

# Use of data in safe havens: ethics and reproducibility issues

Louise Corti  
Service Director Data Publishing and Access  
UK Data Service

UKRIO Research Integrity Webinar: Data Sharing and Ethics  
Online  
11 November 2020

# The UK Data Service

- Social science data service, funded by the UKRI ESRC
- Curates and provides access to data for research and teaching
- Trusted Digital Repository (TDR), accredited to ISO27001 Info Security Management standard, Digital Economy Act Processor application in process
- Work closely with research funders and key data producers /institutions: research centres; UK NSIs, govt departments, British Library etc
- Around 8000 data collections, from open to secure



# Types of data held

## Social survey

- One off large scale survey e.g. Health Survey for England

## Longitudinal/cohort

- Detailed survey following people over time

## Qualitative

- Interviews, focus groups, diaries

## Aggregate data

- UK census counts/tables, country level statistics

## Historical data

- Digitised databases

# Spectrum of Access: UK Data Service

## Open

- No personal identification risk
- Open licence; few restrictions on reuse

## Safeguarded

- Zero to low risk of personal identification
- Authentication and authorisation

## Controlled

- Risk of personal identification
- Authentication and authorisation
- Added safeguards

# Safeguards for data access

- ✓ Data access involves **reduction of risk** in a manner acceptable to the data owner
- ✓ Risk mitigated by **legal gateways** and **appropriate safeguards**
- ✓ **5 Safes Framework** offers a portfolio of safeguards

# Legal gateways to access

- Government Departments, Local Authorities, agencies, public bodies face different **legal restrictions on the nature of the access** they can provide for research
- In some cases, **specific legal gateways** have been drafted to facilitate data sharing
- This can have an impact on the **feasibility of some research projects and data linking** between various data sources e.g. health and social data is often hard to link

# Legislation and legal gateways

- Digital Economy Act (DEA)
- Data Protection Act /General Data Protection Regulation (GDPR)
- Statistics and Registration Service Act (SRSA)

## Gateways:

- ✓ Person-specific ( ‘fit and proper’)
  - ✓ Project-specific
  - ✓ Time-specific
  - ✓ Dataset-specific
- 
- Disclosing the identity or identifying information to someone who should not have access is a breach of the law. Civil as well as criminal sanctions for breaches

# Digital Economy Act, 2017

- The DEA covers: [Access to digital services](#), [Digital infrastructure](#), [Online pornography](#), [Intellectual Property](#) and [Digital government](#)
- Has a useful **Research strand** that broadly enables de-identified information held by a public authority to be **disclosed for the purposes of research in the public good**
- The research framework is underpinned by the [Research Code of Practice and Accreditation Criteria](#), approved by the UK Parliament in July 2018
- The UK Statistics Authority is the statutory accrediting body
- [Research Accreditation Panel](#) to oversee the independent accreditation of processors, researchers and research projects



# Digital Economy Act data example

## Quarterly Labour Force Survey, 1992-2020: Secure Access

[Details](#)[Documentation](#)[Resources](#)[Access data](#)

### Details

Title:	Quarterly Labour Force Survey, 1992-2020: Secure Access
Alternative title:	QLFS
Study number (SN):	6727
Access:	These data are <a href="#">controlled</a>
Persistent identifier (DOI):	<a href="https://doi.org/10.5255/UKDA-SN-6727-21">10.5255/UKDA-SN-6727-21</a>
Series:	<a href="#">Labour Force Survey</a>
Principal investigator(s):	Office for National Statistics, Social Survey Division Northern Ireland Statistics and Research Agency, Central Survey Unit

### Sponsors and contributors

### Citation and copyright

The citation for this study is:

Office for National Statistics, Social Survey Division, Northern Ireland Statistics and Research Agency, Central Survey Unit. (2020). *Quarterly Labour Force Survey, 1992-2020: Secure Access*. [data collection]. 20th Edition. UK Data Service. SN: 6727, [http://doi.org/10.5255/UKDA-SN-6727-21](https://doi.org/10.5255/UKDA-SN-6727-21)

# What makes this controlled data?

- Low level of geography
- Date of birth, including day
- Education and training: detailed training and qualifications
- Household and family characteristics: detail on family and extended family units
- Employment: industry code of main job, employment details
- Unemployment and job hunting: details
- Temporary leave from work
- Accidents at work and work-related health problems
- Nationality, national identity and country of birth
- Occurrence of learning difficulty or disability
- Benefits, including detail on type of benefits claimed

# Controlled data

## Data access/ availability

- 5 SAFES framework (ONS)
- Access only via an accredited (safe haven)
- High access bar
- *Process for reproduction not set up by journals*



## Secure Lab services

Enabling secure research access to data while protecting confidentiality

### The Five Safes Framework



Safe data



Safe projects



Safe people



Safe settings



Safe outputs

As promoted by the Office for National Statistics

[ukdataservice.ac.uk](https://ukdataservice.ac.uk)

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



# 5 Safes framework for Safe Data Access

Is this an appropriate use of the data?

**SAFE PROJECTS**

+

How trustworthy are the researchers?

**SAFE PEOPLE**

+

Does the environment prevent misuse?

**SAFE SETTING**

+

Is the data detail appropriate?

**SAFE DATA**

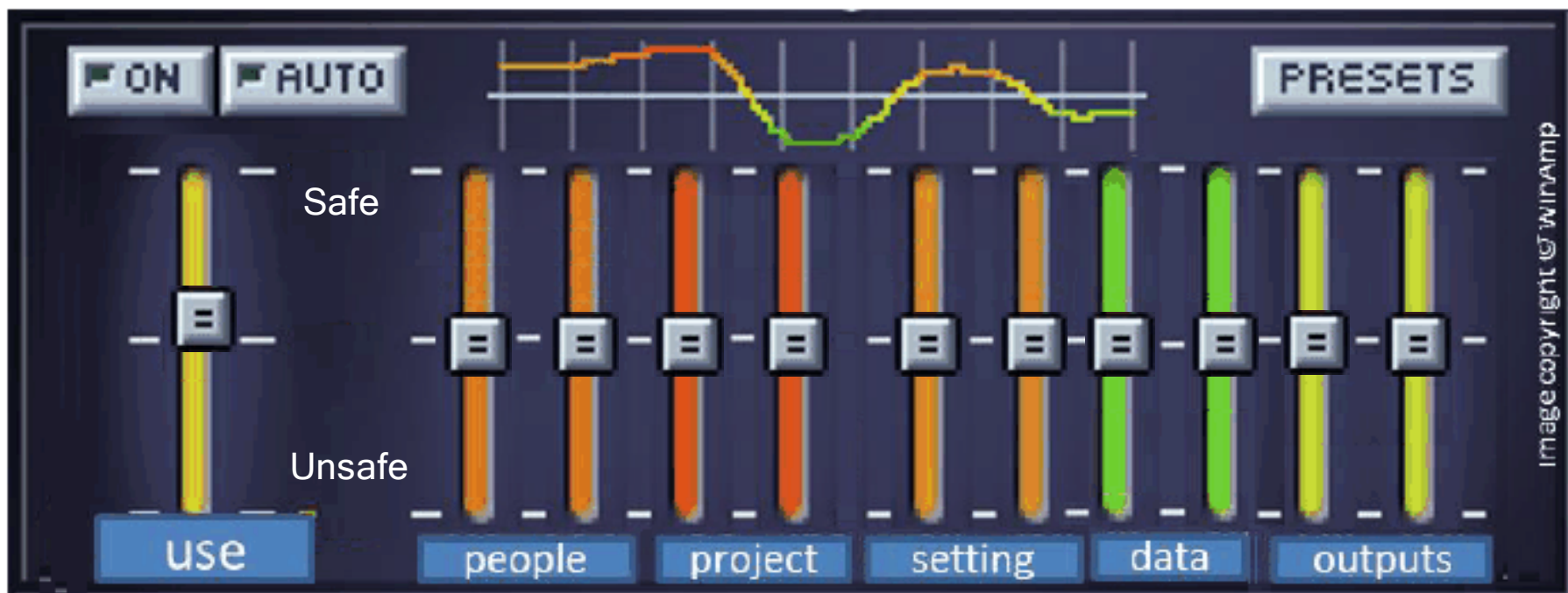
+

Is there any confidentiality risk from publication?

**SAFE OUTPUTS**

**= SAFE USE**

# Balancing the Five Safes: Example



Adapted by the Australian Bureau of Statistics from 'Five Safes: designing data access for research', Desai, T, Ritchie, F and Welpton, R, 2016



# DEA Processors



Accreditation process =  
Information Security + Capability

Northern Ireland Statistics and Research Agency Research Support Unit



Office for National Statistics



ONS Secure Research Service



Secure Anonymised Information Linkage Databank



UK Secure eResearch Platform



Northern Ireland Statistics and Research Agency Census Office



UK Data Archive



Edinburgh Parallel Computing Centre



<https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/better-access-to-data-for-research-information-for-processors/>



# DEA Accredited Researchers

## List of accredited researchers and research projects under the Research Strand of the Digital Economy Act

Digital Economy Act (DEA) researchers are individuals to whom de-identified data is made available for research. DEA research projects are projects that have been approved by the Research Accreditation Panel.

The table below presents a list of accredited research projects and the accredited researchers who are associated with each research project.

Show  entries

Search:

DEA Accredited Project Name	Accredited Researchers	DEA Accredited Data Sources	DEA Accredited Processor	Project Accreditation Date
The fall of the labour share and rise of the superstar region? Labour share and earning in UK regions	Carolin Ioramashvili, London School of Economics	Office for National Statistics: Annual Respondents Database (1998-)	Office for National Statistics Secure Research Service	25 October 2019
The changing nature of the HR and training practitioner profession	Jonathan Boys, Chartered Institute of Personnel and Development	Office for National Statistics: Annual Population Survey (2004-17)	Office for National Statistics Secure Research Service	25 October 2019
Analysis of victimisation data from the Crime Survey for England and Wales	Julian Molina, Office of the Victims' Commissioner for England and Wales	Office for National Statistics: Crime Survey for England and Wales (2013-18)	Office for National Statistics Secure Research Service	14 October 2019

Accreditation process = Safe Researcher Training + Test + Application

Understanding of disclosure risk + producing Safe Outputs

Good attitudes and behaviour = Trust





## Accessing Data via the Digital Economy Act Research Strand

The Research strand of the Digital Economy Act broadly enables de-identified information held by a public authority to be disclosed for the purposes of research in the public good. The research framework in the [Digital Economy Act 2017](#) [↗](#) is underpinned by the [Research Code of Practice and Accreditation Criteria](#) [↗](#), which was approved by the UK Parliament in July 2018. As the statutory accrediting body, the UK Statistics Authority has established a [Research Accreditation Panel](#) to oversee the independent accreditation of processors, researchers and research projects.

### Submitting a research project application to the Research Accreditation Panel

The Research Accreditation Panel meets regularly. Research project applications are submitted to the Research Accreditation Panel by [accredited data processing environments](#) at least one week prior to each Research Accreditation Panel meeting. Researchers are advised to liaise with support teams in their chosen accredited data processing environment to understand the requirements and timelines involved with submitting a research project to the Research Accreditation Panel.

Guidance for researchers completing research project applications to access data via the Digital Economy Act Research strand is available in the 'Documents for Download' section. Researchers are advised to use the project application example guidance to understand the information that is required for the Research Accreditation Panel to consider a research project application.



## 9 Research Methodology:

*Provide details of the research protocol or methodology (e.g. data linkage or matching, web scraping etc) and how you intend to use the data, in no more than 1000 words.*

Clearly **justify why you require access** to legally protected data. If applying for access to multiple datasets, explain how these will be used together to address your research aims.

The data from the small business surveys that we request will complement our use of administrative tax records and firm accounts data. Administrative records are informative because they contain information on the population (rather than a sample) of interest, and well-measured information on incomes and the use of tax allowances. However, they do not measure other useful information contained in surveys, such as type of financing used, plans for future investments, reasons why certain decisions are taken etc. By combining information from the two sources, we will be able to get a richer picture of UK small business owner behaviour.

We will bring analysis from the different data sources by **matching aggregate research outputs together** through the use of **an economic model**, and not by data linkage. We are in the process of building a dynamic model of business owner decisions, including whether to start a business, whether to incorporate, how much to invest, and when to exit. This methodology will allow us to explicitly account for the dynamic aspects of individuals' decisions.

We will estimate the model by matching 'moments' in the model with the same **moments observed in the data**. 'Moments' are features of the data, for example, the mean, standard deviation, or other transformations of the data. In this application, we will use moments that are informative about the behaviours that we are trying to analyse e.g. the amount of investment, propensity to incorporate, distributions of profit and taxable income. We can use moments from multiple data sources (including the business survey data requested here) to estimate the model, without having to link the data at the observation level. Moments that we anticipate will be useful from the business survey data include: plans to launch new products, R&D expenditure, type of finance used, among others.

We also plan to exploit policy features and reforms over our period of study – for example, the introduction of the additional rate of tax on incomes above £150,000, the withdrawal of the personal allowance for those earning more than £100,000, changes to investment allowances, and corporate tax reforms. **We will use this variation to help identify how small business owners respond to various parts of the tax system.**

Once we have estimated the model, we can perform counterfactual analysis to look at the effect of various policy reforms. For instance, what would be the implications of taxing capital income at the same rates as labour income, or removing the flexibility for business owners to shift taxable income across years. **This will allow us to provide important evidence on the potential costs and benefits of various policies that target small business owners.**

Specify the **methods of analysis** you plan to use. If you are proposing a non-standard methodology, provide methodology references. The **level of detail provided should enable an experienced researcher to understand** the type of model, how bias is being addressed and what is likely to constitute an important effect.

Include an explanation of **how your methodological approach will answer the research questions** set out in the project.

Indicate the **starting point** for the **research and/or modelling process**. Include statistical tests that are likely to be used and the types of models to be employed.

### 13 Public Good:

13.1 Please describe how your research project will provide a public good. Complete all the sections that apply.

Public Good	Describe how this research project will provide this public good
Provide an evidence base for public policy decision-making	Small business owners are an increasingly important part of the workforce, and often attract preferential tax treatment and are targeted by a range of other policies. These are often justified by the desire to promote economic growth and boost productivity, but can lead to costly tax avoidance and a misallocation of people and capital. This research project will provide vital evidence on how policy affects a range of business outcomes, and the potential impact of various policy reforms.
Provide an evidence base for public service delivery	
Provide an evidence base for decisions which are likely to significantly benefit the UK economy, society or quality of life of people in the UK	We will be able to quantify the costs and benefits of various existing policies and potential reforms. These have important implications for the government finances and the wider economy. Favourable tax treatment of small business can be costly – for example, <b>Entrepreneurs' Relief</b> (a reduced rate of capital gains tax applied to business assets) is estimated to cost the government in excess of £2.4 billion a year.
To replicate, validate or challenge Official Statistics	
To replicate, validate or challenge existing research	
To significantly extend understanding of social or economic trends or events by improving knowledge or challenging widely accepted analyses	Small business owners have accounted for 40% of growth in the workforce since 2008, so understanding how policy affects both this trend, and the economic impact of their growth is vitally important.
To improve the quality, coverage or presentation of existing statistical information	

Provide as much detail as possible about how your analysis will provide a public good, including **references to specific policies** or decisions that you anticipate the research will provide an evidence base for.

Ensure you **consider the wider impact of your research outputs** and explain how these outputs will inform policy and decision-making processes.

Only complete the public good statements that **apply to your research project**.

# Ethics component

- University REC submission for projects undertaking secondary analysis of data often not always complete enough to pass ethics test; often focussed on data collection
- Clear guidance designed by the National Statistician's Data Ethics Advisory Committee's (NSDEC) to support researchers and statisticians to complete an ethical self-assessment form

# Ethics Self-Assessment

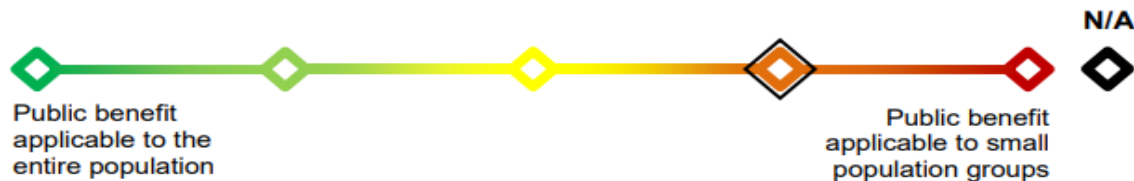
- UK Statistics Authority Ethics Self-Assessment Tool
- Easy-to-use framework to review the ethics of a project
- Helps identify & mitigate any ethical issues
- 6 main principles
  - ✓ Public good, potential harm
  - ✓ Identification, data security
  - ✓ Training, technologies
  - ✓ Legal gateways, frameworks
  - ✓ Public view, engagement
  - ✓ Access, use & sharing of data



# The Ethics Self-Assessment form

- For each item – provide a score
- Enter the score in the spreadsheet
- The spreadsheet calculates an overall score of risk level

## 2. Population coverage



			Please type "1" in the appropriate box below to indicate your rating for each question. Please add a comment in the far right column to briefly justify your selection						Score per principle	Please use this section to give a short justification of your selected rating
Item	Principle 1	Score	1	2	3	4	5	N/A	1.3	Score Justification
1	Public benefit	1	1							
2	Population coverage	2		1						
3	Potential harm	1	1							
4	Biases	1	1							



# Reproducibility factors in safe havens

- ✓ **Impact** of the research outputs is a key requirement
- ✓ Code can be taken out, but is **reviewed for risk**
- ✗ **Hard to reproduce** as data and code behind a big gate
- ✗ Reproducer needs **Accredited Researcher status**
- ✗ **Requires mechanism** for reproducer to be added to an accredited project
- ✗ **Difficult provenance chain** for 'research-ready' admin data (creation, cleaning and versioning)
- ✗ **Journals not ready** to enable this review process

# Using the 5 Safes network

- ✓ Strong **5 Safes framework** for access to unsafe data
- ✓ Very simple ethics assessment – easy to mark
- ✓ Utilise **Accredited Researchers** to undertake reproducibility work
- ✓ Better training on **how to be reproducible**
- ✓ Advocate cleaning, value-added, final code available in safe haven to enable **reproduction** and **reuse**

# Questions

Louise Corti

UK Data Service

[corti@essex.ac.uk](mailto:corti@essex.ac.uk)

