DEPP@scribe project: Better structuring and documenting education data & taking action for research and innovation

Axelle Charpentier[1], Alexis Lermite[2], Thierry Rocher[3], Ronan Vourc'h[4]

[1, 2, 3, 4] *Directorate of Evaluation, Forward-Planning and Performance-Monitoring, Ministry of National Education & Youth, France*

Abstract

This article presents DEPP@scribe, a project that was born out of a desire to better structuring and documenting French education data, which constitute one of the richest national statistical information systems on education in the world. This project led by the DEPP, the statistical department of the French Ministry of National Education & Youth, relies on a unique partnership between central administration and researchers. It involves a large-scale innovative IT project designed to meet the needs of the official statistical system, particularly with regard to data quality and accessibility issues, and those of the scientific community in terms of reducing data-collection costs and strengthening the impact of research in education. To date, DEPP@scribe has resulted in the publication of an online catalogue of DEPP data sets, displaying the broad range of available data to researchers who are less familiar with it and the many opportunities for study and research that they offer. A remote data access platform is currently being developed and will be tested shortly by users' committees. There are still a number of challenges to be overcome, including how to deal more effectively and quickly with the many requests for access to data from the scientific community and from institutional players producing studies and evaluations of the education system or how to facilitate and promote data linkage between education data and data from other areas of public policy. This experience as a whole potentially offers a source of inspiration for other national statistical offices that would be interested in the approach underpinning DEPP@scribe.

Keywords: education, administrative data, research support, statistical system

1. Introduction

Improved access to high-quality administrative data for researchers can dramatically reduce the cost of research and speed up the building of scientific knowledge to better inform policy-making. In this sense, it is an invaluable resource for public good. Administrative data for research serve multiple purposes: researchers can use them to find fields of observation, to build samples, to document short-term to long-term outcomes, and/or to contextualise results produced from specific samples. In the field of education, the availability of these data avoids the substantial costs for both researchers and schools associated with the collection of survey data. Fully aware of these challenges, the Directorate of Evaluation, Forward-Planning and Performance-Monitoring of the French Ministry of National Education & Youth (DEPP) has

embarked in an ambitious partnership with the research community in education to design and provide a sustainable and innovative set of research tools and resources.

With the financial support of the Innovation, Data, and Experiments in Education (IDEE) programme[1], the DEPP aims at providing transparent, coordinated, secure, timely, and last but not least, free access to administrative data and support services for a diverse community of researchers in education. However, before thinking about making data available, it is necessary to think about how it should be organised to maximise the potential offered by such data and address effectively the needs and demands of researchers as well as to help DEPP to carry out its priority tasks as a ministerial statistical service. The IDEE programme is meant to be absorbed gradually by the DEPP, under the name of the DEPP@scribe project.

The DEPP has had close links with the world of educational research for many years, but the research it supported until recently was not very diversified in terms of the subjects studied, the disciplines and the research teams and institutions carrying out the research projects. The recent increase in the quality and quantity of administrative data on education in France paired with the implementation of yearly exhaustive assessments (from Grade 1 to Grade 10 next year) has the potential to increase and diversify empirical scientific work. For this purpose, additional resources must be allocated to document all the data sets of interest to researchers (to this day, over a thousand) but also to facilitate access to them, particularly when an enrichment of survey data collected by researchers is needed.

The rest of the paper is structured as follows. Section 2 discusses the project's origins and the needs it meets, Section 3 describes the organisation of the DEPP@scribe project, the progress already made and the challenges ahead, and Section 4 concludes.

2. The project's origins: quality education data offering great potential for research, but difficult to access

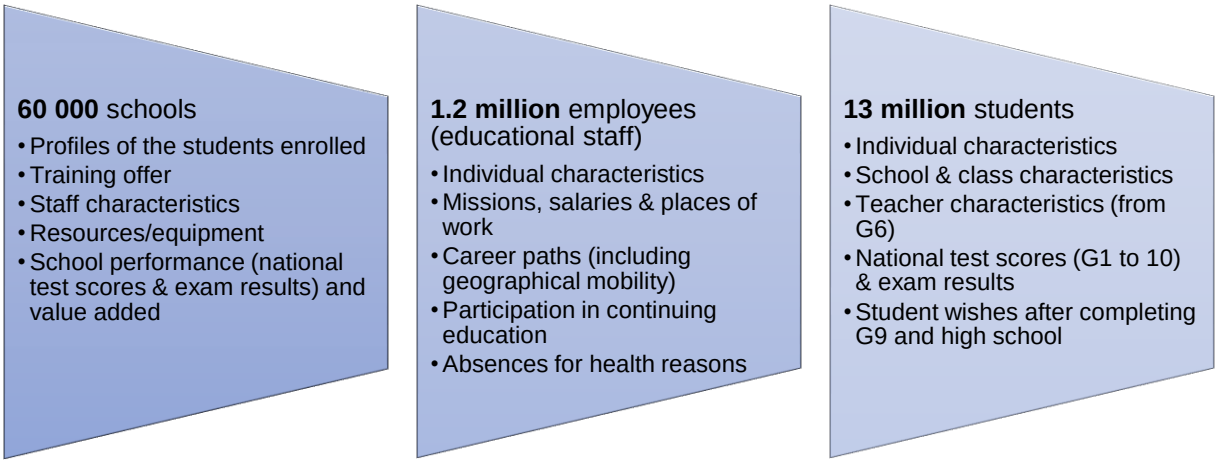
2.1 Overview of the French statistical information system on education

As described by Figure 1, the French national statistical information system on education is extremely broad and offers a myriad of opportunities for study and research on schools, educational staff and students. Available administrative data on primary and secondary education in France cover every school (N≈60,000), employee (N≈1,200,000) and student (N≈13,000,000). For instance, for all pupils starting Grade 1 in September 2024, the DEPP can provide data relating to their test scores in the annual back-to-school national assessments in French and Mathematics, those of their classmates, and data describing the primary school

[1] https://www.idee-education.fr/

they attend (students' profiles, staff characteristics, resources, equipment, various performance indicators, etc.). Once they reach Grade 6, in addition to their gender and age already documented in previous grades, existing data sets contain precise information on the socioeconomic status of their family (parents' occupation, financial aid recipient), students' country of birth (and department and city for French students) and their geolocated home addresses. The statistical information system also allows to identify and describe the secondary school teachers responsible for the students in the various subjects. In other words, data on schools, teachers and students can all be linked thanks to unique identifiers for the various statistical units. Lastly, administrative data enable tracking courses and options taken in secondary education, participation in special educational arrangements, student wishes after completing Grade 9 and high school, and results at national exams at the end of middle school and high school. Higher education data can also be easily obtained to track educational pathways after secondary school.

Figure 1: Overview of comprehensive data on primary and secondary education in France



Despite the richness of administrative data available, it does not cover all the subjects of interest for reflecting on and evaluating the direction of education policies. As a result, the DEPP also conducts periodic surveys based on representative national samples on a variety of topics, such as school climate, the violence felt by pupils and school staff, the well-being and working conditions of staff or teaching practices. To build this type of survey, the DEPP co-constructs questionnaires with different actors: we set up partnerships with researchers from different disciplines (educational sciences, sociology, economics, psychology and didactics) and we involved teachers, teachers' trainers and other key players in the field of education to draft appropriate questions. The DEPP also takes part in numerous international

surveys that enable to compare our education system as a whole[2], the skills of our pupils[3] and the teaching profession[4] with what can be observed in other countries in Europe and elsewhere.

2.2 Issues of access to and use of administrative data for research purposes

Some of the data described above are collected directly by the DEPP, notably national back-to-school assessments and most survey data. However, most administrative data are reprocessed from “management files”, which serve initial purposes other than the production of public statistics. Their re-use for statistical purposes and making data available to researchers involves substantial statistical processing time for the DEPP teams. As analysts and producers of official statistics know all too well administrative data are anything but research-ready data. In many cases, the source administrative data cannot be used as such for statistical purposes (Cotton & Haag, 2023). When they are deemed to be of sufficient quality to produce official statistics, administrative data sources must be integrated into the statistical information system by drawing on existing raw data and metadata from administrative systems and registers developed for operating purposes. To do this, statisticians must be able to exchange with data producers in order to check several points, especially that data administrative sources can be restructured in to statistical units to measure statistical concepts, that they are complete with regard to the target population and that they are documented.

According to Mc Grath-Lone & al. (2022), research-ready administrative data are defined by five key characteristics: accessible, broad, curated, documented and enhanced for research purposes. Data users need to learn about and understand what the variables measure and what they do not, how and when each variable is collected, population scope covered by data, any recoding that could have been implemented, potential changes in definitions or any other parameters across time, etc. The various data sets must be provided with identifiers or variables (e.g. student, class, teacher or school identifier) enabling the necessary linkages to be made. In addition, data pseudonymisation procedures must be able to be implemented to respect the confidentiality of personal data while allowing the linkage of different data sources, including survey data directly collected by researchers. Lastly, there is a need for clear and transparent procedures and prerequisites for obtaining authorisations to access administrative data for research purposes, and these must be communicated to the scientific community.

---

[2] Eurydice (https://eurydice.eacea.ec.europa.eu/).

[3] Programme for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMMS), International Civic and Citizenship Education Study (ICCS).

[4] Teaching and Learning International Survey (TALIS).

Until recently, only a small number of researchers used education data held in the DEPP databases. In 2019, for example, there were around 20 data access agreements for researchers from a dozen institutions (most of them being economists from leading Universities in Paris). Many researchers thought that you had to be an “insider” to be able to access and use this data, mainly due to inadequate data documentation and communication on the datasets available to researchers. There was also the question of physical access to the data, since even today it is necessary to go to the DEPP's office in Paris in order to use the data securely.

2.3 IDEE: A mutually beneficial partnership between central administration and researchers

To advance evidence generation and use in the French education system, J-PAL Europe[5] launched the IDEE programme in 2022, funded for a duration of 8 years by the French National Research Agency under the program *Investissements d’avenir*. IDEE is a long-term investment to develop the infrastructure for experimental research in the French education system and to ensure that scientific knowledge informs policy-making. It is based on three main strands of activity: improving access to administrative data (axis 1), developing and sharing innovative tools and protocols for experimental research (axis 2), and building partnerships and strengthening capacities for experimental research (axis 3). To implement the first priority of the IDEE programme, a partnership was set up with the DEPP from the outset, as this objective coincided with the European quality commitments incumbent on the Official Statistical Service, namely to have documented, optimised and secure processes and to provide better access to official statistical data.

In recent years in France, public policy evaluation has aroused growing interest from public institutions, decision-makers, researchers and civil society in general (Baïz, 2022). This has resulted in considerable efforts on the part of the DEPP to produce new data useful for the production of studies on the education system, and new requests to use these data for steering and evaluation purposes. Faced with an increasingly rich statistical information system and the myriad of opportunities it offers, the DEPP has launched DEPP@scribe, a project aiming to better structuring and documenting French education data and ultimately to promote educational research. To this end, the DEPP is receiving substantial but temporary financial support from IDEE (at 31 December 2023, this represented a total amount of €184,000 in staff costs and €152,000 in (mainly IT) services).

[5] One of the seven regional offices of J-PAL (MIT), a global research centre founded by Abhijit Banerjee and Esther Duflo and focused on conducting randomised impact evaluations to answer critical questions in the fight against poverty.

3. The DEPP@scribe project

3.1 Objectives and implementation arrangements

DEPP@scribe has several objectives:

- Increasing the value of data and make better use of statistical production by facilitating access to the richness of data produced by DEPP's units;
- Meeting quality requirements in line with European statistical best practices and standards;
- Encouraging studies on education carried out by the DEPP and other players such as researchers thanks to easier access to documented data;
- Responding more effectively to requests from the scientific community and thus save time for the production of studies by DEPP's teams;
- Helping to inform public debate on education and political decision-making.

To achieve these objectives, the project plans to develop three new digital tools which will be integrated into a single portal:

1. *Depp@logue*, the DEPP's catalogue of education data, to showcase the richness of statistical information systems, provide a better understanding of study and research opportunities, and build innovative and relevant study protocols;
2. *Depp@thèque*, the DEPP's education data library, a digital platform for accessing data and carrying out studies based on the connection of administrative sources, or even on the enrichment of survey data collected independently;
3. An application for creating and monitoring legal agreements governing data access for researchers.

The project's most emblematic tool is the platform, which can be seen as a data warehouse equipped with calculation servers, providing secure remote access to data and allowing different confidentiality rules to be set. The modularity of the data warehouse is a key aspect of the project since statistical secrecy and data confidentiality may vary depending on the recipient/user of the data, the purpose of the data processing and the nature of the data made available (Figure 2). Access to it will be free of charge and open to any research teams that have signed an agreement with the DEPP.

Project governance was set up to ensure that the tools developed were relevant to the needs identified by DEPP's teams (statisticians, IT specialists, quality experts, etc.) and representatives of the scientific community. In the end, better structuring and documentation of DEPP datasets should benefit all staff involved in producing, storing and making available data and carrying out studies (and not just external users of the data, such as researchers). The DEPP is increasingly called upon to carry out multidimensional impact evaluations of education policies. To meet these growing demands, in 2023 the DEPP created an office

dedicated to this purpose. In this respect, detailed knowledge of DEPP data sets is necessary in order to build relevant evaluation protocols[6].

Figure 2: Types of data made available and related confidentiality rules

3.2 Progress on the project & Future challenges

Two years after the launch of DEPP@scribe, the results are very positive, both in terms of the quality of the working relationship between teams from IDEE and the DEPP and the progress made in achieving the project's objectives. In the partnership's first year, the teams from both partners carried out investigative work to further refine previously identified needs, while DEPP's IT team began to design technical solutions for setting up a secure remote data warehouse and what this would mean for future IT development. In November 2022, the DEPP published its online data catalogue[7], the content of which is gradually being enhanced. In the short term, we will continue to handle requests from researchers for them to access data in an ad hoc manner. These requests are becoming more complex and are increasing sharply as a result of greater awareness of all the study and research opportunities offered by DEPP data. We need to be careful and communicate with researchers about the processing times, which may seem long and contradict the communication from the DEPP and its partner IDEE on the objectives and expected benefits of DEPP@scribe.

Committees made up of researchers, staff from IDEE and the DEPP have been set up to ensure the digital tools developed through the partnership meet the needs of all stakeholders. They met frequently throughout 2023 to assess and provide feedback on the first mock-ups of digital tools. A beta version of *Depp@thèque* should be available before the summer of 2024,

[6] Vourc'h et al. (2022) provides an illustration for such evaluation protocol.
[7] https://catalogue.depp.education.fr

and it will be tested collaboratively with research partners to further inform its development. The platform should be launched and operational by 2025.

Implementing DEPP@scribe is not without its challenges, the first being a financial one, since IT development of the tools and resources developed as part of the DEPP@scribe project was made possible by temporary funding from IDEE. The DEPP needs to find ways of internalising the costs associated with the maintenance and future development of the project, which is one of the reasons why it was essential to align the objectives of the IDEE programme with those of the DEPP's missions. To date, the central administration has recruited one of the IDEE developers and is in the process of hiring for the permanent positions of data engineer and contract support, which until now have been paid by IDEE.

Another major challenge for the DEPP is to maintain the virtuous ecosystem that has existed for many years between the DEPP and the scientific community, and which is essential to enable us to carry out our missions of leading educational research in France and disseminating the results of this research to public decision-makers (Charpentier, 2023). With the development of new digital tools to facilitate access to data for researchers, we run the risk of becoming a mere data access point, or to put it another way, a (free) data supermarket. On the contrary, with DEPP@scribe, we want to develop a community of data users on the DEPP side and on the research side, and equip them with collaborative tools to facilitate the use of data and, why not, develop study and research collaborations.

When it comes to studying the performance of the education system, data needs not to stop at the variables available in the data sets provided by DEPP. Many research projects (particularly evaluations of educational policies/interventions) seek to track the educational (including higher education) and professional trajectories of individuals. For researchers, it can take considerable time to obtain authorizations to access administrative data from multiple statistical departments. What remains to be done is to organize and provide a legal framework for the matching of data from various sources, so that researchers have clear and operational procedures at their disposal when building their research projects. The DEPP is quite proactive in this area, and, for that matter, organized a seminar on quality in official statistics, in March 2024, bringing together various ministerial statistical departments (Justice, Interior, Higher Education, Social Affairs, etc.) and researchers to consider ways of facilitating such data linkages.

4. Conclusions

French education data are rich and offer a wide range of possibilities for research and the evaluation of public policies. In response to growing demand from researchers in recent years,

DEPP has embarked on an innovative process aimed at improving access to documented data. This initiative is part of a structuring approach for the DEPP around the DEPP@scribe project: improvement of the research contracting system, generalization of data documentation, creation of secure online workspaces for researchers.

The work undertaken has led to a continuous improvement in the quality of data availability processes. However, a number of challenges remains. One of the main ones concerns the articulation between data produced in the field of education and those produced by other ministerial statistical services. This work is fundamental, and must serve a major objective: that of supporting public policy through research structured around reliable and documented data.

References

Baïz, A. (2022). « Quelles évaluations de politiques publiques pour quelles utilisations ? », Report, France Stratégie. https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/fs-2022-rapport-bilan_des_epp-juin_2.pdf

Charpentier, A. (2023). « La contribution scientifique de la DEPP à l'évaluation des politiques publiques », *Administration & Education*, n°178. AFAE.

Cole, S., Dhaliwal, I., Sautmann, A. & Vilhuber, L. (Eds). (2020). *Handbook on Using Administrative Data for Research and Evidence-based Policy*. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab.

Cotton, F. & Haag, O. (2023). "L'intégration des données administratives dans un processus statistique - Industrialiser une phase essentielle". *Courrier des statistiques*, n°9. Insee.

Mc Grath-Lone, L., Jay, M. A., Blackburn, R., Gordon, E., Zylbersztejn, A., Wijlaars, L., & Gilbert, R. (2022). „What makes administrative data "research-ready"? A systematic review andthematic analysis of published literature". *International Journal of Population Data Science* 7:1:6. https://doi.org/10.23889/ijpds.v7i1.1718

Vourc'h, R., Charpentier, A., Murat, F. & Rocher, T. (2022) « Élaboration et mise en œuvre d'un dispositif ad hoc d'évaluation de politique publique : le cas concret de l'évaluation de la politique de dédoublement des classes de CP et de CE1 en zones d'éducation prioritaire (2017-2021) », 14es Journées de méthodologie statistique de l'Insee. Insee. http://jms-insee.fr/2022/S22_1_ACTE_VOURCH_JMS2022.pdf