

---

# UKDS Data Skills Framework for working with large scale survey data, census, and macro-level aggregate data

A Discussion Paper from the UK Data Service Training and User Support Directorate

April 2024



---

30/04/2024

Copyright © 2024, University of Manchester. Created by Cathie Marsh Institute, UK Data Service.

Version No.1



The UK Data Service is funded by UKRI through the Economic and Social Research Council (ESRC). This work was supported by the Economic and Research Council grant ES/Y005538/1.

## Table of contents

<b>Table of contents.....</b>	<b>2</b>
<b>Authors and Acknowledgements .....</b>	<b>3</b>
<b>Summary .....</b>	<b>4</b>
<b>The UK Data Service Data Skills Framework for working with large scale survey data, census, and macro-level data. A discussion paper. ....</b>	<b>5</b>
<b>Introduction.....</b>	<b>5</b>
<b>Background .....</b>	<b>5</b>
The UK Data Service fosters data literacy and capacity .....	5
<b>The data environment.....</b>	<b>6</b>
<b>Why these data?.....</b>	<b>6</b>
<b>Why a Framework? .....</b>	<b>7</b>
<b>What can the Data Skills Framework be used for? .....</b>	<b>8</b>
<b>Analysis of the current training landscape .....</b>	<b>9</b>
<b>The UK Data Service Data Skills Framework .....</b>	<b>12</b>
<b>Developing the Framework: what we did.....</b>	<b>12</b>
<b>The UK Data Service Data Skills Framework.....</b>	<b>13</b>
<b>Next steps.....</b>	<b>22</b>
<b>Documents .....</b>	<b>23</b>

---

## Authors and Acknowledgements

The paper is authored by **Dr. Vanessa Higgins**, Service Director for Training and User Support, **Professor Debora Price**, Deputy Director of the UK Data Service, and **Dr. Sarah King-Hele**, Research Fellow and Training Manager, UK Data Service. The UK Data Service Data Skills Framework presented here has been developed collaboratively by the Training and User Support Directorate of the UK Data Service with contributions gratefully received from (in alphabetical order):

**Alle Bloom, Cathie Marsh Institute,  
University of Manchester**

Research Associate

**Dr. Jennifer Buckley, Cathie Marsh  
Institute, University of Manchester**

Research Fellow

**Louise Capener, Cathie Marsh Institute,  
University of Manchester**

Research Associate

**Dr. Rihab Dahab, Cathie Marsh Institute,  
University of Manchester**

Research Associate

**Dr. Nigel de Noronha, Cathie Marsh  
Institute, University of Manchester**

Research Associate

**Professor Oli Duke-Williams, University  
College London**

Service Director for Census and Professor of  
Population Information

**Dr. Jools Kasmire, Cathie Marsh Institute,  
University of Manchester**

Research Fellow

**Nadia Kennar, Cathie Marsh Institute,  
University of Manchester**

Research Associate

**Cristina Magder, UK Data Archive,  
University of Essex**

Data Collections Development Manager

**Dave Rawnsley, Aggregate Data Unit,  
Jisc**

Senior Technical Coordinator

**James Reid, EDINA, University of  
Edinburgh**

Head of Services and Deputy Director of  
Edina

**John Sanderson, UK Data Archive,  
University of Essex**

Deputy Director of UK Data Service (Acting)

**Dr. Pierre Walthery, Cathie Marsh  
Institute, University of Manchester**

Research Associate

To cite this paper:

Higgins, V., Price, D. and King-Hele, S. (2024) *The UK Data Service Data Skills Framework for working with large scale survey data, census, and macro-level data. A Discussion Paper from the UK Data Service, April 2024*. The UK Data Service: Colchester, England. DOI: [10.5281/zenodo.10911305](https://doi.org/10.5281/zenodo.10911305)

## Summary<sup>1</sup>

The UK Data Service (UKDS) Data Skills Framework (DSF) represents a strategic response to the evolving landscape of data analytics, particularly within the social, economic, and population sciences. The DSF has been developed through research and consultation and it addresses the need for a comprehensive skills framework that aligns with rapid technological advancements and the expanding scope of data sources. It is designed to foster robust data literacy and enhance the capacity to engage with large scale survey data, census data, and macro-level aggregate data. It offers a blueprint for training programmes that aspire to equip researchers and data analysts with a balanced mix of foundational and advanced data skills necessary for tackling contemporary research challenges and making the most of emerging opportunities for example those presented by linking data, or by Artificial Intelligence (AI) and machine learning.

The DSF is intended to be flexible and adaptable, allowing for the integration of new teaching methods and learning technologies. It serves not only as a guide for the UK Data Service's own training initiatives but also potentially as a useful framework for other training providers across the data skills ecosystem. By mapping out essential skills and identifying training gaps, the DSF aims to promote efficiency and collaboration among various data training stakeholders, including academic institutions, government agencies, data services, and private sector entities.

As the DSF moves forward, it will continue to evolve in response to feedback from the user community and changes in the data landscape. This ongoing development process underscores the UK Data Service's commitment to maintaining a relevant and forward-thinking training strategy that not only meets current needs but also anticipates future demands.

Our aspiration is that the UK Data Service Data Skills Framework is more than just a training model; it is a strategic tool that enables the fostering of a skilled data analytics community capable of advancing knowledge and innovation in the social sciences. It epitomises a proactive approach to educational enhancement, ensuring that data skills training remains relevant for academic and professional development in an increasingly data-driven world.

---

<sup>1</sup> This Summary was written with the assistance of ChatGPT 4.0.

---

---

# The UK Data Service Data Skills Framework for working with large scale survey data, census, and macro-level data. A discussion paper.

## Introduction

In this discussion paper, we present an analysis of the training landscape in the UK for social, population and economic data, and we propose to the Social Science Data Training Community a Data Skills Framework for working with these kinds of data. This Framework has been developed in the context of a data landscape undergoing unprecedented change at rapid pace. We envisage it as a live and evolving piece of work, and very much welcome discussion on the content.

This initiative aims to establish a robust framework for developing essential data analysis skills for the social, economic and population sciences, focusing on large scale survey, census, and macro-level aggregate data characteristic of the UK Data Service (UKDS) holdings. The framework acknowledges a rapidly advancing data skills landscape. It emphasises continued development of traditional data skills for contemporary research needs, while recognising growing potential for integrating survey or census data with an expanding array of other sources increasingly accessible to social scientists, as well as promising opportunities presented by AI and machine learning for enhancing analysis.

## Background

### The UK Data Service fosters data literacy and capacity

The UK Data Service is the premier repository for social, population, and economic data in the UK. Experts in the collection, preservation, and dissemination of data, we curate datasets from trusted data providers including government, the Office for National Statistics, the ESRC, and diverse other funders and data owners, as well as acting as a repository for researcher-generated data through our ReShare facility. Our digital collection of more than 9,000 datasets comprises nationally and internationally significant data such as the Census, Understanding Society, the UK Cohort Studies, the Labour Force Surveys, the Family Resources Surveys, and many others. With almost 50,000 registered users annually, we facilitate over 200,000 data downloads each year<sup>2</sup>.

Our mission at the UK Data Service is to catalyse impactful social science research by providing access to high quality curated and trustworthy datasets of national strategic importance, fostering data literacy, and contributing to the advancement of knowledge through dynamic data infrastructure and strategic partnerships<sup>3</sup>. One pillar supporting this mission is

---

<sup>2</sup> [ukdataservice.ac.uk/about/strategic-focus/annual-reports/](https://ukdataservice.ac.uk/about/strategic-focus/annual-reports/)

<sup>3</sup> [ukdataservice.ac.uk/about/strategic-focus/](https://ukdataservice.ac.uk/about/strategic-focus/)

the fostering of data literacy and capacity for research data sharing, use and reuse. Through live and on-demand training and support services, the Service helps data creators and data users ethically and legally manage and share their data; and assists users to find, access, evaluate, and use a broad range of datasets, increasing capacity for high quality and impactful data analysis. We have approximately 6,000 people a year attend more than 100 live training and user community events; about 10,000 visits a month to our online learning hub and data skills modules; and we average 7,000 views a month of our YouTube training resources.

This is the context in which we have developed this Data Skills Framework, and we are using it to aid our own gap analysis, training development, and strategic thinking for the next six years. We present a number of potential wider use cases for the framework below and hope it can help in the continued development of a relevant and efficient training ecosystem for data analysis across the social sciences.

## The data environment

Rapid advancements in technology and the availability of data present opportunities and challenges. While the research possibilities are increasing day by day, we are facing widespread societal concern about data privacy and security, and continuous developments in legal and ethical best practice.

In developing this discussion paper, our aim has been to set out a framework for building the multiplicity of skills needed now and into the foreseeable future for data driven analysis of social and economic issues, for those working with large scale survey, census, and macro-level aggregate data typical of the UK Data Service data holdings. Opportunities are unfolding for data linkage to our holdings with newer forms of data including web data, streaming data, and administrative data, while more data is moving into Trusted Research Environments, with data access requiring additional sets of skills to those required for traditional modes of data sharing. Computing power means that new types of analyses are possible, with computational social science and machine learning developing at pace. New conversations are being had and cutting-edge work is being done on synthetic data, and while AI for coding and analysis is bringing new and as yet unknown uses, we have increasing levels of concern in academic communities about verifiability and reproducibility. All this means new skills are needed, and these are constantly evolving.

However, it is very important to keep in mind that while ever more advanced skills are promised and required, foundational skills remain crucial. People still need to start somewhere with data, and they need to have the fundamental skills in place on which to build. The data training ecosystem is therefore, and needs to be, a complex mix of beginner, intermediate, advanced and masterclass level skills, and ideally, this training needs to be located efficiently across the data training community.

## Why these data?

We have focused here on developing skills for the kinds of data that we hold – namely large scale survey data, census data, and macro-level aggregate data, while bearing in mind that

---

these data can increasingly be linked to all kinds of other data with unfolding possibilities for research. Data from large scale surveys, censuses, and macro-level aggregate data are used extensively by researchers, educators and students, central and local governments, policymakers, research funders, think tanks, charities and other civil society actors and the media, to gain insights, knowledge and understanding. As will be seen, this data focus alone has yielded a complex and substantial array of necessary skills. We made the decision that a data skills framework for these kinds of data needed to be contained, and that trying to provide a framework for all types of data across disciplines and data types was too large an exercise for a single framework. We recognise of course that many or even most of the skills we describe will be relevant to many other types and domains of data, and we hope that others with specific training remits for those data might build on this framework, either in high level domains or sub-domains.

The kinds of data that have been our focus here will continue to be crucially important for years to come. They offer representative sampling; the ability to study people longitudinally, the potential for in-depth information into complex relationships; the ability to customise and flex data collection to address a wide range of questions; validation and verifications processes; and adherence to strict ethical standards. While other types of data such as administrative records or smart data offer exciting opportunities for research and policymaking, they are not without limitations. large scale survey data will likely long remain a valuable and complementary source of information, providing researchers with unique insights into the complex dynamics of society and human behaviour.

Although the UK Data Service does hold some qualitative datasets, this Data Skills Framework does not cover specific skills for analysis of qualitative data which would require distinct sector analysis and development across qualitative epistemologies, traditions, and methodologies. We fully understand the growing need for rounded researchers who span the qualitative/quantitative divide, and that advanced quantitative, machine learning and AI/Large Language Model (LLM) techniques can increasingly meaningfully be applied to qualitative data. However, we have focused here on not diluting what we see as essential quantitative data training, from foundational to cutting-edge. These skills are complex to acquire and will remain essential in the world going forward. Many parts of this framework are of course common to all types of data analysis, including qualitative research, and we would welcome groups focused on qualitative methodologies building on this framework.

## Why a Framework?

The aim with this Framework is not to set out training that the UK Data Service will provide, but an overview of the skills that will be needed for social and economic data analysis in the modern era, to be provided across the data skills training ecosystem. We have considered it important to develop a framework at this point in our evolution and the evolution of the data landscape, for several reasons. The volume increase and rapid development of our training offer over the last five years has meant that there is a need to take stock of where we are,



---

what the environment looks like, and to think about the skills that will be needed in the future. A framework enables us to take a holistic approach to data skills development, while also enabling us to consider the detail, who should be providing what to whom, and how the training ecosystem fits together. We also see it as helping us to consider structured learning pathways for participants, which we envisage as an increasing part of the individualisation of learning especially incorporating asynchronous teaching and learning. A framework offers us flexibility and adaptability, to consider specific requirements of different target audiences, and it also offers us potential for the curation of our resources, through our learning hub, online materials, and live training. We hope that it will help to frame a long-term sustainable offering to the data user community, allowing us to evolve with data trends, emerging technologies and changing needs.

## What can the Data Skills Framework be used for?

The Framework has strategic, operational, and descriptive functions, with the potential to aid consensus building on core and essential skills, skill contingencies, and varying skill-levels. We envisage a number of potential uses of the Framework and would also welcome feedback on this.

We see the Framework as potentially being used to:

- Map existing provision within the UK Data service, and across the data skills training ecosystem, and recognise gaps in training needs.
- Ensure efficiencies across the data skills training ecosystem.
- Enable analysis of efficient and optimal training providers, and help to envision how collaborations across trainers could work.
- Help funders, governing bodies, and training providers to prioritise development of new training, across ESRC and non-ESRC investments.
- Contribute to thinking about pathways to skills and expertise in data analysis and aid in their development.
- Assist in framing curriculum development for ourselves, other data services, universities, and other data skills training providers.
- Help people know what it is they do not know – from people who do not know anything to experienced researchers; something for people to refer to across the researcher life-course as data, methods, and technologies evolve.
- Open existing provision to broader audiences including those outside of academia;
- Create a template for considering parallel frameworks for other types of data/skills training [e.g. qualitative methods; smart data].
- Curate training materials across data skills training providers, including aiding in the development of shared vocabularies.
- Evolve along with forms of data, computing, and new methodologies.

There are many providers across the data training landscape, and the UK Data Service provides only a fraction of what is available and needed. We plan to use the framework to help

---



us frame what we do or should provide; how, and to what levels; and to contemplate how to signpost users to other providers. It will also help us to work out optimal collaborations and what optimal training gateways might look like to potential users.

## Analysis of the current training landscape

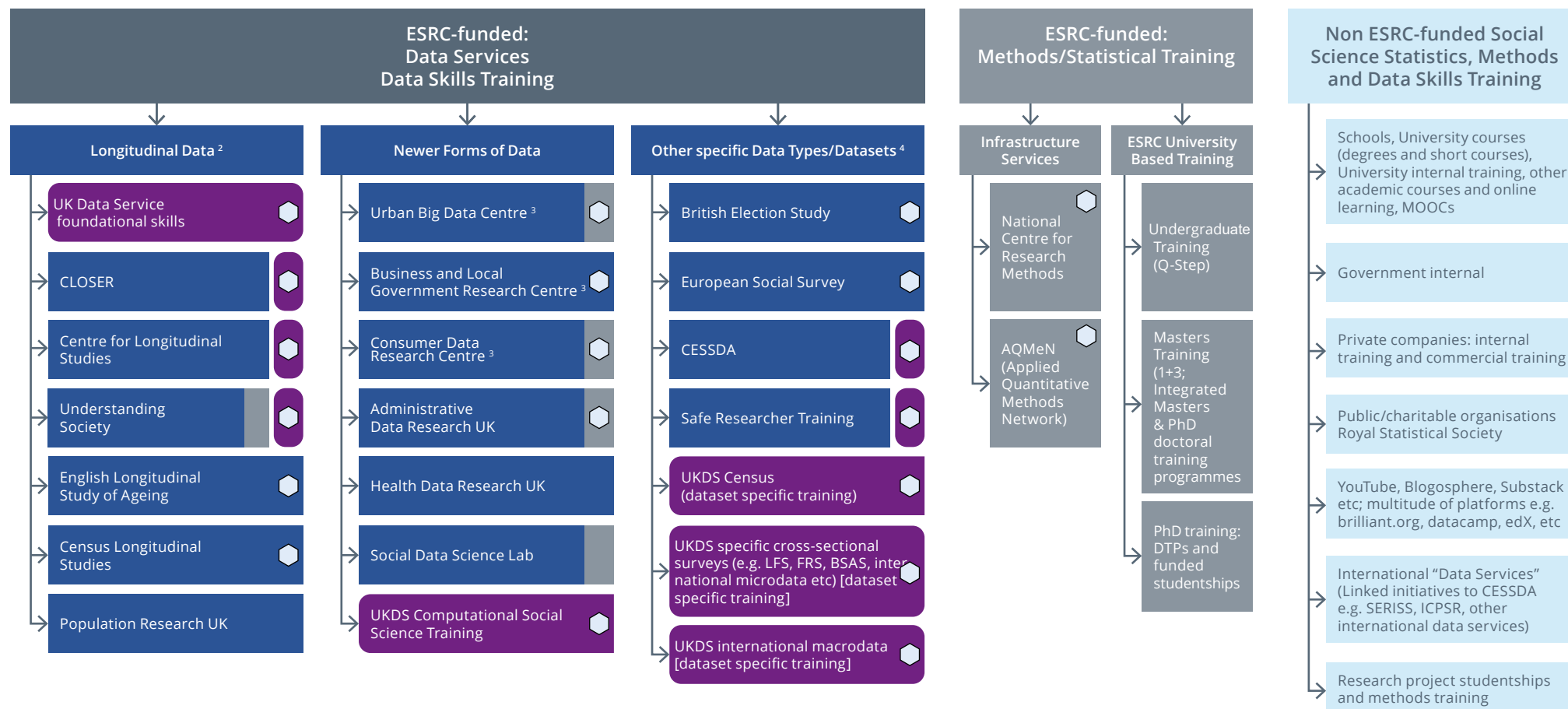
We first present, in Figure 1, our analysis of the UK Training Landscape for Quantitative Data Skills Training in the Social Sciences, considering which investments are providing what training, funded how, and to whom. This analysis is informing three key pieces of strategy for us: our existing and core training; our future training and data skills development; and our collaboration on Training and Capacity Building with partners and other UKRI investments. This analysis is helping us to think about minimising duplication across the system, and the appropriate division of training from foundational to advanced. It has helped to inform our current view that there is little duplication or redundancy across the social science training ecosystem, that advanced methods training is situated in highly expert provision, that the ESRC-funded Data Resources Training Network (DRTN) run by the National Centre for Research Methods (NCRM) forms a useful training network for facilitating efficient collaboration, and that there are useful cross-synergies yet to be developed.

Training is always evolving, as are data and data infrastructure investments. In the UK, the Economic and Social Research Council (now within UKRI) has been the significant investor and to a large extent, therefore, architect of social science training provision over the last 60 years. These investments take several forms – for example, direct investment in infrastructure such as the UK Data Service and many other data services; investment in datasets that provide dataset-relevant and dataset-specific training as ancillary services (such as Understanding Society, or the English Longitudinal Study of Ageing); and investment in doctoral colleges and master's and doctoral programmes within universities. This provision takes place within a much wider ecosystem and mixed economy of training provision for skills for the social sciences, including commercial providers.

Figure 1 is our representation of the UK data skills training landscape in the social sciences, as at December 2023, according to three characterisations: ESRC-funded training for data skills and data services on the left; ESRC-funded methods and statistics training in the middle; and non-ESRC funded social science statistics, methods, and data skills training on the right.

## UKDS builds foundational capacity in 3 core topics:

• Research Data Management • Finding and Accessing Data from the UK Data Service • Using Data (foundational skills)



Source: Analysis by the UK Data Service User Support and Training team at the Cathie Marsh Institute, University of Manchester, December 2023

## Key

UK Data Service providing training in data access and data skills

Non-UKDS ESRC-funded Data Services providing access to data/gateway to access with varying amounts of data skills training

Non-UKDS ESRC funded methods and statistics training

UK Data Services providing training in data access, data skills and methods

Member of the DRTN [Data Resources Training Network]

Non-ESRC funded statistics, methods and data skills training

1. The Centre for National Training and Research Excellence in Understanding Behaviour (CENTRE-UB) has since been awarded funding from ESRC. The training programme is in development and may include data skills training.

2. To note that other longitudinal studies funded or part-funded by ESRC may be part of the Data Research Training network but do not routinely provide training and so have not been included here.

3. Smart Data Research UK (formerly Digital Footprints), may see the introduction of new data services.

4. Some training is provided relating to the UKDS qualitative data archives but that is not included here

---

Although it might be tricky in some cases to separate data skills from statistical and methodological training, we have found it helpful to do so, and this has informed our characterisation of both the training landscape and the Data Skills Framework. This has been driven by a number of factors. As the UK Data Service, we predominantly provide data skills training for research data management, finding and accessing data, foundational skills for using data and dataset-specific skills for using some of our most important national datasets; and we provide introductory training to areas relatively new to social scientists, for example in computational social science. Although we collaborate in other spaces, we do not independently provide advanced dataset-specific training where this is funded and provided via another data service, for example Administrative Data Research UK, or Understanding Society; and we do not provide advanced methods training in statistical or data sciences, which again, is funded in different ways by other providers for example via NCRM or the ESRC Doctoral Training Partnerships with Universities.

As Figure 1 shows, we at the UK Data Service provide foundational capacity in three core topics: research data management, finding and access data from the UK data service, and foundational skills in using data. This training is free at the point of use and open to anyone across the world who wishes to use it. There are then a large number of ESRC-funded data services and investments, shown on the left, that provide, to greater or lesser extents, highly expert investment-specific *data skills training*. This is largely free at the point of use, but also in some cases on a pay-to-access model. Where the UK Data Service has a complementary or collaborative role in skills training, or in a few cases, methods training, this is denoted in Figure 1.

The aim of this graphical representation in Figure 1 is to aid understanding of how data skills and methods training for the social sciences are currently organised, where collaborations and crossovers currently exist, and to highlight the potential for productive and efficient collaboration in the future. Members of the ESRC-supported Data Resources Training Network (DRTN) are marked on this diagram with a hexagon. This is a forum run by NCRM to support collaboration between ESRC-funded data services, coordinating and facilitating opportunities to add value to individual member activities, limiting duplication and sharing intelligence.

---

## The UK Data Service Data Skills Framework

We next present the Data Skills Framework. This represents analysis that we have been undertaking for the last 18 months to inform our training development, data skills modules, data skills pathways, and the development of this Framework.

### Developing the Framework: what we did

We conducted a review of relevant documents and literature including:

- External consultations on data skills and training gaps for social sciences. This included reports based on consultations written by key stakeholders in the field of research methods training such as ESRC, National Centre for Research Methods and CLOSER.
- Policy documents such as the UK Government's National Data Strategy and the report on Quantifying the Data Skills Gap from the Department for Digital Culture Media & Sport and Department for Science Innovation & Technology.
- Data skills competency frameworks developed by organisations such as the Open Data Institute, Organisation for Economic Co-operation and Development (OECD) and the United Nations, as well as other published data skills frameworks.
- Data Skills Training course curricula for example the ONS Data Science Campus and the Data Literacy Project.
- Unpublished reports from two undergraduate students who undertook Data Fellowships at the University of Manchester (Siyi Lu and A'Aishah Patel, supervised by Professor Jackie Carter) to conduct an analysis of academic literature relating to the development of data skills and an analysis of job advertisements seeking “data analysts”. We also reviewed Professor Jackie Carter's Research Skills Framework for SHAPE (Science, Humanities and the Arts for People and the Economy) students.

The full list of documents and literature reviewed can be found at the end of this paper.

We asked the question: what does the modern social scientist need to know, now, and as far as we can tell, into the future? We analysed these documents to collate and tabulate an initial framework of data skills and methods pertinent to the kinds of data that formed the basis for this Framework, as identified by the wide range of scholars, institutions, and data leaders represented in these documents. We then organised these into thematic skills, amalgamating some into higher order skills, and grouping and regrouping skills identified in ways that made sense. Using this initial framework, we then consulted widely and iteratively with experts in data skills and training from across the UK Data Service Training and User Support Directorate (see Authors and Contributors page) over a three-month period. The 16-strong team contributing to the analysis included experts with a wide range of data skills topic and expertise, including survey and census methodology and analyses, secure data, research data management, computational social science, international aggregate data, and many

---

years of experience of delivering data skills training in research and educational environments. In addition, we collected informal information from networks, meetings, and conferences. This led to discussion across the Directorate and several iterations of the skills matrix. The final Framework as presented below was reached by consensus across the Directorate.

## The UK Data Service Data Skills Framework

The UK Data Service Data Skills Framework is set out in Figure 2. It comprises 12 high level skill sets, which we term domains, each of which has then been expanded into its detailed constituent components.

The Framework comprises 12 broadly distinct elements:

1. Understanding what data is and what kinds of data there are.
2. Understanding who collects and owns data, how data are collected, how and where data are archived, and the formats for storing data.
3. Secondary data analysis research project management.
4. Research data management.
5. Legal and ethical issues arising in data collection, storage, curation, access and outputs.
6. Planning research, finding and accessing data for secondary data analysis.
7. Reproducibility.
8. Coding skills for data access, data manipulation and data analysis.
9. Preparing data for analysis and data manipulation.
10. Skills for analysing census, large scale survey and macro-data, including linked data.
11. Specific dataset skills for important data (here we have tabulated those in the UK Data Service collection).
12. Reporting data.

As noted in the introduction, the Framework acknowledges a rapidly evolving data skills landscape. It emphasises continued development of traditional data skills for contemporary research needs, while recognising growing potential for integrating survey or census data with an expanding array of other sources increasingly accessible to social scientists, as well as promising opportunities presented by AI and machine learning for enhancing analysis.

Under each Framework element, we have expanded the skills that have been identified as needing to be required to become competent to work with data in that domain. This has resulted in between 9 and 19 sub-skills tabulated across each of the domains.

We demonstrate in this analysis the deepening and widening array of data skills needed by social scientists in the modern world to comprehend, collect, access, prepare, work with, and archive, these kinds of population, social and economic data. Many of these skills are now complementary and supplementary to the mathematical and statistical skills needed to produce knowledge from the data once it is in condition to be analysed. These analytical skills are constituted under two domains: 10 (skills for analysing these data) and 11 (dataset-

---

specific skills). We also detail an expanding array of skills needed for reporting and reproducibility in research.

Some training needs span domains. For example, AI and other machine learning technologies appear in the Framework under: 5 (ethics), 8 (coding skills), 9 (data preparation), and 10 (data analysis). Linking data appears under: 2 (collection and storage of data), 9 (preparing data for analysis and data manipulation), and 11 (linking surveys with other types of data).

Understanding of linked data will also be relevant for disclosure risk assessment which appears in domains 4 (research data management), 5 (legal and ethical issues), and domain 10 (skills for analysis). Some training needs are specific to domains. For example, diverse methodological expertise in various modelling and statistical analysis appears only under domain 10 (skills for analysing census, large scale survey, and macro-data, including linked data).

It is important to note (perhaps obviously) not everyone seeking to work with UK Data Service or similar types of data will need all these skills - it will depend on their role, which data they want to use, and what analysis they want to do. Also, many skills encompass layers of expertise from what you need to know as a beginner, to what those at the cutting-edge of advancing the field can do. We have not at this stage sought to identify these layers – some will be obvious but others more subtle. Skills may have dependencies – you may need to learn some before others, not only within a skill but also across skills, and we have not at this stage attempted to map or represent these dependencies. Additionally, some skills might be considered 'core' - so fundamental that no-one should be accessing data without them. We have not at this stage attempted to identify these. We envisage that this current Framework might evolve into a framework with these additional layers of sophistication, and we can see that this might be very useful, particularly if it is used for curriculum and pathway development.

The skills required are numerous and complex, and responsibilities for their deployment will inevitably be (and need to be) delivered across a wide array of training providers, across a mixed economy of provision. The Framework can help us to think about who delivers what across the training ecosystem, how this is optimally distributed, and appropriate funding models. It also helps us to think collectively, as a community, about what elements of which domains need to be brought together for specific purposes, for specific people, at specific points in time. Appropriate elements can be selected from the Framework and re-assembled in configurations to support learning pathways for topics for instance a 'Using Longitudinal Data' pathway or a 'Data ethics' pathway. It can also be used for mapping and understanding contingencies, and potentially for thinking about training needs for different roles such as data owners, data producers, educators, researchers, the data skills workforce, and students at all stages of knowledge and understanding. The Framework helps to think about these issues holistically, as well as to understand knowledge and skills gaps.



## A High Level Framework

**AIM**

This initiative aims to establish a robust framework for developing essential data analysis skills for the social, economic and population sciences, focusing on large-scale survey, census and macro-level data characteristic of the UK Data Service (UKDS) holdings. The framework acknowledges a rapidly evolving data skills landscape. It emphasises continued development of

traditional data skills for contemporary research needs, while recognising growing potential for integrating survey or census data with an expanding array of other sources increasingly accessible to social scientists, as well as promising opportunities presented by AI and machine learning for enhancing analysis.

**1**

Understanding what data is and what kinds of data there are

**2**

Understanding who collects and owns data, how data are collected, how and where data are archived, and the formats for storing data

**3**

Secondary data analysis research project management

**4**

Research data management

**5**

Legal and ethical issues arising in data collection, storage, curation, access and outputs

**6**

Planning research, finding and accessing data for secondary data analysis

**7**

Reproducibility

**8**

Coding skills for data access, data manipulation and data analysis

**9**

Preparing data for analysis and data manipulation

**10**

Skills for analysing census, large scale survey and macro-data, including linked data

**11**

Specific dataset skills for important data in the UKDS data collections

**12**

Reporting data



## Detailed framework content

1

Understanding what data is and what kinds of data there are

- 1.1 The philosophy of data (shape, purpose, use, etc.)
- 1.2 Data structures
- 1.3 Different types of data
- 1.4 Personal and sensitive data
- 1.5 Census data
- 1.6 Cross-sectional and repeated cross-sectional data
- 1.7 Longitudinal data and retrospective longitudinal data
- 1.8 Panel and cohort studies; balanced and unbalanced panels
- 1.9 Experiments and natural experiments
- 1.10 Time use data
- 1.11 Location data
- 1.12 Macrodata (including international aggregate data)
- 1.13 Qualitative data in archived holdings and within large scale surveys
- 1.14 Administrative and routinely collected data
- 1.15 Smart data/digital footprint data including wearables, sensors, tracking data, energy and telecoms usage, transactions, consumer data, social media, web footprint and internet of things
- 1.16 Web data
- 1.17 Synthetic data

2

Understanding who collects and owns data, how data are collected, how and where data are archived, and the formats for storing data

- 2.1 Philosophy, epistemology and politics of data collection
- 2.2 Collecting different types of data; methods for data collection; secondary versus primary data collection; mandatory versus consented data
- 2.3 Institutional versus sectoral archives
- 2.4 Sampling and sample designs; non-probabilistic samples; stratification; clustering; survey design effects; attrition on longitudinal surveys
- 2.5 Survey and questionnaire design
- 2.6 Measurement error and total measurement error
- 2.7 Survey data structure including hierarchical data files, long v wide formats, how data from different types of survey questions are stored (including variable types, scales, choice and multi-choice, household structures)
- 2.8 Linking variables
- 2.9 Anonymisation and pseudonymisation
- 2.10 Metadata and data vocabularies
- 2.11 Interoperability

## Detailed framework content

3

## Secondary data analysis research project management

- 3.1 Funders and sponsors
- 3.2 Partners including data sharing agreements
- 3.3 Data management plans
- 3.4 Managing time
- 3.5 Managing budgets
- 3.6 Managing ethics
- 3.7 Managing processes
- 3.8 Managing people
- 3.9 Version control
- 3.10 Co-production approaches to research
- 3.11 Planning for impact

4

## Research data management

- 4.1 Open Science, FAIR data principles and responsible repositories
- 4.2 Research data lifecycle, benefits, challenges and solutions using RDM best practices
- 4.3 Data management planning including risk assessment and risk management
- 4.4 Data collection and acquisition (primary and secondary data RDM considerations - high level)
- 4.5 Legal and ethical issues arising in data collection, storage, curation, access and outputs [see separate skill-set]
- 4.6 Data storage, back-up and disposal (including collaborative research)
- 4.7 Data quality assurance
- 4.8 Data curation
- 4.9 Disclosure risk assessment, pseudoanonymisation and anonymisation
- 4.10 Documenting data and metadata management
- 4.11 Data preservation and archival management; data destruction
- 4.12 Interoperability in archival management
- 4.13 Data governance frameworks (licence and access frameworks including TRES)
- 4.14 Data sharing, compliance and data accessibility statements

## Detailed framework content

5

Legal and ethical issues arising in data collection, storage, curation, access and outputs

- 5.1 Legal and ethical frameworks applicable to research data
- 5.2 Ethical considerations in data collection methods, consent, privacy and transparency
- 5.3 Personal data, applicable legislation and key considerations
- 5.4 Understanding the legal and ethical frameworks for data storage, sharing, access, and researcher accreditation
- 5.5 Ethics in access to and accessing data
- 5.6 Ethical considerations in data analysis: integrity, transparency, reproducibility
- 5.7 The ethics of algorithms
- 5.8 The ethics of using AI
- 5.9 Licensing, IPR and other rights (including data citation and attribution)
- 5.10 Data security, disclosure risk, and dealing with data breaches

6

Planning research, finding and accessing data for secondary data analysis

- 6.1 Understanding the research process
- 6.2 Reviewing what is known and not known
- 6.3 Using secondary data analysis for research questions
- 6.4 Epistemological issues in secondary data analysis
- 6.5 Formulating a researchable research question
- 6.6 Operationalising a research question
- 6.7 Where and how to find data
- 6.8 Conducting catalogue searches
- 6.9 Choosing data and methods; single, multiple and mixed methods
- 6.10 Understanding data quality and data limitations
- 6.11 Applying for data access and data access conditions
- 6.12 Accessing census data
- 6.13 Accessing open, safeguarded and secure data
- 6.14 Trusted and secure research environments; researcher accreditation
- 6.15 Teaching datasets
- 6.16 Downloading data; remote job submissions; scraping data
- 6.17 Sub-setting
- 6.18 Finding and using data documentation (codebooks etc)
- 6.19 Tools and web-tools for browsing and exploring data

## Detailed framework content

7

## Reproducibility

- 7.1 Understanding reproducibility and the benefits of sharing code
- 7.2 Documenting code, processes, decisions and data citation
- 7.3 Capturing insights and ideas
- 7.4 Version control
- 7.5 Storing and sharing code; dynamic documents (e.g. Markdown, Quarto etc)
- 7.6 Accessing stored code
- 7.7 Journal and publication requirements for reproducibility
- 7.8 DOIs and supplemental resources
- 7.9 Acknowledgements and work role attribution
- 7.10 Citation and referencing

8

Coding skills for data access, data manipulation  
and data analysis

- 8.1 Software for accessing and analysing social science data (including SQL, SPSS, Stata, R, Python, specialist software for e.g. multi-level modelling, network analysis, geo-spatial analysis)
- 8.2 Choosing software
- 8.3 Open source programming languages
- 8.4 Open source Jupyter notebooks (and archive specific notebooks like BigQuery to run SQL and Vertex AI Notebooks that run Python and R to access the Integrated Data Service)
- 8.5 Principles of good coding
- 8.6 Importing and exporting data
- 8.7 APIs
- 8.8 Coding for data manipulation; data analysis; data visualisation; simulation
- 8.9 Using AI to write code

## Detailed framework content

9

## Preparing data for analysis and data manipulation

- 9.1 Understanding and using complex surveys
- 9.2 Importing and exporting data
- 9.3 Cleaning data
- 9.4 Anonymisation and pseudonymisation of data
- 9.5 Sub-setting data
- 9.6 Missing data; bias and non-response; weighting; data imputation
- 9.7 Recoding data; transforming data; variable construction; restructuring data
- 9.8 Automating complex data manipulation
- 9.9 Using AI for data manipulation
- 9.10 Linking data and combining datasets
- 9.11 Linking survey data with other types of data  
(e.g. administrative and routine data, streaming, wearables and sensors)
- 9.12 Working with synthetic data

10

Skills for analysing census, large scale survey and macro-data,  
including linked data

- 10.1 Evaluating data and assessing data quality
- 10.2 Understanding variable types
- 10.3 Descriptive and exploratory data analysis; summarising data; understanding denominators
- 10.4 Data visualisation
- 10.5 Analysing data from complex samples
- 10.6 Analysing different types of data including cross-sectional data; longitudinal data; macrodata; streaming, wearables and sensor data; web data; administrative and routinely collected data
- 10.7 Analysing (and analysing with) missing data
- 10.8 Statistical analysis and modelling, including network analysis, geo-spatial analysis, simulation and computational social science models
- 10.9 Analysing change over time; understanding age, period and cohort effects
- 10.10 Machine learning (quantitative and qualitative data)
- 10.11 Using AI for data analysis (quantitative and qualitative)
- 10.12 Using synthetic data
- 10.13 Understanding the limitations of data analysis

## Detailed framework content

11

Specific dataset skills for important data  
in the UKDS data collections

- 11.1 Census Data
- 11.2 Understanding Society and the British Household Panel Study
- 11.3 The Cohort studies
- 11.4 The Labour Force Surveys
- 11.5 The Health surveys
- 11.6 The Crime Surveys
- 11.7 The Social Attitudes Surveys
- 11.8 The Personal Finance Surveys
- 11.9 The English Longitudinal Study of Ageing
- 11.10 Time Use Surveys
- 11.11 The Housing Surveys
- 11.12 Aggregate data (e.g. international datasets)
- 11.12 Linked data linking UKDS surveys to.... [various]

12

## Reporting data

- 12.1 Authorship and acknowledgements
- 12.2 Sharing and disseminating results
- 12.3 Citing data
- 12.4 Creating DOIs
- 12.5 Managing disclosure risks
- 12.6 Copyright [publishing; secondary data analysis]
- 12.7 Presenting data analysis to different audiences
- 12.8 Publishing and open access

## Next steps

While the UK Data Skills Framework was developed for us to better understand the data skills training ecosystem in which we play our part, we hope that it will be of use and interest to the global data skills training community. We plan to use the analysis to develop the UK Data Service training curriculum across the Training and User Support Directorate, to drive our analysis of skills training gaps, to scaffold our thinking about the development of data skills pathways, and to support our ongoing development of our suite of data skills modules. We will use it in thinking about and planning existing and new training collaborations across institutions and investments, including with established and emerging data services, as well as nascent collaboration with the ESRC Doctoral Training Partnerships.

We would very much welcome discussion and feedback from other training providers, from the data user community, from data producers, funders and other population, social and economic data services, as well as anyone with an interest in data skills pedagogy and skills development. We see the Framework as a flexible document that can evolve over time, not only in terms of sophistication – with potential layers of skills levels, dependencies, and the definition by consensus of core elements, but also in terms of the skills themselves, in our every-changing data environment.

We hope that others who operate in adjacent or different skills domains, such as with other forms of data (for example health data, administrative data, smart data) or in other paradigms, such as qualitative research methodologies, might be able to use what is here, but also build new branches, to create a more complex map of data skills needs in the current environment.

We also see the Framework potentially having a role in aiding in consensus building across the training provider community, potentially moving towards harmonised metadata, controlled vocabularies, and increased interoperability across training providers in the social scientific data ecosystem.

If you have any thoughts on the material presented in this Discussion Paper, please share them with us on [ukdstraining@manchester.ac.uk](mailto:ukdstraining@manchester.ac.uk).



---

## Documents

Carter, J. (2021) *Work Placements, Internships & Applied Social Research*. SAGE Publications Ltd.

Central Digital and Data Office (Accessed 2023). [Government Digital and Data Profession Capability Framework](#)

CRAC (2017) *NERC Doctoral Training Partnerships: Mid-term evaluation*. Cambridge: Careers Research & Advisory Centre Ltd.

Data Science Campus (Accessed 2021) [Course catalogue | Data Science Campus](#)

Data Skills For Work (Accessed 2023). [Data Skills Framework](#).

Data Skills Taskforce (Accessed 2021) [Data Skills Taskforce](#)

Davis-Kean, P., Chambers, R.L., Davidson, L.L., Kleinert, C., Ren, Q., Tang, S. (2017) 'Longitudinal Studies Strategic Review', ESRC.

Deloitte & Tech UK (2020) *Digital Skills: Establishing a digital learning pathway*. London: Deloitte.

Department for Digital Culture Media & Sport and Department for Science Innovation & Technology (2021) Quantifying the UK Data Skills Gap.

Department for Digital Culture, Media & Sport (2020) National Data Strategy.

Djumaliev, J., Sleeman, C. (2018) *Which digital skills do you really need? Exploring employer demand for digital skills and occupation growth prospects*. London: Nesta.

Economic and Social Research Council (2021) Review of the PhD in the Social Sciences: ESRC response.

Economic and Social Research Council (2022) Scoping the skills needs in the social sciences to support data-driven research skills across the academic career life course.

Economic and Social Research Council (2023) The skills needed in the social sciences to support data-driven research across the life course: ESRC response.

European Commission (2022) DigComp 2.2.

---

Ezell, S. (2021) *Assessing the State of Digital Skills in the U.S. Economy*. Information Technology and Innovation Foundation.

Government Analysis Function. (Accessed 2023) [Career Framework](#)

Government Analysis Function (2021) Competency framework for the Government Statistician Group.

Government Social Research Profession (2017) The Government Social Research Technical Framework.

Heeley, L., & Wilkinson, S. (2018) *Data Skills Framework: A generic approach to assessing and developing data-related competencies and skills*.

Krčál, A. and Bryan, B. (2018) *Q-Step Benchmarking*. London: Nuffield Foundation.

Jisc (Accessed 2023) [Digital Capability Framework](#)

Luff, R., Durrant, G., Wiles, R., & Crow, G. (2015) Consultation on Training Needs in Advanced Social Science Research Methods. NCRM.

Maddock, J., O'Neill, D., Rainsberry, M., & Hardy, R. (2019) *A review of quantitative analytical training needs of users of longitudinal studies*. London: CLOSER.

Meckin, R. and Elliot, M. (2021) *Computational Social Science: A Thematic Review*. Other. National Centre for Research Methods.

National Centre for Research Methods (2021) Interface Report: Health and social science report.

OECD Science Technology and Innovation (2020) *Building Digital Workforce Capacity and Skills for Data Intensive Science*. OECD.

Office of Human Resource Management, United Nations (2010) *United Nations Competency Development – A Practical Guide*.

Papamitsiou, Z., Filippakis, M. E., Poulou, M., Sampson, D., Ifenthaler, D., & Giannakos, M. (2021) 'Towards an educational data literacy framework: enhancing the profiles of instructional designers and e-tutors of online and blended courses with new competences', *Smart Learning Environments*, 8(1), 18.

---

Pothier, W. G. & Condon, P. B. (2019) 'Towards Data Literacy Competencies: business students, workforce needs, and the role of the librarian', *Journal of Business & Finance Librarianship*, 119, pp. 123-146.

Office for Health Improvement and Disparities (2016) [Public Health Skills and Knowledge Framework](#)

Schüller, K. (2022) 'Data and AI Literacy for everyone', *Statistical Journal of the IAOS*, 38(2), pp. 477-490. [DOI: 10.3233/SJI-220941](#).

Schüller, K. (2020) *Future Skills: A Framework for Data Literacy*. Working Paper No. 53. Berlin: Hochschulforum Digitalisierung. DOI: 10.5281/zenodo.3946067.

Tazzyman, S., Birkin, G., Bowes, L., Hancock, S., Wakeling, P. (2020) *Review of the PhD in Social Sciences: Rapid Evidence Assessment*. Leicester: CFE Research.

The Data Literacy Project (Accessed 2021) [The Data Literacy Project](#)

UK Research and Innovation (Accessed 2022) [UKRI Digital research infrastructure – UKRI](#)

UK Research and Innovation – ESRC (2023) Scoping the skills needs in the social sciences to support data-driven research skills across the academic career life course.

---

[www.ukdataservice.ac.uk](http://www.ukdataservice.ac.uk)

[help@ukdataservice.ac.uk](mailto:help@ukdataservice.ac.uk)

+44 (0) 1206 872143

We are supported by the University of Essex, University of Manchester, Jisc, UCL and University of Edinburgh. We are funded by UKRI through the Economic and Social Research Council.

This work was supported by the Economic and Research Council grant ES/Y005538/1.