Progress on open data publishing in the social sciences

Louise Corti
Research Data Management Team
UK Data Service

Tartu DataCite Conference
23 October 2014

UK Data Service

Overview

- Digital data publishing in the social sciences
- The case of the UK Data Service
- Sharing data ethically
- Spectrum of access for data
- Our citation approach

University of Essex

The Archive

The UK Data Archive

- Based at the University of Essex, since 1967
- 47 years of selecting, ingesting, curating and providing access to social science data
- Data and data support services for higher and further education for research, teaching and learning
- Attained the highest information security standard, ISO 27001
- Concordat with our National Statistical Institute (ONS)

Links with other data archives worldwide

- Qualitative data
- Data archives
- Open data
- Secure datalabs
- Question banks

Data access policy

European data archives | North American data archives | Other worldwide data archives

SHARE

UK Data Service

Open access – a great time for archives!

- Open access and transparency agendas seeking digital curation and preservation knowledge
- Move by data owners considering open licensing
- Huge progress in opening up government data (gov.data)
- Lack of trust in published academic findings – demands for verification
- ….but risks surrounding privacy concerns mean ‘open data’ may be less useful for research

International funder data policies

- OECD Principles and Guidelines for Access to Research Data from Public Funding

- UK funder models
  - Data management plans and sharing recommendations
  - Data sharing mandates and dedicated data centres or institutions taking responsibility

- European Research Council/Horizon 2020)
  - Pilot on open access to research data
  - Data management guidelines for Horizon 2020 (policies)
  - Communication & recommendation on access to / preservation of scientific information (publications, data)

- USA data management plans, e.g.NSF and NIH

Journal / Publisher Data Policies

- Many science journals have data policies relating to data sharing
  - "PLOS ONE will not consider a study if the conclusions depend solely on the analysis of proprietary data" … "the paper must include an analysis of public data that validates the conclusions so others can reproduce the analysis."
  - BioMed Central open data statement
- Data underpinning publication accessible
  - upon request from author
  - as supplement with publication
  - in public or mandated repository (Elsevier uses PANGAEA)
- Citation via unique DOIs
- JORD project: survey of journal policies

Progress in the social sciences (UK)

Good on funder data policy

Good on data centres (funders)

Fair on institutional repositories (funding)

Poor on journal policy

Exceptions: economic journals – verification
Psychology journals – fraud cases

ESRC research data policy

Research data should be openly available to the maximum extent possible through long-term preservation and high quality data management. (ESRC Research Data Policy, 2010)

- ESRC grant applicants planning to create data during their research include a data management plan
- ESRC award holders offer their research data to UK Data Service within three months of the end of their grant

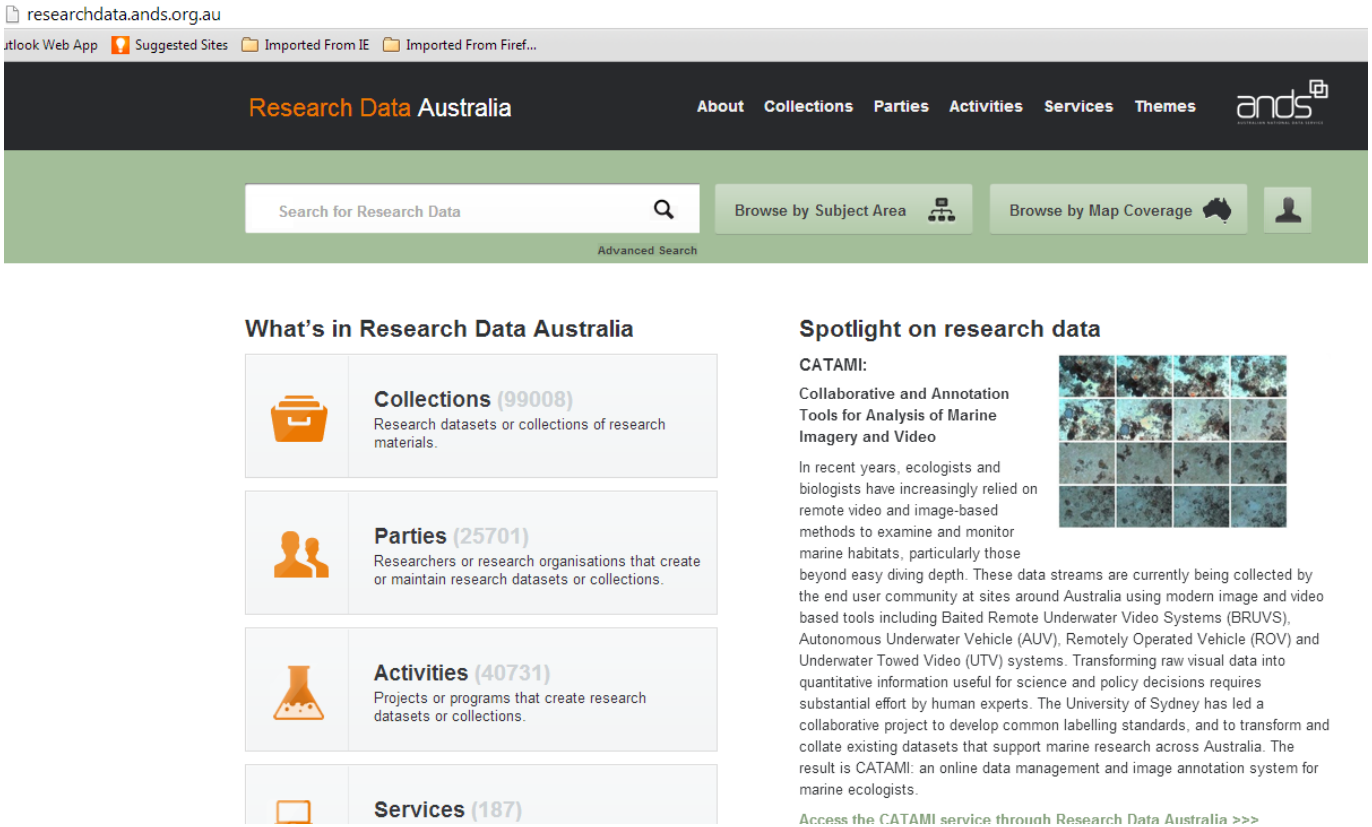
Researchers who collect the data initially should be aware that ESRC expects that others will also use it, so consent should be obtained on this basis and the original researcher must take into account the long-term use and preservation of data. (ESRC Framework for Research Ethics, 2012)

Discipline specific data centres

- Curate and provide access to discipline-specific data:
  - UK Data Service
  - Environmental Information Data Centre
  - Visual Arts Data Service
  - Endangered Language Archive

- UK Data Service publishing pathways:
  - main collection - actively curated reserved for high impact data, primarily from government departments
  - ReShare - self deposit option for most research data offered

Registry of research data



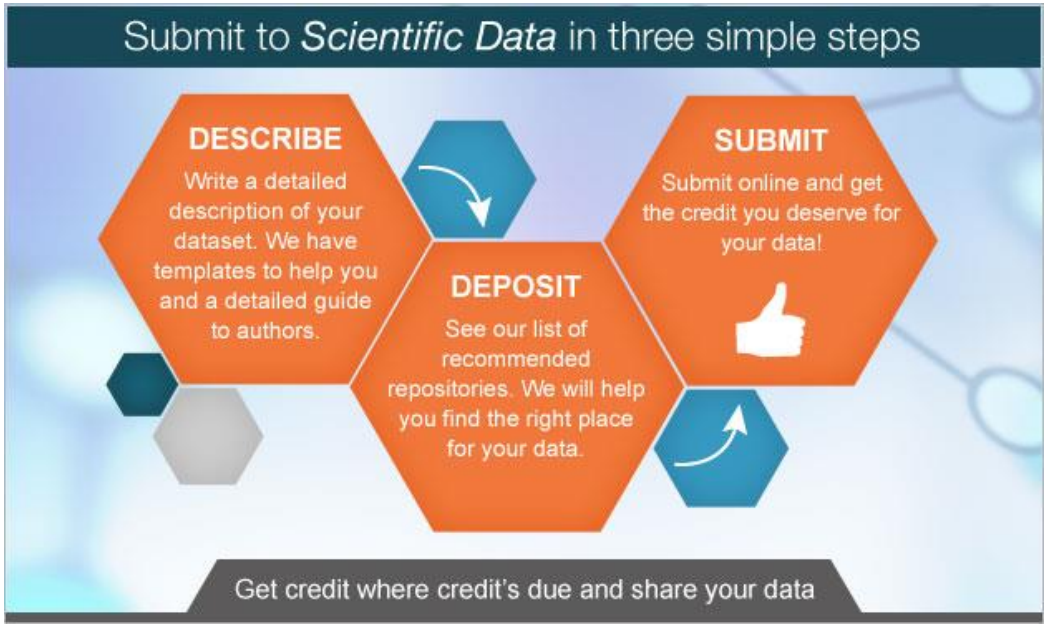
- UK project underway to establish national registry from institutions and data centres

UK Data Service

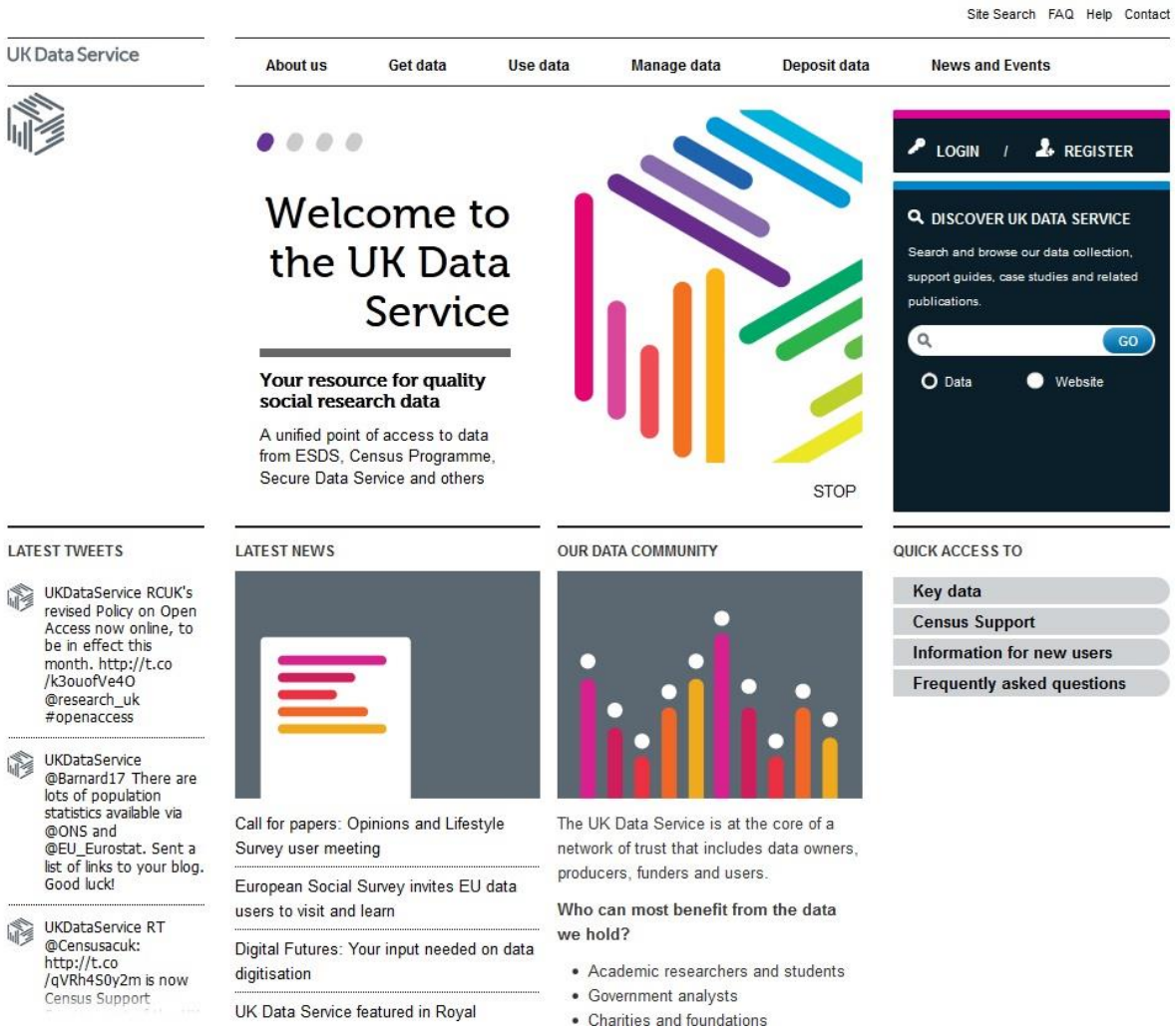
Data journals

- New phenomenon for social sciences
- Publish a detailed journal style article describing the data and its origins
- Recommends or provides a place of deposit

  e.g. Nature Scientific Data (http://www.nature.com/scientificdata/)



UK Data Service



ukdataservice.ac.uk

UK Data Service

UK Data Service

- put together a collection of the most valuable data and enhance these over time
- preserve data in the long term for future research purposes
- make the data and documentation available for reuse
- provide data management advice for data creators
- provide support for users of the service
- information about how data are used
- easy access through website

Our data portfolio

Over 6,000 datasets in the collection

230 new datasets added each year

- Official agencies - mainly central government
- International statistical time series
- Individual academic' research grants
- Market research agencies
- Public records/historical sources
- Access to international data via links with other data archives worldwide

Types of data in our portfolio

UK Surveys

Large-scale government funded surveys

Longitudinal

Major UK surveys following individuals over time

International

Multi-nation aggregate databanks and survey data

Census

Census data 1971 to 2011

Business

Microdata and administrative data

Qualitative

Range of multi-media data sources

Data access at the UK Data Service

- Web access to data (various formats) and metadata
- Data freely available for use. Commercial use charges
- Data available under 3 access levels:

  OPEN

  SAFEGUARDED – End User License

  CONTROLLED – approved researcher /secure environment

- Some data also available via online data browsing

Open data collections and licenses

Census - Open Government Licence

- 2011 and 2001 Census aggregate statistics

Survey data - Open Government Licence

- Various surveys and teaching datasets

Qualitative datasets – Creative Commons 4.0 BY (NC)

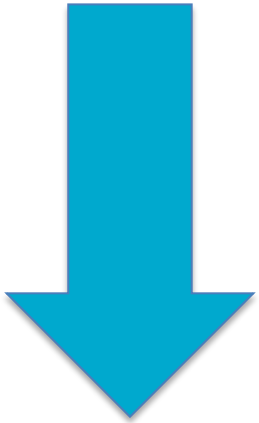
- life story interviews, essays, WWII reports

Aggregate global indicators – bespoke open data license

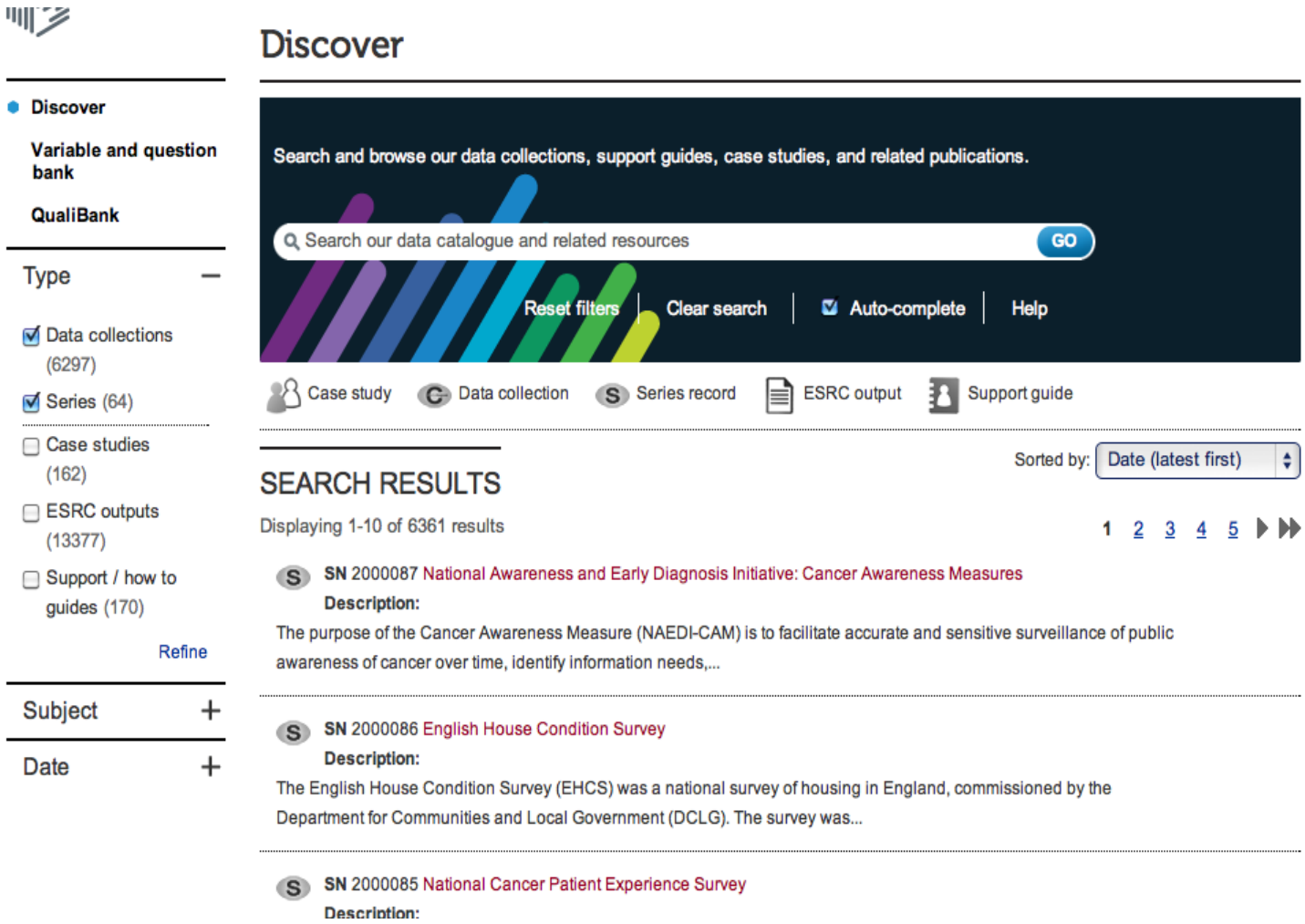
- World Bank Millennium Development goals

Safeguarded data

- Register with us using UK Federation
- Agree to an End User Licence (EUL)
  - ✓ Appropriate data usage
  - ✓ Full citation of data
  - ✓ informing us of re-use
- Select data from the Discover data catalogue using ‘Download/Order’ button
- Specify a project for which the data are to be used
- Download data to local machine in preferred format



Discover at UK Data Service



UK Data Service

Ethical arguments for sharing data

- Store and protect data securely
- Not burden over-researched, vulnerable groups
- Make best use of hard-to-obtain data (e.g., elites, socially excluded)
- Extend voices of participants
- Provide greater research transparency
- Enable fullest ethical use of rich data

In each, ethical duties to participants, peers and public may be present

Data Protection Act and research

- Exceptions for personal data collected as part of research:
  - can be retained indefinitely (if needed)
  - can be used for other purposes in some circumstances
    people should still be informed

The Data Protection Act is not intended to, and does not, inhibit ethical research

Sharing confidential research data

Consider:

- obtaining informed consent for data sharing
- protecting identities not collecting personal data or anonymisation
- restricting / regulating access where needed

Consider jointly and in dialogue with participants

Plan early in research

Do participants consent to share data?

- Timescapes
  - data on personal relationships
  - 95%+ consent rate
- Foot and mouth disease in North Cumbria
  - sensitive community information
  - UK Data Archive consultation; pilot with 4 participants
  - 40/54 interviews; 42/54 diaries; audio restricted
- Finnish research on consent
  - re-contact project: life stores, gender, etc.
  - 165/169 (98%) agreed
  - bereaved relatives wanted others to benefit from their contribution

Why anonymise research data?

- Ethical reasons
  - protect people's identity (sensitive, illegal, confidential)
  - disguise research location
- Legal reasons
  - not disclose personal data (DPA)
- Commercial reasons

Identity disclosure

Direct identifiers

- e.g. name, address, postcode, telephone number, voice, picture
- often NOT essential research information (admin)

Indirect identifiers

- possible disclosure in combination with other information
- e.g. occupation, geography, unique or exceptional values (outliers) or characteristics

Anonymising quantitative data

- remove direct identifiers

  e.g. names, address, institution, photo

- reduce the precision/detail of a variable through aggregation

  e.g. birth year vs. date of birth, occupational categories, area rather than village

- restrict upper lower ranges of a variable to hide outliers

  e.g. income, age

- combining variables

  e.g. creating non-disclosive rural/urban variable from place variables

Anonymising qualitative data

- Plan or apply editing at time of transcription

  except: longitudinal studies - anonymise when data collection complete (linkages)

- Avoid over-anonymising - removing/aggregating information in text can distort data or make it misleading

- Avoid blanking out; use pseudonyms or replacements. Identify replacements, e.g. with [brackets]

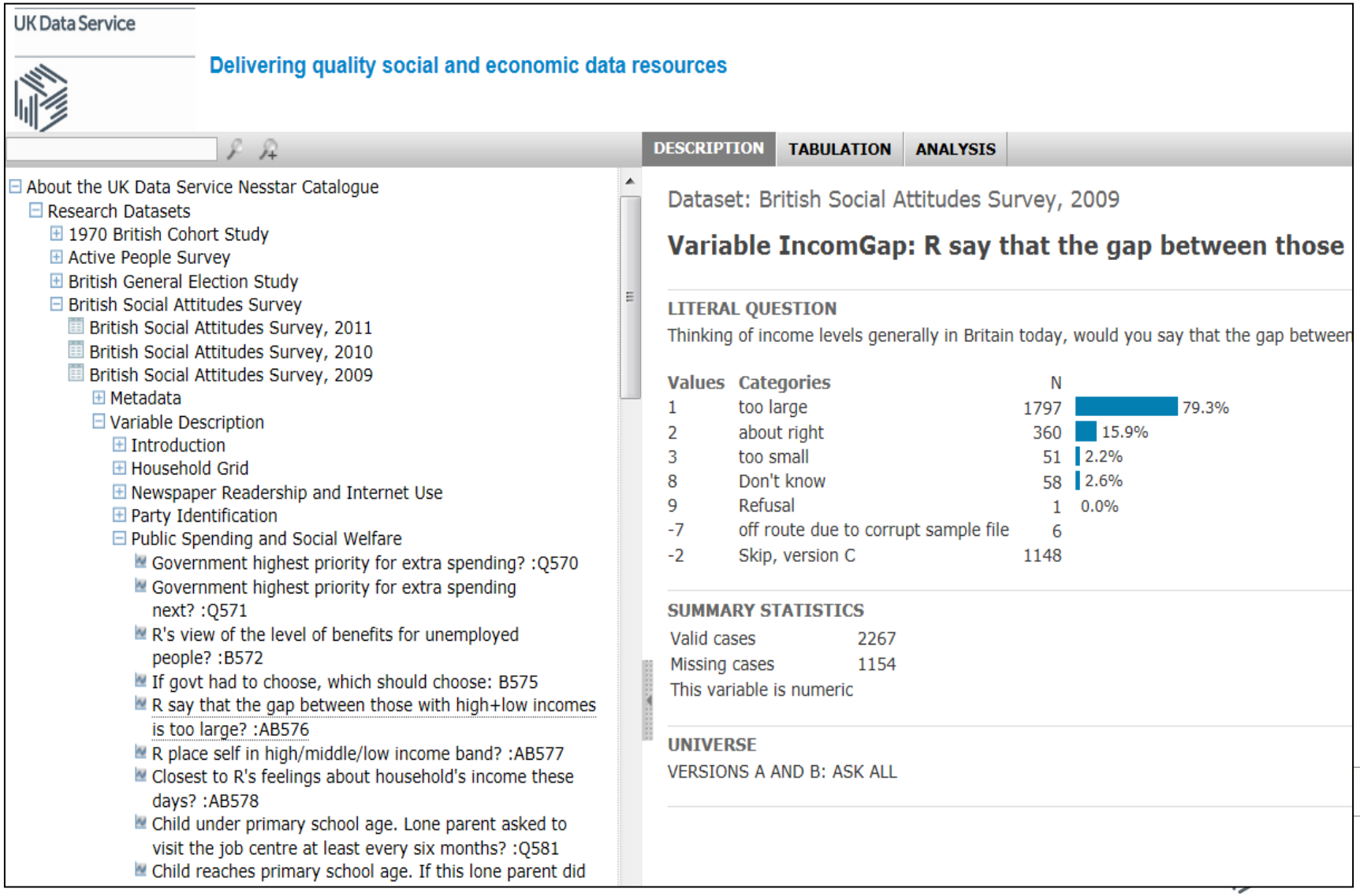
- Consistency within research team and throughout project

- Keep anonymisation log of all replacements, aggregations or removals made – keep separate from anonymised data files

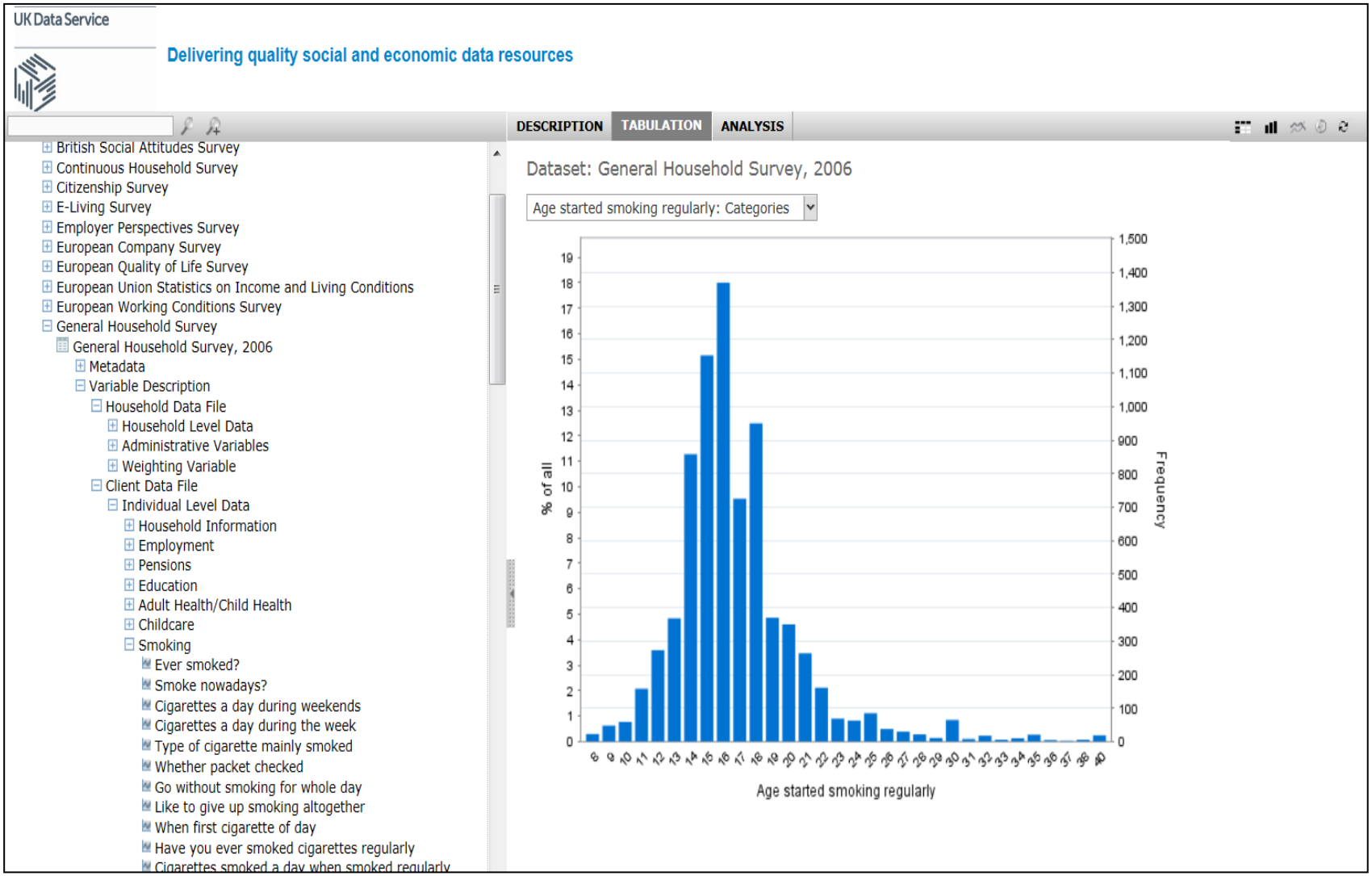
Online instant data browsing

Nesstar	social surveys
UKDS.stat	aggregate global indicators (some restrictions to HE/FE)
InFUSE	aggregate census data
QualiBank	qualitative data
APIs are coming soon!	

Nesstar: British Social Attitudes - Pay gap



Nesstar: GHS - Age started smoking



Nesstar: GHS - time series

UK Data Service

Delivering quality social and economic data resources

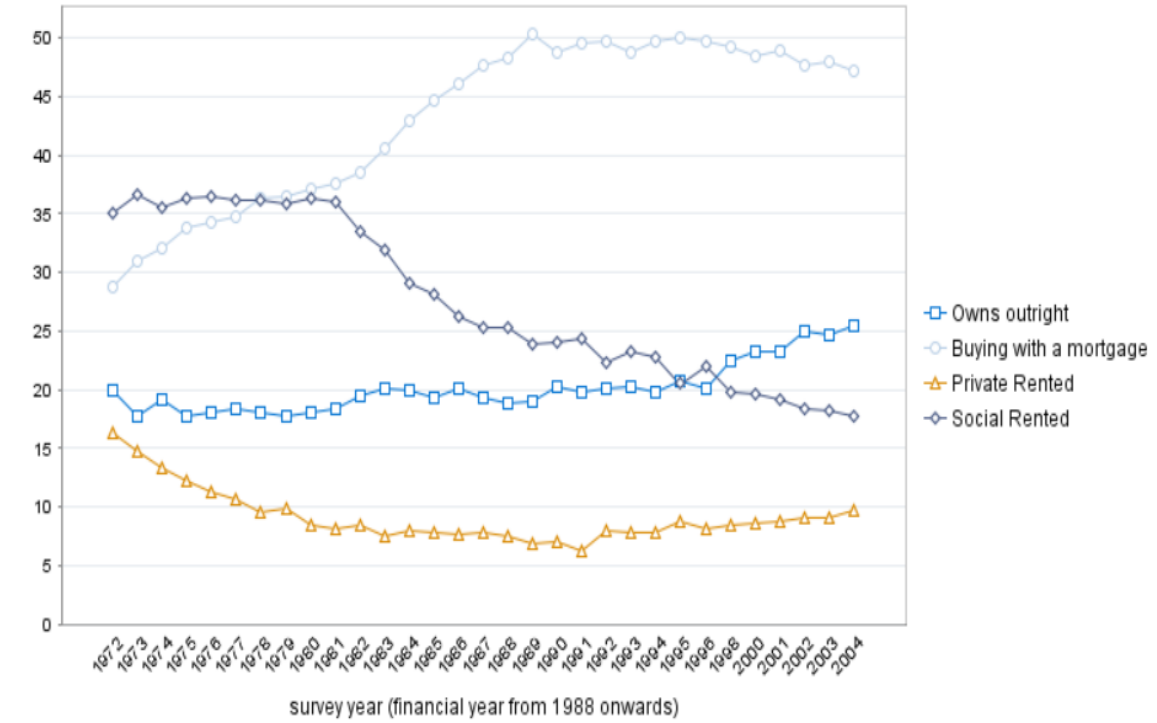
DESCRIPTION | TABULATION | ANALYSIS

- SEX
- Estimated Year of Birth
- 5-year estimated birth cohort
- Marital Status
- Household Type A (Grouped)
- Household type F
- Household type F (Grouped)
- No. of persons in the household
- No. of children aged <5 in household
- No. of children aged 5 to 15 in household
- No. of adults aged 16 to 59 in household
- No. of adults aged 60 and over in household
- No. of adults aged 65 and over in household
- No. of teenagers in household
- No. of children aged <16 in household
- Youngest age of person in the household
- No. of males in household
- No. of females in household
- No. of adult males in household
- No. of adult females in household
- AGE GROUPED (5 YEARS)
- AGE OF ADULTS GROUPED (6 BANDS)
- AGE OF ADULTS GROUPED (7 BANDS)
- AGE OF ADULTS GROUPED (4 BANDS)
- Marital Status (4 Groups)
- Marital Status (3 Groups)
- sex of HRP/HOH
- marital status of HRP/HOH
- age of HRP/HOH
- Number of cars

+ Individual Information
+ Administrative Variables
+ Weighting Variables
- Time Variable
  - survey year (financial year from 1988 onwards)

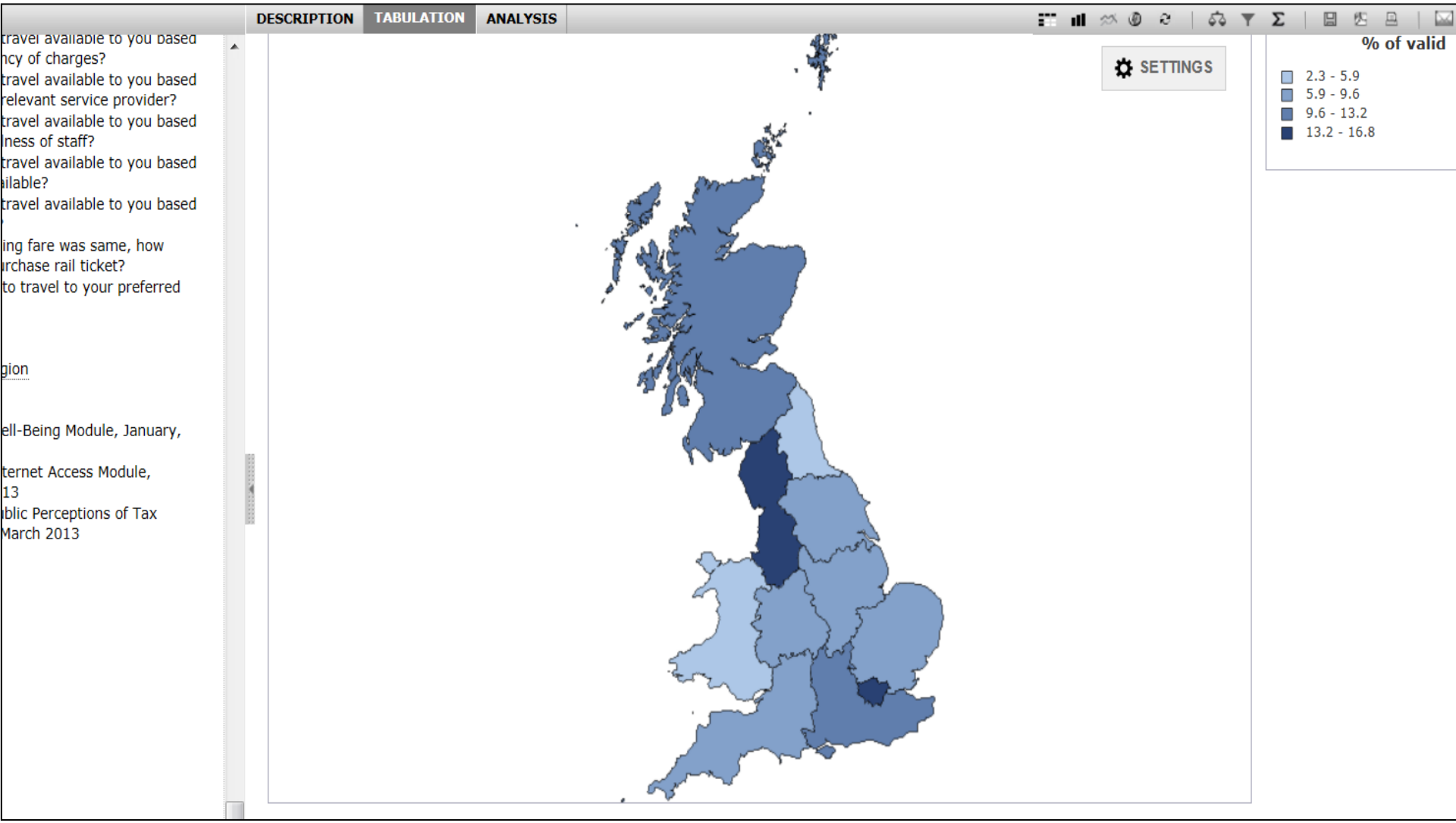
Dataset: General Household Survey: Time Series Dataset, 1972-2004

Tenure: Categories | survey year (fina...rom 1988 onwards): Categories | Type: Column percentage



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

Nesstar: mapping OLS 2013



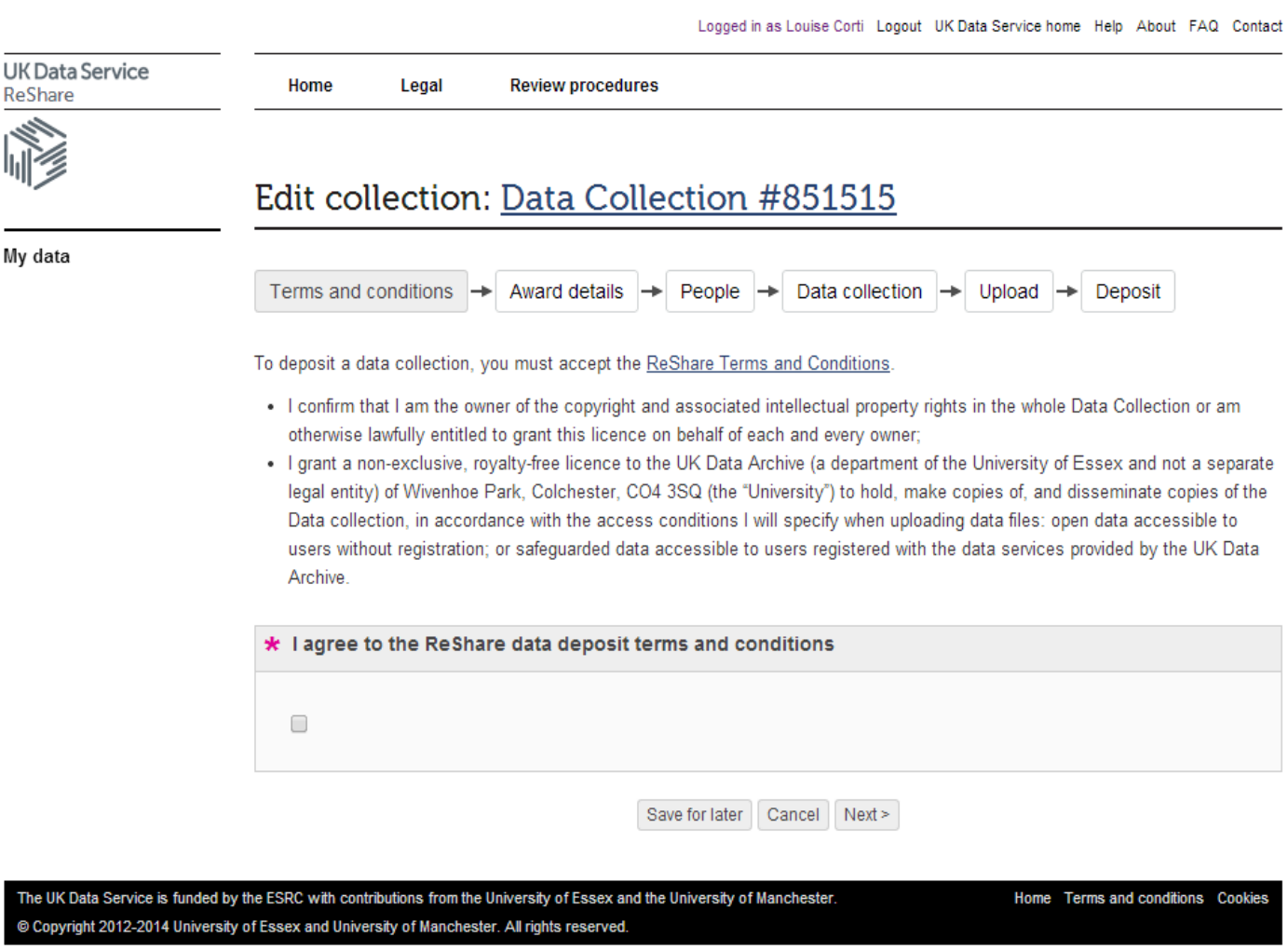
UK Data Service ReShare

http://reshare.ukdataservice.ac.uk/

A platform through which to get research data published

Primarily data produced by UK social science researchers, but also other social science research data of interest

Easy to publish and upload data



UK Data Service

Citing data

- Citation a fundamental part of research and academia
- We have recommended a human readable citation for many years
- Certain attributes of data need to be identified, e.g. versions
- Granularity of citation is important
- Duplication of data sources being cited can be problematic. Which version?

Our persistent identifiers approach

- Our data collections are not digital objects
- Need to capture changes made to data
  - Versioning data in a commonly understood manner
  - Needed rule-based definition of a ‘significant’ change
- Integrate processes with digital preservation activities & work flows
- In 2011 we assigned 6000+ Datacite DOIs for our collections
- Use APA citation style for data

Catalogue

UK Data Service data catalogue record for:

Understanding Society: Waves 1-2, 2009-2011

TITLE DETAILS

SN:	6614
Title:	Understanding Society: Waves 1-2, 2009-2011
Alternative title:	United Kingdom Household Longitudinal Study; UKHLS
Persistent identifier:	10.5255/UKDA-SN-6614-4
Series:	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 NatCen Social Research
Data collector(s):	NatCen Social Research Northern Ireland Statistics and Research Agency. Central Survey Unit
Sponsor(s):	Economic and Social Research Council Department for Work and Pensions Department for Education Department for Transport

Recording significant change

- Approx. 15% UKDA data collections are altered within first year after first publication
- We distinguish between major and minor changes to a data collection = high impact vs. low impact
- DOI allocated to a metadata instance of a data collection
  - DOIs resolve to jump page with version history pointing to all external instances
  - New DOI = High Impact change, with explicit logging
- Provided access only to most up-to-date version of data

Major changes – high impact

- New variable added
- New labels/value codes added
- Weighting variables reconstructed
- Wrong data supplied (e.g., March not April)
- Mis-coded data (e.g., Don't know/Refused confused)
- Change in format (file migration)
- Significant changes in documentation
- Change in access conditions

UK Data Service

A new Digital Object Identifier (DOI) is assigned to the data collection each time there is a major change to data, documentation or metadata. The new DOI will resolve to an updated version of this page containing a log of changes to this data collection since the allocation of its first DOI. The DOI system supports resource discovery and simplifies citation for users of data collections. Data producers benefit directly through increased visibility of their work.

10.5255/UKDA-SN-6614-4

Citation:
University of Essex. Institute for Social and Economic Research and NatCen Social Research, *Understanding Society: Waves 1-2, 2009-2011* [computer file]. *4th Edition*. Colchester, Essex: UK Data Archive [distributor], January 2013. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-4

Change log:
7 January 2013: For the fourth edition (January 2013) finalised data and documentation from Wave 2 were deposited, along with updated data and documentation for Wave 1. See documentation for full details of revisions and updates.

Minor changes
22 August 2013: Minor changes to catalogue record

Full catalogue record >

This is the latest release

10.5255/UKDA-SN-6614-3

Citation:
University of Essex. Institute for Social and Economic Research and National Centre for Social Research, *Understanding Society: Wave 1, 2009-2010 and Wave 2, Year 1 (Interim Release), 2010* [computer file]. *3rd Edition*. Colchester, Essex: UK Data Archive [distributor], February 2012. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-3

Change log:
20 February 2012: For the third edition (February 2012) data and materials for the first year of Wave 2 were added to the study.

A more recent release exists ^

10.5255/UKDA-SN-6614-2

Citation:
University of Essex. Institute for Social and Economic Research and National Centre for Social Research, *Understanding Society: Wave 1, 2009-2010* [computer file]. *2nd Edition*. Colchester, Essex: UK Data Archive [distributor], November 2011. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-2

Change log:

A more recent release exists ^

Raising awareness in the social sciences

- ESRC funding for short-term project on citation
- Advocacy for best practice in citing research data
- Audiences
  - Professional organisations
  - Publishers and journal editors
  - Researchers
- Key activities
  - Data citation principles for social sciences
  - Personal communications
  - Events with BL DataCite, JISC and wider PI community
  - Outreach through Doctoral Training Centres

Demonstrating impact with citation

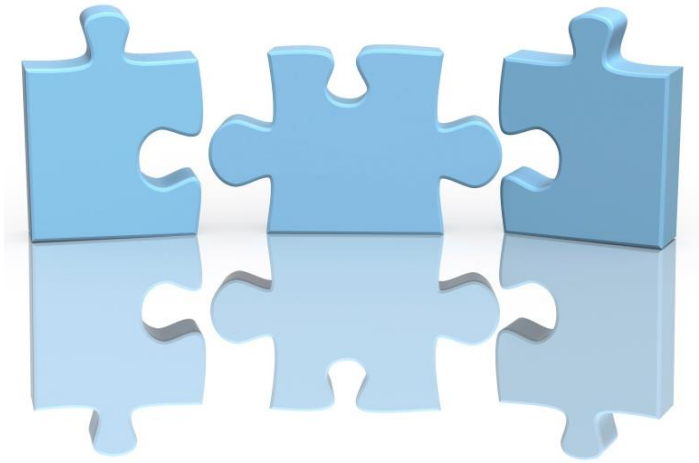
- Hoping for use of our data DOIs!
- Starting to search for use of our DOIs – Google
- Gather data citation statistics from Thomson Reuters Data Citation Index
- Work with BL Datacite and ODIN to gain connectivity between identifiers & outputs – early adopters of ORCID

ORCID
Connecting Research and Researchers

UK Data Service

Dynamic data citation

- Citing parts (fragments) of data collections
  - single files
  - subsets of quantitative data
  - extracts of textual data



- QualiBank system enables extract level citation
  - Citation has rich highly structured XML metadata
  - GUIDs to identify subsets citation database
  - Human reference references the ‘mother’ DOI

UK Quali Bank

Interview with Mr. Hennessey

Study documentation | Web resources Access whole collection

DETAILS

SN:	2000
Title:	Family Life and Work Experience Before 1918, 1870-1973
Sex:	Male
Age group:	65-74
Socioeconomic status:	Lower supervisory/technical
Region:	North West

FULL INTERVIEW TEXT

Interviewer: Can you tell me where your father came from originally?

Subject: I can't, no. From Manchester, I expect. Although he tells me about ... well, me father, he tells me about me grandfather having a little barrel of brandy sent from France.

Interviewer: As if he might have come from France?

Subject: Yes. Hennessey's brandy.

Interviewer: You didn't belong to the Hennessey brandy family?

Subject: I think he did. I think it is. But he said he couldn't pay the duty on it and they had to send it back. That's all I know about it, I mean to say, it might be hearsay, and I don't know.

Interviewer: You don't know where your mother came from?

Subject: No, I haven't the faintest like that.

UK Data Service

Cancel

4 extracts selected

RETRIEVE CITATION

INTERVIEW TRANSCRIPT

Create citation

Select an extract below and then use the ▲ ▼ arrows to expand your selection. When ready, click 'Retrieve Citation' to view your unique citation reference.

Alternatively you may **cite the whole text**.

I: Can you tell me where your father came from originally?

R: I can't, no. From Manchester, I expect. Although he tells me about ... well, me father, he tells me about me grandfather having a little barrel of brandy sent from France.

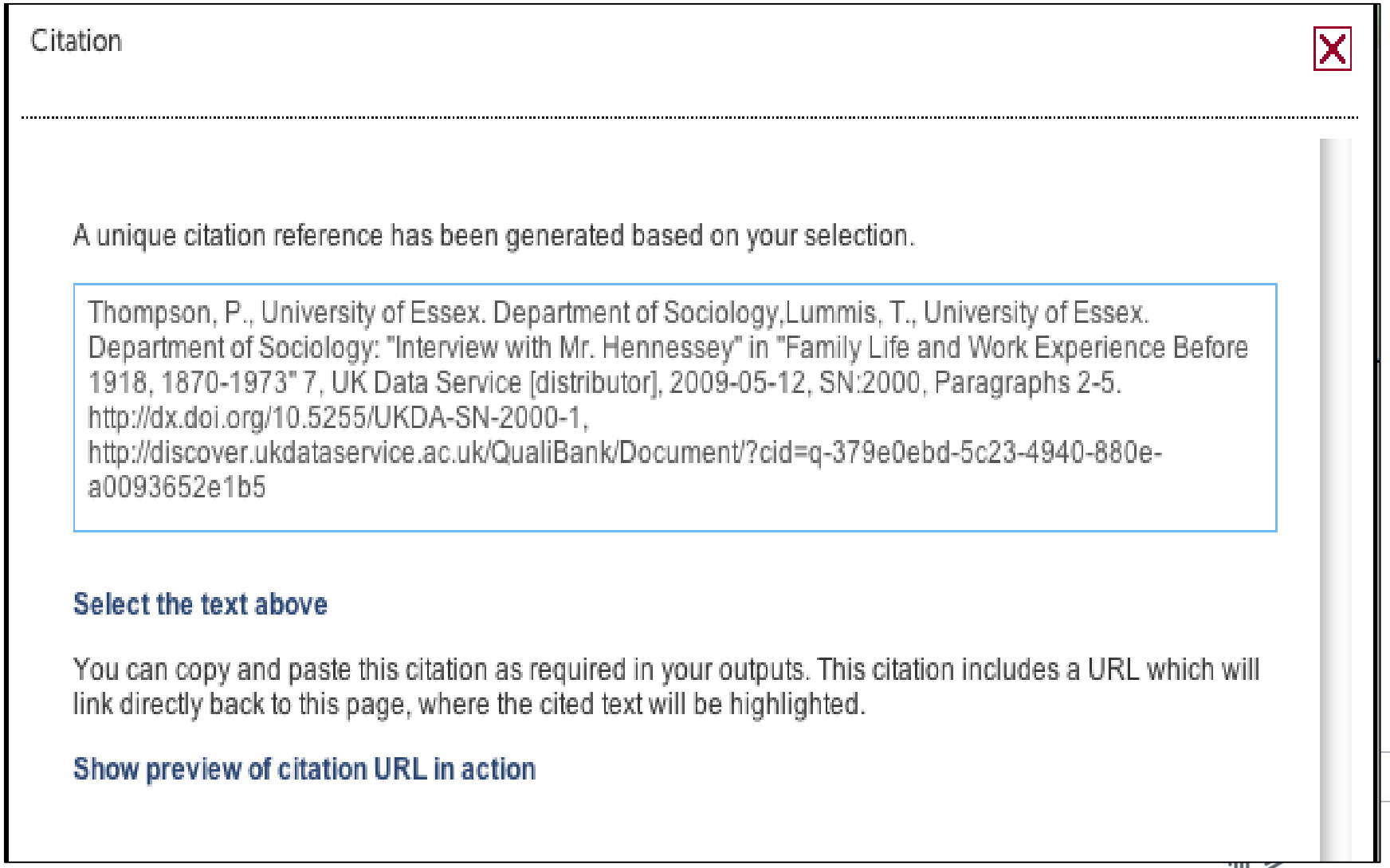
I: As if he might have come from France?

R: Yes. **Hennessey**'s brandy.

I: You didn't belong to the **Hennessey** brandy family?

R: I think he did. I think it is. But he said he couldn't pay the duty on it and they had to send it back. That's all I know about it. I mean to say, it might be hearsay, and I don't know.

APA citation format



More data - big data investments

- ESRC Big Data Network
  - Administrative and business
  - Third sector and local government
  - Social media

- Working with new data producers, some commercial sector
- Will need new data access environments
- New discussions about privacy and digital footprints

Image source: digitalfamilysummit.com

Keep connected

- Subscribe to UK Data Service list: www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE
- Follow UK Data Service on Twitter: @UKDataService
- Facebook
- Youtube: www.youtube.com/user/UKDATASERVICE

Contact

UK Data Service
University of Essex
ukdataservice.ac.uk/help/get-in-touch.aspx