# Longitudinal Education Outcomes dataset

## *Key messages from stakeholders*

**March 2023**

Economic and Social Research Council

ADR ENGLAND

# Abstract

**This report records a roundtable discussion between third-sector stakeholders and Longitudinal Education Outcomes (LEO) project partners on the potential and use of the current iteration of the LEO Iteration 1 Standard Extract dataset. It captures the main findings of the discussion with the intention to inform the future use of the dataset, including research priorities, the scope for growth of the dataset, and how to facilitate the voices of the public in shaping the direction of research, and more.**

**Areas of research focus identified by stakeholders include accounting for the experiences of individuals missing from key datasets and a greater understanding of the impacts of adverse childhood experiences. These ideas will inform research priorities for the current iteration of the dataset.**

**This meeting was held on Friday 16 September 2022. A full list of attendees and areas for future research can be found in Appendix 1.**

# 1. Introduction

Collaboration between stakeholder groups is essential to ensure that research using public sector data is truly in the interests of those it hopes to benefit. Stakeholders in this context include all the intended users of research findings (including policymakers, service providers and public advocacy groups). All of these groups have an interest in knowing whether the focus and methods of research using public sector data are ethical, robust and useful, and that any potential negative consequences have been considered and mitigated.

On 16 September 2022, a group of relevant stakeholders met to discuss the current iteration (Iteration 1 Standard Extract) of the English Longitudinal Education Outcomes (LEO) dataset, a newly linked data resource for research in the public interest (see Box 1). The purpose of the meeting was for stakeholders to find out about the data that is being linked and to discuss the potential of this data resource. The discussion covered:

- details of the current LEO dataset and plans for future iterations of LEO
- priority research areas for current and future iterations of the LEO dataset
- what, if any, are the ethical considerations of this data and how to mitigate them where possible.

Stakeholders included members from third sector organisations working with, or on the behalf of, disadvantaged or vulnerable children, some with a focus on health, social care and/or education. See Appendix 1 for a list of attendees, including stakeholders, the LEO project partners and those from ADR UK. This report captures the key themes arising from discussions at the stakeholder meeting.

***Box 1: Overview of the Longitudinal Education Outcomes linked dataset***

The LEO Iteration 1 Standard Extract dataset links individuals' de-identified education data with de-identified data on employment, earnings (including self-employment income) and benefits. LEO includes the following datasets:

- **National Pupil Database** - which includes data on pupils attending state schools and national achievement test results for all pupils who sat them.
- **Individualised Learner Records** – which includes data on individuals in further education and apprenticeships.
- **Higher Education Statistics Agency data** – which includes data on individuals in higher education.
- **Data from HMRC and DWP** – which include data on employment, earnings and benefits.

The de-identified LEO linked dataset is now available to external researchers via the Office for National Statistics (ONS) Secure Research Service. Analysis of LEO can facilitate research on the longer-term labour market outcomes at person level. This analysis will in turn, allow the evaluation and improvement of education policy and provision. Researchers need to be accredited and submit a successful application outlining the public benefit of their work to access the data.

# 2. The research value of the LEO dataset as identified by the LEO project partners

LEO is comprised of four data sources: the National Pupil Database, Individual Learner Records, data from the Higher Education Statistics Agency, and data from HM Revenue and Customs (HMRC) and Department for Work and Pensions (DWP), as shown in Box 1. The dataset contains records for roughly 38 million individuals focusing on children and young people who have attended public education settings in England or have taken public exams (e.g. GCSEs or A-levels) in privately funded institutions.

The data provides de-identified information on which educational institutions they attended, their qualifications and attainment, attendance (in schools), and characteristic information (gender, ethnicity, etc.). When this data is linked to de-identified data from HMRC and DWP on employment, earnings and benefits, it provides national coverage for individuals from primary school to their early thirties. There are LEO datasets for each of the four nations. Figure 1 shows a summary of the datasets within LEO, detailing the period from which the data extracts begin. Coverage of LEO will expand as the data is refreshed with timelier records, therefore end dates have not been provided.

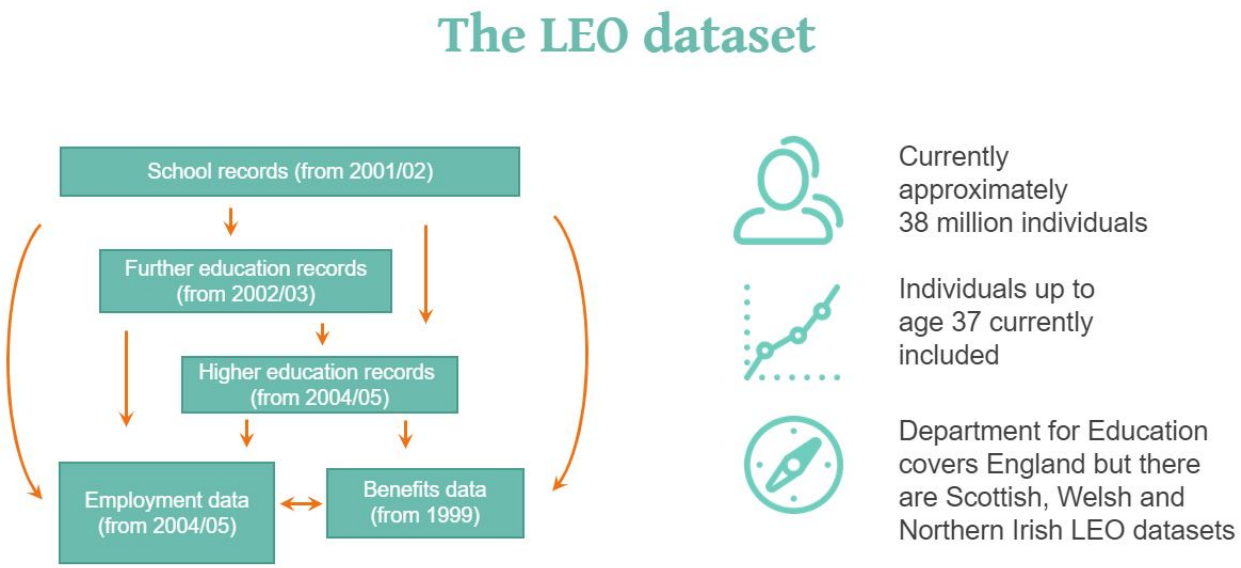


*Figure 1: A summary of the LEO dataset.*

## 2.1 *Why LEO was created*

Education institutions generate a vast range of education data, from the records of pupil registration to their attendance, attainment, and more. Much of this information is in the public record, and available via the Department for Education (DfE) in England. However, without the LEO dataset, analysis is limited to understanding the determinants of educational attainment and other outcomes in the education system. The LEO dataset enhances this analysis with data about individuals once they have left education. This enables analysis of the impact of education and education policies and how they affect labour market outcomes, including the likelihood of being in work and how much an individual might earn. Therefore, the LEO dataset enables researchers to generate transformational insights about pathways within and after education. Analysing this data at a population-level strengthens the evidence available for developing better education policy and practice. Better policy and practice will in turn enhance the life chances of current and future learners.

# 3. Key messages from stakeholders

As LEO project partners delivered a presentation on the research value of the dataset, they invited discussion from stakeholders on priority areas of research and any ethical considerations arising from the availability and use of the data.

## *3.1 Representation (and missing data) in the dataset*

Missing data is not unusual in data collections. In administrative data, this can arise because of the populations on which data are collected, meaning that not all individuals may be represented, or may not have data included at all points in time. Stakeholders were interested in several issues related to representation in the LEO dataset:

- whether a complete education record was needed for individuals to be represented in LEO
- whether individuals would still be represented in the dataset during periods of exclusion or if they were to leave education before completion of qualifications
- how to account for underrepresented groups, including the need to understand their concerns, account for their experience and articulate their needs in evidence-based decision-making.

The LEO dataset re-uses administrative data to draw statistical insights. LEO is intended to be a full-population dataset comprised of data that was originally collected for operational purposes within education, employment and for the purposes of administering the tax and benefit system. Therefore, it includes the full population of individuals who have engaged with public services which contribute to the data within the LEO dataset. This also means the LEO dataset has some limitations in representation of some individuals. For example, if someone were to move from one of the devolved nations to England during primary or secondary education, they would begin to appear in LEO from the moment they engage with the relevant public services in England. Similarly, if an individual migrated to England from outside the UK and did not earn enough to pay tax and did not claim benefits, they would not be represented in the data within LEO.

The UK does not have national identification of individuals. This is in contrast to some other countries, such as Norway, in which there is a system enabling the tracking of individuals across public services over time. It is for this reason that the analysts who produce the data must match individuals across different administrative data sources using a range of information, such as name, address, date of birth and gender. These identifiable data are only accessible to the very small number of individuals involved in the matching process. Once the matching has been done, the identifiable characteristics are removed, and the vast majority of individuals accessing the data – including all external researchers – only have access to de-identified pseudonymous data.

Match-rates have been considered for LEO to ensure accuracy. This means that when the data is matched, if LEO analysts are not confident they are seeing the same individual across multiple data sources, then that individual will be excluded from the dataset that is shared with external researchers. In practice, to date, this type of incidence has been comparatively rare and LEO match-rates across the constituent datasets are currently strong. LEO project partners feel

confident that the number of people missing from the data who are represented in at least one of the constituent datasets is low, both at a societal and local level. But as coverage is paramount, LEO project partners will work alongside DfE to ensure the dataset is as representative and accurate as possible.

A stakeholder from the Traveller Movement highlighted the low educational attainment rate and high dropout rate between primary and secondary school for children and young people within the Gypsy, Roma and Traveller communities. This means they fall out of DfE datasets when they withdraw from formal education. It was asked whether LEO is able to account for the experiences of these children and young people in relation to the linkage with DWP, HMRC and Census data. LEO project partners confirmed that pupils who appear at least once in education record should appear in LEO. They are also represented via their tax and benefit records if they have engaged with those services.

## *3.2 Enriching LEO*

Stakeholders identified particular information as important to enrich the analysis of LEO, including the need for data on:

- school exclusions
- ethnicity
- Special education needs or disability (SEND)
- provision of education and healthcare plans (EHCPs)
- experiences of care and eventually involving crime
- crime and justice data from the Ministry of Justice.

LEO project partners confirmed the current iteration of LEO contains information on school exclusions, ethnicity, SEND status, provision of EHCPs and experiences of care. It is hoped that future iterations of LEO will include crime and justice data (see Figure 2). LEO project partners confirmed the dataset already includes data from the National Client Caseload Information System (NCCIS), a survey annually returned from local authorities which can account for the status of every individual who has left education in the last year. This can help fill in gaps for individuals who aren't present in data on education, employment or training, e.g. those who are homeschooled.

Central to the potential public benefit of the LEO dataset are insights into individual longitudinal outcomes such as employment and wages. A stakeholder from Youth Futures Foundation noted the value in understanding the outcomes of those who experienced government-funded schemes, such as the Kickstart Scheme. There are many government-funded schemes available to individuals, some with similar aims and others that shut down and are launched again over time. A greater understanding of their long-term benefit can help other organisations determine where gaps may lie and what services might be needed to fill them to better work for young people. While this isn't an area currently available to analyse in the present iteration of LEO, onward linkages may give rise to more insights in this area.

Currently, the LEO dataset is being developed in four key ways:

- improving accredited researcher access to the data
- enabling analysis by making more research funding available
- developing and growing a community and knowledge base of accredited researchers
- via new data linkages, as summarized below in Figure 2.

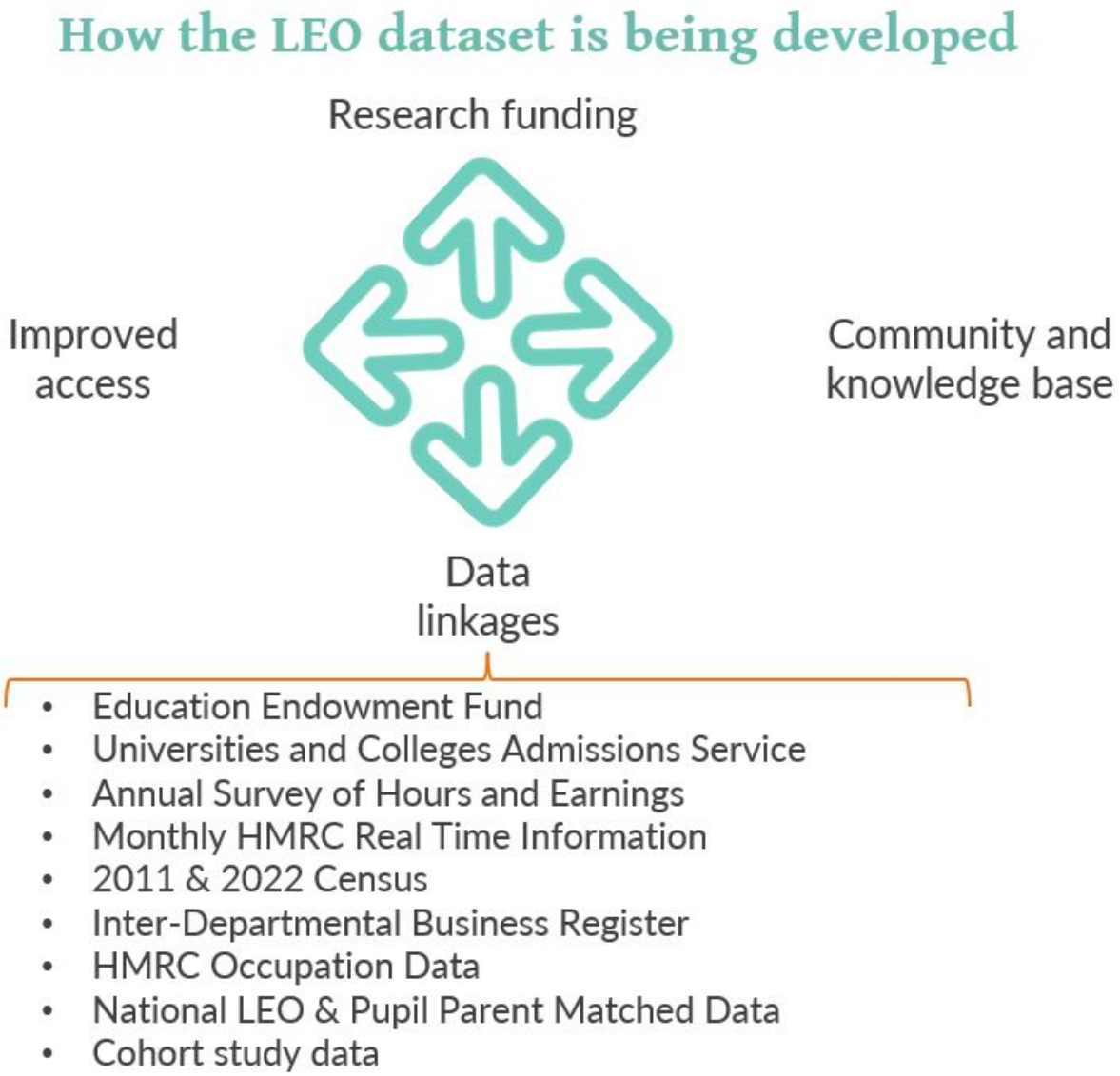


*Figure 2: A summary of how the LEO dataset is being developed.*

LEO project partners aim to enhance the LEO dataset with the addition of new data, as detailed in Figure 2. Below is some information about the sources of the new data and what is to be included in future iterations of LEO:

- The Education Endowment Foundation (EEF) is a non-governmental organization that trials interventions aimed at understanding what works to support the educational attainment and wider outcomes of disadvantaged pupils. The inclusion of this data will allow trial data to be linked to later education and labour market outcomes, to understand whether the benefits of these interventions are long-term. and outcomes
- The Universities and College Admissions Service (UCAS) is administrative data providing information on undergraduate applications from across the UK.
- Survey data from the Annual Survey of Hours and Earnings (ASHE) run by the Office for National Statistics will provide more information on the occupation and hours worked by individuals represented in the survey.
- Real Time Information from HMRC will provide monthly information, rather than annual information, on an individual's income.
- The addition of the 2011 and 2021 Census will enable analysis of an individual's household composition and demographic characteristics in both childhood and adulthood.
- Data the Inter-Departmental Business Register (IDBR) will enable individuals to be located within businesses. It can identify those who work together and can enable analysis of links between education and training and business outcomes.
- HMRC Industry data indicates the broad sector that an individual works in, while occupation data – if it were to be collected (it is currently being consulted on) - will allow analysis of which jobs individuals are doing.
- The Pupil Parent Matched Data (PPMD) links parents' tax records to their children's education outcomes. Linking PPMD to LEO will enable much richer analysis of the links between the circumstances in which individuals grew up and their circumstances later in life, providing a detailed picture of social mobility both at national and local level.
- Cohort study data is rich longitudinal survey data on a group of individuals, often born around the same time. Examples include the Millennium Cohort Study, which follows a cohort of children born in the academic year 2000-01. The inclusion of cohort study data will enable a richer understanding of the mechanisms underlying patterns observed in administrative data. For example, it may shed light on the reasons why particular groups underperform in education or the labour market.

The addition of these new data sources can develop the LEO database as a resource that can generate insights in line with needs identified by stakeholders and more.

## *3.3 Ethical consideration concerning LEO*

A representative from the Magpie Project, a charity that supports mothers and under-fives in temporary accommodation, flagged those individuals subject to the policies of the 'hostile environment', whose parents do not have recourse to public funds, can experience unique levels of destitution and interrupted education which can affect their life outcomes.

She noted the value the LEO dataset can have in highlighting the impacts of adverse childhood experiences linked to insecure immigration status, homelessness, or destitution. However, she cautioned that findings must be contextualised, to avoid information simply recording poor outcomes - without causes or explanations - and then being used against certain communities.

## 4. Next steps

The feedback from this discussion is being carefully considered by LEO partners and will inform the scope of ADR UK funding opportunities related to the LEO dataset.


## Acknowledgements

The Longitudinal Education Outcomes dataset is a collaboration between the Department for Education and the Office for National Statistics (ONS). UK is funded by the Economic and Social Research Council (part of UK Research and Innovation).



## Author

Shayda Kashef, Senior Public Engagement Manager, ADR UK Strategic Hub, shayda.kashef@esrc.ukri.org.


**Visit the ADR UK website.**

@ADR_UK


# Appendix 1: List of attendees, ordered alphabetically by organisation

***Chair:*** Shayda Kashef, Senior Public Engagement Manager, ADR UK Strategic Hub

| ***LEO project partners*** | **Representatives** |
|---|---|
| ADR UK | Emily Oliver, Head of Training & Capacity Building |
| ADR UK | Dr Saba Mir, Senior Manager – Strategic Research & Capacity Building |
| ADR UK | Gregory Meredith, Senior Research Manager |
| ADR UK | Shayda Kashef, Senior Public Engagement Manager |
| Department for Education | Dave Burnett, LEO Programme Manager |
| Department for Education | Phillipa Norgrove, LEO Programme and Project Support Manager |
| Department for Education | Dr Alaster Smith, Head of Research Evidence and Engagement |
| University College London | Claire Crawford, Associate Professor of Economics |
| ***External stakeholders*** | **Representatives** |
| British Association of Social Workers | Dr Luke Geoghegan, Head of Policy and Research |
| Edge Foundation | Kat Emms, Senior Researcher |
| The Magpie Project | Jane Williams, Founder and CEO |
| National Foundation for Education Research | Jenna Julius, Senior Economist |
| NSPCC | Pam Miller, Head of Research |
| TASO | Rain Sherlock, Evaluation Manager |
| The Traveller Movement | Greg Sproston, Policy and Campaigns Manager |
| Youth Futures Foundation | Emily Preston-Jones, Impact and Evidence Manager |