Written evidence from UK Research and Innovation (UKRI)[1] (TEB11)

Public Administration and Constitutional Affairs Select Committee
Transforming the UK's Evidence Base inquiry

Introduction

1. UK Research and Innovation (UKRI) welcomes the Committee's inquiry into this important topic. Our evidence submission has been drafted by the Economic and Social Research Council (ESRC), with contributions from the Natural Environment Research Council (NERC) and our other Research Councils. UKRI's work is relevant to the scope of the inquiry in several ways:
2. ESRC funds world-leading research, data and postgraduate training in the economic, behavioural, social and data sciences to understand people and the world around us. Funding includes support for data infrastructures centred on both traditional and newer forms of data covered by this Call.
3. ESRC's primary investments relevant to the UK's policy evidence base are the Administrative Data Research UK (ADR UK) and Smart Data Research UK (SDR UK) programmes, along with a range of social survey investments, the largest of which is Understanding Society (see Box 1).
4. ESRC also funds many researchers who produce policy-relevant work based on data produced by Government, including the Institute for Fiscal Studies and social scientists at all the UK's leading research universities.
5. Environmental data also provide useful insights that inform decisions. NERC funds world-leading research, data and postgraduate training on environmental science which helps us sustain and benefit from our natural resources, predict and respond to natural hazards, and improve our understanding of environmental change.
6. NERC data is openly available to non-researchers and NERC is working to augment the exploitation of environmental and other data through its investment in the Digital Solutions Programme.
7. UKRI also funds other data infrastructures - such as Health Data Research UK (HDR UK), which was pivotal in accelerating population-level digital health research during the

[1] Launched in April 2018, UKRI is a non-departmental public body sponsored by the Department for Science, Innovation and Technology (DSIT).
UKRI is nine councils working individually and collectively across all disciplines and sectors. We use our broad reach and deep expertise to support a dynamic investment portfolio with aligned incentives. UKRI work with our stakeholders to understand the opportunities and requirements of all the different parts of the research and innovation landscape, supporting a research and innovation system in which people and ideas thrive, to which everyone can contribute and from which everyone benefits.

pandemic and continues to help leverage the UK's unrivalled public health data from primary and secondary healthcare.

8. As UKRI expenditure represents a large part of public R&D spend, the UKRI submission to the ONS to compile statistics on Research and development expenditure by the UK government is important. This provides a national, high-level estimate of research expenditure, which is then included in OECD-level international analyses. UKRI also regularly respond to requests for more detailed analysis of spend relevant to particular industry sectors, research challenges, or other topics. However, it is worth noting that apart from the analyses that UKRI publish, sometimes in partnership with other public and charity funders of research (for example, www.hrcsonline.net) there is no systematic collation and openly accessible dataset of project-level R&D spend data.
9. In this submission we focus on questions relating to the ethical and secure use of government-derived data (administrative, survey and Census) and new data sources (smart data) for public good research.

Summary

10. There is an **opportunity** for Government to make a major investment in the use and accessibility of data, to deliver economic and public policy benefits, especially in the context of emerging applications of AI. This would build on major investments already underway by UKRI (such as ADR UK and SDR UK) and by the Office for National Statistics (notably the Integrated Data Service, IDS). While good progress is being made in anonymising and linking a range of administrative and smart datasets through these programmes, we are currently lacking in investment to develop AI and analytic skills among researchers and data scientists to make best use of this data. A data moonshot of this sort would form a valuable part of technology stack relating to AI, and would help maintain and develop the global position of British technology firms and researchers.
11. Through programmes such as ADR UK, UKRI has **demonstrated** that partnerships between government (UK and devolved) and academia result in more research-ready, linked data being available to both government analysts and the wider research community, leading to more data informed decision-making.
12. A **key gap** identified by UKRI for the Committee to explore is access to datasets that exist within government departments, but which departments do not consider to be a priority for research use, over and above their use in official statistics. If these datasets were accessible to researchers, government analysts and policymakers would be able to benefit from the extensive talent and skills base that exists within the external research community to provide policy-relevant advice, a significant addition to the inevitably constrained resources of government analysts.
13. UKRI **recommends** that:

a. Government departments maximise the value of the limited resource they have available to invest in the use of their data by working in partnership with UKRI, to open secure access to this data to the external research community to deliver wider, policy-relevant benefits.
b. Government departments also work collaboratively with the external research community to understand how to maximise the value of all new forms of data, for both statistics and research, and carry out collaborative public engagement activities to maintain the social contract to use this data.
c. Consideration be given to investing in a major data moonshot, in collaboration with UKRI and the ONS, to provide the basis for new applications of AI and the provision of high-quality policy-relevant advice. Such an investment could build on the success of existing data investments, to make best use of this data for economic and public good.

Box 1: A summary of ESRC's primary data investments relevant to the UK's policy evidence base

In 2018, ESRC launched Administrative Data Research UK (ADR UK). This £149m programme opens up access to administrative data to researchers and analysts within and outside Whitehall to enable improvement of policy and public services. ONS is the major data infrastructure partner within the ADR UK programme.

Through ADR UK, ESRC also works directly with UK government departments and devolved nations to support data sharing activities, where these result in government administrative datasets being made accessible through secure and accredited trusted research environments (TREs) for research use. As part of this collaboration with Government, ESRC also supports:

- public engagement activities related to the use of administrative data for research, to ensure the social contract to carry out research using this data is maintained, and

- training and capacity building activities, to ensure to social and economic research community can benefit from the wide range of government data sources being opened up for research use.

In 2022, ESRC launched Smart Data Research UK (formerly known as 'Digital Footprints'). This £59m programme will deliver a step change in the research use of new forms of data generated through everyday interactions with digital devices – including mobile apps, digital marketplaces, social media, wearables, satnavs, sensors, and smart technology. Such data, sometimes referred to as 'digital footprints data', can provide policymakers with new insights about major social challenges such as energy security, the cost of living, climate change, emergency response, health, wellbeing and social mobility. However, there are considerable commercial, methodological, and technical challenges in using such data for research. Smart Data Research UK will build on the strong foundations laid by previous investments in this field to deliver a step change in the use of smart data for research and innovation.

New forms of data are often valuable in their own right, but their analytical power is increased when researchers combine smart data insights with those from traditional sources including administrative data. Smart Data Research UK is informed by, and will work to promote interoperability with, other forms of data served by UKRI investments.

Additionally, ESRC funds a wide range of social surveys, which create research ready datasets that are then utilised by the wider research community for public good research. ESRC also funds researchers to utilise survey data collected by other organisations, including Census data from ONS (England and Wales), National Records of Scotland (Scotland) and NISRA (Northern Ireland) from across the decades, including Census Longitudinal Study data, which takes a sample of 1971 Census data and links it to birth, death and cancer registration data, as well as subsequent Census data.

Section 1: Data analysis in government:

How are official statistics and analysis currently produced?

14. UKRI is part way through an extensive renewal and transformation of its operational systems, as was commented on by the recent independent review of UKRI. This work will fully and effectively join up data management across the councils brought together in 2018 to form UKRI (which previously all had different approaches to the capture, retention and analysis of their management information), will put data and evidence at the centre of strategy development, and introduce simpler and better processes for UKRI and its stakeholders. As part of this work UKRI intends to establish a single professional data management function under the leadership of a new Chief Information Officer, that will drive improvements in the maturity of data management including a single data warehouse, effective governance, robust data quality processes, and a range of services including cutting edge methods (including the use of AI approaches) for the classification of research projects and research outputs. Within this work UKRI is following good practice such as ONS published guidance on data principles. One objective is more regular publication of details of the research portfolio UKRI supports, its progress, quality and impact to provide our stakeholders in Government, the research community and the public with a clear view of what UKRI supports.
15. UKRI produces two main statistical publications that underpin the evidence around UKRI investments. 1) Gateway to Research is a publicly accessible website that publishes all our on funded projects since 2006. Gateway to Research allows users to search for information on project dates, financial contributions, funded organisations and outcomes. 2) The annual Investment and Outputs Data publication which summarises applications received, funding decisions made, people and organisations supported and outcomes.
16. Based on these two publications' data, UKRI published additional statistics focusing on key areas of interest: Geographical distribution of funding and equality, diversity and inclusion (EDI).
17. It should be noted that although the recent DSIT commission has the ambition of generating a whole public sector view of R&D spend in specific areas, and the ONS statistics on Research and Development expenditure by the UK Government are compiled each year, plus large funders such as UKRI and NIHR (within DHSC) publish project-level details of their portfolio, there is no regularly updated Government-wide dataset on research projects that can be used to analyse public sector supported R&D spend to assess potential gaps and opportunities at a detailed topic-level. Although such a dataset would be useful, the number of national and international funders, across the public, charity and private sectors supporting research in the UK is very large. To provide a detailed view of the research landscape it may be more effective to examine whether administrative data could be harvested from research performing organisations that receive this funding, as the number of these to achieve near complete coverage is manageable.

How successfully do Government Departments share data?

18. The need for ADR UK arose because it was clear that government departments have limited capacity to invest the time and effort required to share data across departments, link this and create a research-ready dataset[2] that is of sufficient quality to allow statistical outputs to be derived from it. Previous experience suggests departments can only justify investing the required resource to share and link data when there is a sufficiently high priority policy imperative. An example of this is the collaboration between DfE, DWP, HMRC and HESA to create the Longitudinal Education Outcomes (LEO) dataset,[3] which is used as the data source for a range of statistical outputs[4] and dashboards.[5] There also needs to be appropriate legal gateways to support cross-government data sharing.[6]
19. The benefits of departments investing the resources required to create a linked dataset such as LEO go far beyond just the publication of statistics. If the research benefits can also be tapped, for example by LEO being made accessible for wider research use through a Digital Economy Act (DEA) accredited TRE, then the case for future maintenance of LEO and the creation of similar linked datasets becomes much easier to justify. This is where the benefits of ADR UK partnering with government departments comes into play, as this programme facilitates secure research access to datasets such as LEO. ADR UK also funds the creation of new administrative datasets for research use, which can also be used for the creation of new statistical outputs; for example, the Ministry of Justice Data First datasets.[7] A summary of all linked datasets now available for research use that include data supplied from across UK government departments is available on the ADR UK website.[8] Details of linked datasets that are under development and soon to be made available to researchers are also published on the ADR UK website.[9]
20. This new model allows ADR UK to work with UK government departments to reach across traditional policy boundaries to link data from multiple sources, creating research resources to inform decision making across a breadth of disciplines, and answering policy questions that would otherwise 'fall down the cracks' between departments.

---

[2] What makes administrative data research-ready? A systematic review and thematic analysis of published literature. IJPDS, Vol. 7 No 1 (2022): What makes administrative data research-ready? : A systematic review and thematic analysis of published literature | International Journal of Population Data Science (ijpds.org)

[3] Summary of the data linked to create the LEO dataset: The Longitudinal Education Outcomes (LEO) dataset is now available in the ONS Secure Research Service - ADR UK

[4] LEO Graduate and Postgraduate Outcomes statistics: LEO Graduate and Postgraduate Outcomes, Tax year 2020-21 – Explore education statistics – GOV.UK (explore-education-statistics.service.gov.uk)

[5] LEO dashboard to illustrate graduate outcomes by industry: LEO Graduate Industry dashboard (shinyapps.io)

[6] Summary of the legal gateways used to support the creation of the LEO database: Longitudinal Education Outcomes (LEO): privacy notice - GOV.UK (www.gov.uk)

[7] Further details of the Ministry of Justice Data First datasets, funded by ADR UK: Ministry of Justice: Data First - GOV.UK (www.gov.uk)

[8] Details of ADR England flagship datasets: ADR-England-flagship-dataset-brochure.pdf (adruk.org)

[9] Details of ADR England linked datasets under development: Browse All Projects - Filter by Type: Data linkage programmes - Filter by Partner: ADR England - ADR UK

21. ADR UK has been able to work most effectively with government departments to open up research access to data when ONS have not also wanted to gain access to the same data for statistics purposes. This is because the ADR UK approach to opening up research access to data is based around understanding the needs and risk appetites of individual departments, and ensuring that in all cases they retain full control. This includes data owners only providing access to pre-linked datasets that only include attribute information from the administrative records, and not to any identifiable information.
22. To date, the approach of ONS towards working with data owners to open up access to data for statistics use is to request access to both the attribute and identifiable information, so this data can be linked to other sources. While this has the potential to reap benefits in the long-term, experience has demonstrated that these negotiations can become extremely long and protracted, taking many years. Only after these negotiations have concluded do ONS want to pursue negotiations over research access to de-identified versions of the data via the ONS Secure Research Service. An unintended consequence of this is that access to some data, such as pre-linked data from DWP and HMRC related to benefits and income, which would be hugely valuable to the UK economist community, have not been made available for research use. Greater recognition of the untapped research value of the pre-linked (but not linkable) datasets that exist within government would maximise their potential.
23. There is an opportunity for ONS to rethink their approach to data access negotiations with their new Integrated Data Service (which will eventually replace the Secure Research Service) if the research value of data is considered alongside the statistical value.
24. Sharing and facilitating access to data will increase the potential for integration of data that can be used to improve the ability to make informed decisions that are of benefit to the future prosperity of the UK. This is the primary aim of the NERC funded digital solutions hub.
25. In summary, government departments have a limited resource to invest in data sharing activities. If this can be directed towards activities that can be delivered in partnership with organisations such as UKRI, which can open up access to this data to the external research community, there is increased potential to deliver wider benefits and reduce the need for additional data collection exercises. ADR UK can point to many examples of where government administrative and health data has been reused for policy-relevant research via the ONS Secure Research Service, for public benefit. Examples range from publications related to: gestational age at birth, chronic conditions and school outcomes;[10] an assessment of how common low pay is in Britain;[11] the outcomes of serious and organised

[10] Gestational age at birth, chronic conditions and school outcomes: a population-based data linkage study of children born in England. Nicolás Libuy, Ruth Gilbert, Louise Mc Grath-Lone, Ruth Blackburn, David Etoori, Katie Harron. International Journal of Epidemiology, Volume 52, Issue 1, February 2023, Pages 132–143, https://doi.org/10.1093/ije/dyac105

[11] How common is low pay in Britain and is it declining? New findings from linked data. John Forth, Alex Bryson, Van Phan, Felix Ritchie, Carl Singleton, Lucy Stokes, Damian Whittard. ADR UK Data Insights publication,

crime cases appearing before the criminal courts;[12] and ethnic inequalities in sentencing in the Crown Court.[13] Increasing research access to data further will increase the opportunities to deliver policy-relevant research.

Section 2: The changing data landscape

Is the age of the survey, and the decennial Census, over?

26. ESRC is constantly assessing the question of whether the age of surveys is over, given the scale of our investments into surveys and alternative data sources. This is in the context of making decisions around further investments into long-standing population surveys such as Understanding Society[14] and whether to invest in the set-up of new surveys, such as the COVID Social Mobility and Opportunities Study (COSMO)[15], and a new UK birth cohort study[16] . Similar discussions underpin the MRC's new Adolescent Health Study, which will combine survey data with other forms of data about study participants.
27. ESRC's conclusion is that investments into survey data collections still offer value for money, because of the richness that insights from these data collection can offer policymakers, over and above those derived from government administrative data and smart data like loyalty card, transactional, social media, and smart device data. Generating robust findings from novel data sources often requires linkage with other sources like survey data[17] to account

---

https://www.adruk.org/fileadmin/uploads/adruk/Documents/Data_Insights/Data_Insight_How_common_is_low_pay_in_Britain_and_is_it_declining.pdf

[12] The outcomes of serious and organised crime cases appearing before the criminal courts in England and Wales. Tim McSweeney. ADR UK Data Insights publication, https://www.adruk.org/fileadmin/uploads/adruk/Documents/Data_Insights/Data_Insight_The_nature__extent_and_outcomes_of_serious_and_organised_crime_cases_prosecuted_in_England_and_Wales.pdf

[13] Ethnic Inequalities in Sentencing in the Crown Court - Evidence from the MoJ Data First Criminal Justice datasets. Kitty Lymperopoulou. ADR UK Data Insights publication, https://www.adruk.org/fileadmin/uploads/adruk/Documents/Data_Insights/Data-Insight-Ethnic-Inequalities-Sentencing-Crown-Court.pdf

[14] Further details of Understanding Society, the UK household longitudinal study: Understanding Society – The UK Household Longitudinal Study

[15] COVID Social Mobility and Opportunities Study (COSMO): COVID Social Mobility and Opportunities Study (COSMO) | IOE - Faculty of Education and Society - UCL – University College London

[16] Further details of the Early Life Cohort Feasibility Study: CLS | Early Life Cohort Feasibility Study (ucl.ac.uk)

[17] For example, see the Cancer Loyalty Card Case-Control Study, which combined consented loyalty card data access from a high street retailer with a self-reported ovarian cancer risk questionnaire: https://www.clocsproject.org.uk/identifying-early-signs-ovarian-cancer-using-loyalty-card-data-case-control-study. See also Andrew K. Pryzbylski's research on mental health and gaming behaviours, which combined a survey of 40k gamers who agreed to share real-time gaming data: https://www.ox.ac.uk/news/2022-07-27-gaming-does-not-appear-harmful-mental-health-unless-gamer-cant-stop-oxford-study#:~:text=Przybylski%2C%20OII%20Senior%20Research%20Fellow,sense%20of%20well%2Dbeing%5D. And Al Baghal, T., Wenz, A., Sloan, L. et al. 'Linking Twitter and survey data: asymmetry in quantity and its impact'. EPJ Data Sci. 10, 32 (2021). https://doi.org/10.1140/epjds/s13688-021-00286-7

for a range of challenges[18]. For example, a 2022 case study about the measurement of financial and consumer behaviour notes, '*The combination of rich survey information and transaction data would be ideal for both research and policy intervention purposes*.'[19]

28. Access to large-scale, inclusive and representative data capturing this detailed information over time gives researchers the opportunity to make sense of this complexity and generates huge policy and practice insight. The UK's science and policy communities benefit from our hosting the world's most comprehensive series of cohort studies that track people over the course of their lives.
29. The availability and use of administrative records for research has grown considerably in recent years, in large part due to UKRI's investment in ADR UK. However, while these sources offer huge opportunities to enhance survey studies via data linkage, they lack the breadth and depth of insights and observations that surveys can gather. For example, birth cohort surveys are a rich source of parents' subjective assessments of their wellbeing, parenting and family life circumstances, while the kinds of biological measures that can be included in a well-funded study are not available in routine health records (e.g., genetic data). Outside of the health system, routine comprehensive administrative data about children's outcomes is not available until they start school.
30. ESRC is continually assessing how these different data sources complement each other. If linkage to administrative or smart data means we no longer need to ask survey respondents certain types of questions, then we can limit surveys to asking just those questions for which there is no other source of that data. As an example, the COVID Social Mobility and Opportunities Study (COSMO) surveys young people and their parents about their lives, to better understand the lives of young people impacted by the Covid-19 pandemic. The COSMO study also links education outcomes administrative data to these survey responses. By doing this, researchers have been able to publish a range of insightful policy-relevant briefings and publications, which would not have been possible without surveying these young people and their parents.[20]
31. In conclusion, there continues to be a need to talk to people in some form by conducting surveys, even if not to ask them about every aspect of their lives. The same goes for the collation of administrative data sources. One very basic example of this is asking people about their ethnicity. There may be very good reasons why people do not want to disclose their ethnicity in some situations (for example, as part of an interaction with HMRC about their income, or DWP about a benefit claim). However, being able to link self-disclosed ethnicity data from a Census to tax and benefits data to create a de-identified research-

[18] Smart data are often non-representative, *Bit by Bit: Social Research in the Digital Age*, by Matthew J. Salganik (2017, Princeton University Press) details both the unique opportunities and challenges presented by new forms of data like smart data: https://www.bitbybitbook.com/en/observing-behavior/characteristics/bad/algorithmically-confounded/#:~:text=The%20ways%20that%20the%20goals,concern%20among%20careful%20data%20scientists.

[19] https://healthpolicy.usc.edu/evidence-base/when-to-use-survey-and-administrative-data-for-financial-health-measurement-lessons-from-the-financial-health-pulse/

[20] A range of publications and briefings from the COSMO study: Publications | COSMO (cosmostudy.uk)

ready dataset[21] can tell us much about ethnic inequalities in income, and changes over time.

32. The comments above also suggest that, while there is considerable potential for administrative, health and smart data to enable a **different** approach to conducting the Census, they are not yet at a standard that means it should be abandoned altogether. Currently, generating findings that are sufficiently precise and robust from novel data sources requires linkage with other sources such as survey or Census data. We also note that the England and Wales Census in 2021 included questions on topics which are poorly covered by administrative data, including ethnicity and unpaid care. How data on these topics could be collected without survey-based methods would be a challenge for the UK's statistical agencies to resolve. Social science researchers, including those supported by ESRC-funded investments such as the UK Data Service, can make a valuable contribution to the development of any new approach to the collection and use of this data.
33. We note that a significant divergence in methodology between the three UK censuses could make undertaking UK-wide or comparative research analyses more challenging, observing the challenges that already exist as a result of the difference in timing between the Census in England and Wales, Northern Ireland, and that in Scotland. The high population coverage of Census data and suitability for linkage has enabled a very broad range of research uses. For example, Census data is also used to assess the representativeness of other data collections such as surveys. Also, indices derived from the Census, such as the Townsend Deprivation Scores, are widely used in the analysis of other datasets, by both academic and government researchers. As set out above, linkage of administrative data to Census data is also proving a valuable way to fill gaps in administrative records, such as ethnicity.

What new sources of data are available to government statisticians and analysts?

34. New sources of data available to government statisticians and analysts include administrative data sources covered above, and smart data from technologies such as mobile apps, navigation systems, digital transactions, social media, sensors, consumer platforms, and wearable devices like fitness trackers.
35. Smart technologies are also applicable for the collection of environmental data through sensor networks, some of which provide data in near real time.

What are the strengths and weaknesses of new sources of data?

36. Our response to this question is based upon the types of data in scope for Smart Data Research UK (described in Box 1) which tend to be far more complex and varied than data

[21] Details of the 2011 Census linked to Benefits and Income dataset for England and Wales: MDX Browser > 2011 Census linked to Benefits and Income - England and Wales @ 213775 (metadata.works)

from survey, experimental, and administrative sources. They encompass a wide range of data types existing interdependently: language combined with video and location data, for example, often requiring a combination of methods for analysis. Many exist at a scale that vastly exceeds traditional data sources. These are rich but largely untapped resources for understanding society, with the potential to improve lives, generate knowledge, and inform policy. They can be harnessed to understand and address key research, business and policy questions about our increasingly digital world. The Government's National Data Strategy, for example, highlights significant opportunities for such data to increase the speed, efficiency and scope of research, alongside its overall value for society and the economy. There are already clear demonstrations of the promise of smart data use in research, covering areas as diverse as ovarian cancer detection (using loyalty card data), the impact of gaming on mental health (gaming data), trends in labour demand (online job postings) and passenger movement and social distancing (through the creation of a 'digital twin' based on real time passenger movement at St Pancreas station during the pandemic)[22].

37. These new forms of data do of course also present distinct challenges. The following brief examples[23] illustrate just three of the distinct challenges that may arise in using such data for research:
    a. **Representativeness:** Datasets are susceptible to both under- and over-coverage because only people interacting with a service are represented, potentially multiple times, rather than all individuals of research interest.
    b. **Algorithmic confounding:** Many systems use algorithms to select content or induce specific user behaviours, which can produce confounding patterns in datasets.
    c. **Unstructured data:** Sources such as social media, image, and sensor data are unstructured, or differently structured to traditional social science data. These datasets cannot be easily analysed in a relational database, and they require specialised techniques and data preparation to allow researchers to extract insights from them.

38. Smart data is typically personal (revealing locations, preferences, opinions) and may also be sensitive in nature. As such, public trust in use of this data for research and statistics is essential and must be earned, to ensure support for the research and statistics derived from it, and subsequent evidence-based policy informed by its use.

---

[22] Further information about the first two projects is provided in footnote 16 (loyalty card data for early ovarian cancer detection, real-time gaming data to understand impacts on mental health). See also 'The Use of Online Job Sites for Measuring Skills and Labour Market Trends: A Review'. Oleksii Romanko and Mary O'Mahony. *ESCoE Technical Report No. 2022-1*: https://www.escoe.ac.uk/publications/the-use-of-online-job-sites-for-measuring-skills-and-labour-market-trends-a-review/ and Real-time passenger data from St Pancras station

[23] Examples taken from a longer list in 'Bit by Bit: Social Research in the Digital Age', by Matthew J. Salganik. 2017. Princeton University Press: https://www.bitbybitbook.com/

39. A wider issue with using smart data sources for research and/or statistics is the data access challenge. Researchers and statisticians have been stymied in developing methods for working with such data because it has been very difficult to access. As a 2020 column in Science noted, access to privately held data is, '*rarely available to academics, and when it is, it is typically granted through a patchwork system in which some data are available through public application programming interfaces (APIs), other data only by working with (and often physically in) the company in question, and still other data through personal connections and one-off arrangements, often governed by nondisclosure agreements and subject to potential conflicts of interest.*'[24]
40. Issues related to licensing, governance and public trust must be tackled concurrently with greater access to such data, to move beyond abstractions and reflect emerging use cases.[25] Smart Data Research UK will build on the strong foundations laid by previous investments in this field to deliver a step change in the use of smart data for research and innovation, enabling new policy insights for official statisticians and analysts.
41. Additionally, environmental data combined with other sources of information may provide new insights and benefits to civil society, however there are challenges, some of which are not dissimilar to those set out above:
    a. The data may be limited in its coverage, only providing for example information from a specific geographic area, rather than national coverage.
    b. Although in general environmental data is less likely to be sensitive, care would be needed if integrating with other forms of data to avoid the possibility of personal identification.
    c. Networks of sensors may result in a deluge of data, which would require processing and preparation before researchers are able to gain insights from the data set and integrate it with other datasets. Automation and for example edge computing, that enables some processing *in situ* may help with this, as could application of technological developments such as machine learning and AI.
42. Building trust in the application of these techniques both with the research community and the public also needs to be addressed before there is full support for the research and statistics derived from it, and subsequent evidence-based policy informed by its use.

[24] 'Computational social science: Obstacles and opportunities.' Science 369 (6507), 1060-1062. DOI: 10.1126/science.aaz8170. https://gking.harvard.edu/files/gking/files/1060.full .pdf

[25] This situation is interlinked with challenges in using data-driven tools in the public sector, where the UK finds itself in 'the nascent 'first wave' stage of transparency and accountability measures.' Public sector use of data and algorithms | Ada Lovelace Institute

Section 3: Protecting privacy and acting ethically

Who seeks to protect the privacy of UK citizens in the production of statistics and analysis? How?

43. In terms of the production of statistics, the UK Statistics Authority Code of Practice for Statistics,[26] to which all producers of official statistics must adhere, is clear that all statutory obligations governing the collection of data, confidentiality, data sharing, data linking and release should be followed. It is therefore the role of all organisations producing official statistics to protect the privacy of UK citizens in the production of statistics. The Office for Statistics Regulation monitors compliance with this code and publishes the outcomes of compliance checks.[27]
44. In terms of the use of data for analysis and research, all organisations holding personal data must comply with the UK's data protection laws, and access to this data for research purposes is carefully controlled. A summary of the legal framework for accessing administrative data for research is published on the ADR UK website,[28] along with further details about how the privacy of UK citizens is protected when data is made accessible for research.[29]

What does it mean to use data ethically, in the context of statistics and analysis?

45. In the context of opening up access to data for analysis and research (including data used in the production of statistics), the UKSA has roles and responsibilities related to the ethical use of data for research purposes. For example, in relation to the Digital Economy Act (DEA, 2017), which was enacted to support lawful data access for research and data linkage across departments, the UKSA is responsible for managing ethics approvals and people, project, and data processor accreditations. These are facilitated through the UKSA Research Accreditation Panel (RAP), and the National Statistician's Data Ethics Advisory Committee (NSDEC). Through these mechanisms, all four ADR UK's trusted research environments have gained DEA accreditation, as well as wider ESRC data infrastructure investments such as the UK Data Service and the UK Longitudinal Linkage Collaboration, so they can all use the DEA as a permissive legal gateway to facilitate access government administrative data for research. In this way, the UKSA is helping the wider research ecosystem to build trustworthiness with the public, ministers and data owning departments about how administrative data can be used for public good research, facilitating high quality research insight of public benefit without the need for further expensive data collection.

---

[26] UKSA Code of Practice for Statistics: About the Code – Code of Practice for Statistics (statisticsauthority.gov.uk)
[27] Office for Statistics Regulation compliance reports: Our Regulatory Work – Office for Statistics Regulation (statisticsauthority.gov.uk)
[28] Summary of the legal framework for accessing administrative data for research: ADR UK The legal framework for accessing data
[29] Summary of how data is made accessible to researchers, ethically and responsibly: Ethics & Responsibility - ADR UK

Are current processes and protections sufficient?

46. An important difference between traditional forms of data such as survey data, and new forms such as administrative and smart data, is that survey participation works on a consent basis, where participants consent to their survey data being used for research and/or statistics purposes, and also consent to subsequent linkage of this data to other forms of data. As administrative and smart data sources are not typically collected for the purpose of either research or statistics, if organisations want to use it for these purposes, in ESRC's view they must go above and beyond what is required of them in law, to ensure they maintain the social contract to use this data for these purposes. As an example of how this can be done proportionately, ADR UK and the Office for Statistics Regulation (OSR) recently collaborated on an impactful, widely publicised public dialogue report[30] to understand what the public mean by the term 'public good' when used in the context of using administrative data for research and statistics. This report builds on the wider approach ADR UK has taken to including the public voice in the use of administrative data for research, as summarised in our public engagement strategy.[31]
47. More widely, the Engineering & Physical Sciences Research Council (EPSRC) also have investments that are designed to provide expertise on the privacy and trust aspects of data use. For example, the SPRITE+ investment brings together people involved in research, practice, and policy with a focus on digital contexts. They are a 'one stop shop' for engagement between academic and non-academic communities, providing a way for these communities to connect and build collaborations across the spectrum of issues relating to security, privacy, identity and trust.[32]

Section 4: Understanding and responding to evolving user needs

How are demands for data changing?

48. As summarised above, the availability and use of administrative records for research has grown considerably in recent years, in large part due to UKRI's investment in initiatives such as ADR UK. However, while these sources offer huge opportunities to enhance survey studies via data linkage, ESRC's view is that they are not yet of the appropriate scale and breadth to allow them to replace the insights and observations that surveys or the Census can gather.

---

[30] Details of the joint ADR UK/OSR public dialogue report: https://www.adruk.org/news-publications/news-blogs/adr-uk-and-osr-publish-research-report-on-public-perceptions-of-public-good-use-of-data-for-research-and-statistics/

[31] ADR UK public engagement strategy: Demonstrating trustworthiness and maximising public benefit: ADR UK Public Engagement Strategy 2021 - 2026 - ADR UK

[32] About SPRITE+ (spritehub.org)

49. As summarised above, there is a huge interest in the research value of smart data sources. However, these data sources tend to be far more complex to analyse than data from surveys, experiments, and administrative data. There are also huge data access challenges, and methodological challenges that are likely to benefit from survey data in complement to novel sources. So again, while we can see a future where these data sources complement and enhance more traditional forms of data, we are not yet at a point where these new forms of data replace survey data.
50. Despite programmes such as UKRI's ADR UK and ONS's Integrated Data Programme, there are still notable restrictions in access to data. As an example, although some HMRC data is now available to researchers via the ONS Secure Research Service (the ADR England trusted research environment), the main route to accessing HMRC data remains the HMRC Datalab. Unlike all ADR UK trusted research environments, which have the facility to access secure data remotely, the HMRC Datalab operates only from a secure room within the HMRC offices in London. This means the number of researchers that can work on the data at any one time has a physical restriction, and the researchers also need to be London-based (or be able to regularly travel to London). The legislation that underpins the setup of the HMRC Datalab (the Commissioners for Revenue and Customs Act, 2005) also requires projects to demonstrate how the outcomes would benefit HMRC. This is much more restrictive that the legislation underpinning the ADR UK trusted research environments. These restrictions mean many UK economists now find it easier to work on non-UK data. This is not just an issue for the researchers themselves, but poses a problem for UK policymakers, since the outcome is that economists are less likely to undertake research with UK policy relevance. Improving access to relevant economic data would not only increase the number of UK-based, UK-funded researchers doing research relevant to UK public policy but increase the number of overseas researchers doing such research, without any increase in UK taxpayer funding.

How do users of official statistics and analysis wish to access data?

51. In addition to being able to access the published data associated with statistical outputs, there is huge public benefit to be gained from also making the underlying data available for research use, to avoid the *missed* use of data – that is, underusing the potential offered within existing data collections.[33] As described above, the benefits of departments investing the resources required to create datasets for statistical outputs go far beyond just the publication of statistics. If the research benefits can also be tapped, then the case for future maintenance of existing and new data sources becomes much easier to justify.
52. ESRC-led digital research infrastructures, including precursor investments to Smart Data Research UK and the Business and Local Government Data Centres[34], have demonstrated

[33] ADR UK blog on the misuse versus the missed use of data: Administrative Data: Misuse vs. Missed Use - ADR UK

[34] These centres focused on making data routinely collected by business and local government organisations accessible for research of mutual benefit to data owners. Two of these, the Consumer Data Research Centre

the benefits of academic-led data centres for policymakers. These centres have produced valuable data services for government stakeholders[35], and proved instrumental in forming responsive partnerships with other stakeholders including government during the Pandemic. The Royal Society's DELVE Initiative points to such collaborations as a key element of preparedness for future emergencies[36].

53. With technological developments, volumes of data being collected are growing, therefore physically moving data will become more challenging and having the compute, for analysis, alongside the data is becoming increasingly important. This is the approach that NERC have taken with JASMIN, the data analysis facility for environmental science, which sits alongside CEDA (Centre for Environmental Data Analysis), one of NERC's environmental data centres.

How can we ensure that official data and analyses have impact?

54. Making this data safely and securely available to trained researchers has the potential to massively increase the impact and efficiency of social and economic research. It also provides direct benefits to government departments as both data owners and policy users. ESRC has an important role to play here, in bridging the gaps between academic researchers, data owners and policymakers to ensure policy relevant research insights are fed into the Civil Service. In turn, the Civil Service needs to be resourced to be able to engage with academics, and to understand the value of this type of engagement to inform decision making.

How do we ensure that users, in the Civil Service, Parliament and beyond, have the skills they need to make effective use of data?

55. The data being considered by this consultation can be distinguished from their predecessors in terms of the speed with which they are generated, their scale and their form (for example, image, audio and video data). Maximising the potential offered by these data has considerable implications for training, and for data storage and access.
56. ESRC's view is that the most effective way of ensuring civil servants have the breadth of skills required to make effective use of the full range of survey, administrative and smart data sources is to work in close partnership with academia. This is in acknowledgement that it takes many years of training before a user of survey, administrative or smart data can react quickly to a new and urgent need for evidence to inform policy. The academic

---

(CDRC) and Urban Big Data Centre (UBDC), are part of Smart Data Research UK phase 1.

[35] UBDC worked with Glasgow City Council to develop innovative methods for monitoring activity on Glasgow's streets: Using spare CCTV capacity to monitor activity on city streets | Urban Big Data Centre (ubdc.ac.uk)

[36] See DELVE: Report on test, trace, isolate and support, 18 May 2020 - GOV.UK (www.gov.uk) and Data Readiness: Lessons from an Emergency (rs-delve.github.io)

community has scalable access to talent and skills - and plays a critical role in training future generations of researchers, some of whom of course will go on to work within the Civil Service.

57. As an example of how effectively such partnerships can work, by supporting ONS to facilitate access to government data to researchers through ADR UK funding, ESRC is supporting ONS and the wider Civil Service being able to tap into this much larger talent base to understand how to make the best use of administrative data. This is both in the context of increasing the flow of talent and skills from academia into the Civil Service, and facilitating learning between the two groups, for the benefit of research and statistics.
58. Examples of how ADR UK funding to ONS supported the government to benefit from academic involvement in research and statistics during the COVID-19 pandemic include:
    d. The government's large-scale virus infection and antibody test study, the COVID-19 Infection Survey (CIS) is hosted in the ONS Secure Research Service (SRS) to support data analysis. Data scientists in No. 10 accessed this data throughout the pandemic via the ADR UK-funded ONS SRS.
    e. The Joint Biosecurity Centre and Public Health England used the SRS to deposit and collaborate on COVID-19 contact tracing data, enabling its use by government for priority work and opening up this data to external researchers. ADR UK worked with the Joint Biosecurity Centre, ONS, and CDRC on a successful pilot programme for 'Local Data Spaces' to enable localities in England and Wales to access detailed COVID-19 area data via the SRS[37]. The evaluation report is published on the ADR UK website,[38] and this informed further ESRC investment into the development of Local Policy Innovation Partnerships (LPIPs) which aim to enhance access to and use of data and expertise to support local decision making.[39]
59. Without this pre-existing relationship between ESRC and ONS, including funding to expand and improve the SRS through ADR UK, it would not have been possible for the SRS to be utilised at the scale it was during the pandemic, where both researchers and civil servants were routinely collaborating to share skills and methods for public good.
60. These types of partnerships between ESRC and the Civil Service have existed for many years in the context of survey design, data collection and reporting. Indeed, many of the large social surveys that are led by academic teams are funded by government departments.
61. With ESRC's new Smart Data Research UK programme, we will be strengthening the links between the Civil Service and researchers across all of UKRI's remit areas in the use of this data – including supporting innovations in data science and AI.
62. ESRC also invests in ways which build skills and capacity in the Civil Service, Parliament and beyond. For example, our Policy Fellowships scheme embeds academic experts into the

[37] Local Data Spaces combined geo-demographic and mobility data provided through CDRC, a Smart Data Research phase 1 project, with de-identified Covid-19 data from Test and Trace to understand the spread of Covid-19 at the local level, as well its impacts on communities.

[38] Evaluation of the Local Data Spaces pilot programme: https://www.adruk.org/news-publications/news-blogs/local-data-spaces-pilot-demonstrates-importance-of-local-level-data-and-analysis-to-inform-local-decision-making-440/

[39] Announcement of funding to support LPIPS programme: https://www.ukri.org/news/ukri-invests-in-policy-innovation-partnerships-for-local-growth/

heart of government departments. This programme has a strong focus on data. For example, we have embedded data scientists within 10 Downing Street[40] and other government organisations to support capability building and analysis of novel data. We have recently confirmed support for a data focused Parliamentary Thematic Research Lead, in collaboration with the Parliamentary Office for Science and Technology, to support scrutiny around data issues. Our Economics Observatory also has a strong focus on economic data and provides tools and training to civil servants to support them accessing and analysing the latest data direct from official sources[41].

63. With regard to users beyond the Civil Service and Parliament, ESRC's recent 'Review of the PhD in the social sciences' flagged the importance of embedding digital methods and strengthening quantitative training within PhD training. These recommendations are being taken forward in the next round of ESRC's Doctoral Training Partnerships, along with additional support for data-driven research skills across the entire career stage.
64. The requirement for digital skills has been recognised throughout UKRI. As a result, developing skills and career pathways is part of the vision for UKRI Digital Research Infrastructure.[42] As mentioned above UKRI plays a critical role in training researchers, some of whom will go on to work within the Civil Service.

August 2023

[40] Announcement of funding for data science fellows in No.10: https://www.adruk.org/news-publications/news-blogs/esrc-and-adr-uk-funded-research-fellows-to-work-with-no10-downing-street-487/
[41] See Economic Observatory Data Hub: https://www.economicsobservatory.com/data-hub
[42] Digital research infrastructure – UKRI