-Luc	National Institute for Statistics and Economic Studies (INSEE) – Directorate-General
TIMBEAU	Xavier	French Economic Observatory (*Observatoire français des conjonctures économiques* – OFCE)

Via video conference:

AUBERT	Magali	National Institute for Agricultural Research (*Institut national de la recherche agronomique* – INRA)
ADAM	Lorraine	PROGEDO
AKIKI	Michel	Ministry of Agriculture and Food – Department of Statistics and Foresight Analysis (*Service de la statistique et de la prospective* – SSP)
ALAZARD	Antoine	Dijon regional council
ALKHOURY	Maria	Remote Secure Access Data Centre (CASD)
ALLAIN	Samuel	Regional Directorate for Food, Agriculture and Forestry – Aquitaine
AMARAL	Philippe	Ministry of Solidarity and Health – Directorate-General for Social Cohesion (*Direction générale de la cohésion sociale* – DGCS)
AMBARD	Julien	Les Morts de la Rue collective
ANDRE	Mathias	National Institute for Statistics and Economic Studies (INSEE)
ANDREANI	Nicolas	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
ANGUIS	Marie	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
ANTUNEZ	Kim	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Dissemination and Regional Action (*Direction de la diffusion et de l'action régionale* – DDAR)
ANXIONNAZ	Isabelle	Private individual

ARCHAMBAULT	Edith	Paris 1 Panthéon-Sorbonne University
BACCAINI	Brigitte	Ministry of Ecological Transition – General Council for the Environment and Sustainable Development
BAGEIN	Guillaume	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
BAILLY	Nathalie	National Institute for Statistics and Economic Studies (INSEE) – Secretariat-General
BAKIA	Halima	Remote Secure Access Data Centre (CASD)
BALAVOINE	Angélique	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
BARRET	Emilie	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
BAULNE	Jimmy	Quebec Statistics Institute (*Institut de la statistique du Québec*)
BAYET	Alain	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Dissemination and Regional Action (DDAR)
BÉGUIN	Jean-Marc	Private individual
BELLOC	Brigitte	French Statistical Society (*Société française de statistique* – SFdS)
BENABDALLAH	Said	Versailles local education authority
BENICHOU	Yves-Laurent	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
BERSON	Clémence	Banque de France (BdF)
BERTHOLON	Raphaëlle	The French Confederation of Management - General Confederation of Executives (*Confédération française de l'encadrement - Confédération générale des cadres* – CFE-CGC)
BESSIERE	Sabine	Ministry of Ecological Transition – Data and Statistical Studies Department (*Service des données et des études statistiques* – SDES)
BIANCO	Emma	INSEE Auvergne - Rhône-Alpes
BLACHE	Guillaume	Pôle Emploi
BLANCARD	Patricia	French Official Statistics Authority (*Autorité de la statistique publique* – ASP)
BLANDIN	Lola	Grenoble Alpes University
BLAVIER	Pierre	National Centre for Scientific Research (*Centre national de la recherche scientifique* – CNRS) – Sociological Change Observatory (UMR 7049)
BOIS	François-Xavier	Kernix
BONDON	Marine	National Institute for Demographic Studies (*Institut national des études démographiques* – INED)
BONNANS	Dominique	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
BONNET	Xavier	National Institute for Statistics and Economic Studies (INSEE) – Internal Audit Unit
BONNETÊTE	Félix	Ministry of Labour, Employment and Integration – Directorate of Research, Economic Studies and Statistics (*Direction de l'animation de la recherche, des études et des statistiques* – DARES)
BOREL	Marie	Ministry of Labour, Employment and Integration – Directorate of Research, Economic Studies and Statistics (DARES)
BOUTIERE	Fabienne	Electricité de France (EDF)
BRIARD	Karine	Ministry of Labour, Employment and Integration – Directorate of Research, Economic Studies and Statistics (DARES)
BRILHAULT	Gwennaëlle	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
BRINGE	Arnaud	National Institute for Demographic Studies (INED)
BRION	Philippe	Private individual
BRUNET	François	National Institute for Statistics and Economic Studies (INSEE) – Internal Audit Unit
BUNEL	Simon	Banque de France (BdF)
BURRICAND	Carine	Ministry of Higher Education, Research and Innovation – Information Systems and Statistical Studies Sub-directorate
CAHOUR	Lisa	French Health Authority (*Santé Publique France*)
CARIOU	Sylvain	National Research Institute for Agriculture, Food and the Environment (*Institut national de recherche pour l'agriculture, l'alimentation et l'environnement* – INRAE)
CARRASCO	Valérie	Ministry of the Interior – Ministerial Statistical Department for Internal Security (*Service statistique ministériel de la sécurité intérieure* – SSMSI)
CARRERE	Amélie	Paris School of Economics (*École d'économie de Paris*)
CASTELLUCCIA	Claude	National Commission on Informatics and Liberty (CNIL)
CAVIER	Bernard	
CAZALE	Linda	Quebec Statistics Institute
CECCI-ANDREANI	Laury	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
CHALEIX	Mylène	National Institute for Statistics and Economic Studies (INSEE) – Information System Directorate (*Direction du système d'information* – DSI)
CHAMBAZ	Christine	National Institute for Statistics and Economic Studies (INSEE) – Business Statistics Directorate (*Direction des statistiques d'entreprises* – DSE)
CHAMKHI	Amine	Pôle Emploi
CHANTEUX	Alice	Departmental Council of Isère
CHAPUT	Hélène	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination

		and International Relations Directorate (DMCSI)
CHARRANCE	Géraldine	National Institute for Demographic Studies (INED)
CHARRIER	Rodolphe	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
CHAUVEL	Brian	Paris Nanterre University
CHAUVIN	Adrienne	Union sociale pour l'habitat, a social housing union
CHAUVIN	Pauline	Paris 5 University – Faculty of Social Sciences
CHEJFEC	Thomas	National Professional Union for Employment in Industry and Trade (*Union nationale interprofessionnelle pour l'emploi dans l'industrie et le commerce* – UNEDIC)
CHEVALIER	Pascal	Ministry of Justice – Statistics and Studies Sub-directorate
CHIN	Francis	French Health Authority (*Santé Publique France*)
CIESIELSKI	Henry	Ministry of Ecological Transition – Directorate of Housing, Town Planning and Landscape (*Direction de l'habitat, de l'urbanisme et des paysages* – DHUP)
CLERC	Marie	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
CLING	Jean-Pierre	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
CLUSE	Margaux	Ministry of Solidarity and Health – Directorate-General for Social Cohesion (DGCS)
COCHET	Paul	National Institute for Demographic Studies (INED)
COLIN	Catherine	Occitanie Regional Council
COLIN	Christel	Ministry of National Education, Youth and Sport – Directorate of Youth, Community Education and Community Life (*Direction de la jeunesse, de l'éducation populaire et de la vie associative* – DJEPVA)
COMMANDEUR	Barbara	Centre for Training Information Resource Management/Regional Centre for Monitoring Employment and Training (*Centre animation ressources d'information sur la formation/Observatoire régional emploi formation* – CARIF-OREF) Pays de la Loire
CORRE	Tifenn	National Institute for Agricultural Research (INRA), Toulouse
COSTENOBLE	Ophélie	Centre for Training Information Resource Management/Regional Centre for Monitoring Employment and Training (*Centre animation ressources d'information sur la formation/Observatoire régional emploi formation* – CARIF-OREF) network
COSTER	Jean-Louis	National Institute for Statistics and Economic Studies (INSEE) – Business Statistics Directorate (DSE)
COTREBIL	Philippe	Oise-les-Vallées urban planning department
COTTET	Sophie	Public Policy Institute
COTTIN	Adeline	INSEE Pays de la Loire
COUDIN	Elise	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
COUDRIN	Caroline	Directorate of the Environment, Urban Planning and Housing (*Direction de l'environnement, de l'aménagement et du logement* – DEAL), La Réunion
COULONDRE	Alexandre	Ecole des Ponts ParisTech, a higher education institute for science, engineering and technology
CROGUENNEC	Yannick	Ministry of National Education, Youth and Sport – Directorate of Evaluation, Forecasting and Performance Monitoring (DEPP)
CRUAU	Natacha	Ministry of Labour, Employment and Integration – General Delegation for Employment and Vocational Training (*Délégation générale à l'emploi et à la formation professionnelle* – DGEFP)
DAMPERON	Alexandre	INSEE Ile-de-France
DE MIRAS	Christelle	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
DEFRESNE	Marion	Ministry of National Education, Youth and Sport – Directorate of Evaluation, Forecasting and Performance Monitoring (DEPP)
DELAHAYE-ADAM	Elisa	Ministry of the Interior
DELAME	Nathalie	National Research Institute for Agriculture, Food and the Environment (INRAE)
DEMOLY	Elvire	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
DEMONSANT	Jean-Luc	Toulouse University
DEROSIER	Alice	Nantes local education authority
DEROYON	Thomas	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
DESPLANQUES	Guy	Private individual
DIAKHATE	Maryama	Ministry of Justice – Statistics and Studies Sub-directorate
DIARD	Karine	National Institute for Statistics and Economic Studies (INSEE) – Business Statistics Directorate (DSE)
DIXTE	Christophe	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
DORNIER	Xavier	French Institute of Horses and Horse Riding (*Institut français du cheval et de l'équitation* – IFCE)
DOURAN	Manal	
DREUX	Cannelle	Departmental council 92
DUBOST	Claire-Lise	Ministry of Labour, Employment and Integration – Directorate of Research, Economic Studies and Statistics (DARES)
DUC	Cindy	National Institute for Statistics and Economic Studies (INSEE) – Business Statistics Directorate (DSE)
DUNNE	John	Central Statistics Office (CSO) of Ireland

DURR	Jean-Michel	CAOS-Consulting
DUSSUD	François-Xavier	National Institute for Statistics and Economic Studies (INSEE) – Business Statistics Directorate (DSE)
DUVERNET	Laurent	Paris 10 Nanterre University
EGHBAL	Sylvie	National Institute for Statistics and Economic Studies (INSEE)
EIDELMAN	Alexis	Ministry of Labour, Employment and Integration – Directorate of Research, Economic Studies and Statistics (DARES)
EL BOUHAIRI	Yacine	Remote Secure Access Data Centre (CASD)
FAIDHERBE	Thibault	Ministry of Overseas France – Directorate-General for Overseas France (*Direction générale des Outre-Mer* – DGOM)
FAU	David	Dijon regional council
FAURET	Camille	INSEE Ile-de-France
FAVARO	Antonin	National Research Institute for Agriculture, Food and the Environment (INRAE)
FERON	Valérie	Regional health observatory of Ile-de-France
FICHE	Dominique	Ministry of Agriculture and Food – Department of Statistics and Foresight Analysis (SSP)
FIRDION	Laëtitia	Paris Region Institute
FONTAINE	Roméo	National Institute for Demographic Studies (INED)
FRANCOZ	Dominique	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
FREPPEL	Camille	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
FRESSARD	Lisa	Provence-Alpes-Cote d'Azur regional health observatory
GALLIC	Gabrielle	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Dissemination and Regional Action (DDAR)
GARBINTI	Bertrand	Centre for Research in Economics and Statistics (*Centre de recherche en économie et statistique* – CREST)
GARCIA	Cédric	Gustave Eiffel University
GAUVIN	Charlotte	Ministry of Agriculture and Food – Directorate-General for Education and Research
GÉLY	Alain	Confédération générale du travail (CGT), a trade union confederation
GÉNIN	Gaëlle	INSEE Nouvelle-Aquitaine
GEORGE	Estelle	Versailles local education authority
GESBERT BOULANGER	Florence	Ministry of Higher Education, Research and Innovation – Directorate-General for Research and Innovation (*Direction générale de la recherche et de l'innovation* – DGRI)
GIFFARD	Quentin	Biomasse Normandie
GILLES	Séverine	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
GIVOIS	Samuel	Ministry of Agriculture and Food – Department of Statistics and Foresight Analysis (SSP)
GODINOT	Alain	Private individual
GOLDBERG	Marcel	National Institute of Health and Medical Research (INSERM)
GOMOT	Eleonore	Ministry of Economy, Finance and Recovery – Directorate-General for Public Finance (*Direction générale des finances publiques* – DGFIP)
GOSSIAUX	Sébastien	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Dissemination and Regional Action (DDAR)
GOURDON	Olivier	National Institute for Statistics and Economic Studies (INSEE) – Directorate-General
GOURMELEN	Julie	National Institute of Health and Medical Research (INSERM)
GRISELLE	Patrick	Official Statistics Quality Label Committee
GUILLAUME	Stéphanie	Institute for Research and Documentation in Health Economics (*Institut de recherche et documentation en économie de la santé* – IRDES)
GUILLAUME	Thierry	Ministry of Agriculture and Food – Department of Statistics and Foresight Analysis (SSP)
GUILLEMOT	Danièle	National Institute for Statistics and Economic Studies (INSEE) – Business Statistics Directorate (DSE)
GUILLOU	Sarah	French Economic Observatory (OFCE)
GUIRCHOUN	Elodie	Ile-de-France Regional Council
HAAG	Olivier	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
HADDAK	Mohamed	Gustave Eiffel University
HAGUET	Laurence	Court of Auditors
HARNOIS	Jérôme	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
HAYS	Olivier	Centre d'étude des supports de publicité (CESP), an advertising and media industry non-profit organisation
HERVIANT	Julie	INSEE Ile-de-France
HERZOG	Judith	PersonalData.IO
HUBERT	Jean-Paul	Gustave Eiffel University
HURPEAU	Benoît	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Dissemination and Regional Action (DDAR)

IDOHOU	Emmanuel	Private individual
ISNARD	Michel	National Institute for Statistics and Economic Studies (INSEE) – Internal Audit Unit
JALUZOT	Laurence	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
JARDIN	Marie	Provence-Alpes-Cote d'Azur regional health observatory
JOUBERT-LECLERC	David	Quebec Statistics Institute
JUDAS	Francis	Confédération générale du travail (CGT), a trade union confederation – Finance Federation
KABLA-LANGLOIS	Isabelle	INSEE Ile-de-France
KARKER	Chourouk	Union sociale pour l'habitat, a social housing union
KOLODZIEJ	Isabelle	Union des industries de la fertilisation (UNIFA), a fertiliser industry union
KOSSI	Dede	Institute for Research and Documentation in Health Economics (IRDES)
KOUMARIANOS	Heidi	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
KURKDJI	Patrick	Provence-Alpes-Cote d'Azur regional health observatory
LABOSSE	Aline	INSEE Auvergne - Rhône-Alpes
LACAILLE	Yves	Union nationale des professions libérales (UNAPL), a professional worker confederation
LAFARGUE	Loïc	Nantes local education authority
LAINÉ	Frédéric	Pôle Emploi
LAMARCHE	Pierre	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
LAMBREY	Serge	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
LAPINE	Malena	National Institute for Demographic Studies (INED)
LAVERGNE	Aurélien	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
LE CAIGNEC	Emilie	Ministry of Justice – Statistics and Studies Sub-directorate
LE ROLLAND	Lucie	Public Policy Institute
LEBRETON	Elodie	French Health Authority (*Santé Publique France*)
LEBUGLE	Amandine	SAMU Social de Paris, an association to help the homeless of Paris
LECOCQ	Marie	FranceAgriMer, the national establishment for agricultural and seafood products
LECOUVEY	François	Centre for Economic Studies and Research into Energy (*Centre d'études et de recherches économiques sur l'énergie* – CEREN)
LEMERLE	Stéphanie	Ministry of the Interior – Department of Studies, Statistics and Documentation (*Département des statistiques, des études et de la documentation* – DSED)
LEON	Olivier	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
LEQUIEN	Matthieu	National Institute for Statistics and Economic Studies (INSEE)
LEQUIEN	Laurent	National Institute for Statistics and Economic Studies (INSEE)
LEROUX	Isabelle	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
LEROY	Claire	National School of Statistics and Economic Administration (*École nationale de la statistique et de l'administration économique* – ENSAE)
LEVI-VALENSIN	Michaël	Ministry of Agriculture and Food – Department of Statistics and Foresight Analysis (SSP)
LEZEC	Florian	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
LIOGIER	Valérie	Ministry of National Education, Youth and Sport – Directorate of Evaluation, Forecasting and Performance Monitoring (DEPP)
LIXI	Clotilde	Ministry of Higher Education, Research and Innovation – Information Systems and Statistical Studies Sub-directorate
LOMBRAIL	Pierre	Paris 13 University
LOONIS	Vincent	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
LORRE	Geoffrey	Ministry of Labour, Employment and Integration – General Delegation for Employment and Vocational Training (DGEFP)
LOSTYS	Emilie	Ministry of Labour, Employment and Integration – General Delegation for Employment and Vocational Training (DGEFP)
LOUPIAS	Claire	Evry-Val-d'Essonne University
LUNGARSKA	Anna	National Institute for Agricultural Research (INRA), Toulouse
MAHIET	Gilles	Orange Lab
MAKDESSI	Yara	Ministry of Justice – Statistics and Studies Sub-directorate
MALAGUTTI	Ornella	Ministry of the Interior
MALHERBE	Lucas	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
MALLÉJAC	Noémie	Private individual
MANDEREAU-BRUNO	Laurence	French Health Authority (*Santé Publique France*)
MARBACH	Léon	Sciences Po university

MAREAU	Quentin	Meurthe-et-Moselle Departmental Council
MARQUIER	Rémy	Remote Secure Access Data Centre (CASD)
MARTIAL	Elodie	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
MATINET	Béryl	Ministry of the Interior – Ministerial Statistical Department for Internal Security (SSMSI)
MEINZEL	Pauline	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
MEJEAN	Isabelle	Sciences Po university
MERCIER	Alice	SAMU Social de Paris, an association to help the homeless of Paris
MERLY-ALPA	Thomas	National Institute for Demographic Studies (INED)
MICHALLAND	Béatrice	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
MICHELOT	François	Paris Region Institute
MILCENT	Carine	Paris School of Economics – Paris 1 University
MISSEGUE	Nathalie	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
MONTAUT	Alexis	INSEE Nouvelle-Aquitaine
MOREAU	Sylvain	National Institute for Statistics and Economic Studies (INSEE) – Business Statistics Directorate (DSE)
MOREL	Claire	Private individual
MOTAMEDI	Kiarash	Ministry of Ecological Transition – Sustainable Development Economics, Assessment and Integration Department
M'PIAYI	Mélissa	Centre d'étude des supports de publicité (CESP), an advertising and media industry non-profit organisation
NAUROY	Frédéric	Ministry of Ecological Transition
NAYMAN	Laurence	Centre for Forecasting and International Information (*Centre d'études Prospectives et d'Informations Internationales* – CEPII)
NGUYEN	Christine	Ministry of Labour, Employment and Integration – General Delegation for Employment and Vocational Training (DGEFP)
NGUYEN HUU CHIEU	Elise	Union nationale des professions libérales (UNAPL), a professional worker confederation
NIAY	Mathilde	Ministry of Ecological Transition – Sustainable Development Economics, Assessment and Integration Department
NICOLAU	Javier	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
NIJARI	Assia	Office of the High Commissioner for Planning (*Haut-Commissariat au Plan*)
OLIER	Lucile	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
OROZCO	Valérie	National Institute for Agricultural Research (INRA), Toulouse
ORZONI	Mathieu	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Dissemination and Regional Action (DDAR)
PAILHÉ	Ariane	National Institute for Demographic Studies (INED)
PALAT	Blazej	Sciences Po university
PAVARD	Clément	National Agency for Housing Information (*Agence nationale pour l'information sur le logement* – ANIL)
PERREL	Céline	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
PETORIN	Elodie	National Centre for Scientific Research (*Centre national de la recherche scientifique* – CNRS)
PICARD	Sébastien	Ministry of Culture – Department of Studies, Forecasting, Statistics and Documentation (*Département des études, de la prospective, des statistiques et de la documentation* – DEPS-Doc)
PIET	Laurent	National Research Institute for Agriculture, Food and the Environment (INRAE)
PIGUET	Virginie	National Institute for Agricultural Research (INRA) – Centre for Economics and Sociology Applied to Agriculture and Rural Areas (*Centre d'Économie et de Sociologie Appliquées à l'Agriculture et aux Espaces Ruraux* – CESAER)
POLLET	Pascale	French Official Statistics Authority (*Autorité de la statistique publique* – ASP)
POMÉON	Thomas	National Institute for Agricultural Research (INRA), Toulouse
PONS	Sébastien	INSEE Bretagne
PORA	Pierre	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
POTREAU	Elisabeth	INSEE Occitanie
POULHES	Mathilde	Ministry of the Interior – Ministerial Statistical Department for Internal Security (SSMSI)
PRÉVOT	Pascal	INSEE Nouvelle-Aquitaine
PROKOVAS	Nicolas	Confédération générale du travail (CGT), a trade union confederation
PRUSKI	Cézane	Ministry of Solidarity and Health – Directorate-General for Social Cohesion (DGCS)
RAIN	Audrey	Public Policy Institute
RAKOTOARISOA	Ifaliana	Remote Secure Access Data Centre (CASD)
RAMAMONJY	V.	Centre for Economic Studies and Research into Energy (CEREN)
RAMBLIÈRE	Lison	SAMU Social de Paris, an association to help the homeless of Paris

RANCOURT	Eric	Statistics Canada
RATEAU	Guillaume	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
RATHELOT	Roland	Centre for Research in Economics and Statistics (*Centre de recherche en économie et statistique* – CREST)
RATHLE	Jean-Philippe	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
RAVEL	Loïc	Ministry of the Interior – Directorate-General of the National Police (*Direction générale de la Police nationale* – DGPN)
RAYNAUD	Philippe	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
REDOR	Patrick	National Institute for Statistics and Economic Studies (INSEE)
REGOLO	Julie	National Research Institute for Agriculture, Food and the Environment (INRAE)
REMILLON	Delphine	National Institute for Demographic Studies (INED)
RENUY	Adeline	National Institute of Health and Medical Research (INSERM)
REY	Grégoire	National Institute of Health and Medical Research (INSERM)
RIBON	Olivier	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
RICAU	Pascale	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
RICHARD	Mélanie	National Housing Agency (*Agence nationale de l'habitat* – ANAH)
RICHET	Jehanne	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
RICHET-MASTAIN	Lucile	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Demographic and Social Statistics (DSDS)
RIEG	Christian	Ministry of Public Sector Transformation and Civil Service – Directorate General for Administration and the Civil Service (*Direction générale de l'administration et de la fonction publique* – DGAFP)
RIMBEAULT	Chloé	Initiative France, a network to support business start-ups
ROBERT-BOBÉE	Isabelle	National Institute for Statistics and Economic Studies (INSEE)
ROBERTI	Vincent	National Professional Union for Employment in Industry and Trade (UNEDIC)
ROBIN	Yoan	National Professional Union for Employment in Industry and Trade (UNEDIC)
ROCHEREAU	Thierry	Institute for Research and Documentation in Health Economics (IRDES)
RODRIGUES	Amandine	INSEE Pays de la Loire
ROSENWALD	Fabienne	Ministry of National Education, Youth and Sport – Directorate of Evaluation, Forecasting and Performance Monitoring (DEPP)
ROTH	Nicole	National Institute for Statistics and Economic Studies (INSEE)
ROUSSEAU	Sylvie	Ministry of National Education, Youth and Sport – Directorate of Evaluation, Forecasting and Performance Monitoring (DEPP)
ROUX	Aliette	Maison des sciences de l'homme, a non-profit research foundation – Nantes
ROVERSI	Aurélia	National Institute for Demographic Studies (INED)
ROY	Delphine	Public Policy Institute
SABOT	Philippe	Ministry of Agriculture and Food – Department of Statistics and Foresight Analysis (SSP)
SALATHE	Manuelle	Ministry of the Interior – National Inter-Ministerial Road Safety Observatory
SAOUD	Ali	HCP Morocco
SAUVEUR	Jean	
SCHUHL	Pierrette	Ministry of Higher Education, Research and Innovation – Information Systems and Statistical Studies Sub-directorate
SÉDILLOT	Béatrice	Ministry of Ecological Transition – Data and Statistical Studies Department (SDES)
SELZ	Marianne-Marion	French Statistical Society (*Société française de statistique* – SFdS)
SERIEYX	Yvon	National Union of Family Associations (*Union nationale des associations familiales* – UNAF)
SILBERMAN	Roxane	National Centre for Scientific Research (*Centre national de la recherche scientifique* – CNRS)
SIMON	Marion	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
SIQUEIRA	Juliana	Paris Nanterre University
SOUAL	Hélène	INSEE Auvergne - Rhône-Alpes
SOULLIER	Noémie	French Health Authority (*Santé Publique France*)
STEHLIN	Anne	Pôle Emploi
SUESSER	Jan Robert	Human Rights League (*Ligue des droits de l'homme*)
SUHARD	Véronique	Institute for Research and Documentation in Health Economics (IRDES)
SULTAN	Joyce	Public Policy Institute
TACHFINT	Karim	INSEE Centre
TARAYOUN	Tedjani	Ministry of Justice – Statistics and Studies Sub-directorate
TCHA	Stéphanie	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
TERSEUR	Bruno	Regional Directorate for Environment, Planning and Housing – Provence-Alpes-Cote d'Azur
TEYSSIER	Geoffrey	National Institute for Demographic Studies (INED)

THÉODOSE	Teddy	Paris 13 University
THOUMELIN	Claire	Ministry of Culture – Department of Studies, Forecasting, Statistics and Documentation (*Département des études, de la prospective, des statistiques et de la documentation* – DEPS-Doc)
TORELLI	Constance	National Institute for Statistics and Economic Studies (INSEE) – International Technical Support Division
TORTOSA	Thomas	INSEE Bretagne
TOULEMON	Léa	Paris School of Economics (*École d'économie de Paris*)
TOURNADRE	Emilie	Permanent Assembly of Chambers of Agriculture (*Assemblée permanente des chambres d'agriculture* – APCA)
TOUW	Alexandre	Paris Dauphine University
TREYENS	Pierre-Eric	INSEE Bretagne
VALLET	Louis-André	National Centre for Scientific Research (*Centre national de la recherche scientifique* – CNRS)
VANDERSCHELDEN	Mélanie	National Institute for Statistics and Economic Studies (INSEE) – Methodology, Statistical Coordination and International Relations Directorate (DMCSI)
VAUTHIER	Mathilde	Permanent Assembly of Chambers of Agriculture (APCA)
VESSILLIER	Delphine	French Construction Company Federation (*Fédération française du bâtiment* – FFB)
VIDAL	Marie	Remote Secure Access Data Centre (CASD)
VIGLINO	Lionel	National Institute for Statistics and Economic Studies (INSEE) – Directorate of Dissemination and Regional Action (DDAR)
VILAIN	Annick	Ministry of Solidarity and Health – Directorate of Research, Studies, Evaluation and Statistics (DREES)
VINCENT	Roseline	Institute for Research and Documentation in Health Economics (IRDES)
VIOLLIN	Guy	Official Statistics Quality Label Committee
VROYLANDT	Thomas	National Professional Union for Employment in Industry and Trade (UNEDIC)
WYCKAERT	Matthieu	Ministry for the Armed Forces – Economic Observatory for Defence (*Observatoire économique de la défense* – OED)
YOUSSEF	Youssr	Private individual
ZOLOTOUKHINE	Erik	PROGEDO

INTRODUCTION

Patrice Duran, President of the National Council for Statistical Information (*Conseil national de l'information statistique* – CNIS)

Hi everyone. In a coincidence of scheduling, our meeting takes place on World Data Protection Day. This day was created in 2007 to inform the public about the collection and processing of their personal data. In particular, it aims to inform the public about the reasons why their data is collected and their rights.

I am pleased to welcome you to the Pierre-Mendès France Conference Centre, although we would have preferred for this meeting to have taken place under better conditions. I hope that the audience will be able to be physically present at the next meeting on 18 May on the topic of the panels and the cohorts of Official Statistics, right here in Bercy.

We have met to discuss the issue of personal data matching. This practice raises questions not only of a technical and methodological nature, but also of a legal, ethical and societal nature.

The main goal of this meeting is to promote consultation between the producers and users of Official Statistics. In addition, the entire purpose for the existence of the CNIS is based around this issue.

I would like to welcome the international experts and colleagues from our partner countries who have agreed to take part in our discussions and will give us the benefit of their experience. We would be particularly pleased to learn more from them about their successes and the problems they face.

It is well known that there is increasing demand at the CNIS for access to administrative data and that the creation of data sets matching multiple sources is progressing, in addition to surveys. Indeed, the increase in the capability of Official Statistics created using administrative data, source matching or, more generally, massive data, accelerated sharply in 2021.

Simultaneously, the number of producers of statistics outside the Official Statistical Service is on the rise, raising questions about the scope of the Service's work. In 2018, a CNIS meeting on the challenges posed by new data sources called on Official Statistics to meet the challenge posed by these new producers by producing high-quality information that would be richer, more current and less costly, while remaining useful for public debate and respectful of privacy.

Furthermore, CNIS Medium-Term General Opinion No 7 for 2019–2023 explicitly calls for the "carrying out of matching between data sources in order to enrich the analysis of links between different topics, while ensuring that confidentiality is strictly respected where the data matching concerns 'personally identifying' information". The issues of transparency and the legal framework are therefore an important part of our concerns.

In addition to the methodological, technical and legal issues, I would also like to reiterate the importance of data matching for directing public policy and more broadly for the management of public action. As Aaron Wildavsky, one of the great masters of policy analysis, put it, public policies are always "combinations of cogitation and interaction". Indeed, public action is not only a matter of reflection and knowledge, but also of collective action.

However, current public issues in relation to managing public problems are largely cross-cutting, affecting administrative nomenclatures and levels of government, which can only increase the need for coordination. Therefore, data matching, with the new information it provides, is a valuable resource for addressing the difficult question of public policy co-ordination. In particular, it breaks the silo mentality that exists within the public administrative authorities.

For example, the Directorate of Evaluation, Forecasting and Performance Monitoring (DEPP) needs to know about aspects relating to employment, while the Directorate of Research, Economic Studies and Statistics (DARES) needs aspects relating to training.

Similarly, the contributions from the file matching project of the Ministerial Statistical Department for Internal Security (*Service statistique ministériel de la sécurité intérieure* – SSMSI) and the Statistics and Studies Sub-directorate (*Sous-Direction de la statistique et des études* – SDSE) of the Ministry of Justice are considerable, in so far as they should make it possible to document the entire criminal chain and thus promote a more rational and less ideological approach to issues that are too often weighed down by prejudice and stereotypes.

At present, the entire management system and style must be shaken up, given that we need to *progress down the path of de-segmenting public interventions and ensuring they are coherent*. So, clearly, data matching corresponds well to the current need for a reinvention of public action, which requires better coordination. Furthermore, the pandemic period that we have been through is very much emblematic of such a requirement. Indeed, this requirement is not limited to the medical field exclusively, as all areas of public action have been affected.

Finally, while data matching may present methodological and technical challenges, it relates more broadly to a critical challenge in directing public action that we cannot ignore and which we will have an opportunity to discuss today.

I hope that you have a fantastic day.

SESSION 1 – SITUATION REGARDING DATA MATCHING PRACTICES

Chair of the session: Mireille Elbaum, President of the French Official Statistics Authority (*Autorité de la statistique publique* – ASP);

Sylvie Lagarde, Director of Methodology, Statistical Coordination and International Relations (INSEE) and Christel Colin, Director of Demographic and Social Statistics (INSEE), for a presentation on Official Statistics matching practices;

Kamel Gadouche, Director of the Secure Data Access Centre (CASD), for a presentation of the data matching performed by researchers.

Mireille Elbaum

I am delighted to meet with you to chair this first session.

I was initially astonished by the phrasing of the title of this meeting, which focused on data matching, because it was not a subject, in my view, that raises no problems. When I star-

ted my studies around the late 1970s and early 1980s, we were already discussing the threats that could be posed by file cross-referencing, as part of questions about the links between IT and freedom.

In addition, I still have questions about the subjects presented from a technical angle rather than a non-thematic angle, based on the needs of the users, as well as about the risk that these technical tools, which provide benefits and substantial savings, lock us into pre-existing administrative data, which is strongly encouraged by our financial constraints. Therefore, although data matching can involve both administrative data and surveys, I would like to point out at the outset that it is not a panacea and that while it does help to answer the questions asked about the data that we want to match, it does not answer all of the questions.

However, a lot of water has passed under the bridge since the 1980s. On 22 September 2021, following a mission by the INSEE Internal Audit Unit aimed at facilitating the matching of personal data within the Official Statistical Service, the ASP published a statement aimed at encouraging and supporting this use of data matching. This statement stresses that the widespread use of this practice can pave the way for particularly innovative and valuable use of the data. It states that Official Statistics, whether constrained by resources or not, must not miss the opportunity to benefit from this enrichment of the uses of data, which is proving to be beneficial for statistical studies and research.

Therefore, we feel that the ability to match “business” data and public policy data with other file types, potentially with very fine granularity, provides interesting applications for directing public policy, as Patrice Duran has pointed out. Data matching thus makes it possible to meet growing demand for information from public authorities and to enrich evaluations of their actions, including in areas that they had not initially envisioned.

In addition, we have made significant progress in respect of ensuring statistical confidentiality and complying with the General Data Protection Regulation (GDPR). In particular, INSEE has made this progress through a set of instruments and structures that we will discuss today, such as the Non-Identifying Statistical Code (*code statistique non signifiant* – CSNS) or the Secure Data Access Centre (*Centre d’accès sécurisé aux données* – CASD). We can also count on trusted third parties, or on the “FOINisation” of health data, i.e., on the personally identifying information occultation function (*fonction d’occultation des informations nominatives* – FOIN), to study the data in a very granular manner.

In a world of “data”, in which we struggle to separate the wheat from the chaff, we need to emphasise that official statisticians have the skills and assets that represent a specific comparative advantage over other data producers, in so far as they provide both comparative and broader insights for social actors and citizens, including through intelligent data matching, for general information purposes. The ASP will soon publish a statement on this point online.

In this respect, for example, the matching of socio-fiscal and health data carried out within the CASD is an essential tool for studying social inequalities in the light of all public policies.

However, there are safeguards in relation to such data matching, as noted by the ASP in its statement. Patrice Duran has just mentioned the legal aspect of these safeguards, but I would like to stress that these are cumbersome operations and lengthy procedures that

slow down the production of information, including in Official Statistics. These procedures require technical, legal and institutional support, which INSEE is particularly called upon to provide, as Christel Colin and Sylvie Lagarde will explain.

Researchers are also experiencing these difficulties. I saw this myself while on an assignment for the Social Affairs Audit Unit (*Inspection générale des affaires sociales* – IGAS) relating to epidemiological cohorts, in which the legal-technical problems and difficulties were all the greater when the cohorts were small and focused on highly granular health data.

Lastly, I would like to note that we are in an intermediate stage in the area of data matching. Indeed, official statisticians and researchers are making great strides, under separate procedures. However, with regard to health data and, in particular, the National Health Data System (*système national de données de santé* – SNDS) as part of the Health Data Hub, progress is not being made at the same pace and is even suffering from setbacks that may pose difficulties for researchers. In this respect, the latter have various options, which are not always clear to them, concerning the framework and the status of their work, depending on whether their access to data is within the framework of the National Institute of Health and Medical Research (INSERM), an agreement with the Department of Statistics and Foresight Analysis (*Service de la statistique et de la prospective* – SSP) or on the basis of an individual initiative.

In this respect, researchers may be required to match health data with socio-fiscal data, in particular, which can be complex and involve many stakeholders. For example, I found that Constances (France's largest epidemiological cohort of 200,000 people) included income data collected from people at the same time as health data through direct questioning. However, information obtained from matching with the CASD's administrative socio-fiscal files would have provided more reliable results, and that is the route down which its managers ultimately want to go. And thus, in France, we have legal corpora that are all legitimate, but which are juxtaposed to each other, which can ultimately result in significant usage problems.

Now, Christel Colin and Sylvie Lagarde will provide an overview of data matching practices in Official Statistics, then Kamel Gadouche will highlight the CASD's contributions in this area.

Data matching practices in Official Statistics

Christel Colin

Hi everyone. We will present a general overview of data matching practices in the Official Statistical Service, namely INSEE and the 16 Ministerial Statistical Offices (MSOs).

Why have a meeting?

The Official Statistical Service has been performing data matching on personal data since the 1960s to answer a variety of questions. It allows new information to be produced at a lower cost than collecting the information from people directly.

This meeting provides an opportunity to present this long-standing practice of data matching through various examples. It also allows us to discuss issues that can arise from such data matching.

We will be asking questions about data matching methods, their legal framework and the information to be communicated to data users or data subjects.

Furthermore, the general context of data matching is evolving, in terms of both technical resources and legal possibilities. The requests being made to the Official Statistical Service are also changing. Therefore, it is a matter of determining what new questions are being asked and identifying which safeguards could be used to answer them.

A long history of data matching

INSEE and the MSOs have long-standing experience in matching personal data. We can go back at least as far as 1956 to find data matching being used for the "Tax Income" survey (*enquête "revenus fiscaux"* – ERF) – later renamed the "Tax and Social Income" survey (*enquête "revenus fiscaux et sociaux"* – ERFS). The survey originally reconciled the population census data with the fiscal data for a sample of people surveyed to calculate the standard of living in France and to establish its distribution. This survey still exists and data matching is still carried out. It has been an annual survey since 1996 and now matches data from the Labour Force Survey with fiscal data, plus data on social benefits since 2006.

We find another precursor in the area of data matching in the Permanent Demographic Sample (*échantillon démographique permanent* – EDP) set up in 1968. It originally entailed the matching of population census data and civil register data for a sample of people selected based on their dates of birth. For each of these people, the EDP is enriched each year with data from various sources, the number of which is gradually increasing. The Directorate of Research, Studies, Evaluation and Statistics (DREES) also recently matched the EDP with health data to create the Permanent Demographic Sample for Health (Health EDP).

During the 1970s, we also saw the development of various panels, which allow people to be tracked over time. In particular, the DEPP panels were introduced from 1973 to monitor academic trajectories. The panels were formed using various administrative databases from the school system. Since the 1990s, these panels have incorporated the results of national student knowledge assessments and surveys of students and families.

In the same period, the "Annual Social Data Declarations" (*déclarations annuelles de données sociales* – DADS) panel has been in place since 1976, linking data on wages and periods of employment reported by employers for a sample of employees. This panel, which was originally limited to the private sector, has gradually expanded into the public sector and now includes all active persons, including non-employees.

In 1988, DREES created the Inter-Scheme Sample of Retirees (*échantillon interrégimes de retraités* – EIR) which, for a sample of retirees, matches data on their pension amounts in the various schemes in order to reconstitute their overall pension amounts. Indeed, the retirement pension paid to an individual may come from several different schemes if they have changed schemes during their career.

A fully-fledged data collection mode, supported by several decades of opinions from the CNIS

These historical examples show that public statisticians have sought to best meet the needs and demands of users for decades now. In particular, they have tried to meet growing demand for increasingly granular and sophisticated data to understand the complexity of situations and trajectories. It has been possible to carry out these operations thanks to the intelligent combined use of multiple information sources, namely statistical surveys and a variety of administrative data. In particular, the aim was to weight the quality of these sources and their respective inputs, as well as integrating cost and expense constraints.

Data matching enables rich information to be produced at a reasonable cost. This is why it would be impossible to collect the equivalent of this information through direct data collection. This is due both to the wealth of administrative data that is often used and to the strength of matching in itself, which reconciles different data sources.

Finally, data matching is a way of collecting statistical information in its own right. Thus, the CNIS has long supported the development of panels and data matching, while stressing the importance of privacy and confidentiality, particularly in its medium-term opinions:

- CNIS medium-term consultation for 1999–2003: CNIS opinion on the inadequacy of monitoring of the social and employment trajectories of people, which has led to the development of panels;
- CNIS medium-term consultation for 1999–2003: “the Council requests that all producers of Official Statistics implement matching between data sources in order to enrich the analysis of the links between different themes, while ensuring strict respect for confidentiality when the matching concerns personally identifying information”.

Uses of data matching

Personal data matching is often proposed for monitoring trajectories or for evaluating or directing public policy. In fact, it allows for a much broader range of uses, which we will attempt to describe.

First, such matching makes it possible to improve the quality of statistical information. This is particularly evident in the case of measuring incomes and standards of living. Indeed, household incomes are better understood through administrative data (income reported to the tax authority, social benefits paid, etc.) than on the basis of direct surveys of people. As a result, since the 2000s, we have seen a widespread enrichment of household surveys using administrative data on various types of income.

In addition, data matching makes it possible to produce information at fine geographic levels, whereas survey samples, which are often too small, are regional in scale, at best. For example, the Localised Social and Fiscal File (*fichier localisé social et fiscal* – FiLoSoFi) measures income and poverty at local level (municipalities, neighbourhoods and square-kilometre grids) by reconciling fiscal files with comprehensive social benefits, as well as incorporating certain income types.

Data matching is also valuable, if not essential, for measuring complex phenomena and covering complete scopes. For example, the “all employees” or “all active people” databases provide a complete view of people’s employment situations over a year: employed in the public or private sector, employed in agriculture, employed by a private individual

and non-employees. In particular, they measure multi-activity. In contrast, the administrative data that track these populations are often distinct and, therefore, provide only a partial view of each population. Similarly, the Inter-Scheme Samples of Retirees and Taxpayers, based on data matching, provide a complete view of pension amounts and future pension entitlements, for all schemes combined. Accordingly, these inputs are essential for knowledge.

Furthermore, the data matching carried out within Official Statistics makes it possible to improve the wealth of statistical information, in particular for studying phenomena affecting more than one domain, which are covered by different sources. For example, this data matching makes it possible to study mortality or fertility by education level, social category or standard of living through the EDP, or to study social inequalities in health through the Health EDP.

In addition, many data matching operations make it possible to describe individual trajectories, particularly trajectories relating to integration in the labour market, employment trajectories and wage trajectories, or even the trajectories of certain populations, such as that of the population in receipt of statutory minimum incomes. They can be used to describe trajectories relating to different domains, such as changing standards of living or residential mobility when entering retirement by using the EDP.

Furthermore, it is well known that the reconstruction of individual trajectories enabled by data matching can be used to evaluate public policies. We find many examples of such use, particularly within the MSOs. Matching is used, for example, in the evaluation of National Education reforms or practices, by using the DEPP panels. DARES and DEPP also use this practice to evaluate the effects of employment support measures on the integration into the labour market of young people having completed apprenticeships or on professional trajectories, using the InserJeunes application which will be presented during the day.

At methodological level, data matching allows for a better understanding of certain phenomena, in particular by helping to analyse the differences between sources concerning the same concepts or similar ones. Thus, the matching of the Labour Force Survey with employer declarations (Annual Social Data Declarations [*déclarations annuelles de données sociales* – DADS] and Nominative Social Declarations [*déclarations sociales nominatives* – DSN]) aims to provide a better understanding of the concept of employment and the measurement thereof. The reconciliation of the Labour Force Survey, which measures unemployment as defined by the International Labour Organization (ILO), with the Pôle Emploi historical file of job seekers, allows for a better understanding of the concept of unemployment and the differences between its different definitions.

<u>How is data matched in practice?</u>

Beyond the examples and uses of data matching, let’s look at the methodology used. Matching personal data involves reconciling data about an individual person from multiple sources. These reconciliations may relate to surveys, specifically either multiple surveys covering an individual period or individual surveys covering consecutive periods. They can also bring in surveys and administrative data. This example makes it possible either to enrich surveys using variables derived from administrative data or to supplement administrative data using surveys conducted on samples. Finally, they can also create links between different types of administrative data on their own.

Reconciling data from multiple sources concerning the same person seems simple, but it is not always easy to implement. It can be difficult to verify that the individuals in the multiple sources reconciled in question are actually the same individual.

Reconciliations are simplest when the sources have a common identifier that enables unambiguous identification of people. This can be a certified identifier, such as the registration number in the National Directory for the Identification of Natural Persons (*numéro d'inscription au répertoire* – NIR) managed by INSEE, otherwise known as the Social Security Number, or the National Student Identifier (*identifiant national des étudiants et des élèves* – INE) managed by DEPP.

These reconciliations can be more complex in other cases, including when we have complete civil register data (surname, forename, date and place of birth and gender). Indeed, in such cases, data matching can be more or less easy or successful, in the sense that few people would be unmatched or mismatched, depending on the quality of the civil register. Typically, it is possible to find different spellings of surnames and forenames in the different sources. The matching can also be done on the basis of incomplete civil register data, for example when not having the surname, supplemented by other information such as an address or a municipality of residence. It can also be carried out using other personal characteristics to reconcile sources (addresses, years of birth, gender or other common variable). Concrete examples presented during this day will help to illustrate this diversity.

Beyond this issue of identifiers and matching keys, various methods can be used to match data, which we cannot discuss further within the limited scope of this presentation.

Which individual identifiers should be used for matching?

The use of common individual identifiers is therefore the easiest way to carry out data matching. Identifiers used for administrative purposes, such as the NIR, which is widely used in social administration, can be used by Official Statistics, under certain conditions, for statistical and research processing, when performing matching. Identifiers used for statistical or research purposes may also be used to match data. To reduce their sensitivity, the principle is to transform the NIR, a personal identification number, into a non-personal identifier that provides no information about people and does not allow linking back to individuals. These are hashing and encryption procedures, used to generate pseudonyms through the FOIN algorithmic procedure for health data, for example. Similarly, the CSNS is obtained through a cryptographic operation using the NIR.

This CSNS is a new feature introduced by the 2016 Law for a Digital Republic. It is a specific identifier to be used by the Official Statistical Service for the purpose of producing Official Statistics, which we will come back to during the day.

The legal framework for data matching has evolved and become more flexible over time, something that Sylvie Lagarde will expand upon.

Sylvie Lagarde

Hi everyone. The long history of data matching within the Official Statistical Service set out by Christel Colin has not been linear. This history has included different periods marked by changes in legislative, technical and international contexts. Therefore, I will try to present

these developments to you, without going into the detail of all the periods in question, simply by highlighting some of their key points.

A legal framework that has become more flexible over time

First, the legal framework has become more flexible over time. This relaxation is undoubtedly linked to a changing context. Indeed, developments in law and in society feed into each other.

Until 2004, this legal framework was mainly governed by the 1978 Law on Data Processing, Data Files and Individual Liberties. All processing of data relating to people that was carried out by the administrative authorities needed to be authorised by law or by regulatory texts, after a reasoned opinion from the CNIL. This framework was therefore very strict and involved rather cumbersome procedures. For example, the use of the NIR in data matching was particularly closely supervised; therefore, it required a decree from the Council of State and an opinion from the CNIL.

In 2004, we saw a first significant development of this legal framework, with an amendment of the Law on Data Processing, Data Files and Individual Liberties. This change transposes the development of the European legal framework, marked by European Directive No 95/46/EC of 24 October 1995 on data protection. This Directive has rendered the post-collection processing of data for statistical purposes or for scientific or historical research compatible with the original purpose of the data collection. Consequently, this amendment to the Law on Data Processing, Data Files and Individual Liberties made it possible to account, by means of a simple order, for data processing for the purpose of statistics or research, retaining a prior authorisation from the CNIL. However, this instance excluded sensitive data and required recognition of the absence of interconnected files of varying public interest. As a result, the matching of files on the basis of the NIR remained closely supervised, requiring a decree by the Council of State and an opinion from the CNIL until 2016.

Important steps were then taken, notably in 2016, with the Law for a Digital Republic, which introduced the CSNS in particular, and then in 2018, with the implementation of the General Data Protection Regulation (GDPR).

Thus, the bodies responsible for processing personal data are primarily responsible for compliance with the GDPR, without systematic referral to the CNIL upstream of such processing. This responsibility is based, in particular, on very strict principles of minimising the amount of data and the length of time it is kept. The GDPR still governs the drafting of impact assessments and the role of the Data Protection Officer. Its implementation therefore profoundly changes the legal context governing data processing.

In addition, the “NIR Framework” (“*cadre NIR*”) Decree provides for all possible uses of the NIR, including in processing for statistical purposes, as well as the conditions for access to the National Directory for the Identification of Natural Persons (*répertoire national d’identification des personnes physiques* – RNIPP). However, processing involving health data is an exception. With the exception of sensitive data, the Official Statistical Service can now use a single identifier for each individual, the CSNS, without a decree from the Council of State, to reconcile different files for statistical processing. This is a major development.

The specific legal framework for health data

Until 2016, there were multiple distinct CNIL authorisation regimes provided for by the Law on Data Processing, Data Files and Individual Liberties in the field of health data. These regimes were broken down in accordance with the purposes of the health data processing. Sometimes, they could require a prior opinion from a Scientific Council. As we have said, a decree by the Council of State also remained necessary to access the NIR.

The law on the modernisation of the French health care system, of 26 January 2016, introduced significant changes to this legal framework. In particular, it established the National Health Data System (SNDS), which was expanded in 2019 and provides for a reconciliation of data from the National Inter-Scheme Health Insurance Information System (*système national d'information interrégimes de l'Assurance maladie* – SNIIRAM) with hospital data from the Information System Medicalisation Programme (*Programme de médicalisation des systèmes d'information* – PMSI), data from the Centre for Epidemiology on the Medical Causes of Death (*Centre d'épidémiologie sur les causes médicales de décès* – CepiDc) and disability data with a history of 20 years. The SNDS is managed by the National Health Insurance Fund (*Caisse nationale d'assurance maladie* – CNAM) and then also by the Health Data Hub, created by Law No 2019-774 of 24 July 2019. The FOINised NIR is used as an identifier within the SNDS and it is not possible to return to the NIR.

In late 2019, the creation of the Health Data Hub provided a single point of access to health data for research, study or evaluation purposes. It also provides the secretariat for the Ethics and Scientific Committee for Health Research, Studies and Evaluation (*Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé* – CESREES) which issues an opinion to facilitate the CNIL's assessment of applications for authorisation to process health data for research, study or evaluation purposes.

<u>The recent context has led to a development in data matching</u>

Beyond the legal context, I would like to address the issue of the evolving technical and international context, which is in favour of the current expansion of data matching.

IT capacities are growing and we are increasingly using comprehensive administrative data. In the past, however, such comprehensiveness was not technically permitted. For example, in the context of using data from the DADS (which has now been replaced by the DSN) we were able to achieve comprehensiveness from the early 1990s, while we were previously limited to a sample of 1/25, then 1/12. Similarly, we have been able to comprehensively use data on income and housing taxes, followed by data on social benefits. In addition, the size of EDP samples or panels has increased.

We are also seeing a proliferation of available sources. In particular, administrative data are more numerous and are becoming more accessible. The sources are gradually expanding. They are being streamlined and centralised. For example, the DSN tends to extend its coverage to all employees, including those in the public sector. The issue of collection at source has also brought about changes to a lot of administrative data, with the introduction of the "Admission of Other Income" (*passage des revenus autres* – PASRAU) system, the Common Social Protection Directory (*répertoire national commun de la protection sociale* – RCNPS) and the Single Career Management Directory (*répertoire de gestion des carrières uniques* – RGCU).

In this manner, important databases are being established and they provide access opportunities for statisticians. This access is based on a legal framework laid down in particular by the 1951 Law, as amended, on Legal Obligation, Coordination and Confidentiality in statistical matters, the 2016 Law for a Digital Republic and European Regulation (EC) No 223/2009 on European statistics.

Lastly, data culture is advancing among the parties in possession of the sources through a more effective statistical methodology or through peer discussion.

Practices of other national statistics institutes

Within other national statistics institutes, there is also an explicit strategy for data matching and establishing statistical directories. In some instances, other countries have been using these practices for longer than France.

The countries of Northern Europe (Denmark, Finland, Norway and Sweden) have had a statistical system based on registers since the 1960s–1980s, with this start-up date varying from country to country. Gradually, this practice has been introduced in the Netherlands, which has had a system of interconnected and standardised surveys and registers in place since the early 2000s, initially with a view to moderating costs. Several of these States cannot legally collect information in a survey if that information is already available in administrative files. These institutes have therefore been required to develop their practices, on the basis of these legal obligations.

Ireland, the case of which will be discussed at the round table part of this meeting, is implementing the PECADO project, which brings together various administrative sources to produce population estimates.

Furthermore, Eurostat is promoting data linkage and the use of administrative data for European statistics in the context of the Vision 2020 Administrative data sources (ADMIN) project. This Directorate of the European Commission thus funds some joint work for European statistics.

Simultaneously, Anglo-Saxon countries and Southern European countries, such as Italy and Spain, are also developing rapidly in respect of these issues. This practice therefore transcends different cultural nuances. In this manner, Anglo-Saxon countries outside of Europe, which are less culturally focused on these issues, such as Canada, Australia or New Zealand, are also working together to examine the challenges associated with linking microdata, as well as the benefits of using administrative data and matched files that would combine surveys and administrative data, in particular:

- Statistics Canada uses a 2017 directive on linking microdata;
- The Australian Bureau of Statistics is directing the Multi-Agency Data Integration Project (MADIP), which involves six agencies and combines health, education and demographic data;
- In New Zealand, a project is being conducted to use administrative data with a “spinal column” and files to be matched.

Interestingly, these issues are being examined closely at international level, driven in particular by the Anglo-Saxon countries. This examination not only concerns the technical or methodological dimensions of data matching. It is a matter of investigating consultation

with users, together with ethical issues, or issues relating to the social licence. This social licence refers to data collection and data matching techniques. This examination goes beyond the legal and technical frameworks, encompassing all areas of society. We will return to these important issues during this meeting's round tables.

What are the safeguards?

We are therefore finding that there are increased opportunities to match data more widely and more easily. However, these new possibilities lead us to think about the definition of safeguards.

First, the question calls for consideration of the creation of a strong legal framework, which would provide a great deal of structure and would be important to fully understand. In particular, this framework is based on purely statistical purposes involving statistical confidentiality. It also means ensuring that personal data cannot be passed back to the owners of the administrative data. Therefore, such data matching must not lead to decisions about individuals.

It is also important to be able to discuss the matching with society representatives ahead of time. In this regard, the CNIS can play a very important role, within the framework of its various commissions. This involves conducting systematic discussions regarding Official Statistics work programmes, which involve a presentation of the data matching work with open-access documents and records. This challenge has been explained further in the CNIS medium-term opinion for 2019–2023. It is also necessary to systematically mention the use of the CSNS in the data matching operations of the Official Statistical Service in the work programmes presented to the CNIS, in order to make such use transparent.

Furthermore, the principle of fair data collection and transparent processing relates to other safeguards. In order to comply with these principles, all processing of personal data is reported on the websites of INSEE and the MSOs. And when survey data are enriched *ex post* with, for example, administrative data on income, the respondent is informed of this in the notification letter they receive before their data are collected.

In addition, the principle of data minimisation is an important safeguard. It is a case of using only the data that is necessary. The issue of how long data is retained is also important. These aspects are incorporated into the impact studies carried out in advance of the statistical processing.

Finally, work on IT security remains an important safeguard. There is a need to be able to restrict and control access to a small number of people responsible for data processing.

Conclusion

In conclusion, the Official Statistical Service is bolstered by a long-standing practice of personal data matching, which was initially rather partitioned by thematic area, with little sharing between peers, both nationally and internationally.

However, the changes to the legal, technical and international context, which have been taking place over the past few years, requires changing the scale and defining a new strategy, in order to facilitate data matching within the Official Statistical Service. It is a matter of increasingly relying on the CSNS and, soon, on the service offered by the Stat-

istical Directory of individuals and Housing (*répertoire statistique des individus et des logements* – RESIL), both of which will be presented to you. In particular, where such support is not possible, the NIR can be used, particularly for data matches performed with social protection organisations, which may be outside the Official Statistical Service.

Finally, it is important to inform and communicate and consult with society in connection with the ethical issues of respect for privacy arising from data matching. In addition, it is particularly important to bear these issues in mind in a society in which digital technology is increasingly important. Today's meeting thus provides us with an opportunity to discuss these challenges together.

Mireille Elbaum

Thank you. I would like to provide two details while waiting for the next speaker to get set up. In its opinion of 22 September 2021 on data matching, the ASP stated that statistical programmes must very clearly indicate "their objectives, their content and the conditions for their execution" so that they can be taken into account in the work of CNIS committees.

I would like to provide a further detail regarding health data. A significant step was taken by Decree No 2021-848 of 29 June 2021, allowing INSEE access to the SNDS. In particular, the previous situation explains why the Health EDP was created by DREES and not INSEE. In fact, this Decree also opens up new possibilities, in particular for all the teams of the National Centre for Scientific Research (*Centre national de la recherche scientifique* – CNRS), which also did not have access to the SNDS. The Decree also allows the Director-General of INSERM to permit access to the SNDS for all its teams.

As a result, this Decree will potentially give rise to important developments. However, with the SNDS data retention limit remaining at 19 years, problems still arise for organisations such as the National Institute for Demographic Studies (*Institut national d'études démographiques* – INED), which does historical work, or more broadly, for any historical research perspective.

Kamel Gadouche will now address the issue of data matching offers made to researchers.

Data matching performed by researchers

Kamel Gadouche

Thanks to Mireille Elbaum for her introductory remarks, as well as to Christel Colin and Sylvie Lagarde for their presentation of the practices used in Official Statistics. I will describe the data matching performed by researchers, for the purpose of research. The dividing line between statistics and research is blurry and these two fields share common points. To a certain extent, research is lagging behind in terms of what is possible through data matching, although progress has been possible in recent years.

We will start by presenting some of the uses for researchers who carry out data matching, while looking at the challenges raised by this practice. Next, we will describe a new secure data matching method based on trusted and secure third parties, in the context of secure bubbles. Finally, we will highlight some concrete achievements related to the CASD, which have been enabled in recent years, notably through legal progress.

<u>Uses for the researchers and the challenges raised by data matching</u>

For years now, researchers have had access to an increasing amount of data through the CASD or other intermediaries. They are more accustomed to manipulating data and they are full of ideas for data matching, allowing them to answer increasingly precise questions.

They use this data matching to study topics such as: intergenerational income mobility between parents and their children; assessing career path reform; links between career and health; monitoring periods of unemployment and career trajectories; or the effectiveness of training, which receives large amounts of public funding.

For example, if we wanted to study the links between career and health, we could rely firstly on health data from the SNDS with its own FOIN-based identifiers and secondly on career data from the DSN, which has another type of identifier, which is also not personally identifying.

However, we face a challenge when we want to reconcile these two types of data so that we can track individuals. Indeed, these two identifiers do not match and it is not possible to combine them directly. This makes it difficult to compare individuals. This problem is the result of a deliberate security measure, with each of these files having their own types of identifiers.

Nevertheless, while this may seem impossible in principle, it is possible to overcome this problem. To do this, you need to be able to work back to the NIR, which is a certified identifier. This NIR then makes it possible to identify individuals and to perform a match.

A new data matching procedure based on secure bubbles

Working back to the NIR, which is a very detailed, and therefore sensitive, piece of health data, requires the availability of a secure environment and procedure. And this issue is a key point of our meeting.

Currently, there are many data sources to which researchers have had access in recent years, through the CASD or other systems. These sources are rich and often exhaustive. The researchers are aware that the French Official Statistical Service produces a lot of high-quality data. This is an important asset for research and the possibility of matching these data opens the door to new possibilities.

The 2016 Law for a Digital Republic made it legally permissible to carry out this type of data matching that uses the NIR. However, this operation is carried out within a specific framework. To do this, there is a specific procedure for research, which shares similarities with the procedure for Official Statistics, which I will not describe.

As Mireille Elbaum explained in her introduction, this procedure is complex and involves several stakeholders. This procedure involves two external organisations that act as trusted third parties and are neither producers nor users of data.

For example, producer A and producer B will each create their own ID1 and ID2 index identifiers associated with the NIR. First, they each send the files that they want to match, with only the NIR and this index identifier. At the same time, they each send the information data associated with this ID1 to the second trusted third party.

Then, the first trusted third party will hash the NIR from the two files provided by the two producers. This hashing is a non-reversible cryptography operation associated with a secret key. By applying a hashing function, we get a very long code of the following type:

99740afc57d67b5879b664681d0f40789cae2109c74fe9d73a7a72c889ab01676aad8d-c8a4731b8d39d36e9da36fa5dfd30c47d12d6547cf9d2033a5a67c6148

The first principle of this algorithm implies that two identical NIRs are associated with a single “H” code. Second, it is necessary to ensure that it is not possible to work back to the NIR using this code, which is made possible by the attributes of the hashing operation. Thus, since two identical NIRs from the two different files provide the same code, it is possible to match the files.

The first “identity” trusted third party then sends the hashed NIRs of the files sent by the two producers to the second third party. The latter then methodically gathers the file information with the hashed NIRs, to finally integrate the matched files into a secure bubble, replacing these hashed NIRs with new identifiers as part of this step.

It is important to note that the first trusted third party has the NIR and only the NIR, while the second third party has the data, but not the NIR. Therefore, the second third party cannot match the information from the data with the identifier codes. None of the parties involved in the procedure, whether the two third parties or the two producers, has access to all the information.

Once this matched file is created, the issue of access to it then arises. Thus, this access takes place within the framework of a secure bubble system, provided by the CASD. In 2019, the Secure Access Data Centre took the form of a non-profit Public Interest Group (*groupement d’intérêt public* – GIP), bringing together INSEE, GENES, the CNRS, the Ecole Polytechnique and HEC Paris. The main mission of this consortium is to establish secure bubbles for research, study, evaluation and innovation purposes. I would like to show you a video summarising the CASD’s missions:

“CASD: A bubble to protect data.” (Le Monde)

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

Many companies, authorities and data producers want to take advantage of their data and may need external expertise to do so. At the same time, many data scientists, researchers and consultants want access to confidential data for their algorithms, studies or research. Sending copies of data to users directly does not offer guarantees of security and traceability. To ensure their security, data must remain contained within a secure bubble. The CASD is a trusted third party providing equipment designed to enable users to work on data supplied by producers under high security conditions.

“The centre (CASD) holds information, including tax and medical data.” (The New York Times)

“After signing a contract with the CASD, the producer securely transmits a copy of their data, which are then stored in the CASD secure infrastructure in France. When making their data available, a producer retains ownership and defines the rules for the use thereof. The CASD provides data owners with an opportunity to take advantage of them

while ensuring the security thereof and compliance with the GDPR. The producer determines who is authorised to access the data for projects. The data remain secure and accessible at all times and become easy to use for users."

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

To ensure the security of the sensitive data entrusted to it, the CASD has created the SD-Box, a remote data access box incorporating biometric authentication and encrypted communications guaranteeing end-to-end security of the system, which includes more than 300 security measures. Once connected to their CASD server via their SD-Box, a user can access the sensitive data that are authorised for their project. The user can then analyse them using a wide range of data processing software available on the server. Permanently accessible servers allow authorised members of a project to work together on the data and share their work. The user can also import or export files and scripts that must be retained and verified by the CASD teams to ensure that they do not contain sensitive data!

These secure bubbles are hosted on servers housed in a former military site in the Paris region, with a security patrol. Within this building, there is a security system with triple-factor biometric access control – card, hand and code.

Some data matching achievements enabled by the CASD

Finally, I will briefly mention some of the achievements enabled by the CASD. First, as part of a study on innovation, it has been possible to compare the income tax file and the qualifications file. In addition, Antoine Bozio, who is in attendance at this meeting, participated in data matching between the corporate share ownership file and the income tax file. In addition, other data matching was carried out based on the DSN with data from Pôle Emploi and data on vocational training as part of the Training, Unemployment and Employment (*Formation, chômage et emploi* – FORCE) system. Therefore, the MIDAS data matching project should make it possible to link data from the DSN, Pôle Emploi and the National Family Benefits Fund (*Caisse Nationale des Allocations Familiales* – CNAF).

I am now going to focus on FORCE data matching, the data of which are currently being used by research teams. This is the matching of data from DARES and Pôle Emploi. The data matching involved DATASTORM, a first trusted third party, as well as the CASD, which assumed the role of the second trusted third party. This data matching followed the procedure we have described in this presentation.

Conclusion

Finally, significant progress has been made in recent years in data matching. Research projects carried out on the basis of data matching performed within the CASD framework and covering multiple subject areas are listed and available on the following webpage: www.casd.eu/en/projects. For the sake of transparency and to ensure good public information, the description of these projects is published on both the CASD website and other websites.

I would like to thank the organisers of this day for enabling this meeting to be held, centred on a practice the challenges of which will remain crucial in the years to come. Indeed, data matching is a source of innovation and advances in knowledge. These advances are

made with research and for research. They also benefit society, through the improvement of the *ex ante* and *ex post* evaluation of public policies and the de-partitioning and cross-cutting nature of studies. Finally, as Patrice Duran pointed out, data matching is used to break the silos.

The CASD is therefore particularly enthusiastic about the idea of participating, in cooperation with Official Statistics, in the development of these data sources that are proving to be innovative for research. I stress that Official Statistics have always been particularly considerate for researchers. Thank you for listening.

Discussions

Mireille Elbaum

Among the members of the CASD, you mentioned INSEE, but not the Ministry of Health or INSERM. And I note that a similar system should be put in place for the SNDS as part of the Health Data Hub.

However, Patrice Duran explained that the coronavirus crisis was not just a health crisis. And, in the context of the Official Statistical Service's missions, it is indeed essential to cross-reference health data with other sources. I hope that the development of this type of data reconciliation will be encouraged by more stakeholders in the future.

Moreover, I note that when statisticians conduct a statistical study or research, they do not always know in advance all the data they will need and ideas can arise as the data is being used. Therefore, the reconciliation of these needs with the principle of data minimisation and with the cumbersome nature of the procedures is not always obvious in practice.

Yvon Serieyx, UNAF

Presentations from this session have shown the complexity of matching health data. Taking a new step should take not only time, but also money, as we can see with the costs incurred by the CASD.

Indeed, we must ensure that privacy and rules on how long data can be retained are respected. Nevertheless, it seems difficult to carry out an undefined number of procedures, in response to the new needs that may arise in the course of the research.

We therefore believe that the issue of the *ex ante* determination of the necessary data matching must be a separate subject which we must discuss in the context of the consultations and the whole comitology process prior to the survey work. Accordingly, a chapter dedicated to this issue should be included in the files presented to the CNIS for opinions on appropriateness. It is a case of working in advance to establish an inventory of the data to be matched and all the knowledge to be advanced, after a review of the scientific literature.

Mireille Elbaum

In particular, the ASP's statement of 22 September 2021 on data matching recalls the importance of this diagnosis of needs which is incumbent upon the CNIS and which requires knowledge of the actual situation.

In addition, in the health and social areas, the mission of the INSEE Internal Audit Unit and the IGAS has led to the diagnosis of a number of information needs related to data matching that can feed into discussions in advance of data processing.

Nevertheless, it is not appropriate to make data matching conditional upon the same type of opinions on appropriateness or other opinions to which surveys are subject. Indeed, they are already giving rise to another “comitology” that is rather cumbersome.

Thus, this ASP statement on data matching simply emphasises the need to mention these projects or initiatives in the programmes sent to the CNIS, so that its thematic commissions can take them into account.

Christel Colin

Mireille Elbaum’s question shows that it is useful to ask the right questions ahead of data processing, even if it is not possible to think about everything before starting data matching procedures.

Furthermore, the new technical or legal facilities do not justify the use of all-out data matching. We need to consider its use based on defined questions, while thinking ahead of time about the data and procedures that are needed. This is important for transparency in the uses of such data matching.

This important *ex ante* phase includes a methodological and statistical consideration of the choice of sources. It also requires consultation with users. It is now requested that the data matching and its purposes be precisely described within the framework of the CNIS work programmes. With this in mind, I would like to point out that the sites of some MSOs do describe the sources and the data matching, thereby participating in this transparency process.

It is not easy to balance the principle of transparency with the sometimes necessary use of additional data during research or statistical studies. Nevertheless, the procedures are simplified and can facilitate the right to change one’s mind.

Sylvie Lagarde

I would add that needs are considered in advance of the implementation of all personal data processing, including data matching, as part of a legal analysis. This consideration is conducted in particular when drawing up the Data Protection Compliance Document (*document de conformité à la protection des données* – DCPOD).

It is a matter of integrating this consideration into the data matching operations, in order to define variables. However, this approach is not so simple, particularly with regard to taking into account the principle of data minimisation, which is checked during discussions with users and the Data Production Officer (DPO).

Louis-André Vallet, CNRS

How is the consultation organised between Official Statistics and public establishments of a scientific and technological nature (*établissements publics à caractère scientifique et*

technologique – EPSTs) such as the CNRS, the National Research Institute for Agriculture, Food and the Environment (INRAE) or INSERM?

Claude Castelluccia, CNIL

Is the CSNS permanent or temporary? Is it the same for all data matching?
Olivier Hays, CESP

Why does the Public Interest Group CASD include the Technical Polytechnic School (X) and HEC, but not the other schools or all engineering schools?

Mireille Elbaum

Regarding the consultation between Official Statistics and EPSTs, it should be stated that the practices are very varied. Indeed, researchers can work on data in multiple ways. Based on the principle of the independence of researchers, they may, in particular, carry out their work free from any dependence on the organisations holding the data, within the framework of the CASD or INSERM.

When researchers want to match the survey data they have produced with other sources, they can conduct joint operations with the Official Statistical Service. This is the case, for example, in the recent survey “Epidemiology and living conditions associated with Covid-19” (*épidémiologie et conditions de vie liées au covid-19* – EpiCov), which included INSERM, INSEE and DREES.

Finally, researchers can work on behalf of organisations such as MSOs on the basis of study and research credits. They can then make use of matched files or surveys, which are made available to them under the responsibility of these organisations. They can still follow adjacent procedures to enable them to complete their studies.

Finally, there is a very broad range of practices centred around the fundamental issue of transparency in Official Statistics. In this regard, we try to promote this transparency among the MSOs, based on the level of transparency provided by certain public bodies that disseminate statistics, such as the CNAM or the CNAF. Making progress in this area should make research easier.

Christel Colin

To answer the question concerning the CSNS, I would like to point out that this code is not permanent and lasts ten years, unless a security breach requires a renewal before that term has expired. Lionel Espinasse will present the CSNS during the third session of this meeting.

Kamel Gadouche

In response to the question concerning the formation of the Public Interest Group, it should be known that the members of this consortium are those who contribute to the operation of the CASD. However, CASD users are part of the much broader community of researchers. For example, almost all French research organisations have access to CASD services.

In addition, the CASD was created as part of the "Equipment of Excellence" (*équipement d'excellence* – EQUIPEX) project, which is part of the Investments for the Future (*Programme d'investissements d'avenir* – PIA) programme. However, the EQUIPEX call for projects, aimed at securing funding, was based on a localisation approach. The choice of the consortium's schools was therefore justified in the context of a campus approach.

In any event, any organisation may apply to join this consortium and discussions have been held with INSERM and the Ministry of Health in particular. The General Assembly of this Public Interest Group decides on the admission of new members.

SESSION 2 – SOME EXAMPLES OF DATA MATCHING WITHIN OFFICIAL STATISTICS

Chair of the session: Antoine Bozio, President of the CNIS Public Services Commission (*Commission Services publics et services aux publics*), Institute of Public Policy (IPP);

Patrick Aubert, Deputy Director of Ensuring Solidarity of DREES, within the Ministry of Solidarity and Health, for a presentation of the national inter-scheme sample of recipients of in-work support and statutory minimum incomes;

Vladimir Passeron, Head of the Employment and Employment Income Department (*département de l'emploi et des revenus d'activité*), INSEE, for a presentation of the data matching between the Labour Force Survey and the historical Pôle Emploi file, to understand the differences between the numbers of unemployed people and job-seekers;

Nathalie Caron, Deputy Director of Summaries at the DEPP, Ministry of National Education, Youth and Sport (*Ministère de l'Education Nationale de la Jeunesse et des Sports* – MENJS), for a presentation of the InserJeunes information system aimed at providing a better understanding of youth integration.

Antoine Bozio

Hi everyone. I am a lecturer at the School for Advanced Studies in Social Sciences (*Ecole des hautes études en sciences sociales* – EHESS), a professor of economics at the Paris School of Economics and director of the Institute of Public Policy, which makes heavy use of the data discussed this morning and makes a modest contribution to data matching.

The aim of this session is to highlight examples of achievements of data matching in Official Statistics, from the perspective of three subject-areas. We will describe the different ways to use such data matching, while promoting them to a wide audience. I will end my introduction here, so as to provide more time for presentations and discussions.

First, Patrick Aubert, Deputy Director at DREES, will present the value of the national inter-scheme sample of recipients of in-work support and statutory minimum incomes (*échantillon national interrégime d'allocataires de compléments de revenus et de minima sociaux* – ENIACRAMS), which allows us to study the entire pathway of recipients of various forms of in-work support and statutory minimum incomes.

Secondly, Vladimir Passeron, Head of the INSEE Employment and Employment Income Department, will demonstrate the value of data matching between the Labour Force Survey, which is designed to measure unemployment as defined by the ILO, and Pôle Emploi

data. This data matching should allow an analysis of the relationship between the statistical concept of unemployment and its administrative concept.

Lastly, Nathalie Caron, Deputy Director of the DEPP, will highlight the value of monitoring the employment integration pathways of young people based on data gained from the data matching of the InserJeunes information system.

The national inter-scheme sample of recipients of in-work support and statutory minimum incomes (ENIACRAMS)

Patrick Aubert

ENIACRAMS makes it possible to draw a link between this meeting and the meeting to be held on 18 May concerning panels. Like other panels, ENIACRAMS is created based on data matching.

The ENIACRAMS panel

ENIACRAMS has existed since 2001. This panel aims to monitor the population of interest composed of the recipients of the so-called working age statutory minimum incomes, monitored through to retirement age. These are recipients of the former Minimum Integration Income (*revenu minimum d'insertion* – RMI), the Single Parent Allowance (*allocation de parents isolés* – API), the Active Solidarity Income (*revenu de solidarité active* – RSA), the Specific Solidarity Allowance (*allocation de solidarité spécifique* – ASS) and the Allowance for Adults with Disabilities (*allocation aux adultes handicapés* – AAH). In short, even though social benefits have developed over this period, the panel covers the area of statutory minimum incomes.

This panel expanded its coverage from 2009 onwards by integrating benefits related to in-work support, namely the employment-related portion of the RSA and then the In-Work Benefit (*prime d'activité*) that replaced it in 2016. Thus, the name of this panel was changed from ENIAMS to ENIACRAMS.

ENIACRAMS is not exhaustive. Indeed, the panel was created at a time when IT capabilities made it difficult to achieve exhaustiveness and the principle of data minimisation also led to the decision to opt for a sample. In spite of the foregoing, this sample is much larger than those studied in the surveys, including around 170,000 RSA recipients or their partners, 90,000 AAH recipients, 25,000 ASS recipients and 410,000 in-work benefit recipients or their partners by the end of 2020. This sample therefore makes it possible to carry out an analysis at a very granular level.

It is a question of monitoring individuals in a longitudinal manner, with an interest both in their trajectories linked to statutory minimum incomes and, where appropriate, following their trajectories once they are no longer in receipt of these benefits. Thus, the panel can monitor different trajectories of benefit recipients, who can leave and then return to the population in receipt of statutory minimum incomes. This panel is therefore of particular interest to help better understand the trajectories of leaving the population in receipt of statutory minimum incomes.

Nevertheless, this panel focuses on situations at the end of the year, as it gathers information on an annual basis. We therefore do not seek to go into all the details of the trajectories, restricting ourselves to monitoring benefit recipients from year to year.

The information contained within ENIACRAMS

ENIACRAMS is created on the basis of matching administrative data from two main types of sources. On the one hand, these data are provided by the social benefit funds, namely the CNAF, the Agricultural Workers' Mutual Benefit Fund (*Mutualité sociale agricole* – MSA) and Pôle Emploi. On the other hand, INSEE provides important additional information concerning knowledge of the individuals, in particular demographic elements that have helped to define the outline of this sample. INSEE also provides data that make it possible to track mortality and, finally, it also provides employment data from the "all employees" and "all active people" panels.

Since statutory minimum incomes relate to the area of social security, the NIR is present in all administrative sources involved. This makes data matching easier.

Thanks to double blind procedures and a non-identifying number, DREES is unable to retrieve identifying NIRs during data matching.

The matched information is very rich. It includes the characteristics of the recipients, as well as details of the benefits received.

Indeed, the funds that pay out the statutory minimum incomes and in-work support also disburse many other benefits, providing additional data on housing support, family benefits or other unemployment benefits under the social security system. When we look at the trajectories, we examine not only the statutory minimum incomes themselves, but also other social benefits that support people. Furthermore, employment data make it possible to observe the entire employment trajectory, both during and after being in receipt of statutory minimum incomes, as well as a multitude of socio-demographic characteristics.

In addition, ENIACRAMS is a coupled system which, in addition to serving as a basis for studies, can also be used as a basis for the drawing of samples for surveys of the population under study.

In this way, it is therefore possible to enrich ENIACRAMS with survey data that provide specific information, which we would not be able to draw from administrative sources. We are therefore able to collect data that is more personal, which can provide us with information on opinions or aspirations, in particular. This is what is done in the context of surveys of recipients of statutory minimum incomes ("BMS survey"), conducted every 6–7 years on average (the most recent one being in 2018).

This point should be emphasised, as we tend to do the opposite. Indeed, we usually start with surveys, before attempting to enrich them by matching them with administrative data. However, such a procedure is often costly, cumbersome, and imperfect, because we never find 100% of the survey population and are forced to perform entries.

Thus, the benefit of ENIACRAMS is that it makes it possible not only to easily find all the corresponding data, but also to monitor the population studied after the survey. For example, we can link all the characteristics observed in the survey to the evolution of the trajectories both during and after being in receipt of statutory minimum incomes, five or ten years later, or even annually. This system is interesting and could act as a model for other fields and panels.

ENIACRAMS, the reference source for the elements of the trajectories of recipients of statutory minimum incomes

In concrete terms, this source is used by DREES to produce the various annual statistics relating to the trajectories linked to statutory minimum incomes, in particular as part of the "Statutory Minimum Incomes and Social Benefits" (*minima sociaux et prestations sociales*) published every year at the start of the academic year. This overview contains various key annual indicators, such as the rates of people starting and ending receipt of statutory minimum incomes, the length of time in receipt of statutory minimum incomes or the rate of people returning to receipt of statutory minimum incomes. These data are also available for research, within the framework of the CASD.

For example, by using ENIACRAMS, we were able to obtain a graph showing, among other things, the rate of people starting receipt of statutory minimum incomes by age of the recipients. These rates are calculated based on the number of new people in receipt of statutory minimum incomes year on year. We find that this rate is the highest among young people, at around 30% for those under 30 and decreasing gradually, falling to around 10% for the elderly.

The graph also shows information about the trajectories of recipients, making a distinction among these new recipients between those who had already been in receipt of statutory minimum incomes over the past decade and those who were "real" new recipients who would not have received statutory minimum incomes during the same period. We see that the bulk of people under the age of 30 are "real" new recipients. Conversely, more than half of those over the age of 30 had already been in receipt of statutory minimum incomes over the previous ten years.

For another example, ENIACRAMS has also enabled us to produce a table setting out the status of former RSA or ASS recipients one year after they stopped receiving those benefits. This data is obtained by matching information linked to the RSA and ASS with employment data and unemployment benefits data. Each year, it is thus possible to monitor the proportion of people employed under permanent or fixed-term employment contracts, whether full-time or part-time, among those ending receipt of the RSA or ASS, or the proportion of recipients of the in-work benefit. This table also includes the proportion of people who died. It is thus possible to create particularly rich indicators.

Data matching possibilities to broaden knowledge of trajectories

This panel is already very rich, but the issue of the trajectories of the recipients of statutory minimum incomes is a very broad one. So let's take a look at the data matching envisaged in the near future to improve knowledge of these trajectories.

We want to avoid a silo mentality and to describe the trajectory and the living conditions of people in a global manner. In addition, we want to participate in this small revolution that affects the world of Official Statistics and makes data matching much simpler.

The benefit of ENIACRAMS is that it is a large sample that allows for matches with multiple sources to be obtained. These matches are facilitated by the NIR, in particular, to which the CSNS can be applied. In addition, it should be noted that individuals are selected in ENIACRAMS based on their date of birth, including the days of the EDP. As a result, this sample shares a common characteristic with many panels, therefore allowing ex-

tensive data matching. We hope to carry out a number of new data matching operations this year.

The first two data matching projects we will present relate to the integration trajectories of recipients of statutory minimum incomes.

First, the system for individual reporting on integration (*Remontées individuelles sur l'insertion* – RI-insertion), which is in development, should make it possible to assemble data on all support and guidance actions carried out by organisations working with these recipients – departmental councils, Family Allowance Fund (*Caisse d'allocations familiales* – CAF) and Pôle Emploi. By matching these data, we would be able to link these support actions with the trajectories. It could then be possible to identify territorial disparities or the most effective actions to help people return to work.

Another data matching operation is planned using data from the Labour Movements Information System (*système d'information sur les mouvements de main-d'œuvre* – SISMMO) created by DARES based on the DSN. This file supplements the INSEE data. It focuses on labour movements, its time scale is very fine and its data are easily accessible. The SISMMO file would make it possible to enrich the indicators relating to people ending their receipt of statutory minimum incomes through employment. It would also help to strengthen indicators on employment trajectories. This contribution would be all the more beneficial given that one in six recipients of statutory minimum incomes is in employment, either in insecure part-time employment or too poorly paid. We therefore seek to cover the whole trajectory of the insecurity of recipients of statutory minimum incomes.

Other reconciliations are being considered and should be outside the traditional coverage of ENIACRAMS. In particular, they are part of this revolution, which tends to break free from the silo mentality, thus breaking with the sealed off development of thematic works.

One of the first major steps that could be taken in this direction could concern the reconciliation of various pieces of data on integration and the world of retirement. It would indeed be possible to match the data on these trajectories with the data from the Inter-Scheme Sample of Retirees (*échantillon interrégimes de retraités* – EIR) and the Inter-Scheme Sample of Taxpayers (*échantillon interrégime de cotisants* – EIC).

At present, we follow the trajectories of recipients of statutory minimum incomes very well, but they disappear from our radar as soon as they enter retirement. Indeed, while we have elements relating to a part of their careers, which would allow us to estimate pension entitlements, we lack information in this area. We will then finally be able to check the proportion of people who would move directly from the RSA to the minimum pension entitlement (*minimum vieillesse*). It should also be conceivable to study the influence that periods without validation of pension entitlements have on the pension amounts of recipients of statutory minimum incomes. In any event, the links between the issue of integration and that of retirement are very rich.

Furthermore, it would also be possible to better understand the issue of disability by matching data from ENIACRAMS with data from the Everyday Life and Health (*Vie quotidienne et santé* – VQS) survey, which is part of the series of "independence" (*autonomie*) surveys.

The link between disability and statutory minimum incomes refers to the AAH. However, with regard to the responses of the recipients during the surveys, we realise that disability also concerns recipients of other statutory minimum incomes, such as the RSA. Accordingly, these questions lead to considerations about public action regarding the targeting of benefits. Reconciliation with the VQS survey, which contains information on the functional limitations of persons, will make it possible to better understand the characteristics of the disabilities of persons in receipt of statutory minimum incomes.

Other projects, which are farther in the future, also aim to move beyond a silo mentality. It is a case of cross-referencing data from ENIACRAMS with those from the DARES Youth Trajectories to Active Labour Market measures (*Trajectoires des jeunes aux mesures actives du marché du travail* – TRAJAM) panel, which concerns the youth guarantee. Data matching will also be possible using data on social housing and on homeless people taken from the SI-SIAO Integrated Hospitality and Guidance Services (*services intégrés d'accueil et d'orientation* – SIAO) software.

There are still plans to carry out reconciliation with the child protection data from the individual longitudinal child protection observation (*Observation longitudinale individuelle en protection de l'enfance* – OLINPE) system. This data matching should help to reflect on the trajectories of people no longer in receipt of children's social welfare measures, particularly in terms of integration into the labour market and careers.

Conclusion

In conclusion, beyond this list of projects, we find it difficult to envisage data matching using SNDS health data. This type of data matching is not impossible, but it is more complex and should take time. I am therefore going to take advantage of my opportunity to speak to pass on DREES's commitment to bringing together the social and health worlds. Within the framework of ENIACRAMS, this reconciliation would contribute to the understanding of integration and, to an even greater extent, to the understanding of disability. Data matching involving health data is one of the major challenges facing the Official Statistical Service.

Data matching between the Labour Force Survey and the Pôle Emploi historical file to understand the differences between the numbers of unemployed people and job seekers

Vladimir Passeron

Hi everyone. Data matching between the Labour Force Survey, conducted among people drawn at random, and the Pôle Emploi historical file was carried out two years ago, under rather more rough and ready conditions compared to the data matching involving ENIACRAMS.

Unemployment and job seekers registered with Pôle Emploi: two measurements that have diverged considerably since 2009

We wanted to perform data matching between the Labour Force Survey and the Pôle Emploi historical file, based on the finding that the number of unemployed persons, as defined by the ILO, measured each year in the Labour Force Survey, began to diverge considerably from the number of job seekers registered with Pôle Emploi and classed as category A (without a job).

Thus, we can see that these two indicators were at an equivalent level around 2008 and that they gradually diverged and now differ by more than a million people. We have therefore concentrated on the period 2013–2017, which saw this difference widen, with nearly 200,000 additional people registered in category A with Pôle Emploi and about 200,000 fewer unemployed people, as defined by the ILO.

We therefore wanted to understand what was going on at the individual level, especially given that, at first glance, these two definitions are very similar. To that end, teams from INSEE, DARES and Pôle Emploi worked together.

Unemployed people as defined by the ILO and jobseekers registered with Pôle Emploi

I would like to remind you that an unemployed person, as defined by the ILO, is a person without a job who is actively looking for work and who is available to work within the next two weeks. This definition of unemployment allows for international comparisons to be made. Thus, very specific questions are asked in the Labour Force Survey to distinguish between unemployed people on the basis of this definition.

Conversely, a jobseeker registered with Pôle Emploi at the end of the month in category A is without a job and is required to search for one. However, there are no criteria or questions that make it possible to measure the active character of their job search.

Thus, by comparing the populations according to these two definitions on a Venn diagram, we find that some of the unemployed people as defined by the ILO are not registered with Pôle Emploi. Conversely, we find that some of the job seekers registered with Pôle Emploi and in category A are not considered unemployed in accordance with the ILO definition.

This data matching, which was intended to enrich the Labour Force Survey with information from Pôle Emploi historical files, was not performed directly. Indeed, we were only able to rely on the forename, date of birth and address, given that we do not retain surnames in the Labour Force Survey as the sample it uses is a sample of dwellings, meaning that it is not necessary to retain them for collection. In contrast, it is important to correctly identify the different individuals in each dwelling by their forenames and dates of birth so that they can be monitored year on year, as the Labour Force Survey is carried out on a panel surveyed quarterly.

However, although we had a forename, a date of birth and an address (e.g. Bruno, born on 1 January 1970 and living at 1 rue de Bercy) in the two databases to be matched and for each individual, the data matching between these two databases has not been easy.

On the one hand, the postal address that Pôle Emploi has to allow it to contact job seekers may sometimes differ greatly from the residential address in the housing tax files, which is the sampling base of the Labour Force Survey.

Forenames that may differ between the two sources

On the other hand, official forenames may differ from the forenames collected by interviewers, which may include spelling variants (Mathieu and Matthieu, Lorène and Lorraine), which raises even greater difficulties for non-French-sounding forenames. In addition, we also accept nicknames such as Cathy for Catherine, as forenames are primarily intended to identify people and monitor them from one quarter to the next.

In the process, we wondered whether it would not have been beneficial to ask identifying questions in the context of this Labour Force Survey from the outset. We could have collected surnames and even NIRs. This would have allowed us to collect the information that we struggle to find today, particularly on incomes, by performing data matching on administrative sources.

In any event, we have implemented different solutions to overcome this difficulty related to forenames. In particular, we compared character strings by accepting minor spelling differences, we calculated the distance between the character strings and we used phoneticised first names.

Addresses that may differ between the two sources

In addition, we still needed to take into account dwellings for which the addresses may differ in the tax files and in the Labour Force Survey. These are mainly dwellings located in buildings overlooking two streets. Corner houses raised this problem more than others. In order to deal with this problem, we geolocated the addresses of the two files to be matched and we then calculated the distances between the two addresses, thereby allowing a reconciliation between individuals in the same dwelling but whose addresses differ in the two sources.

To further improve our matching rates, we have taken another characteristic into account. We found that some groups of forenames shared dates of birth, but were associated with different addresses. Given that they had the same dates of birth in both sources, it was very likely that they were the same people.

Data matching to observe the overlap of two populations

In total, between 2012 and 2017, about 17 million people were registered in Pôle Emploi category A. At the same time, our Labour Force Survey covered 400,000 people. We were ultimately able to perform data matching with a reconciliation rate of 85%. This rate is both low and high. However, we validated the quality of this data matching using a variety of robustness checks, including “manual” checks.

Finally, this data matching enabled us to construct a Venn diagram cross referencing activity categories, firstly as defined by the Labour Force Survey and secondly as defined by the different statistical categories used by Pôle Emploi.

Without commenting on all of the data in the diagram created, let us look at the difference between the numbers of unemployed people registered with Pôle Emploi in category A and of unemployed people as defined by the ILO.

First, we note that the data matching provided a good representation of overall developments in the period 2013-2017, which validated its robustness and quality.

In 2017, 44% of those registered in category A are not unemployed as defined by the ILO...

Finally, we find that in 2017, of 100 persons registered with Pôle Emploi in category A, 44 are not unemployed as defined by the ILO. These figures are rather structural and do not vary significantly year on year.

These 44 people are generally older people close to retirement who, as they are discouraged, do not actively search for a job. Among them, we also find people who are suffering from health problems and who cannot actively search for a job.

... and 33% of those unemployed as defined by the ILO are not registered in category A

Of 100 people unemployed as defined by the ILO, 67 are registered with Pôle Emploi in category A, but one third of them are not. Among this third, we find young people who, as they cannot claim benefits, do not bother with Pôle Emploi.

Reasons for the increase in the difference between the numbers of unemployed people as defined by the ILO and those registered in category A found between 2013 and 2017

Above all, the data matching was motivated by a desire to observe developments and understand the divergence between the two definitions of unemployed people. I remind you that this divergence was around 400,000 people between 2013 and 2017.

Thus, we find an increase of 300,000 people in the population registered with Pôle Emploi in category A, who are not unemployed people as defined by the ILO.

This increase also contains more seniors. We have therefore sought to explain this development. And that explanation can certainly be found in measures aimed at lowering the retirement age and the end of exemptions in respect of jobseeking at Pôle Emploi. This increase in the number of people registered in category A while not being unemployed as defined by the ILO thus concerns those discouraged seniors who are increasingly less inclined to actively search for a job.

In addition, we note that the number of unemployed people as defined by the ILO who are not yet registered with Pôle Emploi has decreased by 100,000. This decrease can be explained by a shift in the labour market, which began in 2015, in favour of young people. However, as young people are often not registered with Pôle Emploi, this reduction does not translate into an equivalent decrease in the number of people registered in category A.

Conclusion

Thus, this data matching, which proved necessary and very useful, allowed us to answer many questions about the differences we had identified between the two measurements. The reconciled data make it possible to put into perspective the proximity between the two definitions of unemployment that seemed very similar, in principle. Indeed, we are measuring fairly consistent differences.

It has been necessary for us to overcome many methodological constraints to carry out this data matching, because we have little identifying information. However, we have been able to prove its robustness (for this, see the online working document on the INSEE website). This reconciliation must now be repeated at regular intervals in order to observe, among other things, divergences after the coronavirus crisis.

Antoine Bozio

Thank you for this presentation, which reiterates that the data matching concerns not only administrative data reconciliations. Linking survey data with administrative data provides

as much richness, both to strengthen a good understanding of the survey data and to provide a better understanding of the information in administrative data.

Better knowledge of youth integration: the InserJeunes information system
Nathalie Caron

What is InserJeunes?

Hi everyone. InserJeunes is a brand new information system, jointly managed by DEPP and DARES, located at the intersection of education and employment. It is a case of matching personal databases from the area of education with a personal database from the area of employment.

The first results were released in February 2021 and were followed by further results published in December 2021. These are multiple data production series. Data matching projects always take time and it took us three years to organise this one, with the help of a Public Action Transformation Fund (*fonds de transformation de l'action publique* – FTAP). This project has been developed with a dedicated team, which has now been dissolved.

First, InserJeunes makes it possible to measure the employment rate among young people finishing vocational training at CAP to BTS level, either in school or through apprenticeships. The system also aims to make use of a very rich database that allows the precise characterisation of this integration through various indicators such as the type of contract, the sectors that young people enter and the correspondence between training and employment, or even wages.

The measurement of the integration rate, which was the initial request, is based on indicators broken down by training at national and regional level, up to establishment level. Indeed, the matching of personal data that is practically exhaustive allows us to go down to very granular levels, i.e. to vocational high schools and Apprentice Training Centres (*Centres de formation d'apprentis* – CFA).

Using these results, we can produce elements of the national and regional context. It is thus possible to make these integration data available to families and young people, to help students ending their third or final years of higher education to make decisions regarding their career paths.

Four data points are provided for each cohort of leavers in relation to their integration. These data points are 6, 12, 18 and 24 months after they leave the education system. Given that two cohorts must be managed simultaneously, data production is rather cumbersome to manage.

The genesis of InserJeunes

The creation of InserJeunes should be viewed as linked with a long-standing desire to reconcile the areas of employment and education, in order to obtain information on integration into the labour market. Until then, we had surveys, which did not provide information at very granular levels.

However, Article 24 of Law No 2018-771 of 5 September 2018 on the freedom to choose one's professional future has raised new requirements.

This article states that “every year, for each CFA and for each vocational high school, the following are made public, when the number of people concerned is sufficient: the academic or vocational qualification obtention rate; the rate of continuation of studies; the rate of interruption during training; the rate of integration into the labour market of people leaving the establishment concerned, following the training provided; the added value of the establishment.” The latter indicator consists in measuring the effect of the establishment itself, taking particular account of the scholastic and socio-demographic characteristics of the pupils.

This Article of the Law further states that “for every CFA, the rate of termination of the learning contracts entered into is also made public every year,” with this indicator being important for integration through apprenticeship.

The situation in 2018, when the Law on the freedom to choose one’s professional future was published

At the time of implementing Law No 2018-771 of 5 September 2018, DEPP and DARES faced the great difficulty of obtaining integration rates at a level as granular as that of the establishments.

In 2018, we were able to rely on two surveys conducted by DEPP, namely the Integration into Active Life (*Insertion vie active* – IVA) and the Professional Integration of Apprentices (*Insertion professionnelle des apprentis* - IPA) surveys. These surveys were exhaustive and therefore concerned all young people leaving education.

However, as part of these surveys, we did not have a database of young people leaving education allowing us to identify and send questionnaires to them. These surveys were therefore very costly and very cumbersome to manage, not only for the heads of establishments, but also DEPP teams and local education authorities, which were involved in the collection. It was necessary to send out questionnaires and issue reminders, which were most often by phone.

Moreover, these surveys were associated with only a single integration point, carried out seven months after leaving the education system, i.e. in February of the year following graduation.

Finally, the response rate was in the range of 50% to 60%, which is quite good, but insufficient to obtain indicators at establishment level or even for certain specialities at national level.

The purposes of InserJeunes

That is why we created InserJeunes, the original purpose of which was to meet the requirements of this 2018 Law. The aim was to create an *ad-hoc* system compiling comprehensive administrative data, which were needed to be able to get down to establishment level. To that end, we have reconciled administrative school databases, based on student and apprentice enrolments, with the DARES SISMMO database, which measures labour movements and is based on the DSN.

It has therefore been possible to calculate the rate of integration into the labour market of the leavers of each establishment and the added value of those establishments. In addi-

tion, we found that these reconciliations of databases also allowed for the calculation of two other indicators provided for by this Law – the rate of continuation of studies and the rate of interruption during training.

InserJeunes also allowed us to go further than what this Law called for, which was limited to establishment level. Indeed, we were able to create indicators at a particularly granular level, at the intersection of the establishment, the training pursued (CAP, baccalaureate, etc.) and the speciality, even at the level of the qualification provided by the vocational pathway.

In addition, as we have described, we measure our indicators at 6, 12, 18 and 24 months after leaving the education system. The time needed to make these rates available is also much faster than that of surveys, as the information from the first point at six months, concerning the situation in January of year n, is obtained from December of year n.

Finally, based on this information system, there is a very rich database that gathers data from DEPP on training together with the data from SISMMO, and which makes it possible, in particular and as I have already mentioned above, to measure the correspondence between training and employment or the level of employment obtained.

Data matching is at the heart of InserJeunes

The InserJeunes information system has about 20 matches. Some are simple and others are more complex. The simplest data matches use the INE, a national identifier specific to each pupil, student or apprentice, which is designed to facilitate the management of the education system and to enable the statistical monitoring of pupils, students and apprentices.

In particular, the INE makes it possible to define a file of those leaving the national education system, who are not found in the registration files for two consecutive years. However, a difficulty appears when matching this file of education leavers with the SISMMO data. Indeed, it is a case of matching files that use different identifiers, with the source in the area of employment being based on the NIR. The data matching is therefore carried out based on surnames, forenames, date of birth, gender and municipality of birth.

In short, InserJeunes is built around twenty or so data matches, ten of which were performed indirectly, including the one carried out with SISMMO, which was crucial to measuring employment. Thus, in 2018, we wondered which tool to choose to perform these matches.

The tool used for data matching

We needed a tool that was robust and fairly fast, while being aware that no manual recovery would be carried out after the fact. We therefore tried several methods and tools.

First, we already had tools to provide the INE to DEPP or to verify it. Secondly, we have also trialled the MatchID tool from the Ministry of the Interior and SAS programs from several sources.

However, none of these tools were satisfactory for performing this indirect data matching. So we built a specific Python data matching tool, using libraries available in open source.

In principle, this tool can adapt to a certain number of data matching operations, in particular because it was built based on libraries available in open source.

First and foremost, data matching requires the starting variables to be of good quality and we are lucky to have variables of extremely good quality. Indeed, the surnames, forenames and other identity traits are of good quality, with both the INE and the NIR.

Dissemination of results by establishment: a dedicated site and data in open data

For example, InserJeunes, the project of which ended successfully in February 2021, made it possible to obtain initial results, at establishment level and at national and regional level. We plan to make more use of them. The data are published online in open data format at https://www.data.education.gouv.fr.

At the same time, a specific site has been created to allow for wide dissemination of the results obtained: https://www.inserjeunes.education.gouv.fr/diffusion/accueil. This site provides easy access to establishment data, providing information for each speciality, the enrolment rate, the rate of people leaving the education system and the professional integration rate of students or apprentices.

Some results at national level

At the national level, InserJeunes also provides initial results that currently cover the 2018, 2019 and 2020 cohorts of people leaving the education system. These results include only the first two data points, at 6 and 12 months, while we have only the first data point for the 2020 cohort. The subsequent data points are expected to be published within the next six months.

In any event, it is already possible to observe the effect of the coronavirus crisis on the professional integration of these generations of people leaving the education system.

We have also published results on the rate of continuation of studies, or the type of employment contract (permanent, fixed-term or temporary employment contract or professional training contract), making a distinction between men and women.

Conclusion

In conclusion, we have the possibility of integrating public sector wages into InserJeunes. Indeed, for the time being, we are relying only on private sector wages, which correspond to the coverage of the employment data source used, which is based on the DSN.

In addition, we want to incorporate training in agricultural colleges, since we have only taken into account the establishments under the Ministry of National Education, Youth and Sports (MENJS) and all of the CFAs.

At the same time, the “Vocational trajectories in higher education as a whole” (*Trajectoires professionnelles dans l’ensemble l’enseignement supérieur*) project carried out by the Statistical Studies and Information Systems (*Systèmes d’information et d’études statistiques* – SIES) MSO of the Ministry in charge of higher education, in partnership with DARES, will soon start. This project is based on an objective similar to that of InserJeunes and concerns higher education students.

Finally, we must also make use of the wealth of available data provided by InserJeunes to better understand the employment conditions of people leaving the educational system.

Discussions

Nicolas Prokovas, CGT

Which source provides information on job characteristics?

Thomas Vroylandt, UNEDIC

ENIACRAMS contains information on job seekers. Does it also contain information on unemployment benefits?

Claude Castelluccia, CNIL

Are these data pseudonymised? If so, how do you supplement these data with surveys?

Elisabeth Potreau, INSEE

Isn't InserJeunes a duplicate of the Entry into Adult Life (*entrée dans la vie adulte* – EVA) survey?

Laurence Haguet, Court of Auditors

Does data matching between already matched data and raw data have a purpose or use in the evaluation of public policy? If so, what safeguards or recommendations could we rely on in this respect?

Stéphanie Lemerle, DREES

Are the very granular results regarding integration rates public? How can we access them?

Patrick Aubert

Data on employment from ENIACRAMS are taken from the INSEE "all active persons" panel, which uses various sources from DADS, the DSN and sources that relate to public-sector employment. ENIACRAMS therefore also wishes to perform data matching with the SISMMO labour movements file.

Regarding Pôle Emploi data, there are indeed data on unemployment benefits. However, we only monitor people once they enter the field of statutory minimum incomes and not earlier.

Finally, at DREES, the data are pseudonymised, so we do not have the identifiers. As I have explained, data matching is carried out on a double-blind basis, involving INSEE. Thus, when we need to conduct surveys, we provide the non-personally identifying numbers to INSEE and the Payment Funds, which then allow us to find the people to be interviewed.

Nathalie Caron

The EVA survey is based on a sample, while InserJeunes is based on comprehensive sources, which makes it possible to obtain data at more granular levels. You will find the link to the site dedicated to the dissemination of these data, as well as the link to the open data at the end of my presentation on InserJeunes.

Mireille Elbaum

As regards the matching of the Labour Force Survey and the Pôle Emploi data, we note that this is an ambitious and difficult operation. However, this data matching is a fundamental, permanent and potentially evolutionary issue, depending on changes in the rules on unemployment benefits or pension benefits. Consequently, what needs to be done to make us view this data matching as a routine operation, rather than permanent data matching? Would it be possible to make it a periodic operation, to meet this ongoing need?

In addition, Patrick Aubert responded that unemployment benefits data are not monitored in ENIACRAMS. Thus, although this is also an ongoing issue relating to the French social welfare system, the division of our ministries and fields of competence means that the data are produced in a fragmented manner and, on the basis of different concepts, in the field of social benefits.

Furthermore, you tell us that InserJeunes does not monitor higher education students. However, I wonder about people who are not in employment and who are not registered unemployed, who may be inactive or who may have left the country. On this basis, how would it be possible to complete the overview on the integration of young people?

Finally, I note one last difficulty affecting DEPP. Are the different data involved in the various data matching operations covered by statistical confidentiality? Are some of these data available to the establishments and administrative stakeholders of the department concerned following processing? In fact, the processing operations performed by DEPP based on the INE, including when it comes to correcting existing data, are not always considered statistical, at all stages of the processing.

Patrick Aubert

As regards unemployment benefits, I would point out that the MIDAS project is a data matching project that will not capture solely the recipients of statutory minimum incomes. Furthermore, the data matching project involving the EIC, which contains information on individuals' entire careers, and thus also on unemployment benefits, will make it possible to better understand the trajectories of recipients of statutory minimum incomes, including prior to receiving such benefits.

Vladimir Passeron

As I have said, we expect to repeat the data matching relating to the Labour Force Survey. Nevertheless, I am not sure that it is necessary to undertake it every year, bearing in mind that the results are rather structural. However, the difference between the total numbers of unemployed people as defined by the ILO and those registered with Pôle Emploi in category A is monitored more regularly, every quarter.

In any event, at personal data level, I said earlier that we were wondering whether or not to add more identifying questions to the surveys. Such an operation would facilitate future data matching and questioning. Indeed, it is currently difficult to collect accurate wage information: to ensure that the required data on wages is properly collected, it is preferable for respondents to take the time to find their payslips. The use of identifying questions in these surveys, and subsequent data matching to enrich the surveys, would therefore be beneficial. However, this practice would entail a cultural change: it may seem contradictory to go to respondents to submit an anonymous survey, while asking for their surnames, forenames and social security numbers, in addition to the very specific questions that the survey asks. This matter therefore requires genuine consideration.

Nathalie Caron

In answer to the other questions, it is true that the issue of people who are not in employment in the private sector remains open. We do not know whether they are unemployed, inactive or outside the country.

In addition, we realised that support was necessary to ensure that the data we produce is able to help families and young people in deciding on their careers. To that end, the public must be able to grasp the data correctly, so that they can be interpreted. We are therefore working on this issue within the Ministry with the Directorate-General of School Education (*Direction générale de l'enseignement scolaire* – DGESCO), in order to raise awareness among all the stakeholders concerned.

Finally, the data matching carried out within the framework of InserJeunes is performed using highly sensitive data and the list of recipients of the personal databases was therefore very clearly defined from the outset, in accordance with the GDPR and the Data Protection Impact Assessment (*analyse d'impact relative à la protection des données* – AIPD).

SESSION 3 – FUTURE PROJECTS

Chair of the session: Xavier Timbeau, President of the CNIS Environment and Sustainable Development Commission;

Lionel Espinasse, Deputy Head of the Department of Demography, INSEE, for a presentation of the Non-Identifying Statistical Code (*code statistique non signifiant* – CSNS);

Olivier Lefebvre, Project Manager of the RESIL programme, INSEE, for a presentation of the Statistical Directory of Individuals and Housing (*répertoire statistique des individus et des logements* – RESIL).

Xavier Timbeau

I am pleased to chair this session on future projects. I think it is clear in everyone's mind that the issue of data matching methods is absolutely crucial for the Official Statistics of today and tomorrow.

Lionel Espinasse will present the CSNS, which was recently introduced within the framework of the Law for the Digital Republic. He will share with us the different challenges of this CSNS. There is a long history behind the fears that led to the generation of this non-

identifying identifier. I am thinking in particular of the very heated issue of René Carmille's introduction of the identification number. Indeed, this number, which aimed to facilitate the use of successive age classes and better combat the Nazis, was later accused of serving the purposes of the Vichy regime.

Finally, Olivier Lefebvre will present the Statistical Directory of Individuals and Housing (*répertoire statistique des individus et des logements* – RESIL). He will highlight important issues, in particular the envisaged solutions to respond to the abolition of the housing tax. Indeed, the abolition of this tax is sometimes poorly received within the statistical community, since the housing tax file is used for the samples of many surveys.

The Non-Identifying Statistical Code (*code statistique non signifiant* – CSNS)

Lionel Espinasse

Hi everyone. The CSNS is a new offering from INSEE, for use by the Official Statistical Service. Without going back over the details of the objectives of the CSNS, which have been discussed during this meeting, I would like to clarify that it is aimed in particular at facilitating data matching and thus drawing more value from the data sources available to us.

What is CSNS?

The CSNS service is based on a technical component, with the generation of a matching key (the CSNS code itself), as well as an organisational aspect.

A CSNS code is assigned to each individual. Calculation of this key must produce the same CSNS value, regardless of the source in which the individual in question is found. This key can therefore be used to match information from different files.

The CSNS is non-identifying. It must therefore not contain any information about the person in question.

Moreover, the CSNS lasts for ten years. If we had calculated a CSNS for an individual in one source in 2020, we would get the same CSNS for that same individual in another source in 2025. At the end of those ten years, a renewal procedure may be carried out. In case of a security failure, we can also shorten this period.

The organisational component of the CSNS entails the proposal of a usage protocol that includes a phase in which an agreement is reached between the different parties. In addition, a dedicated computer application allows users to deposit the files for which they want a CSNS. In return, they will find the results of the CSNS calculation generated by INSEE on this platform.

A robust legal framework

Without going back over the details of the issue of the legal framework, which were discussed in the first session, I would point out that the CSNS resulted from the application of Article 34 of the 2016 Law for a Digital Republic, which defined it very precisely. The CSNS is also backed up by two pieces of legislation, specifically a 2016 Decree and a 2020 Decree.

This 2020 Decree sets out the technical security requirements of the CSNS, particularly regarding cryptography, encryption, storage conditions and its renewal.

At the same time, the CSNS is also governed by Article 30 of the Law on Data Processing, Data Files and Individual Liberties and by the GDPR, in particular by its principle of data minimisation, which rather strongly guides its usage protocol.

A new service

This service assigns to each individual in an administrative or survey file, regardless of the nature of the original statistical operation, a code calculated for each source and the result of which must be unique for each individual.

There are two ways to perform this calculation. If we have a statistical source featuring an NIR, it is fairly easy to calculate the CSNS and the calculation is based on hashing and encryption. There is already an open service dedicated to this operation, which has been in operation since October 2021. It already works for users, mainly for DREES. A few examples of data matching operations involving this CSNS have been mentioned during this meeting and those operations work well.

The second way to perform the calculation makes it possible to obtain a CSNS from statistical sources that do not feature an NIR. The aim is to identify individuals based on identity traits (civil register data). These traits consist of surnames, forenames, gender and dates and places of birth. These traits then make it possible to go back to the NIR, which is then hashed and encrypted to obtain a CSNS. A service dedicated to the application of this second method is expected to open in the second quarter of 2022.

Like Kamel Gadouche, I will also show you an example of the CSNS. This code is 80 characters long and does not allow people to be identified:

v1:3zUYVUpJFX9g7z3oi3zKSIKXB0yLIE4oGmxsQi1Z2atWI0n8IfnmVrxWUiJqeKH-HTvcH+i0PvuO2991M.

Using the CSNS: the major steps

In practice, the CSNS usage protocol consists of seven steps. The first four steps are not technical, but are organisational phases:

1. MSOs must sign a subcontracting agreement with INSEE, while INSEE's internal services that would also like to use the CSNS must sign a charter of use. INSEE, as a subcontractor, is not responsible for the processing. It only provides the CSNS calculation service to facilitate data matching.

2. A partnership agreement must then be established between the MSOs and INSEE services that want to match data between them. This agreement mentions all the reasons why the two organisations want to match their files. In particular, it details the conditions for dissemination and the conditions for studies. The CSNS service does not take part in negotiations between the two parties and merely ensures that the CSNS usage protocol is respected. Each partner is therefore free to define the partnership rules that go beyond the scope of the CSNS usage protocol.

3. Then, the controller must declare the processing, quoting the CSNS. The CSNS service does not take part in this step.

4. Finally, the MSOs and INSEE services must each explicitly include their use of the CSNS in their work programmes sent to CNIS.

Then there are the technical steps:

5. Each party submits the file for which it wants to have a CSNS calculated to the CSNS service via a dedicated computer application. This is mainly a case of depositing and withdrawing files. This procedure obtained security approval in September 2021. This approval will need to be renewed when the service dedicated to calculating the CSNS through the identity traits is opened.

6. Once the files have been deposited, INSEE, as a subcontractor, calculates the CSNS, either from the NIR or from identity traits (starting in the second quarter of 2022), before sending it to partners via the same application through which the files were initially deposited.

7. Partners can then perform their data matching and produce study files. They must then adhere to rules related to the duration and method of data retention, with the imperative of properly isolating the CSNS from the variables of interest of the different files.

Calculation of the CSNS

The CSNS is calculated based on the NIR, which is potentially obtained after a statistical identification step based on identity traits. An NIR hashing and encryption operation is then carried out. The hashing process makes the operation irreversible, meaning that it is not possible to find an NIR once it has been hashed.

In contrast, encryption is reversible. It is therefore possible to find a hashed NIR based on the encryption associated with it, by inverting the encryption key. The encryption step then makes the CSNS renewal phase, which is to be carried out after ten years, possible. To renew a CSNS, we do not need to go back to the NIR, just to the hashed signature of the NIR, which is an additional guarantee of security.

The practical processes for using the CSNS

In practice, two partners, partner A and partner B, want to match their files. Partner B is responsible for the processing. After the first four steps that we described, they then calculate the CSNS. Each partner then asks INSEE to calculate CSNSs separately. These partners then provide INSEE with either their NIRs or their identity traits, in order to obtain CSNSs in return. There are a number of different scenarios that can then arise: maybe only one partner has the NIR; maybe only one partner has very good quality files, etc.

Next, the owner of file A, who is not the controller, will prepare a file for owner B, which will include all of its variables of interest vis-à-vis the corresponding CSNSs calculated by INSEE. Owner B will then be able to perform the data matching based on the CSNS. The data matching will then proceed more or less smoothly, sometimes with CSNSs from one file that will not always find correspondents in the other file. Owner B will ultimately have its data, as well as those in file A, properly reassigned to each individual through se-

quence numbers. The conversion tables containing the CSNSs should be kept completely separate from other data.

Retention of the CSNS and ensuring data security

The INSEE service issuing the CSNS does not retain any data as soon as the processing is validated by the requesting party, especially if the files contain identity traits or NIRs. It is imperative that owner A and owner B store the CSNSs outside of the files containing the variables of interest with sequence numbers so that they can be reused. INSEE, as a sub-contractor, then retains nothing, not even the CSNSs, once all of the technical stages of the data matching have been validated.

A focus on statistical identification based on identity traits

Finally, to return to the case of NIRs to be found based on identity traits, it is noted that in administrative or survey files, the quality with which identity traits have been included does not always allow the correct NIRs to be found. This difficulty can be seen, in particular, in cases where people have responded in a paper format subject to scanning. These difficulties may also be accentuated by the different issues related to names people go by or married names and other elements described in the presentation by Vladimir Passeron.

Thus, when you want to assign a CSNS to an identity trait and be certain that you can find the same CSNS for the same person, regardless of the subsequent statistical processing, the methodological process is not that simple. A balance must be struck between the desire to identify as many people as possible and the desire to limit the risk of mistakes.

It is therefore necessary to assign a quality indicator to each individual calculation. The user should therefore ultimately know the quality of the data matching. The project is still ongoing and we are currently conducting tests with four MSOs, namely DREES, DARES, SDES and SIES, to optimise this balance. Various technical principles have been adopted for the time being, including the search for exact data matching using the RNIPP. However, if a person has three forenames and only one of those three forenames is found and that person is the only person with that forename who was born in a given place on a given day, we can legitimately assume that this is the right person. We are therefore considering the degree of tolerance that could be allowed in respect of the variable correspondences.

The Statistical Directory of Individuals and Housing (*répertoire statistique des individus et des logements* – RESIL)

Olivier Lefebvre

Hi everyone. I am going to discuss the RESIL project, a programme that has just begun and is expected to be completed in 2025. It seemed important to mention this programme today, as it is based around the issue of facilitating data matching.

First of all, I will present to you the context and objectives of RESIL. Then I will discuss the contents of the directories and the services we propose to offer, as they are envisaged today, more specifically. Lastly, I will discuss how we will take into account the legal and ethical issues relating to these directories and to data matching practices.

1. Context and objectives

RESIL and data matching

The purposes of RESIL include the desire to secure and facilitate data matching, and to make it reliable, in accordance with the principles of Official Statistics and data protection. We have seen today, through the various examples presented, that data matching is an extremely useful practice and that pooling certain procedures could benefit everyone.

In addition, RESIL is built by performing data matching on several administrative sources, which allows the most complete and accurate data possible. RESIL must also be sufficiently robust to resist the disappearance or transformation of statistical sources.

Context

We must deal with the planned abolition of the housing tax. Indeed, this tax is very useful in the statistical production process as it makes it possible to determine the profiles for households, which are particularly essential for statistics relating to income or standards of living. The housing tax also provides a sampling base for our surveys. Finally, it is one of the key inputs of the population census.

When the abolition of this tax was announced, we had to find short-term solutions, which will be operational from 2023 and will be based on the use of tax data. We also considered it useful to find medium-term solutions by using different data sources, in addition to these tax sources.

There is an underlying trend that is driving not only INSEE, but also Official Statistics overall, both in France and internationally, to make progress in the use of administrative sources and data matching. This dynamic involves the use of technical and methodological tools.

What is a directory?

A directory is an exhaustive list of objects (in this case, RESIL concerns individuals and dwellings) with very few variables. A directory is both very long, since it potentially contains all the observations of a field, but also very narrow because it contains few variables.

Variables entered into a directory must make it possible to unambiguously identify the units they contain, in particular to avoid duplicates and to enable them to be linked with other elements of the information system. The directory thus plays the role of a backbone of the information system.

Directories managed by INSEE

INSEE has a long history and long-standing experience in administrative directories. INSEE has been managing the RNIPP since 1946. More recently, in 2019, the Institute built and took charge of the Single Electoral Register (*répertoire électoral unique* – REU) for the management of the electoral lists. In the area of businesses, INSEE was entrusted in 1973 with the management of the National Identification System and the National Directory of Businesses and Establishments (*Système informatisé du répertoire national des entreprises et des établissements* - SIRENE). In 2013, INSEE participated in the creation of the Legal Entity Identifier (LEI), an international directory of entities involved in financial markets.

INSEE also felt that we should go further, by creating statistical directories downstream of these administrative directories. In addition, business statistics are ahead of demographic and social statistics. Indeed, a few years ago, we created the System of Registration in the Directory of Statistical Units (*Système d'immatriculation au répertoire des unités statistiques* – SIRUS), a statistical directory of businesses and establishments, enriched by information collected or built by Official Statistics (profiles of groups, level of activity of businesses, etc.). Thus, we now wish to create RESIL, a statistical directory specific to the demographic and social domains.

The characteristics of the statistical directory

Switching from an administrative directory to a statistical directory involves, first and foremost, the definition of different purposes. This type of directory is intended to produce statistical information only and its users are members of the Official Statistical Service. In addition, this type of directory makes it possible to relax the management rules. Indeed, these rules will be more relaxed because they will not have consequences for the businesses or individuals in these directories as their purpose is not administrative. In addition, it is possible to add some statistical concepts into these directories, such as the concept of households, used in RESIL.

The purposes of RESIL

RESIL must be able to fulfil four purposes:

- to facilitate and ensure the reliability of the data matching of administrative sources or surveys of individuals, households and dwellings to fulfil a real need of all data users and producers;
- to analyse the coverage of the administrative sources used within the Official Statistical Service, in order to be able to describe shortcomings in coverage and to consider how to deal with these deficiencies, within the framework of an objective of comprehensiveness;
- to create survey bases for household or individual surveys;
- to create a common reference point for the production of even more homogeneous and comparable demographic and social statistics, as the directory is the cornerstone or backbone of the information system.

2. The contents of the directories and the services proposed

What, specifically, is RESIL?

At the time of its introduction, RESIL will be made up of two statistical directories. The first will list individuals and the other will list dwellings. Both will be continuously updated with births, deaths and various administrative sources, retaining only identification data. In fact, the other data will be sent to the business information systems, which will integrate them into their production processes.

These two directories will therefore be active entities and we will take snapshots of them once a year, which is a reasonable frequency with which to start. Therefore, we will create a list of households, in particular, especially as this is a fundamental input to some statistics.

However, for the snapshot of the directory of individuals, it must be possible to verify that the people listed are still resident in the region. Indeed, this question is raised by all statistical institutes that operate based on administrative directories and data. To that end, some of these institutes have implemented the “signs of presence” method. It is a case of verifying that there is no trace of these people in the various administrative sources that populate the directory, which would suggest that they have left the region.

The snapshot of the list of households should be based on two concepts of what constitutes a household. The first of these concepts, which is the most familiar to us, refers to people who share a single dwelling. The second concept, the use of which is growing, focuses instead on the sharing of resources. In particular, the latter allows for greater relevance for income statistics as well as allowing the analysis of certain forms of solidarity that extend beyond the home.

At the same time, the RESIL programme will make it possible to provide two services to users. First, it will provide an administrative data hosting service that will transform the raw administrative data into a statistical database that can be used as part of a statistical production process. This service will also make it possible to select the identification data that will populate RESIL. As in a marshalling yard, this service will distinguish the variables that are of interest to the various users, so as to ensure that each of them is sent only what they need.

In addition, a service to produce files enriched by data matching will make it possible to simplify matching by providing the information required.

A processing manager who wishes to use these services must describe the purposes of the data matching that they wish to carry out, carry out an impact assessment and fulfil the conditions of transparency and other obligations of the GDPR. They will then specify the variables they need in a given field of interest, the statistical units of which would relate to individuals, dwellings or households.

Sources chosen to populate RESIL

We have chosen various sources to populate RESIL. We will use the RNIPP, as well as tax sources describing the land or tax households. We also hope to be able to draw on the future source resulting from the “Manage My Property” (*Gérer mes biens immobiliers*) process – declarations by the owners on the use of their properties and the identity of the main occupant. In particular, this source can help to restore the link between dwellings and their occupants. In addition, we want to use social sources concerning recipients of social benefits, such as sources relating to the benefits provided by the CNAF or the MSA. We still want to use the RNCPS, which compiles the list of recipients or beneficiaries of different social protection schemes. Finally, we should use the DSN and the PASRAU system. Each time, it will be necessary to retain only the identification or address elements that appear in these files. These sources cover a fairly broad range.

We are still considering using other sources to improve the coverage of RESIL, by focusing on certain categories of people. Among these possible sources, we could consider files on enrolments in higher or school education or files relating to residence permits, so as to have an overview of people entering the region. Again, only identification information should be used.

The information contained in RESIL

RESIL will contain identifiers and identification keys to allow it to fulfil the role of a directory and to thereby avoid duplicates. These identifiers and keys should also help to perform data matching. Among these keys, internal identifiers that will never leave RESIL will be assigned to individuals and help manage the directory database. RESIL will also contain CSNSs to facilitate data matching, as well as technical identifiers for the different sources used. For dwellings, RESIL will still assign an internal identification key, which will be different from the address. Finally, it will collect address identifiers from the address directory that INSEE should create.

RESIL should also contain other variables, such as civil register data, stored in a separate database, or data making it possible to link individuals, dwellings and households. In addition, RESIL should incorporate an indicator concerning residence in France based on the "signs of presence" method that I have described. Ultimately, it should contain various elements that are fairly traditional in the management of directories or databases, which make it possible to ensure the traceability of the proposed processing operations: dates of updates; dates of changes in value for certain RESIL variables, etc.

The RESIL project schedule

We hope that the RESIL project, which is still in its initial phase, can be completed in 2025. We carried out an initial exploratory phase in 2020-2021 that allowed us to target objectives. We are now entering a project phase. This phase will begin with important work on the drafting of the legal texts on processing that we want to see published in early 2023. The statistical and technical engineering work is planned for between 2022 and 2024. We will then be able to start up the directory in 2024, leading to RESIL and the services it offers finally becoming operational in 2025.

3. Legal and ethical issues and principles of action

The legal framework of the RESIL project

The legal framework of RESIL is still being developed. We feel that processing must be justified by a fairly high-level text and, therefore, we hope to obtain a decree from the Council of State, which is the same level as that of the texts that govern the NIR or the CSNS. In addition, it is important for this decree to be able to enforce a CNIL opinion. In this regard, we are currently in discussions with the CNIL in the context of a request for advice. We are also conducting a data protection impact assessment, as it is a directory covering individuals that is intended to be exhaustive and allows various data matching operations to take place.

Consultation and public information

In parallel with the legal mandate given to the RESIL programme by the legislation, it is important that the programme also has a "social mandate". The RESIL services could then be mandated and legitimised to carry out processing, based on user expectations and on the purpose of their assignments, while ensuring that they maintain the confidence they inspire. This social mandate is created, translated and maintained through communication and transparency on the programme's missions. This transparency is particularly important when it comes to data matching.

Data protection guarantees

It is key to stress that RESIL has a confidentiality guarantee, since it is a directory for statistical use, covered by statistical confidentiality. Furthermore, the processing performed by RESIL and the processing that it allows is based on the principle of transparency. In this regard, we could draw on practices related to the CSNS. RESIL will also be based on respect for the principle of necessity and minimisation of data. Indeed, it contains very strongly partitioned data. Only information that makes it possible to unambiguously identify individuals and dwellings is retained, while business data is segmented into other bubbles.

Finally, RESIL must offer guarantees regarding data security. We will therefore adhere to the state of the art in this area during the creation and start-up of this directory system. It will be necessary to partition the data, to restrict access to them and to ensure that the data have a high level of protection, as well as to ensure the availability of the security tools that we use widely within INSEE.

Discussions

Marcel Goldberg, INSERM

Can the CSNS be used by State Scientific and Technological Institute (EPST) researchers or non-MSO academics?

Antonin Favaro, INRAE

Can the CSNS service be used as part of a research project for matching purposes, for example with the agricultural census, the population census, or with other surveys, such as the Labour Force survey, or the Household Wealth survey?

Thomas Merly-Alpa, INED

Can the CSNS be used to match a survey from a non-MSO agency with other data, if that survey has been deemed suitable by the CNIS Label Committee?

Claude Castelluccia, CNIL

In your example, files A and B show overlaps between the CSNS and the NIR. Given this, what is the advantage of using the CSNS instead of the NIR?

Laurent Piet, INRAE

Is there or are there plans to introduce an equivalent to the CSNS for SIREN and SIRET numbers, even though the CSNS is not a signifier like the NIR? For example, if we look at the farm business surveys, we see that they often include data on the farmers themselves. Therefore, the farm's SIRET can then be used to access other information about the individuals, some of which is very easily accessible on the Internet.

Lionel Espinasse

I would firstly specify that the CSNS is reserved for the MSOs. Researchers can therefore either use the services offered by the CASD or work for an MSO. These different working options were explained by Mireille Elbaum in her speech before.

Furthermore, the CSNS may be used as part of a research project for any survey, without limitations in the list of sources, as soon as those sources are of sufficient quality. However, they can only be processed by an authorised user. For example, the Department of Agriculture's MSO is fully authorised to use the CSNS service to work on the Agricultural Census survey.

The CSNS can be used for processing surveys that have not been produced by the Official Statistical Service, provided that the body responsible for processing is an MSO.

There are currently no plans to design a service type equivalent to that of the CSNS for SIRET numbers.

Lastly, the advantage of the CSNS over the NIR is that the CSNS simplifies the procedure. It means you do not need a Council of State decree to carry out a matching operation.

Xavier Timbeau

It is currently difficult to match two sources on the NIR database.

Lionel Espinasse

Yes, this matching process is difficult to perform without using the CSNS.

Xavier Timbeau

What if I want to match two databases, in two statistical offices, where the NIR is present in both of these two databases?

Lionel Espinasse

To carry out this matching operation, the eligibility criteria laid down in the 2019 Decree on the use of the NIR must be met. Otherwise, I think this operation would be very complicated.

Xavier Timbeau

I gather that the CSNS would enable us to find a solution in this case. However, does the CSNS prevent you from carrying out the work presented by Vladimir Passeron, where it was a question of identifying individuals using identity traits in a clever way, with, for example, first names not associated with the same address, but with the same date of birth? Does the CSNS also prevent you from carrying out the DEPP work that was presented to us, which was based on identity traits from good quality civil status registry data taken from administrative sources?

Lionel Espinasse

The answer to this question is not clear-cut: firstly, in the sense that the CSNS provides a robust procedure for finding identity traits and assigning them an NIR and then a CSNS; in this case, the CSNS can be a good solution. However, in the study presented by Vladimir Passeron, there were variables based on address, and the CSNS does not process this variable. In short, when carrying out matching operations in areas associated with the civil

status registry, the CSNS method is the more robust one. Otherwise, if you are looking to use other information for matching purposes, such as addresses, the CSNS does not help.

Olivier Lefebvre

And it is in the latter case that RESIL may be able to provide an additional service. Indeed, RESIL will enable you to use other variables such as addresses to carry out an additional step in matching operations. RESIL will therefore offer services that are completely complementary to those provided by the CSNS.

Xavier Timbeau

To return to RESIL, is it possible to make a link between the RESIL internal identifier and the delivery point number used by the energy distributors?

Olivier Lefebvre

This issue relates to our work on the housing register sources, which is still at the very beginning of its exploratory phase. At this stage, our view is that our register will instead be based on the DGFIP's premises register. Indeed, there is the question of using data from energy suppliers and we need to check whether using this source may strengthen the coverage or robustness of our system. This question remains open.

Xavier Timbeau

I would just clarify that my question was more about issues associated with matching than coverage.

Mireille Elbaum

Will parliamentarians be able to ask to use RESIL to find the proportion of social benefit recipients not residing at the address indicated and therefore not living in a household that would grant them access to those benefits?

Xavier Timbeau

You are referring to the fixation on fraud that some parliamentarians may have.

Olivier Lefebvre

It is indeed possible that some parliamentarians might have the idea of using RESIL in this way. However, because RESIL is intended for statistical purposes, we cannot extract any individual characterisations relating to particular households or suspected fraudsters. Furthermore, RESIL is based on statistical considerations relating to the construction of resident populations and households to define primary addresses. Therefore, the fact that a statistical address differs from that of a given file does not necessarily indicate fraud. Technical considerations still need to be taken into account, but the answer to your question remains difficult.

Xavier Timbeau

If I understand correctly, the RESIL identifier remains internal to the service and no one knows their RESIL number.

Olivier Lefebvre

Exactly, no one will know their RESIL number.

Xavier Timbeau

And a priori, the legal arrangements guarantee that this number remains within RESIL.

Olivier Lefebvre

The RESIL number is only used to manage our database and will never go outside of RESIL. In terms of the breakdown of access rights to the register, I would like to point out that very few people will have access to the table containing the identifiers.

Xavier Timbeau

To answer Mireille Elbaum's question, it will still be possible to request a statistical study on fraud by using a matching operation on files containing RESIL information. In this case, the RESIL service will be able to receive files to be matched and then send them back, but without the identification numbers. This statistical study could then reveal fraud estimates based on differences in actual residence addresses and addresses reported to social welfare agencies.

Olivier Lefebvre

If the study is for statistical purposes, that purpose is legitimate and the person responsible for the processing declares the processing, aims and sources used, then the RESIL service may respond to a request, in accordance with GDPR.

Yvon Serieyx, UNAF

Are consultations with civil society and academia planned at any stage(s) of the development of RESIL?

A remote guest

What is the link between the RESIL programme and the work on the unique housing identifier developed by INSEE in coordination with DGFIP?

Olivier Lefebvre

To answer the question about consultations, on the one hand, we considered it essential to provide proper communication about the construction of our register system. The aim is to present the various services that the RESIL programme can offer. On the other hand, our social mandate, which is based on the needs of the users, requires us to conduct a continuous consultation. Therefore, over the next four years, we hope that the CNIS will

regularly ask for our input, during the various major stages of the project, to better meet user expectations. As regards the need for a specific consultation with academia, it remains to be determined when this would be useful and what purpose it would achieve.

In terms of the link between RESIL and the premises identifier projects, the initial intention would be for the RESIL housing register to use the residential premises file managed by DGFIP. If an inter-administrative housing identifier project emerges, we should seize this opportunity to populate our housing register. We are therefore paying particular attention to developments in current projects that we could use, while the RESIL programme is under construction.

Xavier Timbeau

Therefore, this unique housing identifier would be found as a minimum in the RESIL housing register.

Olivier Lefebvre

If such an identifier were created at the inter-administrative level, it would be important to be able to find it in RESIL, as it would facilitate the matching of sources in relation to housing.

François Guillaumat-Tailliet, Deputy Secretary General of CNIS

Drawing this morning session to a close, I would like to thank the speakers, the audience, as well as the interpreters and technicians. Our guests have confirmed the excellent quality of the interpreting.

ROUND TABLE – WHICH MATCHING PROCESS FOR WHICH USE?

Chair of the Round Table: Philomé Robert, journalist and presenter, France 24;

Francesco Avvisati, Director of the Innovation, Data and Experiments in Education (IDEE) Programme, J-PAL Europe/PSE – Paris School of Economics;

John Dunne, Head of the Administrative Data Centre, Central Statistics Office (CSO), Ireland;

John Martin, Chair of the Irish Government's Labour Market Advisory Council, IZA Research Fellow, former Director for Employment, Labour and Social Affairs at the Organisation for Economic Co-operation and Development (OECD);

Jean-Noël Barrot, Member of Parliament for Yvelines, Vice-Chair of the Finance Committee.

Philomé Robert

Good evening and thank you staying with us. First of all, I would like to thank Cristina D'Alessandro and Françoise Dupont for the honour of the opportunity to chair this round table session. I must admit that, before they asked me, I did not even know that matching operations existed. So, this gave me an opportunity to learn more about the subject, but I

am no expert. I may, however, be able to ask our guests a few questions, and will listen attentively to their answers.

Francesco Avvisati, you are the Director of the Innovation, Data and Experiments in Education (IDEE) Programme at the J-PAL Europe office based at the Paris School of Economics, which you joined in 2021. You worked for 11 years at the OECD, mainly on the Programme for International Student Assessment (PISA). You have a PhD in economics and are a former student of the École Normale Supérieure (ENS).

John Dunne, you are a statistician at the Irish Central Statistics Office (CSO), where you began working after obtaining your undergraduate degree. Since the start of 2020, you have headed up a directorate that explores two data frontiers, which we will return to later. You set up and managed the Administrative Data Centre (ADC). You received a doctorate from the University of Southampton. In addition, you have, in particular, developed methods for estimating the population using administrative data.

John Martin, you are the Chair of the Irish Government's Labour Market Advisory Council and a member of your country's National Statistics Board. You were Director for Employment, Labour and Social Affairs at the OECD. You have published many articles in specialist journals and have written several books in your field.

You will both be speaking to us in English, with an interpreting service on hand to ensure you are understood by all.

Jean-Noël Barrot, you are a Member of Parliament for the second Yvelines constituency, Vice-Chair of the Finance Committee, Secretary General of Modem, and a teacher. You have worked at the Massachusetts Institute of Technology (MIT) and are very interested in issues related to public policy evaluation.

I will now pass the floor to Francesco Avvisati as part of our round table.

How do matching processes contribute to public education policies?

Francesco Avvisati

The world of education was already mentioned several times this morning, and it is clear that matching operations make an undeniable contribution here. For example, the Inser-Jeunes application allows us to develop indicators to better understand the transition phase between education and the world of work.

As this is such a transitional issue with students leaving the narrow field of education, we need to use matching operations to highlight relevant information in order to drive policy.

Over the course of today's session, we have also cited panels, especially sector-specific panels. The use of these panels enables us to assess public policies over the long term, which is particularly important in education. We need data that stretches back over many years to enable us to follow very long-term trajectories. Indeed, a large number of educational policies aim at results over very long periods of time.

Moreover, the objectives of educational policies far exceed the sphere of the education system, as school is a particularly important lever of redistribution policies and drives so-

cial mobility. It has the potential to offer a remedy to inequality, breaking down social determinism.

To give concrete examples, with the aim of promoting equal opportunities and improving prospects for people from disadvantaged backgrounds, we may wish to verify the relevance of a reduction in class size within primary school as a whole or more specifically within institutions located in disadvantaged areas. We can even ask ourselves whether it is worth pursuing after-the-fact, targeted policies, such as individual scholarships for higher education, or whether it is better to intervene at an earlier stage by investing in schooling at the age of two. However, it is difficult to compare the effects of these different policies using immediate data.

In the world of education, there is a consensus on the value of early intervention in the early years of primary school. This intervention would be the most effective and could have very long-term effects, helping to prevent the perpetuation of social inequalities from generation to generation.

A reference article, written by Swedish and Dutch researchers, entitled “Long-Term Effects of Class Size” was published in 2013 in the Quarterly Journal of Economics. This article followed Swedish pupils born before 1982, who were between 27 and 42 years of age at the time of the study. It allowed the researchers to compare the wages of these former pupils.

The researchers based their work on districts of the state school map that were relatively similar. Class size was arbitrary, as there was a rule requiring the opening of a second class when a school reached 30 pupils and then a third class when the number of pupils reached 60. As a result, the researchers identified schools of 29 or 59 students or 31 or 61 students to distinguish between classes of very different sizes.

This made it possible to measure the actual effect of class size by comparing academic results. Most importantly, the researchers supplemented the data with information on the wages of these former pupils who had reached adulthood.

This study, like many others, showed that exposure to small classes strengthened academic results, but this effect tends to fade over the course of a pupil’s schooling.

However, above all, the authors of this article were able to show that the effects of class size on academic results persist into the labour market. Former pupils who were exposed to the smaller classes have the highest wages and the lowest use of social benefits.

Therefore, by matching survey data with education records, and then matching those records with tax records, it was possible to show that these policies on class size could be effective in a given context. However, to draw more general conclusions, we would need to see whether the situation in Sweden before 1982 corresponds to the situation to which this class size reduction is to be applied.

Lastly, very few studies have been conducted to prove the consensus around class size. Therefore, we must highlight the fact that the matching operations carried out on the long-term data collected make an essential contribution.

This time line, which refers to periods of about 20 years, makes it difficult to link this type of matching with the steering of public policies. Nevertheless, these data matching operations help to forge a highly influential consensus that helps to imagine future policies.

Philomé Robert

What are the limitations you face in conducting these matching operations?

Francesco Avvisati

In France, studies have already shown the effects of class size on academic results, based on thresholds for opening additional classes and panels. However, the information system for monitoring children's educational trajectories was relatively late to arrive in France. The INE has only existed since 2002 and initially only related to secondary and higher education. Its extension to include primary education did not come until 2017. We cannot therefore use administrative data to study trajectories ranging from primary education to working life.

Furthermore, this type of matching based on administrative files creates technical matching difficulties. As we saw in the second session of this meeting with InserJeunes, the matching of data from the fields of education and employment requires the use of different identifiers, which can only be matched by cross-referencing identity traits. This difficulty becomes even more apparent when matching data from 20 years apart, not least because of changes of name or family situation. We would therefore expect this matching to be less accurate than was seen in Sweden.

Philomé Robert

Thank you. I would like to remind you that we have put in place a mechanism for collecting your questions, to give us an opportunity for discussion. In the meantime, let us look to Ireland and hear from John Dunne.

What led Ireland to develop a register?

John Dunne

I express my pleasure to participate in the event. Ireland does not have population registers as they are commonly understood in Europe. These registers are typically used by local authorities to manage public services and people register where they live. Instead, Ireland has an official registration number, used as part of authentication and identification in its public administration systems. This number is assigned to a child at birth, when the parents start to avail themselves of universal children's benefit payments. The number is then used by that person in his/her engagements or transactions with the state throughout his/her life. These transactions occur from the cradle to the grave as the person engages with education, tax, welfare, health, pensions and finally death. In reality, it is very hard to live in a modern democratic state without engaging with public authorities. For a long time, the CSO has recognised the untapped statistical potential contained in public administration systems. The Irish Statistics Act passed in 1993, makes strong provision for the use of this administrative data for statistical purposes. The CSO has been accessing this administrative data to varying degrees, since 1993 and probably even

before. Back in 2009, the CSO centralised its infrastructure to collect, sort and process this data to make it safe and available for statistical purposes. The Administrative Data Centre (ADC) acts as a clearinghouse, ensuring that appropriate data governance procedures are in place. For example, Irish statistical analysis is only conducted on pseudonymised data, with all identifiers replaced with protective identifier keys (PIKs). A protective identifier key enables safe linkage on that key all times across data sources, ensuring the original identity is protected.

A key part of public sector reform in Ireland is the deployment of official identification numbers for properties, businesses, and purchases on public administration systems. This focus on smarter data has also resulted in enhanced data analytics capabilities not just for the CSO but also within public sector bodies themselves. John Martin may provide and example of this later. The overriding ambition of using administrative data is to provide significantly enhanced statistics to inform decision-making at a far lower cost. This is the idea of living in an informed society. Of course, one of the key goals is to be able to produce sensitised population absolutes every year at a low cost, similar to what has been done in the Netherland and the Nordic countries for many years. Similar goals are being pursued across the EU as member states will be expected to produce annual census-like estimates in the future. This is where Ireland is at the moment with the use of administrative data.

Philomé Robert

What stage is the project you are describing currently at and what benefits have you obtained at this point?

John Dunne

Outside the analytical capabilities, since the inception of the ADC there has also been a fast response, in particular in times of crisis. The use of administrative data enables highly detailed statistical output and the data is comprehensive and respects specific cohorts in the population. Administrative data is even more powerful and this enables longitudinal analysis of population cohorts. For example, it is possible to conduct statistical analysis about where graduates are employed, x years after they graduate without having to undertake the survey. These types of projects take place in our island.
Ireland now has these capabilities and they are now working to deploy a new postcode, which was launched in 2015 in our public administration system. This postcode system is quite novel and different from other postcodes in that it is associated with the letterbox rather than a group of houses in a specific area. For example, if someone has a given postcode, it can be put it into Google Maps to get the directions straight to the address. To roll out this postcode significantly enhanced Ireland's capabilities for detailed geographic statistics.
In terms of combining census-like population estimates, the CSO has undertaken research on how it can composite such estimates in a robust manner. The results to-date suggest that the CSO can be optimistic about implementing a new, robust system of annual population estimates in the future.

To finish, we are continually finding new opportunities to provide enhanced statistical information in various policy areas.

Philomé Robert

Thank you, please stay with us, the audience will have questions for you. In the meantime, let's give the floor to John Martin.

What contributions do the matching operations make to public employment policy and what matching practices have you seen at international level?

John Martin

On the value added of statistical linkage of administrative data for labour market and social policies, he would argue this has made a huge impact in the field of evaluating the impacts of labour market and social policies. This has been amply demonstrated in recent years by the large-scale meta-analyses of the impact of such policies carried out by the most recent Nobel Prize winner in economics, David Card, and his co-authors Jochen Kluve and Andrea Weber. Many of the evaluations included in their papers, cover a large sample of countries and rely on linking administrative data as the essential building block.
He illustrates this by showing how the ability to link administrative data from different sources has had a huge impact on the evaluation of labour market and social policies in his own country, Ireland. Until the creation of the so-called Jobseekers Longitudinal Database (JLD), by the Irish Department of Social Protection in 2013 – 2014, there was essentially little history of linking administrative data sources in Ireland and exploiting them to evaluate public policies. The Jobseekers Longitudinal Database was developed by analysts in the Department of Social Protection, in close collaboration with statisticians seconded from Ireland's Central Statistics Office, the central statistics office. It follows benefit claims, employment, training, and participation in employment policies of all job seekers and sole parents who have made a benefit claim since the year 2004. The primary identifier for these individuals is the Personal Public Service Number, to which John DUNNE referred to in his presentation. It is needed by all individuals in Ireland to access public services. To protect privacy, all individual data is pseudonymised and encrypted and access to the resulting Jobseekers Longitudinal Database is strictly controlled by the Department of Social Protection.
The JLD brings together data from registers held by the Department of Social Protection, by the Revenue, the Irish tax authorities, and SOLAS, the Irish education and training agency. The resulting database contains many millions of individual episodes of benefit claims, labour force status, education and training, and participation in labour market programmes, as well as data on earnings and tax payments. One of the reasons Irish authorities invested in developing this longitudinal database of jobseekers was to use it as a key building block in evaluating rigorously, the range of labour market, education and training programmes that exist in Ireland. Over the past five years, the Department of Social Protection has contracted with outside consultants, research institutes and academics, to produce impact evaluations of many such programmes. The Department of Social Protection staff is also undertaking evaluations using the Jobseekers Longitudinal Database. Resulting evaluations are all published after a rigorous peer review.
Impact evaluations produced using state-of-the-art econometric methods have highlighted programmes that have worked in terms of achieving their stated objectives and others that have not worked. The evidence base produced by the evaluations using the Jobseekers Longitudinal Database have added greatly to the knowledge about the design and

implementation of employment policies in Ireland and have resulted in significant policy changes towards those programmes. Next in line this year is the Community Employment programme, which is by far the largest single active labour market policy in Ireland in terms of both the number of participants annually and the amount of public spending. This evaluation is being undertaken by the OECD and the Joint Research Centre of the European Commission.

Jobseeker longitudinal data has proved its worth, but it does have some limitations that should be mentioned. First, the time horizon for evaluating post-programme outcomes is typically rather short-term from six to 12 months. Ideally, the researcher might want a time window of up to five years or more in order to assess the long-term impacts of the programme. There is a specific Irish feature that makes it more difficult to track the longer-term outcomes, which is the long tradition of emigration from Ireland, particularly among younger-age cohorts. Second, the jobseeker longitudinal data currently lacks data on educational attainment and skills as one of the key observables, which all research suggests is strongly correlated with outcomes in the labour market. Instead, researchers have to rely on previous occupations of the jobseeker, which is a rather poor proxy for educational attainment and skills. Nonetheless, the Department of Social Protection is currently engaged in extending the database to include a wider range of welfare benefits and labour market transitions with a view to making the database more useful. As highlighted by John DUNNE, Ireland is also investing in linking other administrative datasets as proof that this is a high priority for official statistics in the future.

The second question was on linking administrative data to evaluating impacts of employment policies in a comparative respect. Of course, Nordic countries, especially Denmark, Norway, less so Sweden, and Finland, have been pioneers in this field. Their leading role can be explained by their population registers with unique personal identifiers, which enable researchers to follow the same individuals for long periods, as they transit between different labour market states and interact with a wide range of public agencies. Researchers in these countries have then applied various econometric methods, experimental and quasi-experimental, to evaluate ex post the impact of participating in various labour market programmes and/or availing of different welfare benefits, on a wide range of outcomes. These include not just labour market outcomes, earnings and career prospects, poverty, health, family formation and retirement. Researchers in North America, in the US and Canada have long used linked administrative data to analyse the impact of unemployment insurance and social assistance systems on the labour market. Researchers in both countries have also used the gross flows underlying their regular household surveys to address these and other labour market issues.

If Nordic countries are left aside, lacking population registers, and hampered by concerns about previously endangered security, other European countries were slower to perceive the potential of linking administrative data for research purposes. However, the situation has evolved significantly in recent years with countries such as Germany, the Netherlands and Switzerland making major strides in this field. In the German case, the Hartz reforms of 2003 – 2005 included a requirement for authorities to evaluate the labour market and social programmes introduced by the reforms. To do this, it provided a major impetus to link administrative data sources to facilitate this evaluation process. It also enables some innovations, such as Germany and Switzerland being able to link current employment insurance claimants to the public employment case workers handling their files. This helps to establish whether case workers and their different approaches to their clients actually

make a real difference to their outcomes post participation. Perhaps Southern European nations were a bit slower to recognise the potential of linking administrative data for labour market research purposes, perhaps reflecting a history and concern about data privacy issues. However, they are now much more active in this field, as shown this morning in the case of France. On the other hand, Central and Eastern European countries and Greece are lagging behind and clearly missing out on the possibilities offered by linking administrative data from different sources.

Linking administrative data in order to evaluate the outcomes of labour market and social policies has clearly demonstrated its worth and more and more countries have recognised this.

Philomé Robert

Thank you for that explanation. You described the contribution that matching practices have made to public employment policies in Ireland before taking us to Canada, the United States and even Scandinavia. However, you seem to be saying that France is lagging behind in this area. So, let's turn to Jean-Noël Barrot, who is both an elected representative and an economist.

What is the added value of data matching when drawing up and assessing public policies?

Jean-Noël Barrot

It is well known that, in recent years, and especially since the coronavirus crisis, France has taken an important step forward, by mobilising official statistics, with a view to developing a better approach to public policies, upstream and downstream of their design processes. We have been able to use numerical data to examine the subjects we deal with in a very meaningful and rigorous way. This progress was made possible thanks to the increasing availability of data.

As for the subjects of particular interest to me, particularly those relating to business, I could mention the fact that the French Public Investment Bank (BPI) made available a certain amount of data relating to its operations that took place before the health crisis. Before that, France was lagging behind; Small Business Administration, an independent US government agency dedicated to small businesses, for example, had been providing researchers with intervention data for far longer.

Following the first financial laws of the current five-year period, which dramatically changed capital taxation, we saw a certain number of panels, including the Tax Returns File (POTE), made available to researchers. This availability should allow for an assessment of these tax reforms, particularly the introduction of the single flat-rate tax mechanism, and the elimination of the wealth tax (ISF) which was replaced by the tax on real estate assets (IFI).

This was beneficial not only for the evaluation of the public policy in question, but also for the scientific community, as researchers can now access these databases provided they comply with the procedures of the Statistical Confidentiality Committee.

Moreover, during the coronavirus crisis, thanks to INSEE, France was probably in a position to use data to understand the ongoing phenomena earlier than other countries. It was able to use nowcasting (immediate forecasting) by exploiting data that were not normally used by national statistics institutes. Furthermore, as early as April 2020, France implemented the “flash” Labour Force Activity and Employment Conditions survey (ACEMO) in record time, allowing the country to monitor the impact of this crisis on companies on a monthly basis.

This, then, was a significant effort to make data available to help evaluate public policies.

Philomé Robert

As an economist, what needs do you see for research so that it can provide the information needed to steer public action?

Jean-Noël Barrot

To answer you, I will have to don my researcher hat to assist me in my role as a Member of Parliament. First, I believe that researchers are particularly well treated in France. Having carried out research in the US, I found it particularly difficult to access and carry out matching processes with US administrative data.

To conduct my research across the Atlantic, I had to go to Cambridge, Massachusetts, to the National Bureau of Economic Research (NBER) at the Boston Research Data Center (BRDC). I had to swipe my card several times before I was allowed into a windowless room with very little ventilation. When I was in that room, I wasn’t allowed to use my own computer. What’s more, this access came at a significant cost. Some researchers literally lived in this small NBER building, but I couldn't spend all my days there. As a result, I felt my research was making endogenous progress, supported by a group of specialist researchers who focused solely on specific data, while others did not work on the data at all as cross-sectional research was not a preferred option.

In France, the CASD has enabled data matching to be done remotely, in a secure and pleasant environment. The CASD has considered data security as much as the user experience. It has enabled researchers to remain in their offices and switch between public or private data on their own screens. They can work on administrative data using their own work tools. In this respect, we are particularly spoiled here in France.

I would say that we have an administrative data collection that is the envy of the world, at least as far as business is concerned. This is the trade-off for our stringent administrative tax framework. Indeed, the quality of the data on business is so good that many baseline business studies are built using French data.

In recent years, there has been a very significant injection of new data enabling researchers to answer questions that we could not answer before. And, thanks to the CASD, matching of administrative or association databases is possible, a tool that I find more effective than that available, say, to US universities, because it allows me to work in a secure environment from my own office. This excellent tool means that France benefits from increasingly more accurate research insights into the nature and temporality of data in order to manage its public policies.

As a result, this opening up of a number of databases to researchers has established a new practice. Each year, during the budget discussions at the National Assembly, the Paris School of Economics organises a very detailed analysis of the effect of the government's proposed budget on companies and households. This exercise is carried out in just a few days, or sometimes one or two weeks, as the draft budget is not known in advance. These simulations are possible thanks to the opportunities granted by the French statistics system, namely the availability of data and the speed of matching.

This analysis then gives rise to a debate, which is always an uncomfortable experience for the Members of Parliament of the majority party. Shortly after we leave the Finance Committee, the Paris School of Economics tells us the true impact of the budget with a staggering level of detail. This is the opportunity for the opposition members to use this analysis to bolster their arguments and review our budget, as we had not previously been in a position to foresee some of its effects.

In addition, last year, some of the results of the Paris School of Economics differed from the simulations run by the government when drawing up the budget. As a result, economists at the school were then able to take the time to explain where these differences in analysis might come from.

From there, I think that we are entering a virtuous circle around the presentation of the budget. This dynamic is centred around mutual enrichment or control, made possible thanks to the accessibility of data to researchers. I think that this approach is taking us in the right direction and that we are in a relatively fortunate position here in France.

Discussions

Philomé Robert

Thank you, Mr Barrot, and thank you to all four of you. Now let's move on to the discussion section and gather the questions from the room and chat.

Yvon Serieyx, UNAF

As part of the PISA survey, is it possible to use matching in some countries to replace the "parents" questionnaire, which, I believe, is optional, in order to collect socio-demographic information about the pupils' families?

Francesco Avvisati

International surveys are not so simple, as they require the same quality and type of information to be collected in all countries. For the time being, therefore, there is no question of stipulating the collection of administrative data instead of questionnaires, simply because not all countries participating in these surveys are in a position to perform this task.

Nevertheless, an international survey is based on data from national surveys, composed of a national and international component. The data for this second component are transmitted to the body responsible for overseeing the international survey. These surveys are conducted in numerous countries using student or school identifiers. This information can then be used to supplement international surveys with new data. In Chile, researchers used PISA surveys and administrative databases with student IDs, thereby enabling them to carry out follow-up on a longitudinal basis.

Finally, at international level, it seems much more realistic to replace questionnaires on establishments with administrative data rather than questionnaires concerning socio-demographic family information.

Roxane Silberman, CNRS

Does the CSO perform matching operations for research projects?

John Dunne emphasises that for the CSO in Ireland the primary driver for creating this linked data was the potential of the statistical information for enhanced decision-making. They also recognised the research potential of this linked data and they intended to enable researchers to access it for research purposes, which had been done to a limited degree. However, they had taken a cautious approach as they developed their own understanding of what safe and appropriate research was. They wanted to make sure that their approach was safe and going down this road they used what was called a five-stage framework to think about each step that was taken in developing the infrastructure going forward. The five-stage framework was an ICE framework that configures safe projects. It provided a safe setting where research was done; safe data, making sure that the data was properly anonymised/pseudonymised for the purpose; safe persons, credible researchers undertaking the work; and that the output of the research was also safe and not disclosive. There was significant potential within the administrative data, and they just needed to take a cautious approach to ensure that what they had and did was safe and appropriate going forward.

John Martin

I would like to clarify that the Jobseekers Longitudinal Database that I presented was built by the statisticians of the Department of Social Protection in close collaboration with the CSO statisticians. And this collaboration is continuing with a view to enhancing and improving the longitudinal databases relating to Irish job seekers.

Kamel Gadouche

Are there any plans to formalise, legislatively or via other means, the independent contributions and exchanges made with researchers as part of budget development on the basis of the evidence-based policy contemplated by US Congress?

Jean-Noël Barrot

This idea is interesting, but I don't think such a decision is planned in the National Assembly. We have tried to incorporate into law the fact that Parliament can have more expertise in order to better analyse legislation, and budget texts in particular, upstream, but we have not truly achieved this.

However, with this in mind, thanks to INSEE's goodwill, a cell of data specialists created at the National Assembly, composed of three or four people, was able to devise an interface for elected representatives to simulate the effects of a proposed amendment to certain laws: https://leximpact.an.fr/. This interface helps the legislature to reflect on draft finance bills as well as laws relating to income tax, certain social benefits, and government grants to local authorities. The tool was designed for Members of Parliament and is based on a number of researcher databases. It has already enabled elected representatives to an-

swer various questions and plans are in place to extend it to incorporate legislation on other taxes. This is a small step forward, towards the evidence-based policy you are talking about.

Furthermore, the National Assembly recently introduced the *Printemps de l'évaluation* (Spring Assessment Period), incorporating the principle behind this into organic law. This practice takes place over three weeks, between the end of May and the beginning of June. Here, the ministers responsible for their administrations report on the implementation of the budgets entrusted to them in the previous year. They also report on the effectiveness of the public policies that they have been tasked with implementing. These ministers are interviewed by Members of Parliament, who now have internal National Assembly resources available to them, which in particular enable them to make use of the Assembly's administrators, for drafting an assessment report. Not all Members use works based on data matching or administrative data, but I hope they will be able to make greater use of it during this assessment period.

Jimmy Baulne, Quebec Statistics Institute

Does remote data access through the CASD give access to files that are subject to a certain level of masking to reduce the risk of disclosure?

Kamel Gadouche

The data made available by the CASD are pseudonymised. Direct identifiers have been replaced. However, the direct identifier for company data has been retained. Nevertheless, the CASD does not use perturbation or other anonymisation techniques on the data entrusted to it. Therefore, a system ensuring a high level of traceability has been implemented. Everything done by users is filmed and recorded. This morning, I was able to present the security measures that were implemented. But the purpose of the CASD is to make the most detailed data possible available to researchers in order to enable highly accurate studies.

Philomé Robert

Let's hear from John Martin. You highlighted the use of matching operations as part of public employment policies. We mentioned earlier some points of comparison with countries such as Canada, Ireland, and Switzerland. I was therefore wondering if you could provide greater detail on elements of comparison with France. For example, what are the specific characteristics of France or Canada in relation to matching operations?

John Martin

That is a broad question. In Canada, for example, many studies focus on the unemployment benefit system. In fact, Canada has a relatively specific regime that varies from province to province. Many studies that have used data matching have shown the existence of a very significant level of seasonal work that is subsidised by this Canadian benefit system.

For example, in the Maritime provinces (New Brunswick, Nova Scotia, Prince Edward Island), studies show that the unemployment rate is higher because of seasonal workers who receive these benefits. Furthermore, one very interesting study, which used longitud-

inal administrative data, showed that individuals quickly learned the rules of the unemployment benefit system. They were able to obtain these benefits several times in their careers, driving up unemployment rates, particularly in the Maritime provinces. This is a very sensitive subject in Canada, since the rules surrounding unemployment are the responsibility of the provinces.

Roxane Silberman, CNRS

Have countries other than France sought to establish tools similar to INSEE's CSNS?

Sylvie Lagarde

John Dunne, I was of the understanding that the Irish Personal Identification Code (PIC) was equivalent to the CSNS?

John Dunne One was purely based on a random number generator that was designed to ring-fence the original identifier with it. The other was something called salt and hash, whereby the identifier was prefixed with a secret piece of text that only two or three people knew. Then the concatenative text stream was hashed to get the identifier. This was done in a systematic way for each of the identifiers with a different salt and hash routine for each different identifier. If an identifier was on two data sources, then they were encrypted and the same [inaudible] [1:00:41] was put on those data sources when it was being pseudonymised.

Philomé Robert

Let's come back to our Member of Parliament, Mr Barrot. You organise sessions on public-policy assessment. What lessons do you learn from these? Where is there room for improvement?

Jean-Noël Barrot

I feel that a growing number of academic papers are using administrative data. These papers are incorporated into public debate, and even create debate in parliament. I find this approach virtuous, as it makes a contribution to the design of public policies.

The Spring Assessment Period has taken place twice, two years apart, but not this year. We felt that, as we are just weeks away from a presidential election, the climate was no longer conducive to an assessment of public policies.

But it was important for us that the world of public decision-making and research could meet in the National Assembly, which has a mandate to assess public policies under Article 24 of the Constitution. It was useful for these two universes to convene, to get to know one another and develop a common language. Indeed, it is important that the world of public decision-making has a better understanding of how to use and leverage the results of academic work, and conversely, that academia can better understand the expectations of decision-makers, in order to provide them with useful material. I think that this approach contributed to the general virtuous trajectory which has tended towards greater incorporation of researchers' work into public debate.

For example, in recent days, the topic of taxation of inheritance and gifts has emerged in the presidential campaign debate, mainly following a note issued by the Council of Economic Analysis (*Conseil d'analyse économique* – CAE). This debate is not linked to a high budgetary cost, but has become much more important in view of the strong symbolism associated with taxation. The CAE sought to answer some of the questions using administrative data and was able to bring new elements into play. In particular, it looked at the cost of certain tax exemptions. The National Assembly had already asked the government about the issue, seeking a report on the cost to the public finances of various exemptions, including the inheritance tax advantage over life insurance. But the government was not able to answer this question because of insufficient data on transfers, both in France and many other countries. Ultimately, this CAE note led to a resolution of this problem by quantifying these costs in a variety of ways. These new truths have been incorporated into the public debate by the CAE, moving from academia to politics.

Patrice Duran

The issue of effective public policy, which is currently central to government legitimisation, requires an understanding of the social sphere. As a result, the issue of matching becomes a key element as it allows for public action that is more reflexive and thereby an enhancement of knowledge.

Returning to the question of the role of Parliament in the assessment of government-led public policies, I would like to point out a significant weakness in the National Assembly in this regard. Having had experience of the General Accounting Office, now called the Government Accountability Office, the US Congress body responsible for the federal budget public account audit, but also heavily involved in assessing public policies, I cannot help but wonder about the difference in resources between the US Congress and the French parliament when I see that the GAO has resources roughly equivalent to those of INSEE! In the French case, however, making part of the Court of Auditors available to Parliament so as to assist it in assessing public policies will not achieve great progress, given that this institution is not highly specialised in public policy assessment.

Certainly, as both you and I have pointed out, it is true that all the reports emphasise the importance of reconsidering the link between research, education and the world of public action, both in Parliament and within government ministries. But, like my neighbour, Jean-Luc Tavernier, I wonder about the historical importance of official statistics both in education and in public management. At the CNIS, we have seen from this perspective the considerable gains that the pandemic allowed us to achieve in raising awareness among all public stakeholders of the indispensable and decisive role that relevant and quality official statistics can play. Indeed, official statistics have long had a bad reputation in France due, in many cases, to insufficient knowledge about their nature and usefulness. What can we do to ensure official statistics achieve their rightful place in the future? This is a central issue for us and for all public stakeholders.

Jean-Noël Barrot

I think you are right. Not only do we find official statistics to be very useful in understanding current phenomena, there is also an urgent need for the French and other governments to develop the necessary skills to be able to process the data. Otherwise, governments will be subject to any dictates imposed on them by those with data expertise.

There is one answer, from among various, that you may have mentioned implicitly in your question, which is to integrate statisticians, econometricians and doctors more naturally and fluidly into the state civil service. This solution could be phased in gradually and would allow the civil service to adjust to the challenges of statistics. We still have what we call subject-specific ministries, such as the Housing Ministry, which are still far from integrating this approach. If they do not engage with statistics, other stakeholders, such as Google, may take over.

Philomé Robert

How can we bridge this gap?

Jean-Noël Barrot

To address the ineffectiveness of some public policies, which results from a lack of data processing, other actors may step in by offering services that are supposedly free but paid for by advertising.

Without wanting to sound despondent or brusque, I share Patrice Duran's opinion of Parliament's weakness in assessing public policies.

The OECD uses a sort of taxonomy of independent financial institutions (IFIs) that challenge executive powers in their budgetary choices. Among these institutions, it distinguishes between those that are within parliaments such as the Congressional Budget Office (CBO) or the *Ufficio parlamentare di bilancio* (UPB), on the one hand, and other institutions outside parliaments, the best example of which is the British Office for Budget Responsibility (OBR), which challenges the Treasury's assumptions and choices, on the other. I do not know the situation in Ireland.

When I began making recommendations on the subject, I advocated housing an IFI within the French parliament, perhaps influenced by my professional organisation. This proposal raises some questions, as it involves reperforming the government's spending projection exercise.

Then, the Committee's report on the future of public finances, chaired by Jean Arthuis, to which some of you contributed, questioned the effectiveness of public spending. However, that report instead called for the establishment of an IFI outside parliament.

The OECD currently identifies the High Council on Public Finance (HCFP) as the French IFI. And it ranks it among the lowest IFIs in its rankings, regardless of which indicators it uses, whether its membership, budget, or missions. I also welcome its members who are present in this meeting.

The report by Jean Arthuis suggested setting up this IFI at the HCFP. If the institution is the most suitable to fulfil this assessment role, it needs to be nurtured as soon as possible. It would be regrettable to remain in the current situation where there is nothing in the National Assembly except the small team of three or four people that I mentioned, and where the HCFP merely checks macro-level assumptions.

Finally, fortunately, we have an Official Statistical Service and quality laboratories that serve as assessors and make our system more coherent. In this regard, I presented the

example of the budget evaluation carried out by the Paris School of Economics. But it is true that this issue still needs to be resolved at institutional level.

A remote guest

Could different organisations in different countries that are comparable to the CASD link up in order to provide researchers with access to data from different countries?

Kamel Gadouche

The CASD, primarily through Roxane Silberman, coordinates the International Data Access Network with Germany, the United Kingdom and the Netherlands with a view to setting up a core network to enable transnational research.

So, I would like to take advantage of the fact that John Martin and John Dunne are here to ask them whether they see any usefulness in this opportunity to exchange data access and allow these link ups to take place in order to enhance the mutual knowledge of each of these countries.

John Martin

These links exist at European level, notably through the EUROMOD network, which includes Ireland and a number of other countries. It is a micro-simulation model for assessing and comparing the effects of social tax policies. I will also mention the Luxembourg Income Study (LIS), which produces a transnational database of microeconomic data on income. These are international networks that generally share models and methods. But, they do not exchange individual data. To my knowledge, such practices do not really exist at European level.

John Dunne In the first instance they had a standardised approach to statistical reporting across the EU. In the second instance, it was primarily where they did a lot of their standardisation and comparison purposes in terms of statistical information. The next step down was to create safe micro data files, which was done on separate social surveys possibly. It was possible to link those surveys at special research houses. Progressing down towards the sharing of micro data across countries, this raised a lot of sensitivities, particularly on personal data being shared outside of Irish domains where there would be uncertainty around legislation. There was a lot happening at European level in terms of the European Data Strategy and a Data Governance Act near finalisation. This data landscape seen to be changing quite a lot, but national statistical agencies would probably step forward very cautiously into this area. It remained to be seen what would happen going forward.

Philomé Robert

As there are no other questions, that concludes this round table. Thank you very much to all three of you.

ROUND TABLE – TRANSPARENCY AND INFORMATION FROM THE PUBLIC

Chair of the Round Table: Chantal Cases, French Statistical Society;

Eric Rancourt, Director of Statistical Methods and Data Science, Statistics Canada;

Bertrand Pailhès, Director of Technology and Innovation, CNIL;

Mark Hunyadi, philosopher, professor of moral and political philosophy at the Catholic University of Leuven;

Maryse Artiguelong, Vice-President of the Human Rights League.

Chantal Cases

Hello everyone. For this round table, we have two participants in the room and two who are joining us remotely. I would like to thank all four of them. Our round table focuses on transparency and information from the public in relation to data matching operations. Over the course of today's session, we have shown in detail the richness of the register and matching projects, mainly highlighting their contributions in terms of knowledge.

But, we have barely reflected on the social mandate issue that was raised this morning, and I believe it is time to look further into this. Transparency and ensuring data privacy are essential and important factors. They provide real meaning to these projects.

Moreover, I believe that we must communicate the aims of these projects very broadly, including to the general public who provide us with the data, as well as to specialists and to those who are less familiar with the technical aspects, which are difficult to access. As a result, a number of participants in this meeting have realised that matching operations are technically complex and are not as obvious as one might have imagined.

We must therefore ask ourselves how we can ensure this transparency, quality information from the public, and good consultation on these official statistics projects. We will try to answer this question by interviewing four guest experts.

First, we will be giving the floor to Eric Rancourt, Director-General of Modern Statistical Methods and Data Science at Statistics Canada. He has a wealth of external experience that could shed light on the social mandate associated with statistical projects.

I will first ask him about his data matching practices. I know that Statistics Canada, like INSEE and many other statistical bodies, has long used administrative data to develop statistics. So, before we delve into the issue of information from the public, I would like to know Statistics Canada's current position on the practice of file matching. I would appreciate your input on what led you to think more deeply about your relationships with the public in relation to matching operations.

Secondly, it would be great if you could explain to us more concretely the way in which you organise the information from the public. This morning, Sylvie Lagarde referred to the linking of microdata by Statistics Canada. I would like to know the challenges that this linkage entails and get an idea of the data involved. Finally, I would like you to tell us about any topics that are particularly sensitive to the Canadian public.

Eric Rancourt

Hello everyone, I am very pleased to be participating in this round table organised by the CNIS. I would like to thank Chantal Cases and Françoise Dupont for inviting me. I am going to address the issue of matching records in Canada, focusing on the practices in place at Statistics Canada, although this may tip over into the provincial domain. I am, however, not going to discuss the issue of research data centres, but I would still point out that we are beginning to implement a remote data access option.

Firstly, before presenting the events that occurred four or five years ago and that changed the social mandate situation considerably, I will give you a brief history of the use of administrative data in Canada.

Before the 1940s, we only used pseudo-censuses, together with administrative files, to a greater extent. This continued until Jerzy Neyman's famous 1934 article, which introduced probabilistic sampling. Following Morris Hansen's work in the US, we set up a labour force survey in Canada in 1945, thanks to work carried out by Nathan Keyfitz. Surveys then began to flourish to a very large extent, right up to the 1990s and 2000s. We could call this the crystallisation of an information production approach centred on surveys.

But at the same time, we continued to use administrative data, especially for sampling frames for social matters. We also used them as is in many cases to produce economic statistics.

The 1970s saw a rise in the use of administrative data, with a division dedicated to this type of data created around 1979-1980. These data are therefore part of Statistics Canada's DNA, especially as we are governed by legislation that aims to prevent the duplication of statistical collection in different ministries and to encourage the production of integrated statistics. This law only gives us an overall mandate and does not mention file linking or matching. We see this practice as part of the efforts to prevent duplication of collection. In addition, a specific clause specifies that Statistics Canada has access to all files produced at federal, provincial and municipal level, and even in the private domain.

There was a considerable increase in the use of matching operations up to a point a few years ago, mainly with the aim of improving or verifying quality and to reduce survey costs. Gradually and increasingly, these operations were used instead of survey data. We then created secure matching environments and established a centre of expertise dedicated to this practice within the Directorate of Methodology. We established fairly binding and complex file access processes. We also set up a committee that reviewed all matching projects up to 2017 and conducted an in-depth assessment of privacy indicators. Then, once the matching operations had been completed, the information could then possibly be posted on the Statistics Canada website.

We operated in this way for several years, processing thousands of files and conducting all kinds of matching operations. In fact, the 2010s saw a massive increase in data production in both Canada and elsewhere.

At the same time, people became more aware of the existence of these data, increasing demands for very detailed information. But, this also came with a steady yet considerable decrease of the survey response rate.

In response to this phenomenon, Statistics Canada implemented a modernisation programme with a view to reversing this information production paradigm. The aim was to start with administrative data, supplement them with surveys and anchor this process in a system characterised by a valid chain of inferences. In doing so, we used a few new administrative data sources in addition to the existing ones. But we used them in a different way.

In 2017, the Statistics Act was amended for specific reasons not related to the use of administrative data. Various clauses were added, such as clauses relating to the appointment of the Chief Statistician, or the decriminalisation of certain cases of non-response. But, as we are using administrative data and updating our statistical approaches, some people have begun to associate the amendment of this act with the development of the use of administrative data. From there, some people began to show their reluctance at the use of these data.

For example, in 2018 there was uproar following our discussion on the possibility of using data from individual bank transactions by obtaining information from banks. The project did not go ahead, but the controversy went as far as the Prime Minister, who faced questions in Parliament on the matter. This episode had a lasting impact, changing the perception of the social mandate of public institutions, in general, and Statistics Canada, in particular.

We then took this episode into account in our ongoing development work. This led us to significantly increase transparency. We share information on our website before, during and after matching operations. This information is therefore communicated to the public and to the Minister at least thirty days before we proceed with any matching. This makes the operation more complex, but it is very transparent.

We then developed a process or framework for acquiring administrative files and, by extension, for matching operations, based on necessity and proportionality. This framework has been made much more robust and incorporates ethical considerations that take into account privacy, transparency, fairness, trust, and even non-maleficence precautions. It focuses on the benefits of using administrative data and matching operations, not only for the statistical system, for decision makers or for statistics partners, but, first and foremost, for the public. It is a case of asking ourselves how the collection of administrative data and matching operations can be in the public interest, and then implementing an approach in line with that need.

Let’s take a closer look at the 2017 directive on matching operations. This includes all types of data matching and works on the assumption that the data have or can be acquired. As I mentioned, the acquisition of any file is provided for in the Statistics Act. The legal mandate is therefore well established. Nevertheless, in around 2018 or 2019, the acquisition of the social mandate was impeded. The framework of necessity and proportionality was therefore introduced to help us obtain this social mandate. We had to provide a very detailed justification for why we use data, for what purpose, and how their use would meet the principle of proportionality, even if the Statistics Act fully allowed us to use them.

Therefore, as this falls within an administrative data approach and not just a survey approach, rather than assessing the factors associated with privacy at each matching request (given that we conduct thousands each year), we created a general authorisation and evaluation procedure. This is to avoid the procedure for requesting matching opera-

tions for all routine activities, such as the use of administrative data in the construction of survey bases, quality assessment, or coding assessment.

By contrast, adding new variables to an old or new match requires a request, which involves a form, an assessment of the factors associated with privacy, and in some cases must go through a committee. The compliance of these requests with Statistics Canada's mandate is then examined to ensure they are in the public interest and respect confidentiality. The committee is still verifying the potential effects of data processing on Canadian values such as fairness or non-maleficence, while ensuring compliance with the need for proportionality and verifying the scientific and methodological interest of the matching. The data collection process then takes place in a secure environment, using anonymisation and a linking key.

At the same time, the information is published upstream on the Internet. And at the end of the year, a report is presented to Parliament with all the matching operations carried out. In all cases where privacy-related factors are assessed, the matching operations are presented to the Privacy Commissioner, a senior official reporting directly to Parliament.

Before conducting a matching operation, the files must be acquired. The acquisition of any new file is subject to a summary assessment of the potential sensitivity of the data. This assessment is then followed by an assessment of the necessity, effectiveness, proportionality and possible alternatives. This procedure is transferred to the Ethics Committee if the assessment highlights ethical considerations. Following this, the Chief Ethics and Scientific Integrity Advisor determines whether the project can start and makes a recommendation to the Programme Manager. The necessity and proportionality framework applies to all file acquisitions and matching operations.

Furthermore, we are currently working to establish virtual research centres to enable researchers to access data from home, based on the five security levels described by John Dunne.

We are also working to establish a sensitivity scale based on the analysis of non-responses in surveys as well as on focus groups and meetings specifically aimed at gathering information on Canadian sensitivities to different subjects. For example, survey data or public utility data (electricity, cable subscriptions, etc.) are much less sensitive than financial transaction data, for example. I cannot show you this scale today as we are still in our preliminary work phase. We want to develop a scale that can help matching operation managers from the outset to determine the level of security and transparency needed, enabling them to adequately estimate the work required to document and justify their programmes.

Finally, we are developing a central register system. In Canada, we have a business register and an address register that has been converted into a real estate register. We are now working on a register of individuals, using a new secure infrastructure dedicated to data integration. This project is part of an agenda similar to that of the RESIL programme. We should therefore have a comprehensive system to facilitate all matching operations by 2025.

I have now given you an overview of the current work being undertaken by Statistics Canada to answer your question. Thank you for listening.

Chantal Cases

This perspective from across the Atlantic that you have shown us is interesting. Thank you for that, and for getting up before sunrise in Canada to attend the whole meeting; it is very gallant of you. I will now give the floor to Bertrand Pailhès, Director of Technology and Innovation at the CNIL. In the recent past, he was Chief of Staff to Axelle Lemaire, Secretary of State for Digital Affairs. In this role, he led efforts for the adoption of the 2016 Law for a Digital Republic, which specifically made it easier to conduct matching operations for official statistics and research.

To take stock of compliance with confidentiality and information from the public, from whom personal data are collected, what safeguards have been developed by the CNIL to perform file or register matching operations while maintaining privacy? Could you remind us again how the risks and potential benefits of these operations are assessed within the CNIL? We found that it was not always easy to understand all these projects in detail.

Secondly, I hope you can answer a more controversial question. Safeguards and confidentiality are essential. But, these safeguards can become too restrictive. So, how can we build safeguards that do not prevent the production of matches and analyses that are useful for the collective good? Perhaps Mark Hunyadi will tell us more about this issue, but this matter is close to my heart and I think this is also the case for the President of the ASP. For example, matching health and social data is still extremely complex, which prevents us from learning much about the social inequalities in health and access to care. This information would be particularly useful in addressing these inequalities.

Bertrand Pailhès

Hi everyone. I apologise for not being there with you, but the health situation forces me to be remote. To answer your first question on safeguards for matching operations, I think that the different countries mentioned today have cultures with a focus on personal data protection. France has also historically had a high level of protection for these data, which can be seen in particular in the stringent framework surrounding the NIR.

First, it is important to clarify that matching operations are subject to the GDPR framework, which requires each study or survey, each “processing” operation as defined by the GDPR, to evaluate the principle of proportionality and the purpose justifying the processing of personal data. This general principle applies to matching operations in particular.

It should also be noted that the GDPR, which is a European text on personal data protection, has also introduced a specific regime for statistical processing, with the aim of lifting certain information obligations or rights of opposition or access that people may have with regard to the statistical processing of data relating to them. It is paradoxical to mention this in a round table associated with the issue of transparency. Finally, the intention of the European legislator is to broadly allow the use of data for statistical purposes, including for purposes other that those behind the collection of those data. This intention is part of the idea of data processing for the common good.

In addition, as part of our assessment of risks and potential benefits, the GDPR provides for the use of a Data Protection Impact Assessment (DPIA). It is a document that aims to lay down the risks to people and the measures to be taken to limit those risks. The DPIA is

not required for all processing operations, but in practice, large-scale processing involving sensitive data requires a priori use of this tool; this applies in many cases of matching projects.

At the CNIL, we worked with our European colleagues to develop a method for this purpose, which aims to provide the highest number of elements in an attempt to assess the impact of this processing on the rights and freedoms of individuals. This approach is undoubtedly similar to the work done by Statistics Canada on the sensitivity scale described by Eric Rancourt. This method is part of an approach that is certainly taken, in part, from an approach centred on information system security. But this method is intended to be more multidimensional, adapting to the type of data, the party performing the processing and the type of threats that could affect the data. In some cases, data that may not seem to be sensitive in one processing operation may be more sensitive in another context, for example where there is a chance they may be disclosed and thereby have a direct effect on people.

In particular, this risk analysis exercise invites us to reflect on human rights and freedoms and ask ourselves how data processing can affect these fundamental rights.

There are also other safeguards. For the CNIL, matching is not an end in itself: any data matching operation addresses a given purpose. For example, data matching can be performed to study career success in terms of educational attainment, as Francesco Avvisati described. We will always look at these matching operations in terms of their purposes.

I would also point out that regulatory acts still allow us to perform matching operations involving health data, and that there is much more to the discussion about this purpose. To do this, we have the Health Data Hub, which is intended specifically for performing matching operations.

Moreover, governance provides a further safeguard. Eric Rancourt mentioned various committees that monitor data processing and we also have such committees in France. For example, the CNIL is part of the Official Statistics Quality Label Committee for official statistics survey projects and the Statistical Confidentiality Committee for Research. A small but constant part of the CNIL's activities has for a very long time aimed at helping to assess the timeliness of data processing in the context of studies and surveys. The CNIL continues to work to identify risks that would not have been identified by the research and statistics communities, by taking a deeper look.

In this context, the CNIL pays particular attention to sensitive data. This is the subject of regular debate in France. In particular, this debate relates to ethnicity data, which are sensitive data under legislation and require special attention. The proportionality check is particularly important here.

The opinions issued by these committees may, for example, push for the collection of sensitive data to be made optional, so that respondents have the option of not responding, in order to protect their freedoms. In practice, however, this option does not seem applicable to all surveys.

This governance also works in health research, but under a far more binding legal framework than elsewhere. This framework is characterised by authorisation schemes, which may be issued by the CNIL, or by the Ethics and Scientific Committee for Health Re-

search, Studies and Evaluation (CESREES), which are questioned on the public interest of data processing operations.

I will now move on to your second block of questions, which I think is particularly interesting. In France, there is a tradition of personal data protection. Complaints against the automated system for administrative files and the register of individuals (SAFARI), which was set up to create large files and interconnect data, were lodged back in 1974. As a result, conditions for access, in particular, were tightened on the grounds of two main fears. The first was the fact that the NIR is an identifier as it corresponds to a birth order number in a given municipality on a given date that allows for direct identification of individuals. The second was associated with concerns about large volumes of identifying data.

Two main approaches were developed to avoid the strangleholds associated with these safeguards. The first solution relates to the legal system. It is a question of changing the law on the basis that the preventive measures introduced in the 1970s are no longer necessary and that we would do better to move closer to the system used in what we call “register-based” countries that have detailed registers of populations and housing and thereby create a more flexible system for using certain identification numbers.

The Irish case that was presented is indeed interesting in this respect. I note that in other countries such as Estonia, which has experienced a communist dictatorship, having a large database with ID numbers does not seem so bad in comparison with the Soviet period. But France, which has not had the same history, is still very committed to such protections. This solution is a political choice that must go through Parliament, and the role of the CNIL is simply to apply the legislation in force.

The second solution, which seems to me the most interesting given that I am the head of the Directorate of Technology and Innovation at the CNIL, calls for technical assistance. The CSNS really falls under this approach. When I initially began working at the CNIL in the 2010s, I noticed that the CNIL was often questioned about bottlenecks. To perform an NIR-based data matching operation, you had to obtain a Council of State decree, which was a complex procedure for researchers. The CNIL then launched a reflection process in relation to a non-identifying code derived from the NIR.

When I worked at the State Secretariat for Digital Affairs, there was talk of integrating this idea into the Law for the Digital Republic, in order to simplify the process for researchers to access data. Indeed, data matching is crucial to guiding the conduct of public policies and promoting their effectiveness. We had to find a balance between the two that would protect us against the risks associated with personal data protection. We then introduced a clause into that law, proposing the CSNS solution, which is more complicated than direct use of the NIR in matching operations, but which allows us to cover those risks. The number is no longer identifying, and if a search base disappears, it will not be possible to easily match these numbers with the real identities, thanks to the cryptographic mechanisms in place.

These technical solutions also include building large, secure data hubs and accepting broader data access that is nevertheless still subject to a stringent security framework. And the CASD, whose recent development has been well received by the world of research, is part of this strategy. In particular, the CASD has helped us to improve the ergonomics of researchers’ work, which is also a very good point. Also as part of this strategy, we are following up on the Government’s Health Data Hub project. In terms of research

data, the legislator really wanted to collect data in a very secure place, as evidenced by the 2016 and 2019 acts, which were mentioned during this meeting. For the CNIL, which plays the role of regulator, it is ultimately somewhat reassuring to have such a system, managed by professionals, rather than having databases that are taken to research laboratories, where security is not always guaranteed.

As for the cross-referencing of health and social data, I'm afraid I do not have an answer at this time. Perhaps the solution will come from infrastructure such as the CASD, which retrieved much of the SNDS data last year with CNIL approval. I think that we now have the possibility of performing data reconciliations that seemed relatively impossible 10 years ago. I therefore hope that we will be able to maintain our pace in the context of international competition, while maintaining a high level of protection.

Chantal Cases

We will return to your comments during the discussion section. I will now give the floor to Mark Hunyadi, professor of moral and political philosophy at the Catholic University of Leuven and member of the newly established Orange Ethics Committee, as well as the Joint Ethics Committee of the National Institute for Agricultural Research (INRAE), the French Agricultural Research Centre for International Development (CIRAD), the French Research Institute for the Exploitation of the Sea (IFREMER) and the French National Research Institute for Sustainable Development (IRD). He has also published various works, including *Au début est la confiance* [In the beginning is trust] published in 2020, which will probably be mentioned in our discussions.

First, we have been shown various register and matching projects this morning. And there are yet further international projects that have not been presented. Fortunately, these projects provide an opportunity for statisticians to deepen their thinking about their professional ethics. I would therefore like to hear your views on the ethics of official statistics stakeholders.

In addition, since you have conducted a significant amount of work on the concept of trust, I would like you to offer us your advice and share your views on how official statistics can maintain the public's existing confidence, particularly in terms of matching operations involving personal data.

Mark Hunyadi

Hi everyone. I would like to thank Chantal Cases and Françoise Dupont for inviting me to this round table. I am not going to be speaking as a statistician, but as a moral philosopher. I therefore feel I am sneaking into a club whose codes and assumptions I do not share.

When it comes to data matching, statistical stakeholders, such as INSEE, are faced with an ever changing context, characterised by the emergence of new data sources and players. This context requires a reassessment of the practices and identity of official statistics bodies.

However, this reassessment of identity is in itself an ethical issue. It does not relate to technical problems, but to ethical ones. It is a question of reassessing your role, your func-

tions and your usefulness, as well as the principles and values that govern these new practices and guide your actions.

I think you can all sense the existence of this identity issue. At the very least, I have sensed the existence of this concern in my interactions with you. This identity question relates to an ethic of identity. But I will specify that this is not an identity ethic, where identity relates to a redefinition of oneself, given that a universalist identity is also an identity.

To answer your question on ethics more precisely, it is always interesting to distinguish between two types of extraordinarily different ethical issues.

The first refers to what I call “little ethics”, and I apologise at the outset for this size descriptor, which may seem pejorative. This little ethics is centred on the individual and is at play in the questions we all ask ourselves about data security or privacy. The CNIL is interested in exactly this type of ethics.

In our societies, this ethic is embodied in the ethics of human rights, the highest element of the normative architecture of our advanced Western societies. “Little ethics” therefore refers to this set of problems that relate to the defence and protection of individual rights and freedoms. Bertrand Pailhès provided a beautiful illustration of how these problems can be solved in principle through law and technique, for example, through the establishment of safeguards.

As part of this little ethic and our reflections on data matching, I was rather worried this morning by the content of the different presentations. I really wonder about our ability to reconcile the principle of data minimisation with the notion of maximising matching. So what I heard is quite spine-chilling, especially in relation to RESIL, irrespective of the safeguards that we might imagine. I have a good understanding of the principle of the different legal and technical safeguards that you presented, but I would like to draw your attention to the fact that these safeguards can also be removed with the stroke of a pen, as a result of changes in laws, normative context or government. So, I can only worry about this possibility.

However, in my view, this remains an issue of little ethics, centred on the individual; it is not a global ethic. “Global ethics” or “big ethics” are the terms I use to refer to the second level, which arises when we ask ourselves about the purpose of the matching operations, their role, and their overall use as part of a societal project designed by the use of statistics. In short, this ethic is apparent when we reflect on the meaning of this vast statistical project.

In general, we ask ourselves a “big ethics” question when we ask ourselves about a world in which we seek to fully quantify everything. This desire establishes a relationship with the world shown through numbers.

And we know the importance that this number-based mediation occupies in the political sphere. The importance of this mediation is indeed quite normal and understandable, given that our modern and complex societies quite logically require management through statistics. That is why statistics were created in the 19th century, not from mathematics, but from a branch of social policy, on the basis of a demand from governance and society.

Statistics are governed by a supreme value, which is the production of a certain type of knowledge, namely what is measurable and objectifiable. Yet, these interests for knowledge are constantly being threatened by an instrumentalisation of power. Indeed, if there is an interest for knowledge in research, these interests also trigger the emergence of a governmentality project. There is a lot of talk at the moment about the concept of algorithmic governmentality driven by Tech Giants and I think here we can also speak of statistical governmentality.

This statistical governmentality project uses matching operations and reinforces the security problems linked to little ethics. But the project also falls under big ethics as it is a societal project that uses statistics to monitor the direction of public policies and to assist in the planning of those policies.

Although I do not want to give a purely negative view of this project, I think that, in our complex societies, it is inevitable. And this inevitability may be distressing. However, as part of this statistical policy steering, we are addressing citizens not as free subjects with inventiveness, imagination, and political will, but as beings who can be steered and programmed. I am exaggerating a little bit, but within this statistical governmentality of politics, statistics serve to perform this kind of programming.

I am convinced that there is a legitimate interest in statistics. However, I see it as a very ambivalent phenomenon. For this reason, I want to show that while there is indeed the need, required by the complexity of our society, to respond to this interest in statistical knowledge, this interest places us in very problematic proximity to the taming and, above all, domestication of reality and the people that statistics help to govern.

There is therefore a great responsibility associated with statistics, which shape people's social views as they seek evidence in numbers. This morning, the French Finance Ministry was happy at the news that INSEE had just shown that growth reached 7% in 2021. In France, we are addicted to numbers that shape social views and the way we represent the individuals who are the subject of those statistics. In this way, we then see these individuals as beings to be steered, rather than as autonomous people of free will. And these considerations are not without problems in a democratic context.

These questions do not refer to little ethics. In speaking about this societal project arising through this statistical governmentality project, I am referring to a more general concept of society. It seems to me that these questions relate to real ethical problems, given that those that fall under little ethics can in principle be solved more or less easily through legal and technical means. By contrast, when it comes to "big ethics", we face a very different problem, for which no technical or ethics committee is envisaged. Indeed, no ethics committee agenda includes these questions on the societal project that is emerging around digital technology and statistics. They are not interested in these big movements that are currently sweeping us up like a tidal wave over which we have no control. No ethics committee exists to regulate these projects, which is a real problem. All of our ethics are centred around little ethics and we are trying to solve them without understanding that we are also being swept away by the problems of big ethics.

To answer your question on trust, I would say that the first thing we need is to know what trust is. As you pointed out, my last book focuses on this concept. We all hear this word without necessarily knowing what it refers to.

In a nutshell, trust is always associated with expectations we may hold in relation to something. We sit on chairs and expect them to support us. To trust is to bet that these chairs, which may well collapse, will support us.

The same is true of people, about whom we have expectations in relation to behaviour. Trusting someone is to expect them to behave in a certain way. And the same applies in relation to institutions. For example, we often hear the phrase “we trust in the justice of our country”, which means that we trust that justice and those who represent it will act in accordance with the ideals of justice, such as fairness and the multiplication of viewpoints.

Therefore, trusting in an institution means that we expect that institution to be up to the expectations that it itself has produced. Taking the example of INSEE, we expect it to produce sound knowledge. To verify that this knowledge is sound, we must verify that it conforms to standards developed by statisticians. So, to trust an institution is to bank on a normative expectation that meets what we expected. And it is up to you to know what is expected of INSEE.

Chantal Cases

Thank you for offering up an alternative point of view. I think your presentation is very useful in our reflections on matching operations.

I will now hand over to Maryse Artiguelong for the last presentation of this round table. She comes from the world of IT and is here as Vice-President of the Human Rights League and the International Federation for Human Rights. She leads the Human Rights League’s “Freedom and Information and Communication Technology” working group. She is also a member of the Freedoms and Digital Technology Observatory. So, we are particularly interested in her point of view.

First, what are your thoughts about the data matching and register projects presented today in relation to official statistics? What benefits and risks do you see for the people affected by these data? What is your position on the points relating to “big ethics” issues?

Finally, as we are trying to organise transparent information and consultation on these projects, it would be good if you could provide us with your suggestions on this, in particular on the procedures that would be necessary upstream and downstream of the projects. In addition, the CNIS is a typical venue for consultation of this type and there are also others. I would therefore like to know whether you feel that this framework is sufficient and whether you have any ideas to improve it.

Maryse Artiguelong

Hi everyone. I would like to thank INSEE for inviting the Human Rights League. It is true that we have a long tradition of establishing ad hoc partnerships with you to address certain questions. I would just like to mention that the International Federation for Human Rights is commemorating its 100th anniversary this year. It was set up back in 1922 by the French and German leagues who were experiencing the emergence of difficult times without being able to prevent their arrival.

While we can admit that the Human Rights League’s role in tackling discrimination and protecting personal data and privacy is not as significant as we would have liked, it is still a

part of our daily struggles. I have learned a lot today and I would like to thank my colleague for encouraging me to accept this invitation. As a result, I have learned that my day-to-day work falls under the concept of “little ethics”.

More seriously, having worked on the issue of protecting privacy and personal data for a long time, I have personally found that matching operations relate more specifically to links between police and justice files. In fact, we recently took action against the matching between the alert processing files for the prevention of terrorist radicalisation (FSPRT) and files relating to people hospitalised without consent for psychiatric reasons (HOPSYWEB). We felt this matching project was really rather awful as it relates to health issues and generates significant, shameful suspicions relating to the individuals interned by court order.

Nevertheless, I was impressed by the ethics, rigour and monitoring of the doctrine, which govern all the projects that have been presented. However, the amount of data being processed seems staggering. In fact, I do not know if everyone realises that these huge administrative files include each and every individual.

So, while many safeguards have been presented today, I worry that they may be circumvented, especially, as we have seen, as adding just one question to a survey is not necessarily a neutral operation. Although I know that the European Union monitors the official statistics institutes of each EU member state, I worry about the possibility that a government could make a change to the ethical framework of these data matching operations.

Furthermore, I agree with Jean-Noël Barrot’s remarks regarding the possibility of Tech Giants replacing statistics institutes and quickly obtaining more or less reliable data from their big data systems. If this were to happen and funding for these institutes were to stop, we would risk losing the reliability of the information and no longer be able to answer our questions with truths.

What’s more, I note that the RESIL presentation raised many questions that are consistent with some of my own concerns.

The matching operations and registers that you have described have advantages. It is useful for the population to benefit from public policies based on official statistics as well constructed and reliable as the ones you have presented. In this regard, I am reassured. But, we might also ask ourselves whether statistics should be the sole basis of public policy.

Indeed, official statistics data provide results relating to a given moment. But they cannot predict the unpredictable. Who would have said that we would be locked down for two years and that many would be confined to remote working?

I will now answer your second question on consultations and exchanges that might take place upstream of the matching projects. Ultimately, we have learned a lot today, thanks to the information shared during this meeting, and we will all leave today better for it. I think that we should organise shorter and smaller meetings, but with lawyers and technicians who do not come from statistics institutes, and also open them up to civil society, whether organised or not.

In this regard, I note that the National Consultative Commission on Human Rights (CNCDH) is currently working on a statement on artificial intelligence and human rights. In

this context, opinions from many legal experts as well as other actors have been heard. Of these, we heard ATD Fourth World speaking about the organisation of workshops with people in need whose level of digital knowledge was not very high. It was very interesting to hear this side and I think we have everything to gain from listening to everyone. This operation may be complicated, but it is fundamentally positive.

Finally, I believe that we must be bold and communicate with the general public. They do, of course, know that censuses are conducted periodically, but they do not know the details of the work conducted by the Official Statistical Service. We all hear about the INSEE figures, but we don't know where they come from.

Discussions

Chantal Cases

I would point out that ATD Fourth World and other civil society actors participate in CNIS committee meetings, as part of the consultations carried out upstream of the statistical operations.

It is time to take a look at the questions from the chat and the room.

Remote guests

Could you provide details on the statistical processing sensitivity scale presented by Eric Rancourt? In particular, I would like to understand the link between this scale and non-responses to surveys, and its link with opposition to matching operations.

Are there any plans to conduct a survey or meta-analysis of the available data, so as to know what the main parties affected, namely those surveyed, think about it? What do we know today about their trust or distrust of the privacy guarantees that are advertised by the official statistics stakeholders? Finally, have the reasons for refusing to respond been analysed? What do we know about the reasons behind this worry or about what in particular makes them feel secure?

Some experiments have been conducted on establishing trust, particularly in the Nordic or Anglo-Saxon countries. These involve committees of citizens linked to large surveys that integrate matching operations or sensitive data, or rely on large data centres. Could this be seen as a way to build public trust? Where would these committees fall between little and big ethics?

Eric Rancourt

I will start with the first question, but also provide a few details in response to the other two questions.

I referred to the link between non-responses to surveys and the planned sensitivity scale currently being developed by Statistics Canada. This scale should allow us to grade a variety of subjects in order to guide our efforts in terms of transparency, ethics, fairness, security, and privacy. This is not, however, about unsafe or unethical data processing operations. Nevertheless, we do not have the markers we need to ensure that these principles are respected. As an example, depending on the level of sensitivity measured, we may, in some cases, prefer to collect aggregated data instead of micro-level data. We may also

reduce the size of a file by using a small sample rather than a complete administrative file. It would even be possible to reduce the number of variables.

Among other things, this planned sensitivity scale follows the discussions regarding our plan to study banking transactions and credit data in an attempt to improve our understanding of households in insecure situations. The project had a very important social purpose, but society was not prepared to grant a social mandate to enable the collection of comprehensive microdata. Taking this sensitivity into account therefore leads us to adjust the method or the nature of a project.

I would also point out that, in order to develop our sensitivity scale, we will seek to understand why non-respondents do not respond to surveys. We are also interested in the views of the interviewers who carried out the survey with these individuals. In parallel, we organise focus groups with non-respondents to obtain qualitative elements. Psychological researchers specialising in social participation help us understand these non-responses.

Finally, our openness to society is more than simply publishing information on our website. It also involves trying to contact citizens in different ways. In particular, we have set up an external ethics committee. We are still trying to develop satisfaction or trust surveys, asking the public about the factors that seem to have the greatest impact on their trust in Statistics Canada, as well as ways to improve it.

Chantal Cases

There may be other participants in this round table who wish to answer these questions. I would probably give the floor to some of the INSEE members to address questions relating to satisfaction surveys or even questions relating to the analysis of non-responses. But as the term “trust” was mentioned, I think Mark Hunyadi would like to respond.

Mark Hunyadi

It is certainly true that citizens’ committees increase the political relevance of statistics. I am not aware of these experiences in the Nordic countries that you mentioned and we should, in particular, check to see to what extent these citizens are taking part in the decisions.

Furthermore, in this new context of increased data sources, particularly from private actors such as the Tech Giants and where the official statistics institutes are constantly competing, I would like to stress the existence of an important card that official statistics can play.

The Tech Giants simply record behaviours that have actually occurred, and they do not do so with any cognitive interest, but in a predictive sense. Indeed, their aim is to predict future behaviour, mainly for commercial purposes. They merely draw *ex post* conclusions from recorded behaviours as this is the way the digital tool works.

So, in the face of Tech Giants, official statistics bodies have an extraordinary asset, given that they are not limited to simply recording facts, such as administrative facts, but can also carry out surveys. That is why it would be interesting to reconsider the survey, because it enables us to understand not only what people are doing, but also what they want to do and achieve. No Tech Giant digital device can obtain this information.

In this way, a more survey-oriented methodological paradigm could increase the democratic relevance of statistics. In some ways, this would truly allow us to reconcile the citizen's perspective to the overriding perspective of the person simply measuring the phenomena. It is a question of asking people what they want in different areas, such as work mobility. Only the survey can find out these aspirations. It seems to me that this could be seen as an important key to giving statistics their full place in this context dominated by digital data recording.

Chantal Cases

Thank you for rehabilitating the survey and its questioning of behaviours and aspirations. I would also like to remind you that from the start of this meeting and throughout, we have stressed the complementarity between surveys and administrative files.

Maryse Artiguelong

I would like to point out that the Tech Giants also study the past to influence people's behaviour. As Etienne Klein said, algorithms can be used to predict the future provided they have a strong resemblance to the past.

Chantal Cases

Perhaps I can provide some clarification on how we measure public trust or satisfaction, although Jean-Luc Tavernier should probably address this point in his conclusion. In this regard, I would firstly point out the existence of various satisfaction surveys. What's more, it seems to me that non-responses are generally analysed upstream of surveys and downstream when preparing analyses.

Christel Colin

As an addition, I would like to return to something mentioned by Eric Rancourt. Statistics Canada has been forced to respond to the sharp drop in the response rate in household surveys. But, clearly, we are not seeing this phenomenon in INSEE surveys. There may be a slight drop in the response rate, but it has not plummeted. I guess this is due to efforts made to convince households to respond, including by providing questionnaires that are appropriate and limited in length.

In addition, we have qualitative feedback on respondents' reactions to questions addressed to them. They show that those not wanting to answer most often do not answer due to time constraints and not for reasons related to lack of trust or mistrust. By contrast, some respondents invited to take an online survey ask how their email address has been found, indicating that there is some sensitivity to this issue.

Lastly, you have just mentioned the existence of satisfaction surveys on indicators produced by INSEE and on INSEE in general. However, I am not in the best position to provide clarification on these surveys. In any case, these satisfaction surveys, which are generally aimed at Internet users, are conducted every year to monitor public trust in the INSEE indicators.

Jean-Luc Tavernier, Director-General of INSEE

I must say that I did not plan to discuss the issue of trust in my conclusion. There are, of course, satisfaction surveys aimed at measuring trust in our institution. On the one hand, we interview Internet users who visit the INSEE website, but we are aware that this means this audience is already favourable towards us. But, on the other hand, we also interview the general population.

And in this second case, we don't expect any great results. As an official public-domain expert, we are affected by two stigmas: one associated with expertise and one associated with belonging to the public, official sphere.

But, the trust that we measure more generally, indicator by indicator, is not associated with these stigmas. Indeed, when we ask "do you have faith in official public statistics indicators?", the answers are not very positive. However, when we contextualise our question by specifying that INSEE produces a particular indicator each month and by asking the public about its trust in these indicators, the answers are far more favourable.

So, in my role, I see a real difficulty in maintaining trust as, throughout the world, and especially here in France, there is a general tenancy to adopt a defiant attitude to anything that looks like expertise or an official word.

A remote guest

Within the framework of this comprehensive "statistical governmentality" project, how can we interpret the absence of examples of matching operations concerning the energy transition, which would allow us to better monitor the renovation of housing or the vehicle fleet, or even household or business energy consumption, as well as the associated aids? Achievements made through data matching operations seem to indicate a hierarchy in political priorities, which are long-standing and linked to the resistance of the actors involved.

Bertrand Pailhès

I am going to present a point of view that is not the one held by the CNIL, but comes from my experience, as well as from the discussions held in 2016 as part of the drafting process for the Law for a Digital Republic. Here, we were primarily guided by the idea of granting access to administrative databases for any researchers, not just "official researchers". I believe that the CASD recently signed an agreement with Banque de France. However, five or six years ago, in order to conduct research on Banque de France, it was necessary to obtain the institution's approval.

I believe that trust is also based on a certain plurality of research and scrutiny, in the context of targeted questions and matching operations. Indeed, historical structures or powers may lead to many studies on growth and very few on housing renovation. That is why we clearly need to build a framework that enables others to explore the field, access data, and conduct research on other issues. In this way, we could combat the two stigmas referred to by Jean-Luc Tavernier, linked to the embodiment of public authorities.

Finally, I also believe that the solution lies in being able to challenge the public sector through independent research from public universities or other sources and that could develop a multi-faceted view of the world.

Chantal Cases

Do you think that broad-based open data, under good conditions, will foster trust?

Bertrand Pailhès

I cannot speak for the CNIL, but I can say that this is indeed one of the ideas raised by the Law for a Digital Republic, and by the platform data.gouv.fr launched a while back by François Fillon. But I am also thinking of access to data by means other than open data. Here, I am thinking in particular of work on taxation carried out by certain researchers. So, I wonder about ways researchers can access data, via the CASD or through other channels, so that they can perform analyses that are not performed elsewhere.

Chantal Cases

Official statisticians are already accustomed to being challenged by research, and are in full agreement with this principle. The question of data openness beyond the sphere of research does indeed need to be asked, and its dangers and its benefits outlined.

Bertrand Pailhès

Exactly, at the CNIL, we are rather cautious about open data. I think that this openness has both benefits and dangers. It would allow us to achieve some form of contextualisation. It would enable us to explain how the figures were obtained and thus to restore trust. Moreover, opening up data signals to the public that they can always use the numbers and produce their own analysis if they do not believe in the reliability of the analysis produced by researchers or statisticians. However, this approach is not possible for many data types, especially in cases where personal data are used.

Chantal Cases

I also think the pandemic has made it clear how important this issue of trust truly is.

Patrice Duran

I would like to share a series of reactions with you. Matching operations complicate statistical issues. For this reason, it is important, if not vital, that INSEE and CNIS play an educational role. We are doing what we can in this regard, but we must go further.

Indeed, too many mistakes are being made, mistakes that are sometimes published under the names of scholars. Take, for example, the book *La gouvernance par les nombres* [Governance by numbers] by Alain Supiot, a member of the Collège de France. This brilliant legal expert in labour law develops an attack on official statistics without having any precise knowledge of it, which leads to a number of mistakes obstructing his prose, which Jacky Fayolle, a former INSEE colleague, has, rightly, severely criticised for the complacency that this thesis promotes with ideological arguments about the role of “numbers” in human government! The widespread lack of knowledge of what official statistics really are

has now become a real problem. Too often people confuse data, facts, and statistics, and unfortunately this ignorance is still far too prevalent in public administration as a whole, whether at state level or within local authorities. We must therefore be careful when we talk about governance by numbers, and we must not confuse the statistics with the way we use them. The real problem here is therefore one of training on this subject, which raises questions of both national education and the training of civil servants.

As part of this educational effort by INSEE and the CNIS, we invited different actors such as ATD Fourth World. In this context, we are also organising working groups, with the next one, linked to a request from the rights defender, set to focus on discrimination.

Moreover, it is clear that there is insufficient knowledge surrounding the reality of official statistics, even among those who might think they are better informed than others. For example, we were asked about child health by members of the French National Health Authority (HAS). However, while they were aware that the issue was not just a strictly medical subject, they were surprised to learn that it was far more documented than they thought, when we presented the various data from the different MSOs. Ultimately, it is the ignorance of what official statistics are and what they do that we can no longer tolerate and that we absolutely must take into account.

If we are to combat the distrust in public action that many citizens express, we need to know how to explain and provide information about what we are doing in terms of official statistics. Today, trust is even more crucial because the world is more complex. The German jurist and sociologist Niklas Luhman published a book entitled *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität* (Trust: a mechanism for reducing social complexity), in which he rightly argued that trust was a way to manage complex problems. Where we do not have full control of an issue simply because we do not have the sufficient training to understand the policy in place to manage it, not having trust becomes problematic. The *Gilets Jaunes* movement, for example, showed just how distrust of politics can have serious consequences for social functioning. Trust in institutions is also at stake here.

But this question falls under the “little ethics” that you mentioned. The problems posed by this little ethic have arisen in the context of the pandemic. For example, the refusal to wear masks or to be vaccinated was based on an assertion of individual freedom. It is the whole question of the relationship between human rights and citizens’ rights that is at stake here, because it reveals issues with very different impacts.

It is indeed interesting to see how French administrative law has encountered this problem of little ethics and incorporated the defence of the individual inherent in political liberalism and the preservation of the public interest in its own way. For example, French administrative law distinguishes between liability ‘without fault’ and “fault-based” liability. Fault-based liability conventionally corresponds to the production logic of public organisations where users can seek compensation for damage caused by an organisational malfunction. It is clear, of course, that the increased intervention of the state, and thus of its administrations, increases the opportunities for litigation, whether arising from mismanagement or from failure to apply laws and decrees. The administration has an obligation, which, if not of performance, at the least is one of effectiveness. Liability without fault is both interesting and complex. Its logic could be a forerunner to a modern notion of liability from public action.

And liability without fault is, to a certain extent, a way to combine state power with the defence of human rights. The issue here is not reparation, but compensation. Public action of general interest can have damaging consequences for individuals. As such, they deserve compensation without the policy implemented by public authorities being rejected. That was exactly the aim of the Council of State decree, known as “Ville Nouvelle-Est” (New Town-East), in which the Council of State defended the balance theory in a situation in which a town had expropriated property to create a university complex.
The issue of public action therefore requires that the individual and the collective be combined in a reflection on both the aims of the action and their consequences. We can probably lament the fact that the law on administrative liability does not include, or at least represent, a modern doctrine of public action. This is now the primary aim of ex ante impact assessments, even if the Council of State has not been satisfied with their quality!

I was Chair of the Scientific Interest Group (SIG) for Democracy and Participation, where I also saw the extremely complex relationship between these two terms. The recreation of the political power justification registers creates new constraints in the exercise of that power itself. The claim of legitimacy of our leaders can no longer be satisfied through the mere legality of their actions regardless of their impact. Where efficiency and performance have become the watchwords of a doctrine of public action in most modern states, the evidence of the results of public action must always be highlighted, and official statistics are one of the important pathways for this. As Pierre Rosanvallon put it, it is the “democracy of exercise” that needs to be invented. Unfortunately, the topic of participation as embodied in the notion of pluralism is not free of significant ambiguities. But this is not the problem that we are addressing here.

As such, the issue of little ethics is of tremendous complexity. Yet, while it is clear that we need knowledge to act effectively on public policy, considerations relating to little ethics may potentially impede the most relevant knowledge. This is the primary theme of the debate on ethnic statistics. While we need to work to prevent harmful applications of these statistics, we must also know the people who live in our environment. As Jean-Noël Barrot pointed out, if we are to benefit from effective policies, we need to know what the world is made of.

Chantal Cases

I cannot close this round table without giving Mark Hunyadi a little time to speak if he wants to respond, given that Patrice Duran has just raised the issues of big and little ethics.

Mark Hunyadi

I agree with everything Patrice Duran said, except for his comments on Niklas Luhman and Alain Supio, but these are really points of detail.

Chantal Cases

You will be able to discuss these details after the seminar has finished. I can also see that there is a guest who wants to ask a question.

A remote guest

Little ethics includes elements that give society guarantees that are internal to INSEE, i.e. the set of internal decisions that are made within the institute. It also includes guarantees from external processes that are incorporated into the CNIS or Council of State framework. But when we switch from NIR to CSNS, we switch between these internal and external guarantees. So, how is this link seen today, and how does it account for the strengths and weaknesses of these two types of processes?

Chantal Cases

I am not sure that this question is directed specifically at our guest experts at this round table.

Mark Hunyadi

I do not understand the technical issues of this question, though they do still seem significant. In any case, I can say that when I speak of “little ethics”, I do not mean that it is not important, but simply that it is small.

Sylvie Lagarde

The CSNS, which is internal to official statistics, has a legal framework and allows matching operations to be simplified. Legally, we could work almost entirely internally, while ensuring compliance with GDPR. But I think it is important to take the issue of the CSNS outside INSEE, hence the interest in talking to the CNIS about it, making it visible, and discussing its use, beyond the statistician community, with all users. This is a very important aspect of our work.

Chantal Cases

I also think that this point is essential. During this round table, we have heard about different committees, ethics, improved communication to the general public and greater openness to civil society. So we have a lot of work to do. Thank you very much for participating in this round table to discuss topics that you were not always familiar with, at least for two of you. I would like to thank all the participants.

CONCLUSION

Jean-Luc Tavernier, Director-General of INSEE

Good evening everyone. It is now my task to close this meeting. I will summarise my observations in four points.

<u>The interest in matching operations to further knowledge of social facts</u>

Historically, the 1951 Act defined two pillars in official statistics, administrative data and surveys. Recently, we have talked a lot about private data, and now we are talking about matching administrative data, a particularly important practice supported by administrative data.

INSEE began using this matching practice, rather tentatively, in the “Tax Income” survey almost a decade after its creation, before using it more widely. In parallel, the DEPP was the first MSO to use matching, from 1973 onwards, in relation to a panel of students. The Official Statistical Service has been using this practice now for many years.

However, a more systematic approach to matching was adopted just recently, and I would like to reiterate its interest. As Jean-Noël Barrot pointed out, we do not lack for tax data, as our country has a broad tax domain. So, we have no shortage of administrative data.

However, these data represent only snapshots. Indeed, while these data relate to people, they are pictures of their situations at a given time and with different statuses (pupil, student, job seeker, employee, trainee, etc.). In practice, it is therefore not surprising that phenomena that are part of any dynamic require either surveys or observations at different times. But this absolutely requires matching operations, and for this the statistical information itself is not sufficient.

John Martin explained that this practice has been well developed to document labour market policies, and to better understand the impact of educational policies, unemployment insurance reforms, workfare, and social mobility. One participant raised the interesting possibility of using matching to better understand the issues concerning thermal renovation. This would require data collected before and after these renovations to be compared.

In theory, it is possible to produce information about these social facts from surveys, but their use raises some problems. Before describing these difficulties, I want to specify that we must continue performing surveys and that there is very strong social demand for them. Here, when Chantal Cases stated that we should rehabilitate the use of the survey, I think she had a slip of tongue, because the surveys are and remain very widely used. However, the problem in using surveys lies in their cost. In addition, surveys rely on the respondent’s memory. And that memory can be quite short, making it difficult to obtain details relating to distant chapters of their lives. Finally, surveys have limits on granularity and cannot provide data on small territories, whereas the development of many public policies requires access to territorial information.

This shows us the role that matching operations can play here. These are particularly useful for obtaining information on social facts such as intergenerational mobility, which spans a significant amount of time. For example, Emmanuel Saez’s work has shown mechanisms of social mobility in the US. His work was made possible thanks to the existence of an identifier associated with the children in their parents’ tax return, which they then retain. He was then able to match the tax returns of the children and the parents.

In this regard, based on the streetlight effect (or the drunkard’s search principle in which a drunk person is searching for their keys at night on a wide street but limits their search to the small space lit by a streetlight), I am almost inclined to think that we, in France, attach such importance to the static redistribution of wealth because of this difficulty in measuring social mobility throughout working life or across generations. Due to this lack of data, research focuses on reducing inequality through a system of levying and transfer under a static method. Therefore, the effect of the lack of data on how we approach phenomena makes it vital that we develop matching operations across as large a time span as possible.

However, compared to surveys, data matching is limited by the fact that it makes international comparisons more difficult. This is because matching operations are based on administrative data, which varies from country to country. In addition, John Martin noted that Ireland is a small country that is highly open to migration, and that movements of people are giving rise to grey areas in parts of some individual trajectories. Indeed, observation systems based on administrative data remain eminently national.

The development of knowledge through the multiplication of matched databases and respect for the rights of individuals

I was surprised to see that we have made progress on both matching and respecting “little ethics”, to use the terminology shared earlier. A few decades ago, we were lagging far behind, but are now moving close to the cutting edge of technological progress. This meeting shows that more data are being matched and opened up to research and that this is positive for knowledge. Bertrand Pailhès, whom I welcome and thank for our cooperation in the drafting of the 2016 Law for a Digital Republic, tells us that the CASD has made considerable progress in the security and protection of data. I would also like to thank Kamel Gadouche for his contribution to the development of the CASD. We have no doubt taken the best approach in this arbitration between the multiplication of databases and the respect of little ethics.

Moreover, I was struck by the fear surrounding data protection in the event of a potential shift to an undemocratic regime. I think that this fear is unfounded. If we look at the infrastructure of the CSNS, we see that we can destroy it just in a few minutes. It is entirely possible to burn our boats should there be a regime change. Conversely, an undemocratic regime could very quickly establish a totalitarian system.

Nevertheless, we have seen that conducting matching operations to improve knowledge while respecting individual liberties is a lot of work, especially in France. But, while the entire European Union is covered by the matrix of the Global Privacy Assembly (GPA) and the European GDPR, not all of its members employ the same concerns in their arbitration. I have cited the example of Estonia, which has a recent democratic history and is less concerned with data protection. But it is surprising that Spain, too, is widely using the Padrón identifier and the national identity card in its matching operations. In addition, the Nordic countries, which have a democratic past, use identifiers on a massive scale, making their matching processes much easier.

It is not my place to judge these differences; they are the result of political choices. Nevertheless, I must stress that our matching operations require a great deal of work, because of the multiplicity of identifiers and the strict limitation of the use of the NIR, a story with which we are familiar. Kamel Gadouche showed us that the CASD matches required a two-thirds participation and were quite complex. Lionel Espinasse also showed a similar complexity with the CSNS. These matches require effort and may even be a challenge. For example, Vladimir Passeron showed that it was possible to conduct matching operations using different identity traits, despite some losses.

The link between matching operations and the world of research

There is a statistical stream and a research stream. The Law for a Digital Republic describes both. In this context, I wonder whether we should imagine matching operations already made by officials statistics and made accessible in particular to researchers, as

we are beginning to do, or whether we should, conversely, recycle the work of researchers as essential facilities for official statistics. In fact, I wonder whether the matching operations for these two streams should be done once and for all. In particular, this issue poses a problem of confidentiality, as, in this case, the matched data is more durable and accessible to more users.

In my view, until now, the general trend in France has been that matching operations were part of a research project in an economic laboratory, and that it was up to the researcher to describe limit criteria and quality criteria and to define the number of matching steps required to achieve maximum reconciliation. But a reflection that has yet to mature could lead us to develop a portfolio of ready-to-use matching operations for researchers. Should we move in this direction? Is it lawful and justified to ask researchers who have conducted matching operations to potentially make them available to other researchers or to official statistics?

The issue of little ethics

Finally, without going back over “big ethics”, which could be the subject of much discussion, or trust, which was discussed in the previous round table, I will return to the question of “little ethics”. I think that, at the moment, the public, which provides administrative data, is not aware that these data may give rise to matching operations. They are not aware that, in this context, it is possible that their data is processed such that they can be identified from it. Nor are they familiar with the ethical challenges of matching. We must therefore make a special effort to inform the public.

Official statistics are not subject to controversy in this regard. However, we should anticipate this. If matching operations emerge, we need to reflect on the framework of transparency that we owe to respondents. Here, we have already decided that the CNIS should be aware of all the planned work programmes involving matching operations. In addition, we must ensure compliance with the minimisation and proportionality of data processing.

We should also undoubtedly consider the possibility of external validation of these principles, which seems to be lacking. I think we have to think about these issues in advance rather than waiting to address any controversy. It is not just a matter of self-reliance, but of having an external validation attesting to the effectiveness of our efforts as part of a specific process to enforce these principles of minimisation and proportionality, as the label committee does for surveys.

Finally, I would like to thank all those who have had the stamina to stay with us right up to the end of this day. More than two hundred listeners were connected this morning and perhaps more than one hundred are still connected. I think that the interest in the theme of this meeting goes beyond the producers of official statistics. I would like to thank all the contributors and give a particular mention to Eric Rancourt, who has truly earned his breakfast given the time now in Quebec. I would like to thank all those who agreed to participate today. In particular, I want to highlight the quality of their slideshows. Finally, I would like to thank Françoise Dupont and her colleagues in particular, who were key to this meeting. Thank you to everyone and have a good weekend.