

**Recueil du Symposium de 2022 de Statistique Canada :
Désagrégation des données : dresser un portrait de données plus représentatif
de la société**

**Vers un système de données statistiques
intégrées sur l'éducation et la formation**

par Giovanna Brancato, Claudia Buseti et Donatella Grassi

Date de diffusion : le 25 mars 2024



Statistique
Canada

Statistics
Canada

Canada

Vers un système de données statistiques intégrées sur l'éducation et la formation

Giovanna Brancato, Claudia Busetti et Donatella Grassi¹

Résumé

L'éducation et la formation sont reconnues comme des éléments fondamentaux du développement d'une société. Il s'agit d'un phénomène multidimensionnel complexe, dont les déterminants sont attribuables à plusieurs conditions sociales et socioéconomiques interdépendantes. Pour répondre à la demande de renseignements statistiques à l'appui de l'élaboration des politiques et de leurs processus de suivi et d'évaluation, l'Institut national italien de statistique (Istat) renouvelle le système de production de statistiques sur l'éducation et la formation en mettant en œuvre un nouveau registre statistique thématique. Ce dernier fera partie du système intégré de registres de l'Istat, ce qui permettra de relier le phénomène de l'éducation et de la formation à d'autres phénomènes pertinents, comme la transition vers le marché du travail.

Mots clés : intégration des données; données longitudinales; suivi rétrospectif et prospectif; éducation et formation.

1. Introduction

L'éducation et la formation sont reconnues comme des facteurs clés de la croissance sociale et économique d'un pays. Elles constituent un des objectifs fixés par le Programme de développement durable à l'horizon 2030 des Nations Unies, qui vise à garantir l'accès de toutes et tous à l'enseignement préprimaire, secondaire, technique, professionnel et tertiaire, à éliminer les disparités entre les genres et à garantir l'accès aux personnes vulnérables. De plus, il s'agit de l'un des secteurs aux plus forts investissements dans le Plan national pour la reprise et la résilience (PNRR) de l'Italie, établi pour remédier aux dommages économiques et sociaux causés par la crise de la pandémie de COVID-19. Le plan a pour portée générale d'améliorer les services éducatifs, en passant de la petite enfance aux universités, en augmentant l'offre de services de garde, en réformant la profession d'enseignant, en améliorant les compétences, à savoir la connaissance des langues numériques et étrangères, en augmentant le nombre de diplômés des programmes de STIM (sciences, technologie, ingénierie et mathématiques), en renforçant la formation professionnelle par des investissements dans le système d'apprentissage, en renouvelant les infrastructures scolaires, et en réformant les programmes de doctorat.

Afin d'établir les politiques les plus efficaces et de permettre le suivi et l'évaluation de l'effet des investissements sur le terrain, il est nécessaire de posséder des données détaillées, structurées et hautement désagrégées. De plus, compte tenu des interactions du phénomène de l'éducation avec plusieurs facteurs socioéconomiques, il est également opportun de construire un système d'information statistique qui dépasse les données sur l'éducation et la formation.

À l'heure actuelle, l'Institut national italien de statistique (Istat) produit des indicateurs sur l'éducation et la formation au moyen de plusieurs sources non intégrées. Plus précisément, les statistiques sur les étudiants inscrits et les diplômés sont dérivées des données administratives; le niveau de scolarité de la population est estimé au moyen des données du recensement permanent de la population et des ménages et de l'utilisation de données administratives (Di Zio et coll., 2019); l'Enquête sur la population active (EPA) et l'Enquête sur l'éducation des adultes (EEA) fournissent des résultats comparables au niveau international sur le niveau d'éducation et la participation à l'éducation et à la formation. Le tableau informatif est complété par un ensemble d'enquêtes sur la transition de l'éducation au marché

¹Giovanna Brancato, Istat – Institut national italien de statistique, Via C. Balbo, 16, Rome, Italie, 00184 (brancato@istat.it); Donatella Grassi, Istat – Institut national italien de statistique, Via C. Balbo, 16, Rome, Italie, 00184 (dograssi@istat.it); Claudia Busetti, Istat – Institut national italien de statistique, P.zza G. Marconi 26/C, Rome, Italie, 00144 (busetti@istat.it)

du travail concernant trois populations différentes, à savoir les diplômés du deuxième cycle de l'enseignement secondaire, les diplômés de l'université et les titulaires de doctorat.

Depuis quelques années, Istat appuie de plus en plus sa production statistique sur des systèmes fondés sur des registres, plutôt que des systèmes fondés sur des enquêtes. Dans ce cadre, un nouveau registre statistique, le *Registre thématique sur l'éducation et la formation* (RTEF), est en cours de conception et de mise en œuvre. Il sera le pilier du nouveau système de production statistique sur l'éducation et la formation et répondra à la demande de données désagrégées et intégrées. Le registre fera partie du système intégré de registres d'Istat.

Le présent article est structuré comme suit. La Section 2 décrit le nouveau système de production de statistiques sur l'éducation et la formation et la possibilité qui s'ensuit de renforcer les renseignements statistiques fournis à la société. Le nouveau système repose sur le RTEF et est complété par des renseignements provenant des enquêtes statistiques (section 2.3), de façon intégrée. Afin de mieux cerner le RTEF et ses relations avec d'autres registres statistiques, la section 2.2 traite du système intégré de registres d'Istat. La section 3 fait état de certains des problèmes de qualité et de méthodologie auxquels s'est heurté Istat dans la conception et l'élaboration du nouveau système de statistiques sur l'éducation et la formation. Enfin, la section 4 présente des conclusions tirées du projet.

2. Le nouveau système de données intégrées sur l'éducation et la formation

2.1 Le registre thématique sur l'éducation et la formation d'Istat

Le principal objectif du registre thématique sur l'éducation et la formation est de fournir des données stables et structurées qui suivent de façon longitudinale les personnes dans leur parcours d'éducation et de formation tout au long de leur vie, de la première entrée dans le système d'éducation formelle à leur dernière sortie. Par conséquent, le registre suit chaque élément du cheminement dans l'éducation formelle, du préprimaire à l'enseignement supérieur, en marquant les événements pertinents de la vie étudiante (niveau d'éducation atteint, changements de programme, stages, décrochage, etc.).

Les principales unités cibles du registre sont :

- les étudiants (italiens, étrangers) de tout établissement public ou privé d'enseignement et de formation en Italie; les Italiens qui suivent des cours d'éducation et de formation à l'étranger ne sont pas inclus, à moins qu'ils ne soient enregistrés en tant qu'étudiants dans le Registre des Italiens résidant à l'étranger (AIRE); les étudiants qui passent de courtes périodes d'études à l'étranger sont inclus, à condition que leur absence temporaire soit déclarée dans des sources administratives;
- les diplômés selon les niveaux de la Classification internationale type de l'éducation (CITE);
- les établissements d'enseignement et de formation (les établissements étrangers situés en Italie sont exclus).

Les principales variables du RTEF concernent le niveau d'éducation et les facteurs qui influent sur l'éducation et la formation, p. ex. les compétences d'apprentissage, le rendement, les caractéristiques des établissements d'enseignement et de formation, les caractéristiques du personnel enseignant et les conditions socioéconomiques, tant au niveau individuel qu'au niveau des groupes.

Le calendrier de référence du RTEF commence à partir de l'année scolaire ou universitaire 2010-2011. Certaines sources sont disponibles à des dates ultérieures (Instituts techniques supérieurs, niveau post-baccalauréat, post-maîtrise et doctorat). On étudie actuellement la possibilité d'utiliser les données sur le niveau de scolarité tirées du Recensement italien de la population et des ménages de 2011.

Le registre sera appuyé par la gestion des métadonnées et des indicateurs de qualité (Simeoni G., 2022) ainsi que des classifications statistiques normalisées.

Le tableau suivant décrit les sources administratives qui doivent être intégrées dans le registre.

Tableau 2.1.1
Sources administratives du Registre thématique sur l'éducation et la formation

Responsable	Données	Type
Ministère de l'Éducation	Effectifs, fréquentation et niveau d'éducation atteint dans les écoles primaires, le premier cycle et le deuxième cycle du secondaire	micro
	Bâtiments et matériel scolaires	micro
	Indicateurs au niveau de l'école et de la classe	macro
	Profil du personnel enseignant	micro
Institut national pour l'analyse des politiques publiques (Inapp)	Inscriptions et niveau d'éducation atteint des cours de formation professionnelle donnés par les régions italiennes	macro
Institut national pour l'évaluation du système d'éducation et de formation (Invalsi)	Résultats du test normalisé d'italien, de mathématiques et d'anglais effectué en 2 ^e , 5 ^e , 8 ^e , 10 ^e et 13 ^e année et facteurs socioéconomiques individuels	micro
	Facteurs socioéconomiques au niveau de l'école et de la classe	macro
Institut national de documentation, d'innovation et de recherche en éducation (Indire)	Inscriptions et niveaux d'éducation atteint aux Instituts techniques supérieurs; stages obligatoires; profession 12 mois après le niveau d'éducation atteint	micro
Ministère de l'Université et de la Recherche	Inscription au baccalauréat et à la maîtrise et obtention du diplôme (premier et deuxième cycles)	micro
	Inscriptions et diplômes post-baccalauréat et post-maîtrise; inscriptions et diplômés en spécialisation; inscriptions et diplômes au doctorat	micro
	Inscriptions et diplômés en beaux-arts, en art dramatique, en danse et en musique des premier et deuxième cycles (baccalauréat et maîtrise); inscriptions et diplômes post-baccalauréat et post-maîtrise; diplôme universitaire de recherche avancée	macro
	Examens effectués*	
	Profil du personnel universitaire	micro
Agence nationale italienne pour l'évaluation du système universitaire et de la recherche (Anvur)*	Indicateurs au niveau universitaire	macro

* Les accords officiels concernant les données à collecter n'ont pas encore été établis.

2.2 Système intégré de registres statistiques d'Istat

À l'heure actuelle, les registres statistiques sont largement adoptés aux fins de la production statistique. Dans ce contexte, Istat met progressivement en œuvre un Système intégré de registres statistiques, qui se veut le fondement de la production statistique et des données d'enquête statistique (Istat, 2016). Les registres statistiques sont régulièrement alimentés par des données provenant de sources administratives et sont réorganisés de façon à ce que soient déterminées les unités statistiques d'intérêt et leurs caractéristiques. Conformément à ce que la littérature sur le sujet propose (Wallgren A. et Wallgren B., 2014), le système d'Istat comprend *des registres de base* et *des registres satellites*. Les premiers concernent les populations dites de base, à savoir les personnes et les ménages, les unités productives et les unités territoriales. Les deuxièmes mettent en lumière des domaines spécifiques (p. ex. éducation et formation, handicap, travail, revenu) ou étendent l'information sur les populations de base au moyen de variables supplémentaires (p. ex. registre étendu sur les administrations publiques).

Le système intégré de registres d'Istat est élaboré et tenu à jour au moyen de règles de gouvernance structurées. Les modèles d'ontologie appuient la possibilité de couplage entre registres (Radini et coll., 2018). Des investissements sont actuellement engagés dans la méthodologie afin d'assurer l'exactitude des estimations (Alleva et coll., 2019). L'information est organisée de manière à éviter la redondance (une même variable ne peut pas être gérée dans différents registres) et à garantir la cohérence des renseignements.

L'adoption d'un nouveau système de production de ce type comporte plusieurs avantages, allant d'une réduction du fardeau de réponse à une augmentation du degré de détail des estimations, en passant par l'augmentation des possibilités d'information obtenues au moyen du couplage de données.

Le RTEF sera intégré conceptuellement et physiquement à d'autres registres du système. L'objectif est double : d'une part, compléter les données (c.-à-d. les variables gérées par d'autres registres, les valeurs manquantes ou incohérentes) et, d'autre part, analyser conjointement différents phénomènes.

Dans le RTEF, les renseignements de base sur les élèves et étudiants seront tirés du *Registre de base sur les individus et les ménages*, c.-à-d. les caractéristiques démographiques, les événements démographiques soustrayant les individus de la population d'élèves et d'étudiants (mouvements, décès) et les relations familiales. Les caractéristiques des établissements privés, publics et sans but lucratif offrant des services d'éducation et de formation seront tirées du *Registre de base sur les unités productives*. Chaque variable contenant une adresse dans le RTEF sera référencée avec un *code unique d'adresse* établi par le *Registre de base sur les lieux et les adresses*.

Le couplage du RTEF à d'autres registres thématiques permettra d'analyser l'éducation et la formation en concomitance avec d'autres phénomènes, en particulier la situation professionnelle, les caractéristiques du travail et le revenu.

2.3 L'effet du RTEF sur les enquêtes concernant l'éducation et la formation

Les enquêtes statistiques continueront de jouer un rôle pertinent dans la conception du nouveau système de statistiques sur l'éducation et la formation. Les données sur l'éducation collectées par l'intermédiaire de l'EPA et les données de l'EEA, deux enquêtes relevant de la réglementation européenne, seront fondamentales dans la production d'indicateurs sur l'éducation et la formation comparables au niveau international ainsi que dans la production de renseignements statistiques concernant les résultats sur le marché du travail et l'enseignement formel et non formel. Cependant, la mise en œuvre du RTEF augmentera la possibilité d'analyser conjointement les données d'enquête et les données de registre et, par conséquent, d'élargir l'information statistique produite.

Par ailleurs, les enquêtes sur la transition des études au marché du travail seront remaniées. L'information sera tirée autant que possible du Système intégré de registres, de façon systématique et à une fréquence plus grande qu'auparavant. Des enquêtes visant des renseignements ciblés sur la transition vers le marché du travail seront menées périodiquement.

3. Questions de méthodologie et de qualité dans l'élaboration du nouveau système statistique sur l'éducation et la formation

Dans la transition de processus statistiques indépendants à un système intégré élaboré autour du RTEF, Istat fait face à de nombreux problèmes de qualité et de méthodologie. Les sous-sections suivantes présentent certains des problèmes les plus pertinents en ce qui concerne les données d'entrée, le débit de traitement et les données de sortie.

3.1 Qualité des sources d'entrée

Istat peut compter sur des relations de longue date, solides et collaboratives avec les propriétaires des données administratives destinées à alimenter le RTEF. Dans l'ensemble, la qualité de ces sources est bien plus qu'acceptable. Toutefois, des erreurs de couverture concernant certains segments du système d'éducation et de formation sont présentes dans les microdonnées, c.-à-d. les services de garde préprimaire (0 à 3 ans), les élèves et étudiants suivant des programmes de formation professionnelle proposés par les régions italiennes, les élèves et étudiants suivant tout

niveau des programmes en beaux-arts, théâtre, danse et musique. Comme le montre le tableau 2.1.1, seules des macrodonnées sont disponibles pour certains segments de l'éducation. Le PNRR pourrait cependant favoriser la création d'autres bases de microdonnées qui viendraient appuyer ses missions. En outre, les données relatives à certaines régions italiennes autonomes (de taille limitée) sont parcimonieuses et ne sont pas harmonisées avec les données nationales. Ajoutons que le phénomène se révèle assez dynamique, tant au niveau des universités que des écoles, car des réformes fréquentes modifient la structure du système et introduisent de nouveaux cours. De plus, le degré d'autonomie dans la définition des programmes d'enseignement et des cours est très grand. En collaboration avec les établissements concernés, Istat travaille à l'amélioration de la qualité des données et à l'élaboration d'une nouvelle classification des programmes et des niveaux d'éducation atteints en matière d'éducation et de formation, qui serait harmonisée avec les classifications internationales de la CITE et son extension au domaine d'études (CITE-F).

Les enquêtes statistiques d'Istat sur l'éducation et la formation sont menées depuis de nombreuses années et s'appuient sur une méthodologie solide. Istat continuera de réaliser l'EPA et l'EEA conformément à la réglementation pertinente, tout en poursuivant l'intégration avec les données de registre afin d'exploiter toute l'information disponible, et ainsi réduire le fardeau statistique. La situation des trois enquêtes sur la transition de l'éducation au marché du travail est différente. Comme cela a été mentionné plus haut, elles seront remaniées à la lumière du nouveau registre, ce qui nécessitera un investissement dans la méthodologie pour la révision du plan de sondage, du questionnaire et de la production d'estimations en fonction de besoins d'information ciblés et précis.

3.2 Méthodologie et qualité du processus

Afin de mettre en œuvre le RTEF, un certain nombre de procédures sont en cours d'élaboration. De façon générale, elles visent : l'harmonisation des données des sources d'entrée; des vérifications de la qualité des renseignements manquants et de la cohérence des données dans les sources d'entrée uniques; la micro-intégration des sources; le contrôle de la cohérence des données intégrées; le calcul de nouvelles unités statistiques pour le registre; la macro-intégration; l'établissement de relations conceptuelles et fonctionnelles avec les unités et variables d'autres registres aux fins de la portée du traitement et la portée de la production de données de sortie; l'estimation statistique (statistiques descriptives, estimations de modèles, prédictions).

Le RTEF emploie des données anonymisées. En effet, pour chaque source administrative collectée par Istat, les *unités de base* sont identifiées et, si elles ne sont pas déjà saisies dans les sources précédentes, un nouveau code d'identification leur est attribué. Dans le cas contraire, si elles sont déjà présentes dans les archives d'Istat, le même code d'identification leur est attribué. Ce processus est géré au niveau central à Istat et n'est pas sous le contrôle direct des gestionnaires du RTEF. La possibilité de coupler les données au moyen des *unités de base* comme clé de couplage est assurée et facilitée par l'existence de ces codes d'identification. Cependant, comme dans tout processus statistique, cette procédure peut être sujette à des erreurs. Ces erreurs peuvent avoir une incidence sur la capacité de calculer correctement les unités statistiques d'intérêt pour le registre sur l'éducation et la formation (poste en éducation, individu, unité productive). Par conséquent, pendant la vérification de la cohérence des données d'entrée et la création des unités statistiques, certaines règles de cohérence peuvent être enfreintes en raison d'erreurs qui se sont produites pendant la phase d'anonymisation des unités. D'après les essais effectués jusqu'à présent, cette situation semble rare. Il est toutefois prévu de mettre en place un système de signalement rapide des erreurs, pour qu'elles puissent être évaluées et résolues. Il faut souligner que les changements apportés aux codes d'identification des *unités de base* ont une incidence sur tous les registres du Système intégré de registres contenant ces unités, et c'est pourquoi il faut faire preuve de prudence.

La construction des unités statistiques nécessite des règles de gestion des renseignements manquants et des incohérences dans les données, de sorte que les erreurs de modèle et d'hypothèses peuvent être introduites au cours de cette étape du processus. Istat estime que, compte tenu de la nature du phénomène de l'éducation et de la formation, l'utilisation de données longitudinales fournira un cadre informatif vaste, qui permet de construire correctement les unités, y compris en cas de données annuelles incomplètes.

À l'étape de l'estimation, les questions méthodologiques les plus pertinentes sont liées à la macro-intégration et aux erreurs de modèle dans les prédictions.

3.3 Qualité des données de sortie

L'absence de microdonnées sur les segments déjà mentionnés du système d'éducation, p. ex. les programmes de formation professionnelle, qui sont très importants dans les politiques actuelles, pourrait diminuer la pertinence des statistiques produites.

La production d'estimations intégrant des données provenant de différents registres impose une réflexion sur la cohérence. En effet, ce problème se pose de façon générale pour le Système intégré de registres statistiques. De plus, la production statistique d'Istat concernant l'éducation et la formation devra assurer la cohérence des données de sortie avec les statistiques diffusées par d'autres producteurs.

L'actualité des estimations reflète les délais du calendrier d'acquisition des sources administratives et du traitement par Istat aux fins de vérification des données, d'application des procédures d'anonymisation et de chargement de l'ensemble de données dans le système interne. On utilisera la méthode des prédictions afin d'obtenir des résultats statistiques plus actuels.

4. Conclusions

La conception et la mise en œuvre du nouveau système de production de statistiques concernant l'éducation et la formation, qui s'articule autour du *Registre thématique sur l'éducation et la formation*, constituent un projet exigeant. Toutefois, son élaboration peut bénéficier de la vaste expérience acquise jusqu'à présent à Istat dans l'établissement d'autres registres statistiques.

La diffusion du registre se fera progressivement, de façon à mettre à la disposition des utilisateurs internes des modules des segments de l'éducation avant que la base de données ne soit terminée.

Le registre et la possibilité de coupler l'information avec les données d'enquête et les données d'autres registres statistiques accroîtront fortement la capacité d'analyser en profondeur le phénomène de l'éducation et de la formation au moyen de tous les facteurs socioéconomiques et contextuels qui l'influencent, mais aussi d'adopter une perspective longitudinale.

Des analyses approfondies seront possibles grâce au couplage des données du registre sur l'éducation avec celles des registres du travail et du revenu. L'information sur la structure familiale disponible dans le registre de base sur les individus et les ménages permettra d'établir un lien entre le niveau d'études des enfants et celui des parents. Le suivi rétrospectif permettra d'expliquer les conditions de travail par rapport aux parcours et aux performances scolaires.

Le nouveau cadre informatif permettra de mieux définir les politiques et les interventions en matière d'éducation et de formation et de surveiller leur mise en œuvre et leur effet réel sur la société.

Bibliographie

Alleva, G., P.D. Falorsi, O. Luzi et M. Scannapieco (2019), « Building the Italian Integrated System of Statistical Registers: Methodological and Architectural Solutions », document présenté à l'atelier sur l'utilisation des données administratives et des statistiques sociales du Système statistique européen. Valence, 4 et 5 juin 2019, https://ec.europa.eu/eurostat/cros/system/files/building-italia-integrated-system_istat_0.pdf

Baldi, C., C. Ceccarelli, S. Gigante, S. Pacini et F. Rossetti (2018), « The labour register in Italy: the new heart of the system of labour statistics », *The Italian Journal of Economic, Demographic and Statistical Studies*, vol.72, n° 2, p. 95-105.

Di Zio, M., R. Filippini et G. Rocchetti (2019), « An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data », *Rivista di Statistica Ufficiale* n° 2-3, p.143-174 https://www.istat.it/it/files/2021/03/RSU_2-3_2019_Article-4.pdf

Istat (2016), « Istat's modernization program », https://www.istat.it/it/files/2011/04/IstatsModernisationProgramme_EN.pdf

Radini, R., M. Scannapieco et L. Tosco (2018), « The Italian Integrated System of Statistical Registers: Design and Implementation of an Ontology-based Data Integration Architecture », https://www.istat.it/it/files/2018/11/Scannapieco_original-paper.pdf

Simeoni, G. (2022), « A model for documenting and monitoring quality of statistical registers according to GSBPM and GSIM », présenté à l'atelier ModernStats World 2022, à Belgrade, en Serbie, du 27 au 29 juin 2022 https://unece.org/sites/default/files/2022-07/MWW2022_Presentation_Italy_Simeoni.pdf

Wallgren, A. et B. Wallgren (2014), *Register Based Statistical methods for Administrative Data*, New York: Wiley.