Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society

Toward a system of integrated statistical data on education and training

by Giovanna Brancato, Claudia Busetti and Donatella Grassi

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2022
Data Disaggregation: Building a more representative data portrait of society

Toward a system of integrated statistical data on education and training

Giovanna Brancato, Claudia Busetti, Donatella Grassi[1]

Abstract

Education and training is acknowledged as fundamental for the development of a society. It is a complex multidimensional phenomenon, which determinants are ascribable to several interrelated familiar and socio-economic conditions. To respond to the demand of supporting statistical information for policymaking and its monitoring and evaluation process, the Italian National Statistical Institute (Istat) is renewing the education and training statistical production system, implementing a new thematic statistical register. It will be part of the Istat Integrated System of Registers, thus allowing relating the education and training phenomenon to other relevant phenomena, e.g. transition to work.

Key Words: Data Integration; Longitudinal data; Backward and forward tracking; Education and training.

1. Introduction

Education and training are recognized to be a key factor for the social and economic growth of a Country. It represents one of the goals set by the United Nations 2030 Sustainable Development Goals Agenda, which indicators are aimed at ensuring access for all to pre-primary, secondary, technical, vocational and tertiary education, eliminating gender disparities and ensuring access to vulnerable people. In addition, it is one of the sectors with higher investments in the Italian National Recovery and Resilience Plan (NRRP) settled to repair the social and economic damage from COVID-19 pandemic crisis. The broad scope in the plan is to improve educational services, from early childhood to universities, by: increasing the supply of childcare facilities; reforming the teaching profession; improving skills, i.e. digital and foreign languages knowledge; increasing graduates in STEM (science, technology, engineering and mathematics) programs; reinforcing vocational training investing in the apprenticeship system; renewing school infrastructures and reforming PhD programs.

In order to set the more effective policies and to allow the monitoring and the evaluation of the impact of the investments in the field, it is necessary to have detailed, structured and highly disaggregated data. Moreover, given the interactions of the education phenomenon with several socio-economic factors, it is also opportune to build a system of statistical information, beyond education and training data.

Currently, the Italian National Statistical Institute (Istat) produces indicators on education and training using different non-integrated sources. Specifically: statistics on enrolled students and graduates are derived from administrative data; the education level in the population is estimated by means of data from the Permanent Population and Household Census and the use of administrative data (Di Zio et al., 2019); the Labor Force Survey (LFS) and the Adult Education Survey (AES) provide outputs on education level and participation in education and training comparable at international level. The informative picture is completed by a set of surveys on education-to-work transition concerning three different populations: upper secondary graduates, university graduates and Ph. doctors.

In the last years, Istat is increasingly shifting its statistical production from survey-based to register-based systems. In this framework, a new statistical register, the *Thematic Register on Education and Training* (TRET) is being designed

[1]Giovanna Brancato, Istat – Italian National Statistical Institute, Via C. Balbo, 16, Rome, Italy, 00184 (brancato@istat.it); Donatella Grassi, Istat – Italian National Statistical Institute, Via C. Balbo, 16, Rome, Italy, 00184 (dograssi@istat.it); Claudia Busetti, Istat – Italian National Statistical Institute, P.zza G. Marconi 26/C, Rome, Italy, 00144 (busetti@istat.it)

and implemented. It will be the pillar for the new system of statistical production on education and training and will serve the demand of disaggregated and integrated data. The register will be part of the Istat Integrated System of Registers.

This paper is organized as it follows. Section 2 describes the new system for education and training statistics production and the consequent potential boost of the statistical information provided to the society. The new system hinges around the TRET, complemented with information coming from the statistical surveys (section 2.3), in an integrated way. In order to better frame the TRET and its relationships with other statistical registers, sections 2.2 deals with Istat Integrated System of Registers. Section 3 reports some of the quality and methodological issues that Istat is facing in the design and development of the new system of education and training statistics. Finally, in Section 4 some conclusions on the project are drawn.

2. The new system of integrated data on education and training

2.1 Istat Thematic Register on Education and Training

The main objective of the new TRET is to provide stable and structured data following longitudinally the persons in their lifelong education and training paths, from the first entering in the formal education system to the last exit. Therefore, the register tracks each single piece of the individuals' paths in formal education, from pre-primary to tertiary education, marking the relevant events in the student life (attainments, programs' changes, internships, dropouts, etc.).

The main target units of the register are:

- students (Italians, foreigners) in any public or private education and training institution in Italy; Italians attending an education and training course abroad are not included, unless they are notified as students in the Register of Italians living abroad (AIRE); students spending short periods of study abroad are included, as long as there are signals of their temporary absence in the administrative sources;
- graduates by levels of the International Statistical Classification on Education (Isced);
- education and training institutions (foreign institutions located in Italy are excluded).

The main variables in the TRET concern the education level and the factors affecting education and training, e.g. learning skills, performance, characteristics of education and training institutions, characteristics of the teaching staff and socio-economic conditions, both at individual and at group levels.

The timeframe of reference for the TRET starts from the school or academic year 2010/2011. Some sources are available at later years (Higher Technical Institutes, post-bachelor/post-master/PhD). It is under study the possibility to use data on level of education from the 2011 Italian Census on Population and Households.

The register will be supported with the management of metadata and quality indicators (Simeoni G., 2022) and standard statistical classifications.

The following table describes the administrative sources to be integrated in the register.

Table 2.1.1
Administrative sources of the Thematic Register of Education and Training

Owner	Data	Type
Ministry of Education	Enrollments, attendance and attainments in primary, lower and upper secondary schools	micro
	Schools buildings and equipment	micro
	Indicators at school and class level	macro
	Profile of the teaching staff	micro
National Institute for the Analysis of Public Policies (Inapp)	Enrollments and attainments of vocational courses provided by the Italian Regions	macro

National Institute for the Evaluation of the Education and Training System (Invalsi)	Results of standardized test on Italian, Mathematics and English performed at grade 2,5,8,10,13 and individual socio-economic factors	micro
	Socio-economic factors at school and class levels	macro
National Institute for Documentation, Innovation and Educational Research (Indire)	Higher Technical Institutes enrollments and attainments; mandatory internships; occupation at 12 months after attainment	micro
Ministry of University and Research	Bachelor and master enrollments and graduations; (first and second level degree)	micro
	Post-bachelor and post-master enrollments and graduations; Specialization enrollments and graduates; PhD enrollments and graduations	micro
	Fine Arts, Drama, Dance and Music first and second level academic enrollments and graduates (bachelor's and master's degree); post-bachelor and post-master diploma enrollments and graduations; Advanced research academic diploma	macro
	Exams taken*	
	Profile of the university staff	micro
Italian National Agency for the Evaluation of Universities and Research Institutes (Anvur)*	Indicators at University level	macro

* Formal agreements on the data to be collected has not yet been settled

2.2 Istat Integrated System of Statistical Registers

Nowadays, statistical registers are extensively adopted for statistical production. In this context, Istat is gradually implementing an Integrated System of Statistical Registers, which is meant to be the foundation of the statistical production together with statistical survey data (Istat, 2016). Statistical registers are regularly fed with data from administrative sources, properly re-organized to identify the statistical units of interest and their characteristics. Following what proposed in the literature (Wallgren A. and Wallgren B., 2014), Istat system includes *base registers* and *satellite registers*. The former concern the so-called base populations, namely, individuals and households, productive units and territorial units. The latter shed light on specific domains (e.g. education and training, disability, labor, income) or extend the information on the base populations with additional variables (e.g. extended register on public administrations).

Istat integrated system of registers is developed and maintained by means of structured governance rules. Ontology models support linkability among the registers (Radini et al., 2018). Investments in methodology are being carried out in order to assure accuracy of the estimates (Alleva et al., 2019). The information is organized as to avoid redundancy – the same variable cannot be managed in different registers – and ensure coherence of the information.

The adoption of such a new production system entails several advantages, from a reduction of response burden, to an increase of the detail of the estimates, to the increase of information potentialities obtained by linking data.

The TRET will be conceptually and physically integrated to some of the other registers of the system. The objective is twofold: to complete the data (i.e. variables that are managed by other registers, missing or incoherent values) and to jointly analyze different phenomena.

In the TRET, basic information on the students will be drawn by the *Base Register on Individuals and Households*, i.e.: demographic characteristics, demographic events subtracting the individuals from the population of students (movements, deaths) and familiar relationships. The characteristics of private, public and non-profit institutions providing education and training will be drawn by the *Base Register on Productive Units*. Every variable containing an address in the TRET will be referenced with an *address's unique code* set by the *Base Register on Places and Addresses*.

The linkage of the TRET with other thematic registers will allow to analyze education and training together with the other phenomena, first of all, occupational status, work characteristics and income.

2.3 The impact of the TRET on the surveys on education and training

In the design of the new system of statistics on education and training, the statistical surveys will continue to have a relevant role. Data on education collected through the LFS and data from AES, both surveys being under European regulation, will be fundamental to produce the indicators on education and training comparable at international level, and to provide statistical information on work outcomes and non-formal and informal education. However, the implementation of the TRET will increase the possibility to jointly analyze survey data and register data and thus to widen the statistical information produced.

On the other side, the surveys on education-to-work transition will be redesigned. Information will be derived as much as possible from the Integrate System of Registers, in a systematic way and with a higher frequency respect to the past. Surveys on targeted information on work transition will be periodically conducted.

3. Methodological and quality issues in the development of the new statistical system on education and training

In the shift from independent statistical processes to an integrated system developed around the TRET, Istat is facing many quality and methodological issues. The next subsections reports some of the more relevant issues with reference to the input, throughput and output.

3.1 Quality of the input sources

Istat can rely on long, solid and collaborative relationships with the owners of the administrative data meant to feed the TRET. Overall, the quality of these sources is far more than acceptable. However, coverage error concerning some segments of the education and training system is present in the microdata, i.e.: pre-primary (0-3 year) care services; students attending vocational programs offered by the Italian Regions; students attending any level of the programs in Fine Arts, Drama, Dance and Music. As shown in Table 2.1.1, for some education segments only macrodata are available. However, the NRRP might foster the building of other microdata databases supporting its missions. In addition, data relative to some autonomous Italian regions (limited in size) are sparse and not harmonized with the national data. It is worthy to add that the phenomenon results to be quite dynamic, both at university and school levels, since there are frequent reforms changing the structure of the system and introducing new courses. Moreover, there is a high degree of autonomy in the definitions of education programs and courses. Istat, in collaboration with the relevant institutions, is working on the improvement of data quality and the development of a new classification on education and training programs and attainments, harmonized with the international classifications Isced and its extension with the field of education (Isced-F).

Istat statistical surveys on education and training have been carried out since many years and are supported with solid methodology. The LFS and AES will continue to be conducted according to the relative regulations, while pursuing integration with the register data to exploit all the available information reducing statistical burden. Different is the situation of the three surveys on education-to-work transition. As already mentioned they will be re-designed in the light of the new register, thus requiring an investment in methodology for the revision of the sampling design, the questionnaire and the production of the estimates on specific targeted information needs.

3.2 Process methodology and quality

In order to implement the TRET a number of procedures are being developed. They can be broadly listed as: harmonization of the input sources' data; quality controls on missing information and coherence of the data in the single input sources; micro-integration of the sources; control of coherence of the integrated data; derivation of the new statistical units for the register; macro-integration; establishing conceptual and functional relationships with other

registers' units and variables for processing and outputs production scopes; statistical estimation (descriptive statistics, model estimates, predictions).

In the TRET, anonymized data are used. Indeed, for each administrative source gathered by Istat, *base units* are identified and, if not already captured in previous sources, a new identification code is assigned; on the contrary, if already present in Istat archives the same identification code is attributed. This process is managed at central level at Istat and is not under the direct control of the managers of the TRET. The possibility to link the data using the *base units* as linkage key is ensured and facilitated by the existence of these identification codes. However, as any statistical process, such a procedure can be subjected to errors. These errors can have an impact on the capacity to correctly derive the statistical units of interest for the education and training register (education position, individual, productive unit). Therefore, while checking the coherence of the input data and building the statistical units, some coherence rules may be violated due to errors in the units' anonymization phase. The tests performed so far have shown that this situation seems to happen rarely, however it is planned to have a system of timely reporting of possible errors, for further assessment and resolution. It has to be underlined that changes in identification codes of *base units* have impact on all the registers in the Integrated System of Registers containing those units, thus caution has to be adopted.

The construction of the statistical units requires rules to manage missing information and incoherencies in the data, thus model and assumptions' errors might be introduced during this process step. Istat believes that, giving the nature of the education and training phenomenon, the use of longitudinal data will provide a wide informative framework allowing the correct construction of the units, even when yearly data are incomplete.

In the estimation phase, the most relevant methodological issues are related to macro-integration and model errors in predictions.

3.3 Output quality

The lack of microdata on the already mentioned segments of the education system, e.g. the vocational programs, which are very important in the current policies, might result in a loss of relevance of the statistics produced.

The production of estimates integrating data from different registers imposes a reflection on the coherence. Indeed, it is a general problem being faced within the Integrated System of Statistical Registers. Besides, Istat statistical production on education and training will need to ensure coherence of the output with statistics released by other producers.

The timeliness of the estimates reflects the scheduling of administrative sources' acquisition and Istat processing times for checking the data, apply anonymization procedures and loading the dataset in the internal system. Predictions approach will be used in order to have more timely statistical results.

4. Conclusions

The design and implementation of the new system of production of statistics on education and training, hinged around the *Thematic Register of Education and Training*, is a demanding project. However, its development can benefit from the wide experience gained so far at Istat in the setting of other statistical registers.

The release of the register will be incremental, so as to make available to the internal users modules of education segments before the whole database is completed.

The register together with the possibility of linking the information with survey data and data from other statistical registers will strongly increase to capacity to analyze the phenomenon in depth of education and training with all the socio-economic and contextual factors influencing it, also adopting a longitudinal perspective.

Extensive analyses will be possible by linking the data of the education register with that of the labor and income registers. Information on the familiar structure available in the base register on individuals and households will allow to relate the level of education of children with those of parents. Backward tracing will permit to explain work conditions in relation to the education paths and performance.

The new informative framework will allow to better define education and training policies and interventions and to be able to monitor their implementation and the real impact on the society.

References

Alleva, G., Falorsi P.D., Luzi O., Scannapieco M. (2019), "Building the Italian Integrated System of Statistical Registers: Methodological and Architectural Solutions" paper presented at the ESS Workshop on the use of administrative data and social statistics. Valencia, 4-5 June 2019, https://ec.europa.eu/eurostat/cros/system/files/building-italia-integrated-system_istat_0.pdf

Baldi, C., Ceccarelli C., Gigante S., Pacini S., Rossetti F. (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, pp 95-105.

Di Zio, M., Filippini R., Rocchetti G. (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, pp.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/IstatsModernistionProgramme_EN.pdf

Radini, R., Scannapieco M., Tosco L. (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", presented at ModernStats World Workshop 2022, Belgrade, Serbia, 27-29 June 2022 https://unece.org/sites/default/files/2022-07/MWW2022_Presentation_Italy_Simeoni.pdf

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