
How metadata drives 'The Archive'

Louise Corti
Functional Director, Collections Development
and Producer Support
UK Data Service

Knowledge Exchange Workshop: Better survey
data management with metadata
British Library, London
21 may 2015

UK Data Service



Surveys – an ‘end of life approach’

- The end product and how to get there
- Preparing and documenting
- Deposit Guide
- The art of packaging surveys
- Online browsing needs



UK Data Service acquisition

- We **proactively acquire data** for use in research and teaching
- Data are deposited by:
 - National statistical institutes (contractual)
 - UK government departments
 - Intergovernmental organisations
 - Research institutes
 - Research companies
 - Individual researchers including ESRC Data Policy
- **Criteria for selection** are set out in our Collections Development Policy



Our data 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



UK survey series

- High quality repeated cross-sectional surveys
- Individual or household level data
- Cover topics including health, work, crime, social attitudes, family expenditure, living costs, housing etc.

Examples:

- Labour Force Survey
- British Crime Survey
- Health Survey for England
- British Social Attitudes
- Annual Population Survey

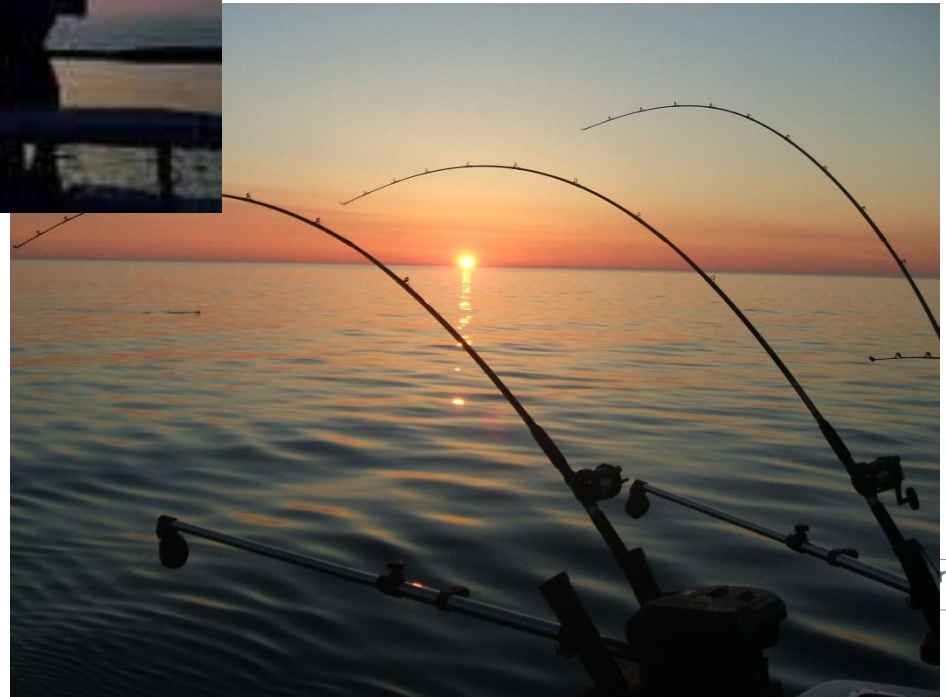


Collections Development work



Trawling

Line-caught



Adapted OAIS Functional Model (ISO 14721)

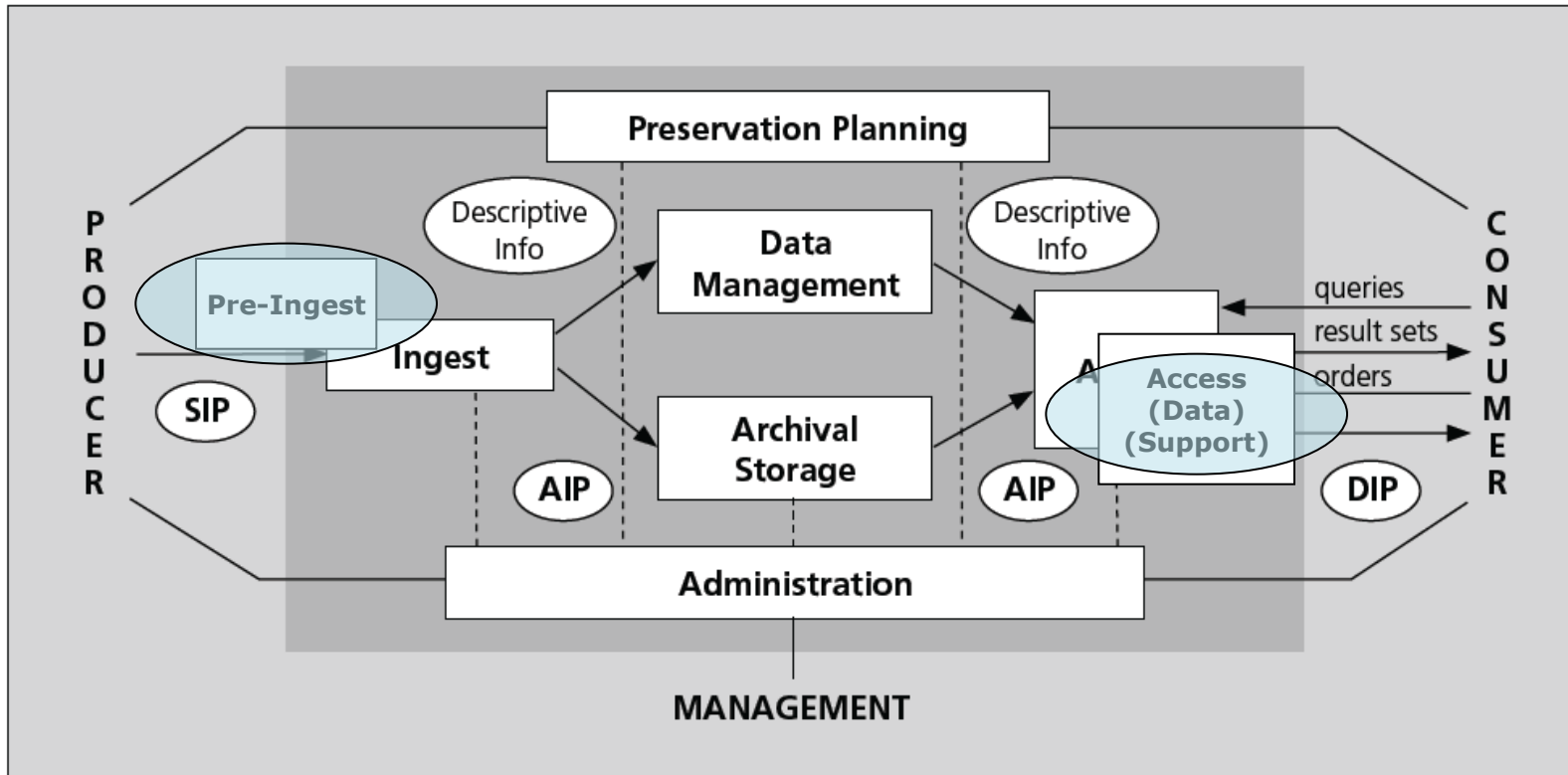


Figure 1. The OAIS Functional Model.

Assessment for new deposits

- Our Data Appraisal Group assesses data according to our Collections Development Policy
- Decision will usually be one of the following:

accept the data into our main collection

deposit in ReShare repository

deposit in an institutional repository

use an alternative place of deposit



Accepting into the main collection

Complete a data deposit form

- used to populate a data catalogue record

Submit data and documentation files

- via the [University of Essex ZendTo Service](#)
- on CD, DVD or memory stick

If data files contain sensitive information

- ensure data are encrypted and sent securely

Provide a licence agreement

- where required if not under a concordat

Access conditions

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

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

Common issues with depositing surveys

- Choice of licensing and access pathways
- Many organisations are overly risk averse and choose restrictive access
- Work underway to draw up bench marks for objective and transparent disclosure review
- Huge loss of questionnaire metadata, which could be improved...



Data publishing - when good metadata becomes vital

- Documentation systems and question banks
- Data exploration systems
- Currently hard to match up **Question and Variable** information
- So much manual work
- Must do better.....

UKDS: Online instant data browsing

Nesstar social surveys

UKDS.stat aggregate global indicators

InFUSE aggregate census data

QualiBank qualitative data

APIs coming soon!

Nesstar: British Social Attitudes - Pay gap

UK Data Service

Delivering quality social and economic data resources

DESCRIPTION TABULATION ANALYSIS

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

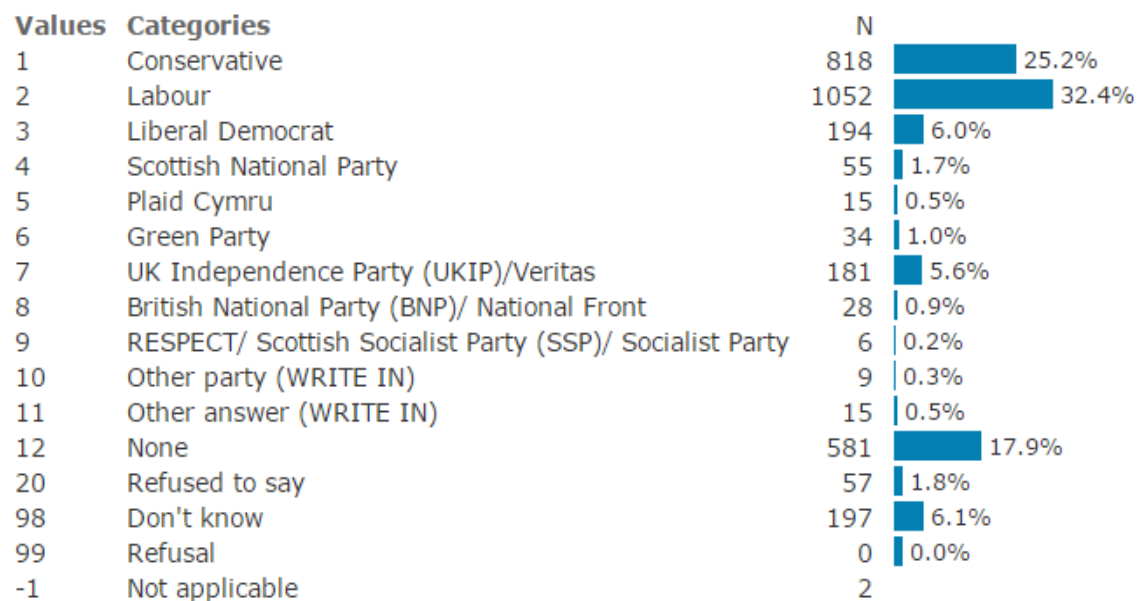
- About the UK Data Service Nesstar Catalogue
 - Research Datasets
 - 1970 British Cohort Study
 - Active People Survey
 - British General Election Study
 - British Social Attitudes Survey
 - British Social Attitudes Survey, 2011
 - British Social Attitudes Survey, 2010
 - British Social Attitudes Survey, 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, 2013

Variable PartyIDN: R s political party identification :Q251

LITERAL QUESTION

IF 'yes' AT [SupParty] OR AT [ClosePty]: Which one? IF 'no' OR DON'T KNOW AT [ClosePty]: If there were a general election tomorrow, which political party do you think you would be most likely to support?



SUMMARY STATISTICS

Valid cases 3242
 Missing cases 2
 This variable is numeric

INTERVIEWER INSTRUCTIONS

DO NOT PROMPT

UNIVERSE

IF 'yes' AT [SupParty] OR 'yes', 'no' OR DON'T KNOW AT [ClosePty]

Nesstar: GHS - Age started smoking

UK Data Service

Delivering quality social and economic data resources

DESCRIPTION TABULATION ANALYSIS

Dataset: General Household Survey, 2006

Age started smoking regularly: Categories

Age started smoking regularly	% of all
0	0.2
1	0.5
2	0.8
3	1.2
4	2.0
5	3.5
6	4.8
7	11.2
8	15.2
9	18.0
10	15.2
11	9.5
12	12.5
13	4.8
14	4.5
15	3.5
16	2.2
17	1.0
18	1.0
19	0.8
20	0.5
21	0.5
22	0.5
23	0.5
24	0.5
25	0.5
26	0.5
27	0.5
28	0.5
29	0.5
30	0.5
31	0.5
32	0.5
33	0.5
34	0.5
35	0.5
36	0.5
37	0.5
38	0.5
39	0.5
40	0.5
41	0.5
42	0.5
43	0.5
44	0.5
45	0.5
46	0.5
47	0.5
48	0.5
49	0.5
50	0.5
51	0.5
52	0.5
53	0.5
54	0.5
55	0.5

Nesstar: GHS - time series



UK Data Service

Delivering quality social and economic data resources



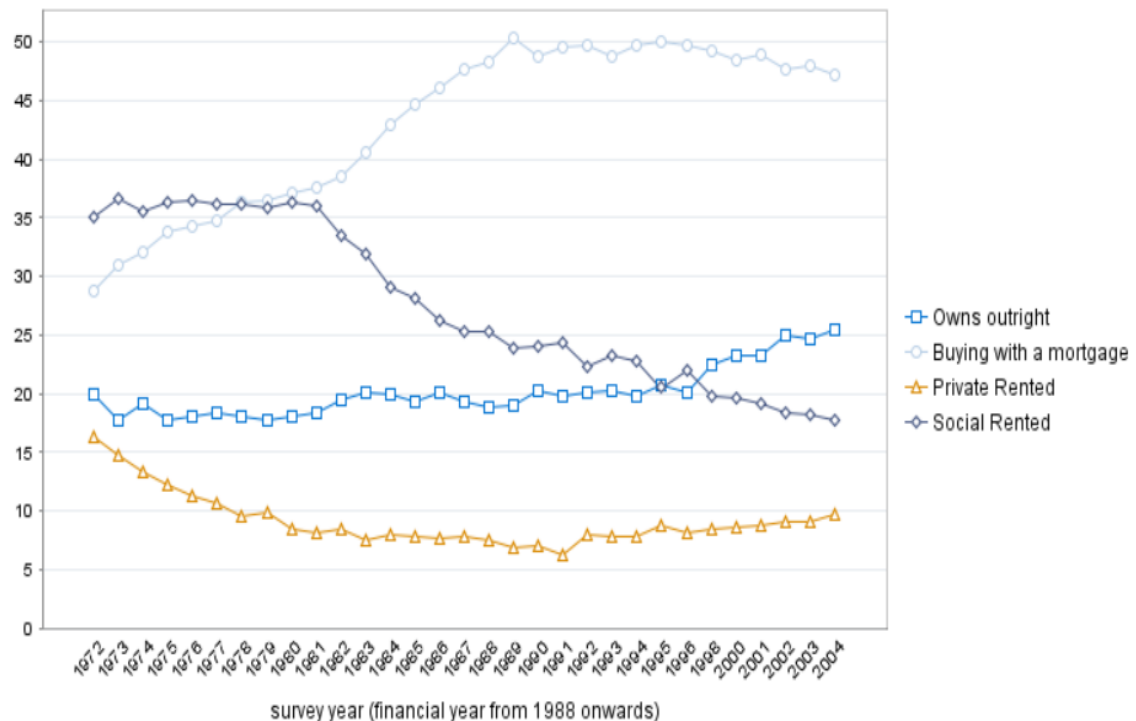
DESCRIPTION TABULATION ANALYSIS



- 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 (financial year from 1988 onwards): Categories Type: Column percentage



Question banks

Date + Results per page: 10 Sorted by: Relevance


Displaying 1-10 of 4107 results

1 2 3 4 5 ▶▶

Data collector +

Question: Harmonised set +

Question: Intent +

 **marital : Legal Marital Status**

Question Text: Are you (is he/she)... single, that is never married, married and living with (husband/wife), married and separated from (husband/wife), divorced, or, widowed?

Health Survey for England, 2003

— **Close responses...**

Add to My Variables

View all instances of this variable

1	...single, that is never married,	3842
2	Married and living with husband,	8029
3	Married and separated from husband,	402
4	Divorced,	1266
5	Or, widowed?	1290
-9	No answer/refused	5
-8	Don't know	2
-7	Refused/not obtained	0
-6	Schedule not obtained	0
-2	Schedule not applicable	0
-1	Item not applicable	3717



How to deposit data

Different steps for different depositors:

UK Data Service


Site Search FAQ Help Contact

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

Home > Deposit data > How to deposit

How to deposit

"Depositing data is straightforward and rewarding"



SHARE ↗

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](#)

LOGIN / REGISTER

DISCOVER UK DATA SERVICE

GO

Data Website

QUICK ACCESS TO

How to prepare your data

ESRC ReShare depositors

ukdataservice.ac.uk/deposit-data/how-to.aspx

UK Data Service



Some common metadata issues

- What do we get, typically?
 - SPSS or STATA file
 - Word documentation; questionnaire
 - Excel sheet of variables – if lucky
 - Word deposit form (our fault)
- Variable ordering in SPSS files does not often meet questionnaire flow
- Lack of consistent variables naming over time or data series
- Partially documented changes to variables over time

Just tell me how to.....



Short brochure for survey products

- Worked closely with data owners and producers
- Existing information too complex
- What is really expected!
- Transferrable information
- Not a bible



Contractor mandates

- Specify data documentation requirements in the commissioning tender for fieldwork

Example

The Centre for Longitudinal Studies (CLS) currently commissions the national cohort studies: the Millennium Cohort Study; the National Child Development Study; the 1970 British Cohort Study and the Longitudinal Study of Young People in England. CLS has started to expect as a deliverable the Computer Assisted Interviewing (CAI) implementation as a Data Documentation Initiative (DDI) compliant XML file, and a file that maps the CAI question to the data variables.

- Mapping between questions and data outputs
- Improved readable questionnaire for end users

Documentation - practice makes perfect

DOCUMENTATION

Title	File Name	Size (KB)
Lists of Variables and Derived Variables	7649datadocs.pdf	2429
Questionnaires, Showcards, Coding Frames and Consent Booklets	7649interviewingdocs.pdf	3494
Interviewer, Nurse, Coding, Measurement and Editing Instructions	7649supportingdocs.pdf	3173
User Guide	7649userguide.pdf	666
Study information and citation	UKDA_Study_7649_Information.htm	23
READ File	read7649.htm	10

Documentation

RELATED STUDIES AND GUIDES

[+ View related studies and guides...](#)

ervice



Survey producers and data publishers

- ✓ Brochure a start
- ✓ Great work via CLOSER on questionnaires
- ✓ Making survey metadata reusable across the lifecycle will support archiving end points



The screenshot shows the DDI website header with the logo <ddi> and the title "Data Documentation Initiative". The navigation menu includes "What is DDI?", "DDI Alliance", "DDI At Work", "Resources", "Specification", and "RDF Vocabularies". The main content area features a "Home" link and a featured article titled "Survey Metadata Reusability and Exchange: A Call to Action for Questionnaire Documentation". The article is dated "Last Updated: Wed, 2014-11-05 15:06" by "ddiadmin". Below the title are two buttons: "View Endorsements" and "Add Endorsement". At the bottom of the page, it states "Produced by the 'Survey Metadata: Barriers and Opportunities' Meeting June 26, 2014, London".



Data preparation and QA tools

Common tasks

- Disclosure review
- Shape of data
 - Variable and value labels
 - Missing values
 - Out of range values

In-house tools

- In-house Bespoke python scripts
- Nesstar 4 Publisher
- R Tools



Online browsing - Nesstar cleaning station

- Publish SPSS file
- Uses DDI 1.X
- Focus on enhancing variable metadata
 - Question text, routing, summary statistics
- Group variables to reflect questionnaire
- Quite a lot of additional manual work
- From word, pdf questionnaires, and if we are lucky, those excel sheets

```

</var>
- <var ID="V77" intrvl="discrete" dcml="0" files="F1" name="LegStat3">
  <location width="2"/>
  <labl>And what is your legal marital status? :Q153</labl>
  - <qstn>
    <qstnLit>And what is your legal marital status?</qstnLit>
    <ivuInstr>CARD A2</ivuInstr>
  </qstn>
  - <valrng>
    <range max="9" min="1"/>
  </valrng>
  - <invalrng>
    <range max="-1" min="-9"/>
  </invalrng>
  <universe clusion="I">IF 'Living with a partner' AT [MarStat6]</universe>
  <sumStat type="vald">315</sumStat>
  <sumStat type="invd">2929</sumStat>
  - <catgry>
    <catValu>1</catValu>
    <labl>Married</labl>
    <catStat type="freq">2</catStat>
  </catgry>
  - <catgry>
    <catValu>2</catValu>
    <labl>In civil partnership</labl>
    <catStat type="freq">10</catStat>
  </catgry>

```



MIDLIFE STUDY IN THE US

 **A1PD1**



8 REFUSED/MISSING

9 INAPP

Variable Details

Concept

Perceived satisfaction





qstnLit

At present, how satisfied are you with your LIFE? Would you say A LOT, SOMEWHAT, A LITTLE, or NOT AT ALL?

preQTxt

And now a few questions about you.

Appears Within

- ⊕  Studies (1)
- ⊕  Series (1)
- ⊕  Variable Groups (3)
- ⊕  Data Files (1)

Ratings

Rating:

Be the first to rate this item.

My Rating



Surveys at a Glance	Topics at a Glance	Graphs and Tables	Download Data and More	Create Analysis Data	Read Literature	Help	About
--	---------------------------------------	--------------------------------------	---	---	------------------------------------	----------------------	-----------------------

[Home](#) » [Create Analysis Data](#) » [ELSA](#) » [ELSA 2002](#) » [HE. Health](#) » [HEHELP](#)

Please **Login** for Carts
or [Register](#) to get an account.

HEHELP

Location: [ELSA](#) » [ELSA 2002](#) » [HE. Health](#)

Description: Health status

Item type: Question

Question text: Would you say your health is...

Answer type: Enumerated

Answer choices:

1. Excellent,
2. Very good,
3. Good,
4. Fair,
5. Or, poor?

Harmonized measure: This item is used in a [Harmonized survey](#).

Associated variables: [hehelp](#) ⓘ

Flowchart: [locate in flowchart](#)

Topics: [health in general](#)

Concurrent items: [HEHELP \(ELSA 2008\)](#), [HEHELP \(ELSA 2006\)](#), [HEHELP \(ELSA 2004\)](#)

« [previous item \(HEGENH\)](#) | [next item \(HEILL\)](#) »

User Comments:

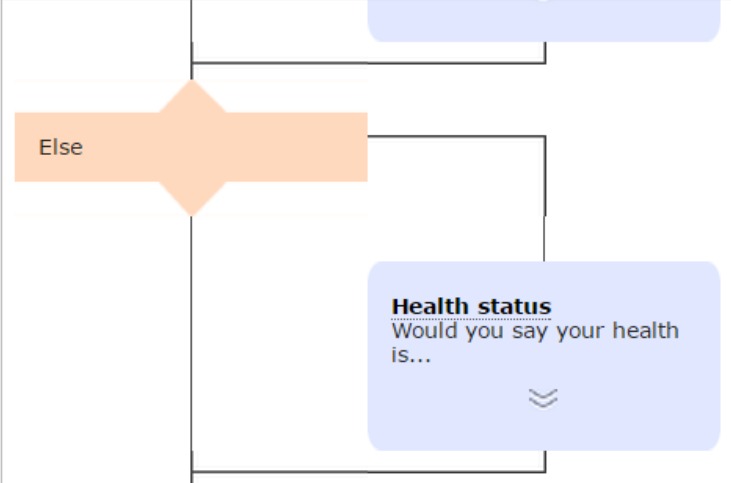
Be the first to comment!

[Log in](#) or [register](#) to comment.

HE. Health

The Health module concerns the respondent's state of health, functional limitations, and health behaviors. The main subsections are including self-reported general health, longstanding illness and limiting longstanding illness, eyesight, hearing, walking, awareness of disease, chronic diseases and psychiatric problems ever diagnosed by a doctor, falls, fractures sustained as a result of falls, joint replacement, dizziness, respiratory symptoms, urinary incontinence, disability and functioning, receipt of formal and informal care in connection with IADLs, health behaviors including smoking, drinking, fruit and vegetable consumption, and self-report of level of physical activity.

Module items (135) **Flowchart** Codebook



Whether has long-standing illness
Do you have any long-standing illness, disability or infirmity? By long-standing I mean anything that has troubled you over a period of time, or that is likely to affect you over a period of time?

If Whether has long-standing illness = 1. Yes »

Peer review of data



- Increasing in popularity
- Journals doing this - replicability agenda
- No one single standard for 'quality'
- Make metadata quality explicit:
 - Collection description
 - Data description: file and variable names & labels
 - Relationships between tables/files
 - Provenance of data and methods



Open data collections

94 open collections (out of 6553)

Government data - Open Government Licence (OGL)

- Census and survey teaching datasets

Survey data – Creative Commons CC4 BY, some NC

- Academic surveys, some qualitative data, historical data

Global indicators – bespoke open data license

- .STAT - World Bank Millennium Development goals

Open data requirements

- All methods and processes are transparent
- Data delivered via APIs where possible
- Self documenting, so absolute clarity needed about variables
- For time series, a concordance grid is very useful
- UKDS – Nesstar output to API



Keep connected with us

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



CONTACT

UK Data Service

University of Essex

Wivenhoe Park

Colchester

Essex CO4 3SQ

•
T +44 (0)1206 872145

E corti@essex.ac.uk

