

The UK Data Service: managing and sharing large archives and social sciences data

Louise Corti, Director Collections Development
and Data Publishing, UK Data Service

Wellcome Data Managers Group

28 September 2017

London

UK Data Service



 DataFirst



Overview

- ✓ Overview of the UK Data Archive/Service
- ✓ Data sharing **support and access** frameworks
- ✓ Cohort & longitudinal studies with biomedical components
- ✓ **Enhancements** to longitudinal data e.g. HDSS type data

What is the UK Data Archive

- Department of the **University of Essex**
- Established in 1967 by the Economic and Social Research Council (ESRC) as a '**Data Bank**' for science science
- **50 years** of curating and providing access to data
- Data and support services for **research, teaching and learning**
- Speciality in **social survey data, qualitative data, historical databases, some biomedical** and coming...**big data!**
- Leads the ESRC UK Data Service



University of Essex

The Archive



What does the UK Data Service do?

- ESRC supported National Data Service
- Put together a collection of the most valuable data
- Make data and documentation available for reuse
- Preserve data in the long term for future research
- Provide data management advice for data creators
- Provide support for users of the service
- Information about how data are used
- Easy access through website



UK Data Service

[About us](#)[Get data](#)[Use data](#)[Manage data](#)[Deposit data](#)[News and Events](#)

Welcome to the UK Data Service

Your resource for quality social research data

A unified point of access to data from ESDS, Census Programme, Secure Data Service and others



STOP

[LOGIN](#) / [REGISTER](#)

DISCOVER UK DATA SERVICE

Search and browse our data collection, support guides, case studies and related publications.



GO

☐ Data☐ Website

LATEST TWEETS



UKDataService RCUK's revised Policy on Open Access now online, to be in effect this month. <http://t.co/k3ouoFVe40> @research_uk #openaccess



UKDataService @Barnard17 There are lots of population statistics available via @ONS and @EU_Eurostat. Sent a list of links to your blog. Good luck!



UKDataService RT @Censusacuk: <http://t.co/qVRh4S0y2m> is now Census Support

LATEST NEWS



Call for papers: Opinions and Lifestyle Survey user meeting

European Social Survey invites EU data users to visit and learn

Digital Futures: Your input needed on data digitisation

UK Data Service featured in Royal

OUR DATA COMMUNITY



The UK Data Service is at the core of a network of trust that includes data owners, producers, funders and users.

Who can most benefit from the data we hold?

- Academic researchers and students
- Government analysts
- Charities and foundations

QUICK ACCESS TO

[Key data](#)[Census Support](#)[Information for new users](#)[Frequently asked questions](#)

Progressive data sharing in the UK

- **Concordat** with our National Statistical Institute
- Supporting the **ESRC Data Policy** since the 1960s
 - Role in **policing the policy** – compliance /data quality
 - Advise and support for **applicants** and **award holders**
 - Included qualitative and historical data from 1996
- First to publish RDM **guide for researchers** in 2005
- Our approach to making data shareable based on:
 - **everyday challenges** faced by researchers
 - experience in **handling** range of research data

Our data portfolio

UK Surveys

Large-scale government and academic funded surveys

Longitudinal

Major UK surveys following individuals over time

International

Multi-nation survey data

Census

Census data
1971 – 2011

Business

Microdata and administrative data

Qualitative

Range of multimedia qualitative data sources



Who is it for?

- academic researchers and students
- government analysts
- charities and foundations
- business consultants
- independent research centres
- think tanks
- citizen scientists, where skills enable analysis



Some statistics about our UK Service

Data for research and teaching purposes, used in all sectors and by many different disciplines

- **Over 7,237** data collections
- **1086** qualitative /mixed methods collections
- **400** new datasets and new editions added within last 12 months
- **Over 25,000** registered users
- **60,000** downloads worldwide per annum
- **4000+** user support queries per annum



Links with other data archives worldwide

Qualitative
data



Data archives

Open data

Secure
datalabs

Question
banks

Explore online

Data access
policy

Thesauri

European data
archives

North American
data archives

Other worldwide
data archives

SHARE 



Sharing data at the UKDS

UK Data Service



We use 2 data access/curation pathways:

- Our ‘**curated collection**’ - high impact data, e.g. government departments, major surveys
- UK Data Service **ReShare** – self-deposit system for academic and smaller research data

UK Data Archive - digital data preservation

-
-
-
-



curity

tion



UK Data Service acquisition

- Reactively and proactively acquire data
- Data deposited by:
 - UK government departments (contractual)
 - Research institutes and researchers
 - Third sector, NGOs
 - Companies
- Selection: Collections Development Policy
- Appraisal: Criteria and curation pathways

Data collection requisites

We ensure that data collections are:

- ✓ Long-term preserved and usable
- ✓ Self-explanatory for users
- ✓ Non disclosive where promised
- ✓ Rights are in place to redistribute

FAIR DATA:

FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE

Access conditions

Depositor selects, with guidance, the access category most appropriate for the data, under a license

Open

- available for download/online access under open licence without any registration

Safeguarded

- available for download/online access to logged-in users who have registered and agreed to an End User Licence

Controlled

- available for remote or safe room access registered users whose research proposal has been approved by an access committee and who have received specialist training

Safeguarded data – conditions of access

- Most common license choice
- Register with us - UK Federation
- Agree to an End User Licence
 - ✓ Appropriate data use
 - ✓ Full citation
 - ✓ Informing of use outputs
- Click 'Download' from catalogue
- Specify project for which the data are to be used
- Download data in choice of formats

370 Open Collections

Data discovery and access



Discover

Discover

Variable and question bank

QualiBank

Type

☒ Data collections (198)

☐ Case studies (2)

☐ Series (1)

☐ ESRC outputs (24)

☐ Support / how to guides (0)

[Refine](#)

Subject +

Date +

Data type +

Key data +

Country +

Data format +

Spatial unit +

Analysis unit +

Search and browse our data collections, support guides, case studies, and related publications.

India

GO

[Reset filters](#)

[Clear search](#)

☒ Auto-complete

[Advanced search](#)

[Help](#)



Case study



Data collection



Series record



ESRC output



Support guide

[Guide to icons](#)

Results per page: 10

Sorted by: Relevance

Displaying 1-10 of 198 results

1 2 3 4 5



SN 852597 **New urbanisms in India: Urban living, sustainability, everyday life**

Sophie Hadfield-Hill, University of Birmingham



[Full record...](#)

[Download](#) | [DDI XML](#) | [Similar data collections](#)



SN 7478 **Young Lives: School Survey, India, 2010-2011**

Solon, A., University of Oxford. Department of International Development



[Full record...](#)



[Download/Order](#) | [DDI XML](#) | [Similar data collections](#)



SN 5380 **Resisting Subjugation: Law and Power amongst the Santal of India and Bangladesh, 2002-2004**

Shariff, F., University of Warwick. School of Law



[Full record...](#)



[Download/Order](#) | [DDI XML](#) | [Similar data collections](#)

Quality assessment of data

- Assess data, metadata and documentation:
 - Data integrity
 - File, variable and value label metadata
 - Documentation coverage
- Set publishing standard - depends on likely use
 - Online data browsing tools requires high level of data enhancement
- Data anonymisation: disclosure review

Social science: data access strategy

Three pillars

- ✓ Informed consent for long-term data sharing
- ✓ Protection of identities when promise
- ✓ Regulated access where needed (all or part of data) e.g. by group, use, time period

Open where possible, closed when necessary

Spectrum: Open - Safeguarded - Controlled

Protecting confidentiality: the '5 Safes'

FIVE SAFES data protection

Approved Re
Health Founda

- Safe data -
- Safe people
- Safe projec
- Safe setting
- Safe output



t the needs of
transparency

, HMRC,

identity

afely

,

ata

ned

Assessment complete

- Processing level allocated (A*, A, B, C) - 30 days
- Data **enhancements**: edits made in consultation with depositor
- Generate **multiple data formats** for dissemination and preservation
- Assemble documentation for users
- Prepare catalogue record (DDI)
- Archived to **preservation system**
- Released as zip bundles to catalogue
- **DataCite DOI** assigned - versioned

Health Survey for England, 2014

[Documentation](#) | [Related Studies](#) | [Publications](#) | [Syntax](#)



[Access online](#)



[Download/Order](#)

[DDI XML](#)

TITLE DETAILS

SN: 7919

Title: Health Survey for England, 2014

Alternative title: HSE

Persistent identifier: [10.5255/UKDA-SN-7919-2](https://dx.doi.org/10.5255/UKDA-SN-7919-2)

Series: [Health Survey for England](#) [Health Survey for England, 1991-]

Depositor: NatCen Social Research

Principal investigator(s): NatCen Social Research
University College London. Department of Epidemiology and Public Health

Sponsor(s): Information Centre for Health and Social Care

CITATION

The citation for this study is:

NatCen Social Research, University College London. Department of Epidemiology and Public Health. (2016). *Health Survey for England, 2014*. [data collection]. 2nd Edition. UK Data Service. SN: 7919, [http://dx.doi.org/10.5255/UKDA-SN-7919-2](https://dx.doi.org/10.5255/UKDA-SN-7919-2)

Series catalogue record



Discover

Variable and question
bank

QualiBank

Series

UK Data Service series record for:

Health Survey for England

[Abstract](#) | [Access](#) | [Get started](#) | [FAQ](#) | [Related](#) | [Links](#) | [Search](#)

SERIES ABSTRACT

The Health Survey for England (HSE), sponsored by the Information Centre for Health and Social Care and the Department of Health, began in 1991 and has been carried out annually since then. A number of core questions are included in every wave but each year's survey also has a particular focus on a disease or condition or population group, which are subsequently revisited at appropriate intervals in order to monitor change. The survey combines questionnaire-based answers with physical measurements and the analysis of blood samples. Blood pressure, height and weight, smoking, drinking and general health are covered every year. An interview with each eligible person in the household is followed by a nurse visit.

DATA ACCESS

— [GN 33261](#) | [HEALTH SURVEY FOR ENGLAND, 1991-](#)

SN	Study Description	Access Online	Download / Order <input type="checkbox"/>
7919	Health Survey for England, 2014		<input type="checkbox"/>
7649	Health Survey for England, 2013		<input type="checkbox"/>
7480	Health Survey for England, 2012		<input type="checkbox"/>
7260	Health Survey for England, 2011		<input type="checkbox"/>
6986	Health Survey for England, 2010		<input type="checkbox"/>

ADMINISTRATIVE AND ACCESS INFORMATION

Date of release:

First edition: 02 March 2016

Latest edition: 07 November 2016 (2nd Edition)

Copyright: Crown copyright material is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland

Access conditions: The depositor has specified that registration is required and standard conditions of use apply. The depositor may be informed about usage. See [terms and conditions of access](#) for further information.

Availability: UK Data Service

Contact: [Get in touch](#)


Sets out processing summary and any key data issues for users

DOCUMENTATION

Title	File Name	Size (KB)
Changes to Lab Procedures in HSE 2010-2015	7919_hse2010-15_lab_procedures.pdf	457
Lists of Variables and Derived Variables	7919_hse2014_dataset_documents.pdf	2353
Questionnaires, Showcards, Coding Frames and Consent Booklets	7919_hse2014_interviewing_documents.pdf	2706
Interviewer, Nurse, Coding, Measurement and Editing Instructions	7919_hse2014_supporting_documents.pdf	1803
User Guide	7919_hse2014_user_guide.pdf	294
Study information and citation	UKDA_Study_7919_Information.htm	6
READ File	read7919.htm	11

Online browsing for surveys: Nesstar

UK Data Service



Delivering quality social and economic data resources

DESCRIPTION

TABULATION

ANALYSIS

About the UK Data Service Nesstar Catalogue

Research Datasets

- 1970 British Cohort Study
- Active People Survey
- British General Election Study
- British Social Attitudes Survey
 - British Social Attitudes Survey, 2011
 - British Social Attitudes Survey, 2010
 - British Social Attitudes Survey, 2009
 - Metadata
 - Variable Description
 - Introduction
 - Household Grid
 - Newspaper Readership and Internet Use
 - Party Identification
 - Public Spending and Social Welfare
 - Government highest priority for extra spending? :Q570
 - Government highest priority for extra spending next? :Q571
 - R's view of the level of benefits for unemployed people? :B572
 - If govt had to choose, which should choose: B575
 - R say that the gap between those with high+low incomes is too large? :AB576
 - R place self in high/middle/low income band? :AB577
 - Closest to R's feelings about household's income these days? :AB578
 - Child under primary school age. Lone parent asked to visit the job centre at least every six months? :Q581
 - Child reaches primary school age. If this lone parent did

Dataset: British Social Attitudes Survey, 2009

Variable **IncomGap**: R say that the gap between those

LITERAL QUESTION

Thinking of income levels generally in Britain today, would you say that the gap between

Values	Categories	N
1	too large	1797 79.3%
2	about right	360 15.9%
3	too small	51 2.2%
8	Don't know	58 2.6%
9	Refusal	1 0.0%
-7	off route due to corrupt sample file	6
-2	Skip, version C	1148

SUMMARY STATISTICS

Valid cases 2267

Missing cases 1154


This variable is numeric

UNIVERSE

VERSIONS A AND B: ASK ALL

Online browsing for qualitative - QualiBank

UK Data Service



Discover

Variable and question bank

Quali Bank

Collection title

☒ The Edwardians (17)

☐ BOAPAH (0)

☐ Morale and Home Intelligence Reports (0)

☐ Poverty in the UK (0)

☐ School Leavers Study (0)

Refine

Resource type +

Date +

Sex +

Age group +

Socio-economic status +

Region +

Site Search FAQ Help Contact

About us Get data Use data Manage data Deposit data News and Events

Discover > Quali Bank

Quali Bank

Search and browse qualitative surveys, interviews and open-ended questions.

typhoid GO


Reset filters Clear search ☒ Auto-complete Copyright Help

Extract Image File Audio

SEARCH RESULTS


Displaying 1-10 of 17 results

1 2 ▶▶▶

 Interview with Mrs. Omison
SN2000 The Edwardians, 1870-1973
Sex: Female. Age group: 75-84. Socio-economic status: Routine. Region: North West.

... very little? No, I don't think so. I think we'd be getting on, about 7 or 8 before we went on our own. Did you share the bedroom with anybody? With the maid. And then I had typhoid fever, once, and then the maid had to go and sleep at home. She had her family at Plank Lane, but she didn't used to...

[Access this collection from Discover](#)

 Interview with Mrs. Clark
SN2000 The Edwardians, 1870-1973
Sex: Female. Socio-economic status: Semi-routine. Region: Wales.

... lost her you see, she - she got typhoid fever and was - called home. Yes. We were only in England about seven years and then she was - she left us. Mm, it was sad because we - I suppose if - father and mother had gone back - to their work - at the other side of the world, missionary work, I suppose she...

[Access this collection from Discover](#)

CLOSER: browsing for variables & questions

Search

Explore

List 1

Metadata Tools


Appears Within

Information

History

Download

Help

Promoting excellence in longitudinal research

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

MRC
Medical
Research
Council

+

Avon Longitudinal Study of Parents and Children

ALSPAC Pregnancy, Birth and Infancy (pregnancy to 12 months)

ALSPAC Having a Baby Questionnaire Dataset

Variable

Details

Lineage

Name

b651

Label

Age started REG smoking

Dataset

alspac_91_hab

Value	Label	Frequency
-2	Never smoked	6433
-1	Missing	191
99	DK	0

Valid	Invalid	Min	Max	Mean
6566	6624	2	36	16.31



Self upload – supporting ESRC award holders

- Upload data to our [ReShare](#) EPrints repository
- Harvester project information from Funder Gateway
- DataCite DOI assigned
- Records feed into central data portal
- Supports **Nature's Scientific Data journal** as a host repository



Easy to publish and upload data

UK Data Service
ReShare



My data

Logged in as Louise Corti Logout UK Data Service Home Help About FAQ Contact

Home

Legal

Review procedures

Edit collection: [Data Collection #25555](#)

Terms and conditions

Award details

People

Data collection

Upload

Deposit

To deposit a data collection, you must accept the [ReShare Terms and Conditions](#).

- I confirm that I am the owner of the copyright and associated intellectual property rights in the whole Data Collection or am otherwise lawfully entitled to grant this licence on behalf of each and every owner;
- I grant a non-exclusive, royalty-free licence to the UK Data Archive (a department of the University of Essex and not a separate legal entity) at Wivenhoe Park, Colchester, CO3 3SQ (the "University") to hold, make copies of, and disseminate copies of the Data Collection, in accordance with the access conditions I will specify when uploading data files: open data accessible to users without registration or safeguarded data accessible to users registered with the data services provided by the UK Data Archive.

* I agree to the ReShare data deposit terms and conditions

Save for later

Cancel

Next >

Collection period:	Date from: 1 October 2008	Date to: 30 September 2012
Country:	United Kingdom	
Data collection method:	Collection of information from official documentation across several countries	
Observation unit:	Organisations, Text units	
Kind of data:	Alpha-numeric	
Type of data:	Qualitative and mixed methods data	
Resource language:	English	

— Access and Administration

Data sourcing, processing and preparation:	Information from various sources, including the International Bureau of Fiscal Documentation			
Copyright holders:	Name Devereux, Michael	Email Unspecified	Affiliation University of Oxford	ORCID Unspecified
Contact:	Name Devereux, Michael	Email michael.devereux@sbs.ox.ac.uk	Affiliation University of Oxford	ORCID Unspecified
Notes on access:	None			
Publisher:	Economic and Social Research Council			
Last modified:	28 Apr 2014 22:42			

AVAILABLE FILES

Data

+ CBT_Tax_database.xlsx

Documentation

+ CBT_Tax_database_description.pdf



Self deposit: handling queries on deposit

- Full-time repository administrator
- Some hand-holding for depositors prior to upload
- 1 in 10 questions relayed up to more senior staff
 - Ethics and disclosure review; more technical issues
- Query tracking system to manage/log responses
 - ✓ Can see past queries and responses
 - ✓ SLA – mostly answer within 3 working days
 - ✓ Easy to add common issues to your FAQ

Data management guidance and training

Been doing this of 20 years

- Manage and administer **ESRC's Research Data Policy**
 - Includes awareness raising and TCB in RDM
 - Early intervention for large grants
- Co-wrote **MRC data policy**
- Advice data sharing for MRC and funders on request
- Support data sharing for ESRC/MRC CLOSER
- Advice on Wellcome Cohorts policy

UK Data Service

About us

Get data

Use data

Manage data


Deposit data

News and Events

Home > Manage data

Prepare and manage data

"Good data habits from the moment you start planning your research"



Data lifecycle

Plan to share

Legal and ethical

Copyright

Document your data

Format your data

Store your data

Collaborative research

Training

SHARE

DATA CREATED FROM RESEARCH ARE VALUABLE RESOURCES THAT CAN BE USED AND RE-USED FOR FUTURE SCIENTIFIC AND EDUCATIONAL PURPOSES.

Good data management practices are essential in research, to make sure that research data are of high quality, are well organised, documented, preserved and accessible and their validity controlled at all times. This results in efficient and excellent research. Well managed data are easily shared and can thus be used for new research or to duplicate and validate existing research.

Data management needs to be planned early on in research, so that practices can be implemented throughout the research cycle.

Roles and responsibilities of the various players in the research process need to be explicitly established.

LOGIN / REGISTER


DISCOVER UK DATA SERVICE

GO

Data

Website

MANAGING AND SHARING RESEARCH DATA



DOWNLOAD OUR DATA GUIDE

Comprehensive best practice guidance for researchers on managing and sharing data

QUICK ACCESS TO

FAQ about managing data

RELATED LINKS

UK Data Archive

UK Data Service

About us

Get data

Use data

Manage data


Deposit data

News and Events

Home > Deposit data

Deposit data

"Valuable resources that can be used and reused"



How to deposit

Preparing data

Owners and producers

ORCID

Depositor stories

SHARE

DATA CREATED OR GENERATED DURING RESEARCH OR ADMINISTRATIVE PROCESSES ARE VALUABLE RESOURCES THAT CAN BE USED AND REUSED FOR FUTURE SCIENTIFIC AND EDUCATIONAL PURPOSES. SHARING DATA CAN:

facilitate research beyond the scope of the original research

encourage scientific enquiry

avoid duplication of data collection

provide rich resources for education and training

We proactively acquire data that are suitable for use in research and teaching and that fall within the thematic scope of our Collections Development Policy. We offer different pathways for regular depositors of major survey or other data studies, ESRC award holders and other types of data collections. Please use the appropriate sections to help guide you.

How to deposit data

LOGIN / REGISTER


DISCOVER UK DATA SERVICE

GO

Data

Website

DEPOSITING SHAREABLE SURVEY DATA



QUICK ACCESS TO

Guide to preparing data

ReShare self-deposit repository

FAQ on depositing data


Get in touch

Depositing data video

Depositing data

How to deposit

"Depositing data is straightforward and rewarding"



At the UK Data Service we offer different pathways for regular depositors of major survey or other data studies, ESRC award holders and other types of data collections.

New depositors

New depositors, who are not ESRC-funded, can offer their data by sending us a short description of the data collected. We appraise data according to our Collections Development Policy.

Regular depositors

For depositors of large-scale social surveys or government data series we offer guidance on how to prepare data, completing our deposit form and licence agreement, information on the checks we carry out on received data, and the ways in which we safeguard data.

ESRC award holders

LOGIN / REGISTER


DISCOVER UK DATA SERVICE

GO

Data

Website

DEPOSITING SHAREABLE SURVEY DATA




QUICK ACCESS TO


How to prepare your data

How to deposit data video

ESRC ReShare depositors

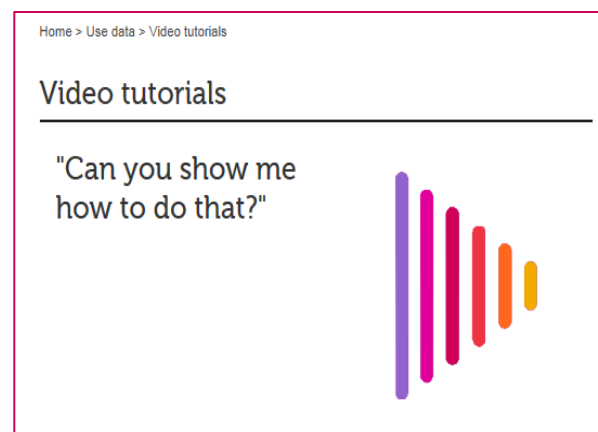
UK Data Service






User support and resources

- ✓ Help desk, individual user support
- ✓ Promotional events/ workshops/webinars
- ✓ User guides/ thematic guides (link from catalogue records)
- ✓ Data Impact blog
- ✓ Case studies of use
- ✓ Teaching & learning resources
- ✓ Undergraduate student dissertation prize



Getting creative with open data

- Used a crowdsourcing projects to generate **innovative uses and outlets** for data
- Produced high quality open data with no disclosure risk and published to an API
- Run an App Challenge 
- UKDS Data Quiz App – Google and Android



Data enhancement and rescue

- Undertake data enhancement or rescue as short projects
- Compiling code books for much older data
- Harmonising data
- Restructuring difficult-to-use data e.g. Ghana Millennium Villages



Wellcome data sharing

- Policy on Data Management and Sharing
- Data management and sharing plan required
- Expect researchers funded to make their research data available with as few restrictions as possible
- Limited data support service or data services – but very good networks emerging
- Increased advocacy and early intervention work
- Best practice borrowing from long-running data archives e.g UK, US and South Africa DataFirst
- New Cohorts Policy now out
- How much policing of policy is desirable?



Working with Wellcome

Various interactions

- Award to look at data discovery
- Ran recent survey on open science and data sharing

Towards Open Research

Researchers funded by Wellcome Trust/ ESRC: biomedical, clinical, population health, humanities, social sciences ((N=842)

- Practices, experiences, barriers and motivations for
 - open access publishing
 - sharing and reuse of data and code

Towards Open Research
Practices, experiences, barriers and opportunities

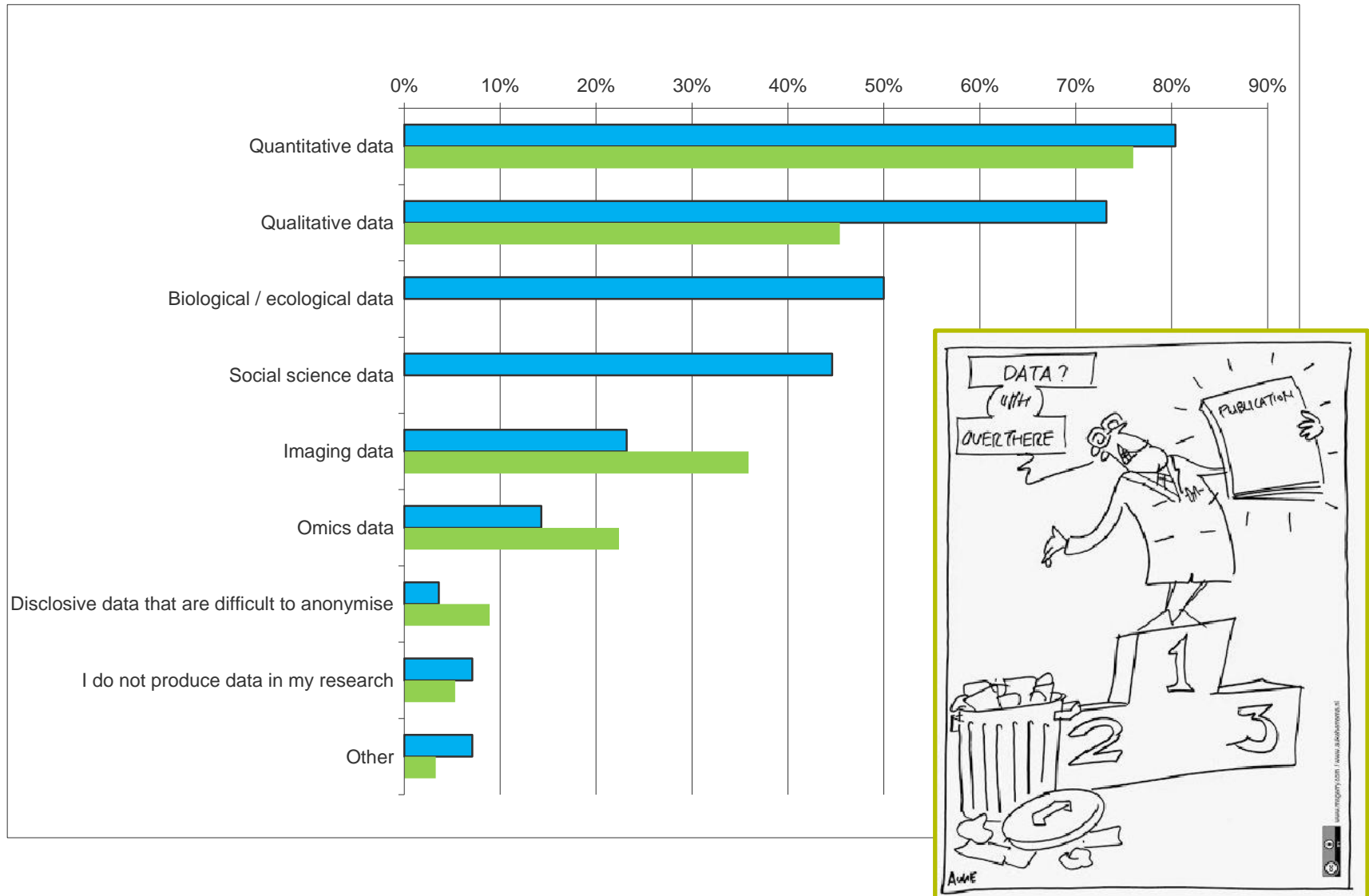
October 2016

Verité Van den Eynden, Gareth Knight, Arca Vlad, Barry Radler, Carol Tenopir, David Leon, Frank Marotta, Jimmy Whitworth and Louise Corti

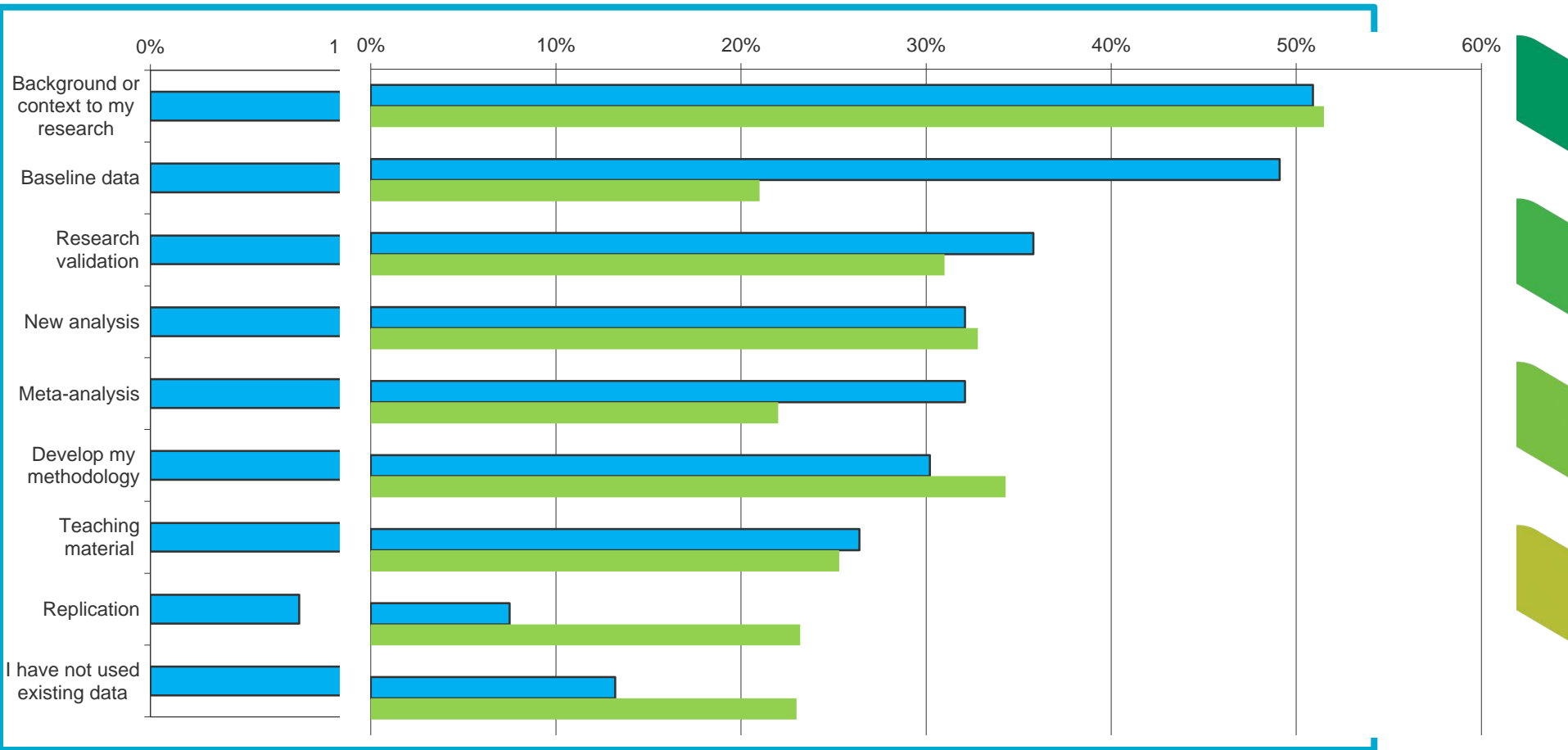


- Contributed to the *Expert Advisory Group on Data Access (EAGDA)* and its reports
- Fed into various data policies

Creating research data

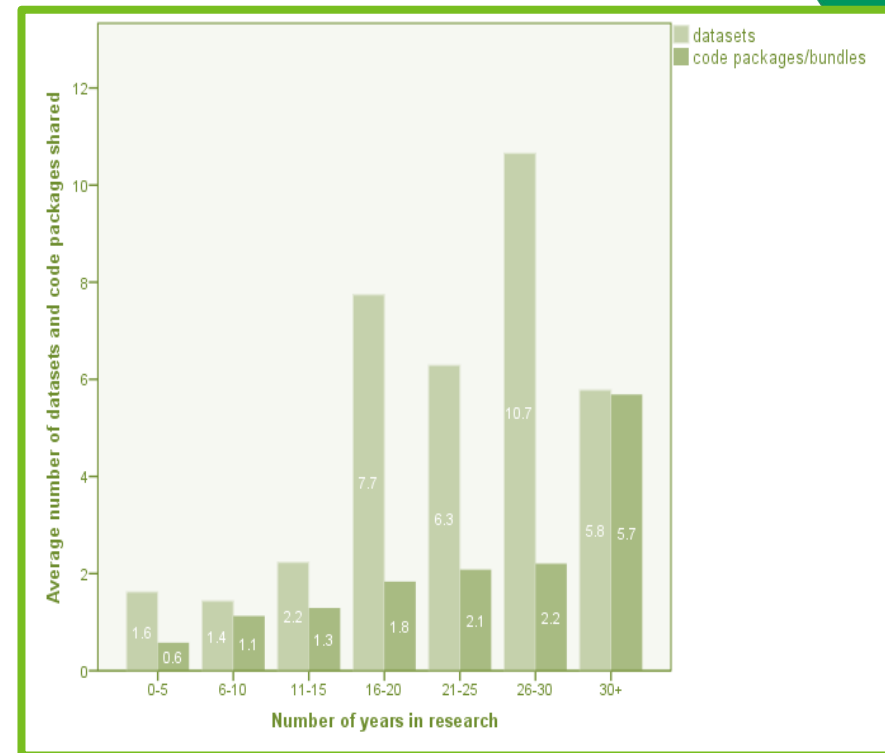


Reuse of data

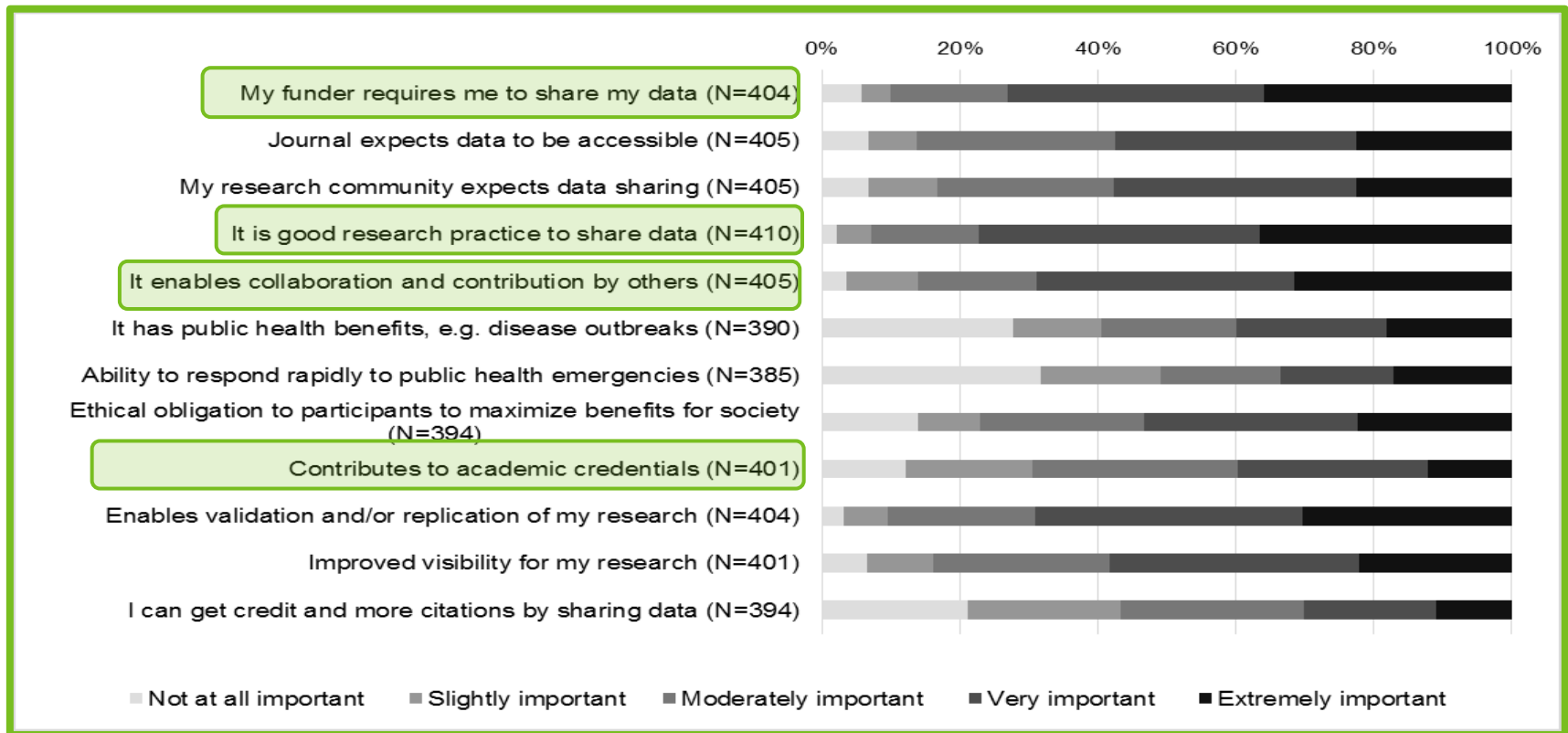


Data / code sharing

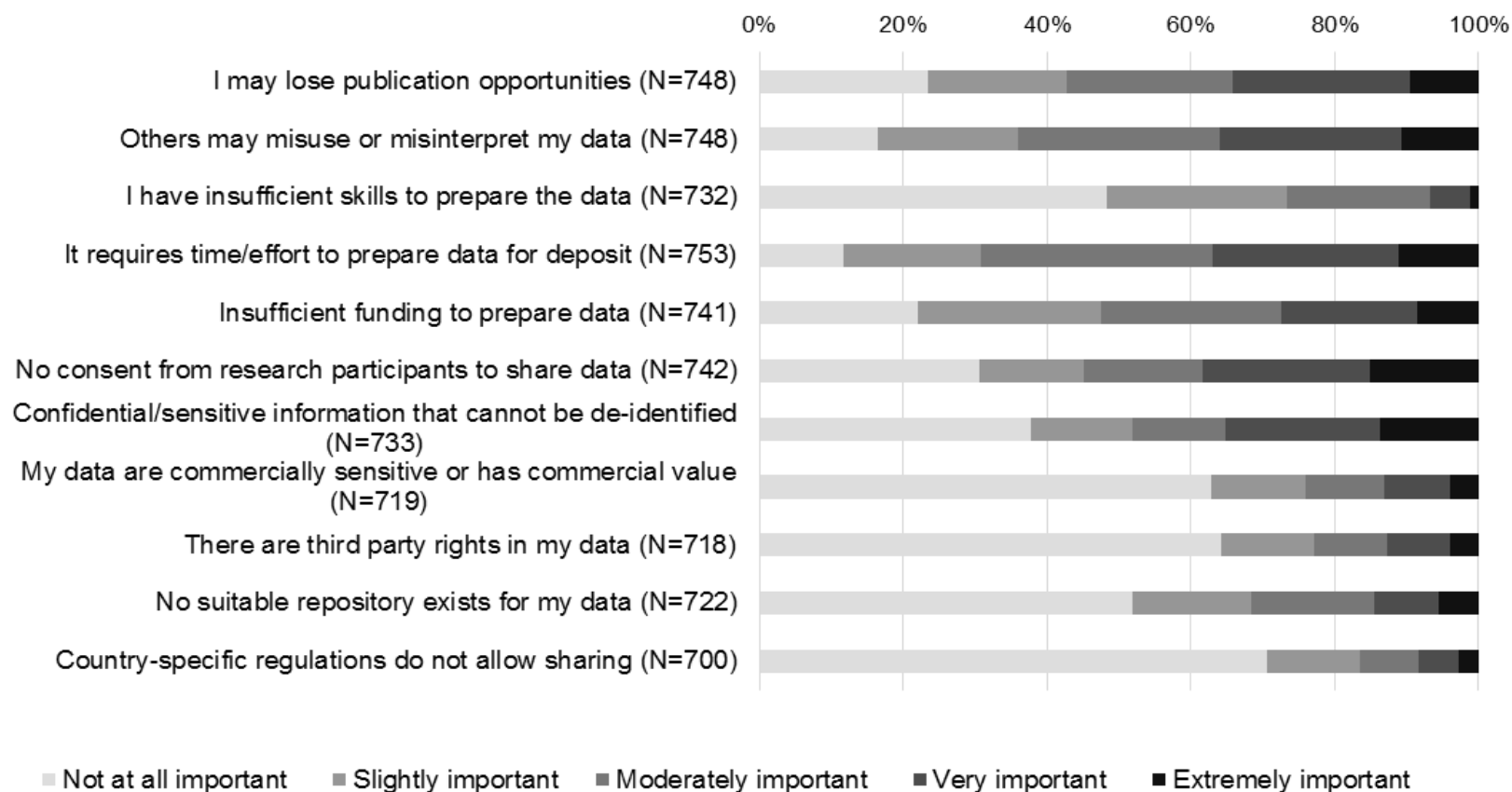
- 95% of respondents generate research data
 - 52 % shared research data last 5 years
 - 3.4 (6.5) datasets on average
 - sharing increases with career length
- 40% of respondents generate code
 - 43% shared code last 5 years
 - 2 (4) code packages on average
 - sharing increases with career length



Reasons to share data



Barriers to data sharing



Difficult to share data widely

- Many complex fieldwork operations already use excellent fieldwork management, enable limited sharing, but rarely longer-term solutions for data access
- Ethical, legal and research integrity challenges
 - Personal, confidential or sensitive information
 - Linkage of data in multi-disciplinary projects
 - social with biomarkers or geo-located data
 - Data licensing, governance and 'trust models' not consistently applied across data access points
- Limited support for shared common infrastructure for data sharing. It is very expensive!

Archived cohort & longitudinal studies

- 1970 British Cohort Study (BCS): 1970+
- English Longitudinal Study of Ageing (ELSA): 2002
- Growing Up in Scotland (GUS)
- Longitudinal Study of Young People in England (LSYPE)/Next Steps: 2004+
- Millennium Cohort Study (MCS): 2000+
- National Child Development Study (NCDS): 1958+
- British Household Panel Survey (BHPS): 1991-2009
- Understanding Society (UKHLS): 2009+
- Ghana Millennium Villages Project
- Girls Education Challenge, Africa
- Whitehall II coming in 2018

Access to different types of data

- Main survey data sit with us as **Public Use or Restricted Use files**
- Some disclosive variables sit under **Secure Access** e.g. Postcodes, detailed SIC and SOC, linked admin data
- We undertake all **frontline support**; complex data queries passed to data owners
- **Genetic data and bio samples** accessed via data owners using a common framework METADAC
- Not as joined up as could be
- Much duplicated infrastructure for sample storage/biobanking and data management

Formal longitudinal data archiving: option 1

Waves/sweeps are archived as separate records

- Each year as a single data collection
- Distinct DOI for each sweep e.g. NCDS

SN:	7669
Title:	National Child Development Study: Sweep 9, 2013
Alternative title:	NCDS9
Persistent identifier:	10.5255/UKDA-SN-7669-1
Series:	National Child Development Study [National Child Development Study, 1958-]
Depositor:	University of London. Institute of Education. Centre for Longitudinal Studies
Principal investigator(s):	University of London. Institute of Education. Centre for Longitudinal Studies
Data collector(s):	National Centre for Social Research
Sponsor(s):	Economic and Social Research Council

Formal longitudinal data archiving: option 2

Sweeps collated into a single collection

- All waves/sweeps integrated into One Study ID
- One single versioned DOI
- Each wave/year a distinct file e.g. Understanding Society

SN:	6614
Title:	Understanding Society: Waves 1-5, 2009-2014
Alternative title:	United Kingdom Household Longitudinal Study; UKHLS
Persistent identifier:	10.5255/UKDA-SN-6614-7
Series:	Understanding Society [Understanding Society: Waves 1- , 2008-]
Depositor:	University of Essex. Institute for Social and Economic Research
Principal investigator(s):	University of Essex. Institute for Social and Economic Research
Data collector(s):	TNS BMRB NatCen Social Research
Sponsor(s):	Economic and Social Research Council

LMIC intervention/surveillance studies

- Much investment by Wellcome, DFID, NIH, World Bank etc. in data collection in LMICs (some longitudinal)
- Face challenges in providing extended data access
- Hard to access without a personal request
- Projects wound up and multiple stakeholders cannot agree on **longer-term** solutions for data access
- **Data structures complex** - migration and demography
- More capacity needed for **data preparation for reuse**

Examples of archived (or to be) studies

- Ghana Millennium Village Impact Evaluation (DFID-Ghana)
- Girls' Education Challenge (DFID)
- Agincourt HDSS (Wellcome, NIH, Mellon and Hewlett Foundations, SAMRC)



Millennium Villages Impact Evaluation Ghana

DFID-Ghana [Intervention study](#) (N=26,000)

- New village created in N Ghana - £11m to meet MDG
- Data sharing clearly not envisaged despite DFID data policy
- 2012-15: Year 0,1,2,3, 4 – longitudinal
- 750 households (treatment group); 1,500 households (control group - 'near' and 'far')
- New data collection put at risk as no data shared
- PIs worked with UKDA to solve stalemate
- Multiple stakeholders with competing interests: US, UK, Ghana

Data sharing issues

- Formats hard to process/ analyse – 130 separate Stata files
- Little metadata in files; complex subfolder structures; poor documentation - lots of time-consuming cross-referencing
- Disclosure risk assessment; post-hoc US IRB approval
- Trust in data sharing procedures by data collectors/owners



Disclosure review

- Identify **potentially disclosive variables** within each dataset as well as between groups of datasets
- Initial screening of data files for **direct identifiers or key variables** to identify individual units
- *sdcMicro* package for R
- Frequency analyses of all variables across all data files to determine **low-frequency responses and extreme outliers**
- Additional **qualitative classification** of potentially disclosive variables carried out

Treatment

- Raw **age, community, and village names** had very small frequency counts
 - excluded from the released dataset
- Other variables for which **local knowledge essential**
 - to indicate risk - implicit or quasi-identifiers
 - E.g. ethnicity, fuel type use, toilet facilities with flushing mechanisms, house wall material
 - **recoded** to reduce the number of available categories

Variables	Disclosure risk	Action
Community	Low frequency counts for all named communities, respondents who gave answers very easily identifiable (especially in combination with other variables)	Exclude variable from dataset
Age	Low counts of older respondents over 75 years old	Top-code age ≥ 75 as '75 and over'
Main occupation during last 12 months	Low counts of very specific occupations	Occupations aggregated into standard occupation codes
Ethnicity of the Household Head	Low counts of specific ethnicities.	Recode the low-frequency responses (all responses but 'Mamprusi' and 'Builsa') into 'Other'.
Household's primary type or energy/fuel used for cooking	Very low counts for 'Gas/LPG' and 'Electricity-solar panel' responses may lead to household identification (especially if combined with other datasets)	Recode all responses into the following main categories: 1 - 'Firewood'; 2 - 'Electricity-based'; 3 - 'Charcoal'; 4 - 'Other', 5 - 'Don't know'; 6 - 'NA/missing'.
Main material of the wall of the house	A number of low-frequency responses; exterior features (households/buildings easily identifiable)	As the main material of the wall refers to the exterior of a building, it may be advisable to recode the low-frequency and 'Other' variables into 'Other (incl. wood-based and stone-based)' and retain the remaining groups
Crops grown on plots	A number of low-frequency specific responses for each variable	Variables are recoded into crop categories

UKDS access solution: Phase 1

- Release 1: Household data only - Special Licence
- Data Access Committee and procedures for decision making about applicants
- Solution discussed for access to more than one dataset to be judged (e.g. household data plus bloods)
- Unfriendly codebook – hard to find variables
- Pretty hard to use; hundreds of stata files
- 500 variables, not asked consistency across time, e.g. malaria
- Problems with IDs - same IDs used for households in targeted and controlled villages

Girls' Education Challenge (GEC)

- £344 million Girls' Education Challenge (GEC) Fund
- Department for International Development (DFID)
- UK's main contribution to the Millennium Development Goal of eliminating gender disparity in primary and secondary education
- Period: funding - four-year cycle, 2013–2017
- Evaluation design: data collected from households and schools in GEC intervention and control areas
- Afghanistan | Ethiopia | Kenya | Mozambique | Sierra Leone | Somalia | Tanzania | Zimbabwe | Democratic Republic of Congo (DRC)

Girls' Education Challenge Data

Baseline and midline data – follow up coming

- Baseline
 - Household Survey: 6,323 cases
 - School-based Assessment: 8,743 cases
 - School Visit Survey: 2,956 cases.
- Midline
 - Household and School Visit Survey: 6,021 cases
 - School-based Assessment/ learning assessment)
6,197 cases
 - School-based Assessment: 4,133 cases.



Data and documentation issues

- Excellent technical report
- Good case studies on gov.uk website
- Restricted access only
- Data eas(ier) to use
- Still inconsistency
 - School IDs not consistent across data collections
e.g. schoolid_x and schoolidx
- Follow up study done by different fieldwork agency – documentation looking poor...

Agincourt HDSS data sharing

North-east of South Africa: baseline census 1992; annual census rounds conducted since 1999

- Collaboration with PIs
- 1 in 10 data training data
- INDEPTH core data - limited event history data covering core demographic events (1992 - 2010: 14 variables)
- Longitudinal datasets not yet formerly 'archived'
- Data formats difficult for those outside the population, migration, epidemiology research domains

Extending HDSS data for other disciplines

Recent work by Collinson & Wittenberg to restructure Agincourt HDSS data - to meet social science needs:
Linked **panel data format** (long-form) at 3 levels:

Individual level data (N=200,000)	Life events from 1992 - every person Educational events from 2000 - most people Labour force events from 2000 - most people
Household level data link to Person ID	Size by year from 1992 Assets and consumption from 2000
Village level information	Various facilities, and linkage to Nightlights satellite data

Uses for restructured Agincourt HDSS

- Traditional **social science panel study methodology** - considering carefully designed **units of analysis**
- Longitudinal data analysis around a constructed **household type** (Unit = Household formation and whether it persists or not)
- Set up panels of individuals as **new panels of households**, through which longitudinal changes in **household electricity access** could be explored
- Novel approach used to **re-present data** from an HDSS
- Will be archived at **DataFirst as FAIR data** under Secure Access (Safe Lab)

Opportunities for further impact through data

Hugely useful as exemplary showcase on how to
prepare panel data to address micro social and
economic change

- ⌘ Time consuming
- ⌘ Methodologically complex
- ⌘ Posthoc work, not core work of the study
- ✓ Could encourage funders to support longitudinal data preparation to agreed standards
- ✓ Wellcome and DFID benefit from case studies like this
- ✓ Authors writing a data paper e.g Scientific Data; Open Health Data

Funders - issues for new data collection



- Consider data sharing requirements in design and fieldwork contracts
- Resourcing for onward data sharing
- Issues
 - Ethical protocols and data sharing
 - Ownership and licensing
 - Future access conditions and maintenance of access
 - Sustainable data documentation
 - Enabling longer-term longitudinal use (variables)
 - Enabling linkage with administrative sources
 - Long term data preservation

Dealing with 'new and novel' data

- Varied Nature of data of interest in social sciences
 - Surveys collecting biomedical/real-time data
 - Impact of urban environments
- Value at High Volumes
 - Smart energy meters
 - Environmental sensing
 - Wearables and geo-health
 - Social Media
 - Surveillance
 - Transactional data
- Varied provenance and often unknown error



Methodological challenges for big data



Scaling up for big data

- Open Source Enterprise Hadoop system now at UKDA
- Data Service as a Platform (DSaaP)
- Open sources tools for semi-automated QA of data and interrogation – e.g. R tools
- Upskilling staff and social scientists in big data
- Project to show smaller data archives what is involved in scaling up
- Major role for institutions as they accommodate multi-faceted research data needs
- Big data summer schools Jan 2017 and summer 2017

UK Data Service resources

- UKDS [webpages](#) and [video](#) on preparing data
- UKDS [webpages](#) on operating the ESRC Data Policy
- UKDS [webpages](#), book and video on RDM issues
- [Depositing Shareable Survey Data](#) brochure
- UKDS ReShare guide/checking guidelines
- UKDS [Collections Development Policy](#)
- UKDS [Selection and Appraisal Criteria](#)
- UKDS [Data Purchase Guidelines](#)
- [Call to action](#): Use of DDI metadata in survey production process

Keep connected

Louise Corti

Director, Collections Development and Data Publishing

corti@essex.ac.uk

Twitter: @LouiseCorti

UK Data Service

University of Essex, Colchester, UK

UK Data Service list:

www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE

Follow UK Data Service on Twitter: @UKDataService, @UKDSRDM

Youtube: www.youtube.com/user/UKDATASERVICE