Understanding and improving data linkage consent in surveys:

USER GUIDE

January 2022 (v2.0)

Annette Jäckle (University of Essex)

Jonathan Burton (University of Essex)

Mick P. Couper (University of Michigan)

Thomas F. Crossley (European University Institute)

Sandra Walzenbach (University of Konstanz)

Understanding Society

UKRI Economic and Social Research Council

Nuffield Foundation

Contents

1. Acknowledgements

These data were collected as part of a research project on "Understanding and improving data linkage consent in surveys", funded by the Nuffield Foundation with co-funding from the Economic and Social Research Council (OSP/43279). The questionnaires were implemented on the PopulusLive online access panel by NatCen Social Research under the excellent direction of Curtis Jessop.

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds research that informs social policy, primarily in Education, Welfare, and Justice. It also funds student programmes that provide opportunities for young people to develop skills in quantitative and qualitative methods. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics and the Ada Lovelace Institute. The Foundation has funded this project, but the views expressed are those of the authors and not necessarily the Foundation. Visit www.nuffieldfoundation.org.

2. Overview

The present data were collected as part of a project on "Understanding and improving data linkage consent in surveys". The aim of this project was to understand how respondents make decisions when asked for consent to link their survey data to government administrative records about them. In this case, we focus on data held by HM Revenue and Customs, the Department for Work and Pensions, the Department for Business, Energy and Industrial Strategy, the Department of Education and the National Health Service. The data include a series of survey experiments that were designed to test how question wording, order, and format of the consent request influence both the probability of consenting and how well the request is understood (informed consent). For further details about the project see Jäckle et al (2021).

The data are from three rounds of data collection that were implemented by NatCen Social Research on the PopulusLive online access panel (see https://www.populuslive.com/), with cross-sectional and longitudinal elements. The data are available to researchers from the UK Data Service. They complement data collected for the same project on the *Understanding Society* Innovation Panel wave 11 (available from the UK Data Services, SN6849).

3. How to cite the data and User Guide

The bibliographic citation for this user guide is:

Jäckle, A., Burton, J., Couper, M.P., Crossley, T.F., and Walzenbach, S. (2022) *Understanding and improving data linkage consent in surveys: User Guide.* Version 2.0, January 2022. Colchester: University of Essex.

The bibliographic citation for the main data is:

Jäckle, A., Burton, J., Couper, M.P., and Crossley, T.F. (2022). *Understanding and improving data linkage consent in surveys, 2018-2019.* [Data Collection]. Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-855036.

4. Fieldwork

The data were collected on two independent samples from the PopulusLive online access panel, which we refer to as the Access Panel (AP). The first sample was surveyed twice, with a one-year interval. The first wave (AP1-1) was fielded in June 2018 and included eleven experimental conditions with n~500 respondents each. A total of 46,206 panelists were invited to AP1-1, of whom 6,532 started the survey and 5,633 completed it (401 broke off and 498 were screened out), for a survey response rate of 12.2%. To track changes in consent over time, four of these eleven experimental groups (consent groups 1,2, 6 and 7 in Table 6.2) were re-interviewed about a year later (AP1-2). Of the 2,053 panelists invited to AP1-2, 1,693 started the survey and 1,630 completed it, for a response rate of 79.4%. As a follow up to the results from these two surveys, a second sample was drawn (AP2) and surveyed, with eight experimental groups designed to address further research questions. This sample was fielded in December 2019. A total of 30,682 panelists were invited to AP2, of whom 6,459 started the survey and 3,850 completed it (301 broke off and 2,308 were screened out), for a response rate of 21.1%. See Table 4.1 for a summary of sample sizes.

Table 4.1: Fieldwork dates and samples sizes

Survey	Fieldwork period	Respondents (N)	Respondents (N) per experimental condition
AP1-1	31/05 - 02/07/2018	5,684	513-523
AP1-2	20/05 – 02/06/2019	1,634	401-416
AP2	02/12 – 13/12/2019	3,850	476-487

The samples were restricted to Great Britain with quotas to match the composition of the *Understanding Society* Innovation Panel (University of Essex, Institute for Social and Economic Research 2019): gender (50% male, 50% female), age (33% 16-40, 33% 41-59, 33% 60+), and highest educational qualification (40% degree or equivalent, 20% A-level or equivalent, 40% GCSE or lower).

All surveys included either a single data linkage consent question or a set of five consent questions, as well as background questions on socio-demographics, understanding of the linkage request, perceived sensitivity of the consent request, trust in data holding institutions, and general data sharing attitudes and behaviours. Dependent on experimental group, median times for completion of the questionnaire ranged between 9 and 12 minutes (in AP1-1).

5. Questionnaire content and additional material

The survey questions can be found in the respective questionnaires. See Section 6 on how to read the questionnaire, and how the questions relate to variables in the data. Most of the questionnaire content was identical in all three surveys. Changes in the questionnaires mainly concerned the implementation of additional consent experiments, some additional questions, and minor modifications of existing questions.

Apart from general socio-demographics, sources of income and housing situation, all questionnaires contain:

- Data sharing attitudes and behaviours
- The consent request to link survey data to administrative records (experimentally varied)
- Questions on how the respondents made the consent decision
- Confidence in the decision made
- Objective and subjective understanding of the consent request
- Sensitivity of the consent request
- Trust in the data holder

In addition, the consent request came with additional material that respondents could click on to access more detailed information about data linkage. Firstly, there was a leaflet that explained the mechanics of the linkage process in text form. Secondly, a flowchart illustrated the process visually. While there was only one version of the leaflet, respondents either received an easy or standard version of the flowchart, dependent on their experimental condition. The easy question wording came with the easier flowchart, the standard question with the more difficult one. Both versions of the flowchart can be found in the Appendix.

6. Survey experiments

The study included a range of experimental variations of the consent request. Respondents were filtered into one of the following experimental conditions:

- single HMRC consent request with easy, standard, opt-in, opt-out wording or additional information / with or without trust priming
- single NHS consent with or without trust priming
- multiple consent sequence of five requests in different orders and formats (sequence of pages, same page, joint request)

Table 6.1 indicates which survey experiments were implemented in which survey. The study was designed to better understand how respondents make consent decisions. The survey data were not actually linked to the administrative records, even if respondents gave permission. Respondents were informed about this in a debrief section at the end of the AP1-2 and AP2 surveys.

Table 6.1. Survey experiments implemented in the three surveys

Consent request	AP1-1	AP1-2	AP2
Wording: easy vs. standard	x	x	
Wording: opt-in vs opt-out	x		
Wording: additional information	x		
Single consent: domain (HMRC vs NHS)			x
Single consent: trust experiment			x
Multiple consents: format	x	x	
Multiple consents: order	x		x

As summarized in Table 6.1, respondents answered different variations of the data linkage request. In all data sets, the experimental conditions that respondents were assigned to are recorded in the variable “ConsentGroup”. Table 6.2 provides a more detailed overview of this variable and helps to identify the comparable groups across waves.

Column 1 indicates the general topic of the experiment. Columns 2 to 4 list the respective experimental conditions that were implemented in AP1-1, AP1-2 and AP2, using the values and labels of the variable “ConsentGroup”. Identical and thus comparable experimental conditions are placed within the same row.

AP1-1 and AP1-2 share the same naming convention, meaning that for example ConsentGroup 2 refers to the condition with difficult wording in AP1-1 and AP1-2, but denotes a different experimental condition in the follow-up data collection of AP2. This also means that, even if label and numbering are not consistent across the AP1 and AP2 surveys, some experimental conditions are identical in AP1 and AP2 and can be compared.

Table 6.2: Overview of experimental conditions by data collection (as indicated by the variable "ConsentGroup" in the survey data)

Experiment topic	AP1-1	AP1-2	AP2
Wording: easy vs. standard*	1 Easy 2 Difficult (Standard)	1 Easy 2 Difficult (Standard)	1 HMRC
Wording: opt-in vs opt-out	9 Consent as default		
Wording: additional information	10 Additional information with follow up 11 Additional information		
Single consent: domain and trust			1 HMRC 2 HMRC with trust statement 3 NHS 4 NHS with trust statement
Multiple consents: order and format	3 Most sensitive first-sequence 4 Most sensitive first-joint request 5 Most sensitive first-single response 6 Least sensitive first-sequence 7 Least sensitive first-joint request 8 Least sensitive first-single response	6 Least sensitive first-sequence 7 Least sensitive first-joint request	5 Order 1 (HMRC, DWP, BEIS, EDU, NHS) 8 Order 4 (NHS, EDU, BEIS, DWP, HMRC) 6 Order 2 (HMRC, EDU, BEIS, DWP, NHS) 7 Order 3 (NHS, DWP, BEIS, EDU, HMRC)

*The "standard" question wording in ConsentGroup 2 of AP1-1 and AP1-2 refers to the standard that has previously been used for consent requests in the *Understanding Society* surveys. The "easy" wording breaks the information into bullet points, avoids passive voice, consists of shorter sentences, and contains more information. The easy wording results in a better readability according to two different readability scores (Flesch-Kincaid Grade level scores: difficult 14.3 – easy 8.8). This easy wording was also used for all other experimental conditions that do not explicitly mention wording in their labels (see questionnaires for exact wordings).

7. How to read the questionnaires

For each question, the questionnaires document the question name, the routing instructions defining which sample members were asked the question, the question wording and response options. Figure 7.1 provides an example to illustrate the questionnaire specification and how this relates to the variables in the data.

Figure 7.1: Example question specification

CSUndstd2 [VARLAB - Subjective understanding of consent request]

Universe - Ask all

How well do you think you understand what would happen with your data, if you allowed us to link it to records held by {IF CONSENTGROUP = 1, 2, 9, 10, 11: HM Revenue and Customs? / IF CONSENTGROUP = 3, 4, 5, 6, 7, 8: government departments}?

Please select one only

1 I do not understand at all
2 I understand somewhat
3 I mostly understand
4 I completely understand

The variable corresponding to the question in Figure 7.1 is called "CSUndstd2". The label for that variable is "Subjective understanding of consent request", and its values (1 to 4) are labelled according to the response categories in the questionnaire specification.

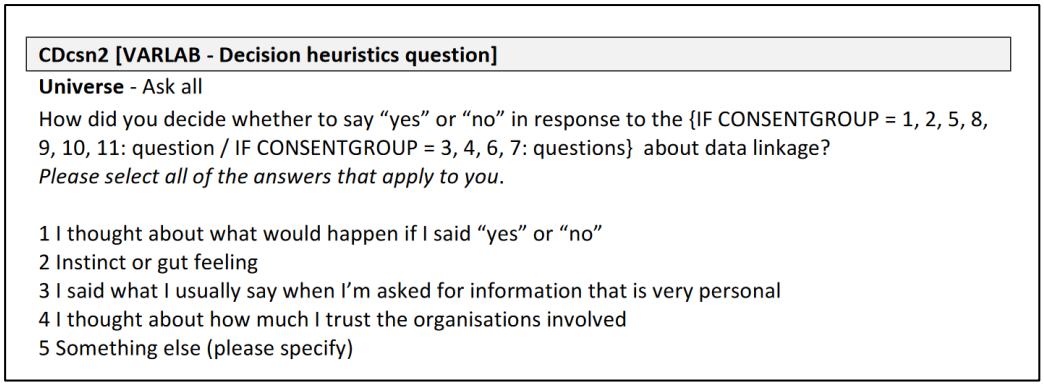
The Universe specifies who was eligible for this question: in this case all respondents. In the example, the question wording itself contains some additional scripting notes in parentheses, because the question wording was not exactly the same in all experimental conditions. If the respondent answered a single consent question (ConsentGroup is 1, 2, 9, 10, 11), they were asked for records held by "HM Revenue and Customs". If the respondent answered multiple consent requests (ConsentGroup is 3, 4, 5, 6, 7, 8), the wording was adapted to refer to records held by several "government departments".

8. Data structure and naming conventions

8.1 Naming of variables from multicode questions

For some questions, respondents are asked to *"Please select all that apply"* from a list of response options. For such multicode questions, the data files include one variable for each response option. The variable indicates if the response option was ticked or not. These binary indicators are named according to the question name documented in the questionnaire, followed by a number corresponding to the response option. As an example, for the question shown in Figure 8.1, the responses are recorded in the variables "CDcsn21", "CDcsn22", "CDcsn23", "CDcsn24" and "CDcsn25".

Figure 8.1: Example multicode question



8.2 Missing values

Missing observations are recorded using negative values rather than system missing values. The code -1 indicates "Don't know, code -2 indicates "Refusal". Respondents were shown these two response options if they clicked "Next" without selecting a response option. In addition, there is a code for questions that were not applicable: -8. This code is used for questions that the respondent was not asked due to the routing in the questionnaire.

8.3 Data files

Table 8.1 lists the ten data sets that are available from the UK Data Service. The prefixes indicate the sample and wave, in which the respective data were collected: AP1-1, AP1-2 or AP2.

Table 8.1 Data sets

Names of data file	Content
AP1-1Survey AP1-2Survey AP2Survey	contains the survey data; one file for each data collection
AP1-1Profile AP2Profile	includes all invited sample members and some info on non-respondents; one file for AP1, one file for wave AP2
AP1-1Paradata AP1-2Paradata AP2Paradata	contains string variables with response latency times; one file for each data collection
AP1.1 clean time stamps AP2 clean time stamps	contains response times derived from the paradata files for each experimental consent question

8.4 How to link data from the PROFILE, SURVEY and PARADATA files

All datasets contain the unique personal identifier “pid”. This can be used to combine the survey data with the profile and/or paradata file of the same wave, and to combine the data from waves 1 and 2 for sample AP1.

9. Content of the data files

9.1 Contents of the SURVEY data files

Most of the variables in the SURVEY files correspond to the survey questions in the questionnaires (see Section 7). The files include some additional variables that are documented in Table 9.1. These include time stamps for the interview start and end time, and the survey “Outcome”, indicating whether the respondent completed the full questionnaire or dropped out part-way through.

The variable “Diagram” represents the version of the flowchart that respondents were offered to help them understand the data linkage process. This was embedded as a link on the consent question page. In AP1-1 and AP1-2 the flowchart was either “easy” or “standard” and sample members were randomly allocated to one of the two groups. (See the Appendix for the flowcharts.) In AP2 all respondents received the easy version and “Diagram” either indicates “HMRC” or “NHS” records, depending on the consent question to which respondents were randomly allocated.

In addition, AP1-2 contains some feed-forward variables, that include the answers to the consent question(s) that the respondent gave in the AP1-1 survey. These feed-forward variables were used to check on consistency of the consent decision within the survey, and to ask respondents about reasons for differences, if the response fed-forward from the first survey was different from the response in the second survey.

Table 9.1: Additional variables in the SURVEY data

Variable	Description	Values	AP1-1	AP1-2	AP2
pid	Unique identifier	numeric	X	X	X
Outcome	Outcome code	110 Fully productive 210 Timed out 310 Screened out	X	X	X
DateStart	Date started survey	%tdD_m_Y (DD m YY)	X	X	X
TimeStart	Time started survey	%tc (HH:MM:SS)	X	X	X
DateEnd	Date ended survey	%tdD_m_Y (DD m YY)	X	X	X
TimeEnd	Time ended survey	%tc (HH:MM:SS)	X	X	X
Diagram	Version of flowchart explaining linkage	1 Version A (easy) 2 Version B (difficult)	X	X	X
ff_Country	Country of residence - fed forward from previous wave			X	
ff_ConsentQ1	Consent to Q1 - fed forward from previous wave	1 Yes 2 No		X	
ff_ConsentQ2	Consent to Q2 - fed forward from previous wave	1 I have read the leaflet and am happy to give consent 2 I do not want to give consent		X	
ff_ConsentQ6a to ff_ConsentQ6e	Consent to Q6a to Q6e - fed forward from previous wave	1 Yes 2 No		X	
ff_ConsentQ7a to ff_ConsentQ7e	Consent to Q7a to Q7e - fed forward from previous wave	1 Yes 2 No		X	

9.2 Contents of the PROFILE data files

There is one profile data file for each sample, AP1 and AP2. These files were derived from the access panel sample database and included observations on each selected sample member, so including respondents and non-respondents. Both files contain the variables shown in Table 9.2.

Table 9.2: Contents of the PROFILE data files

Variable	Description	Values
pid	Unique identifier	numeric
Outcome	Survey outcome code	110 Fully productive 210 Timed out 310 Screened out
Income	Combined income (categorised)	1 Up to £7,000 2 £7,001 to £14,000 3 £14,001 to £21,000 4 £21,001 to £28,000 5 £28,001 to £34,000 6 £34,001 to £41,000 7 £41,001 to £48,000 8 £48,001 to £55,000 9 £55,001 to £62,000 10 £62,001 to £69,000 11 69,001 to £76,000 12 £76,001 to £83,000 13 £83,001 or more
Education	Highest Level of Education	1 No Formal Education 2 Primary Education 3 Secondary school, high school, NVQ levels 1 to 3 4 University degree or equivalent professional qualification, NVQ 4 5 Higher university degree, doctorate, MBA, NVQ level 5 6 Still in full-time education 7 None of these
Children	Number of Children	numeric
Marital_Status	Marital Status	1 Single - Living on own 2 Single - Living with others 3 Single - Living with parents 4 Cohabitating 5 Married 6 Civil partnership 7 Separated 8 Divorced

		9 Widowed
Tenure	Tenure	1 Owned outright (without mortgage) 2 Owned with a mortgage or loan 3 Living with parents/relatives 4 Rented from Council 5 Rented from housing association 6 Rented from someone else 7 Rent free
Ethnicity	Ethnicity	1 White 2 Black or Black British 3 Asian or Asian British 4 Chinese 5 Mixed 6 Other ethnic group
Mobile01	A pay-as-you-go mobile telephone account	0 No 1 Yes
Mobile02	A pay-monthly mobile telephone account that you pay for	0 No 1 Yes
Mobile03	A pay-monthly mobile telephone / blackberry account that is paid	0 No 1 Yes
Mobile04	A SIM Only mobile telephone account	0 No 1 Yes
Mobile05	None of these	0 No 1 Yes

9.3 Contents of the PARADATA files

The paradata files contain the variables listed in Table 9.3: the respondent identifier (pid), the survey outcome (Outcome), a string identifying the browser used to complete the survey (BrowserInfo), and a string identifying the device used to complete the survey (Device). In addition it contains the allocation to experimental treatment groups (ConsentGroup) as documented in the corresponding questionnaire.

The file also contains one string variable for each of the survey questions. Variable names are a combination of the respective question that the paradata refers to and the suffix "Para". The string variables themselves first contain a marker of the question type ("%SinglePunch%" if respondents were asked to select one response option, "%MultiPunch%" for *tick all that apply* questions). Then follows a series of dates and times and the events on the respective page of the online survey. To illustrate, this would be the paradata string variable "ConsentQ1Para" referring to the survey question ConsentQ1:

%SinglePunch%20/05/2019 20:03:32:054-#ConsentQ1#;20/05/2019 20:04:02:878-No:On;20/05/2019 20:04:04:459-!Next!;

It indicates that “ConsentQ1” is a single punch question. The respondent arrived on the page with this question at 20:03:32 on 20.5.2019. Half a minute later, at 20:04:02, the respondent ticked the response category “No” and left the page by clicking “Next” at 20:04:04.

The paradata for the consent questions also contain information on whether and when respondents clicked on the links to additional information: the information leaflet and the flowchart illustrating the linkage process. The following paradata string provides an example:

%SinglePunch%20/05/2019 19:07:50:393-#ConsentQ7#;20/05/2019 19:09:19:827-=leaflet ON=;20/05/2019 19:09:59:844-=leaflet OFF=; …

Table 9.3: Contents of the PARADATA data files

Variable	Description	Values
pid	Unique identifier	numeric
Outcome	Survey outcome code	110 Fully productive 210 Timed out 310 Screened out
BrowserInfo	Browser used to complete survey	String, e.g. Mozilla/5.0 (Linux; Android 5.0.2; HTC One Build/LRX22G) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/67.0.3396.68 Mobile Safari/537.36

9.4 Contents of the CLEAN TIME STAMPS files

These files contain the response times for of the experimental consent questions, derived from the timestamps in the paradata files. The files contain the variables listed in Table 9.4: the respondent identifier (pid), the randomized treatment allocation (ConsentGroup), the time taken to answer each of the single consent questions (TT_ConsentQ*), and the total response time for those asked multiple consents (consent_secs). The file also includes binary indicators for each of the experimental consent questions for whether the respondent clicked on the diagram (D_ConsentQ*) and whether they clicked on the leaflet (L_ConsentQ*). For the multiple consent questions that were asked as a sequence of pages, the links to the leaflet and diagram were only shown on the first page with the first consent question. For this reason, the respective variables only exist for the first question in every sequence.

Table 9.4: Contents of the cleaned timestamps files

Variable	Description	Values
pid	Unique identifier	numeric
ConsentGroup	Randomised allocation to consent treatment group	"AP1.1 clean time stamps.dta": 1 Easy 2 Difficult (Standard) 3 Most sensitive first-sequence 4 Most sensitive first-joint request 5 Most sensitive first-single response 6 Least sensitive first-sequence 7 Least sensitive first-joint request 8 Least sensitive first-single response 9 Consent as default 10 Additional information with follow up 11 Additional information "AP2 clean time stamps.dta": 1 HMRC 2 HMRC with trust statement 3 NHS 4 NHS with trust statement 5 Sequence 1 (HMRC – DWP – BEIS – EDU – NHS) 6 Sequence 2 (HMRC – EDU – BEIS – DWP – NHS) 7 Sequence 3 (NHS – DWP – BEIS – EDU – HMRC) 8 Sequence 4 (NHS – EDU – BEIS – DWP – HMRC)
TT_ConsentQ*	Response time	Total time (milliseconds) taken to answer ConsentQ* [for single consent questions only]
consent_secs	Total response time	Total response time (seconds) for multiple consents
L_ConsentQ*	Whether clicked on leaflet for ConsentQ*	0 No 1 Yes
D_ConsentQ*	Whether clicked on diagram for ConsentQ*	0 No 1 Yes

10.References

Jäckle, A., Burton, J., Couper, M.P., Crossley, T.F., and Walzenbach, S. (2021) "Understanding and improving data linkage consent in surveys", *Understanding Society Working Paper 2021-01*. Colchester: University of Essex. Available at https://www.understandingsociety.ac.uk/sites/default/files/downloads/working-papers/2021-01.pdf

University of Essex, Institute for Social and Economic Research. (2019). Understanding Society: Innovation Panel, Waves 1-11, 2008-2018. [data collection]. 9th Edition. UK Data Service. SN: 6849, http://doi.org/10.5255/UKDA-SN-6849-12.

Project webpage: https://www.iser.essex.ac.uk/research/projects/understanding-and-improving-data-linkage-consent-in-surveys

11. Appendix: Flowcharts visualizing the data linkage proces

Figure 11.1: Standard version

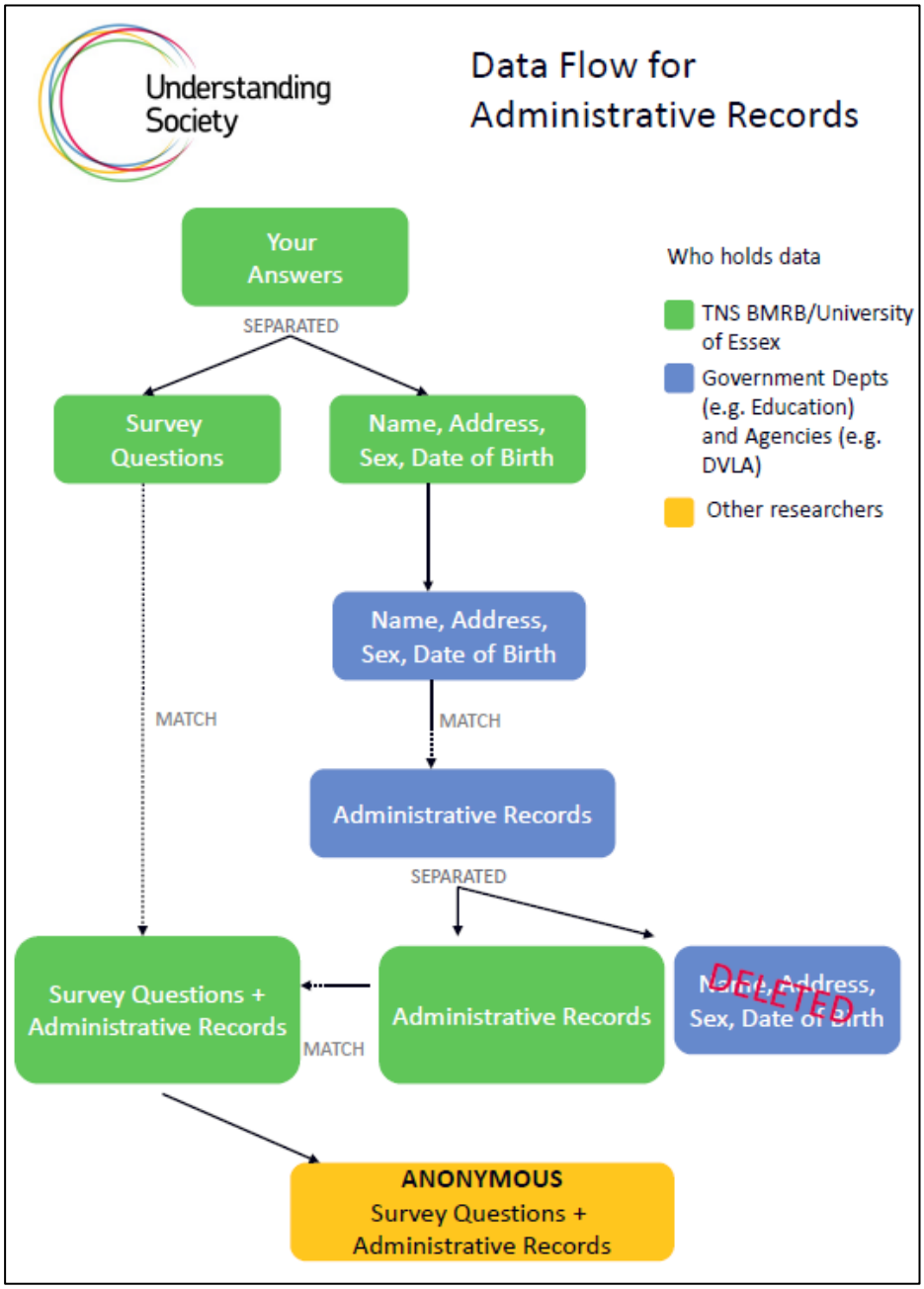


Figure 11.2: Simplified version