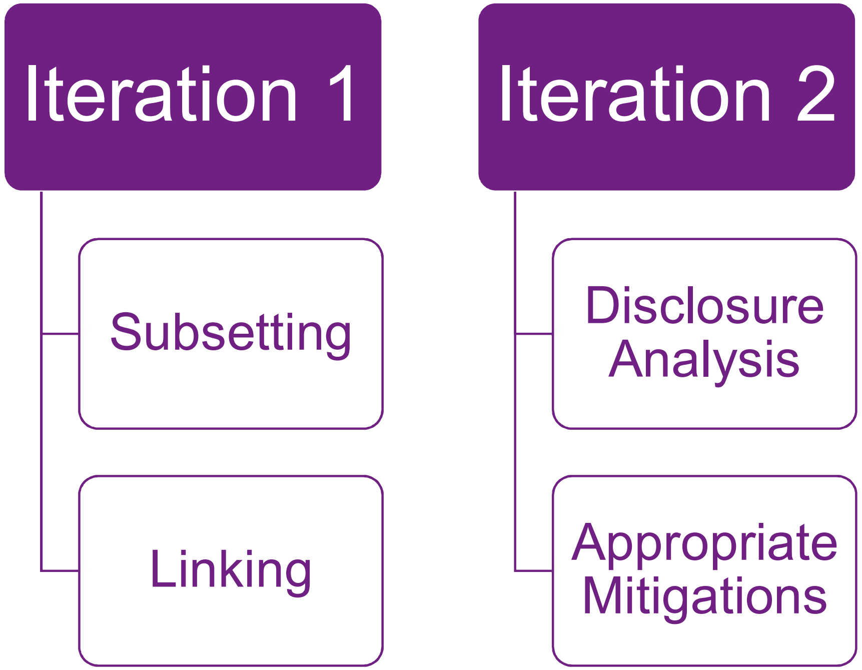
UK Data Service

Powerful DDI-CDI Metadata – the How and the Why?

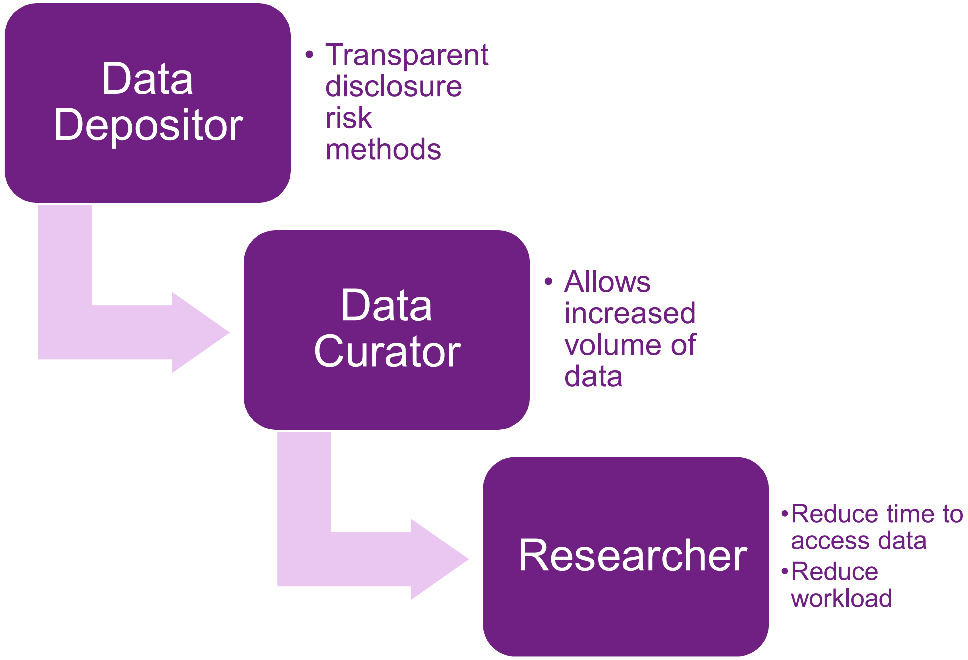
Darren Bell, Associate Director, IT Services
Thomas Gilders, Data Engineer
Deirdre Lungley, Principal Developer
Ivan Evdokimov, Data Engineer
Jon Johnson, Closer

UKRI Economic and Social Research Council

UKDS Data Product Builder



Why?

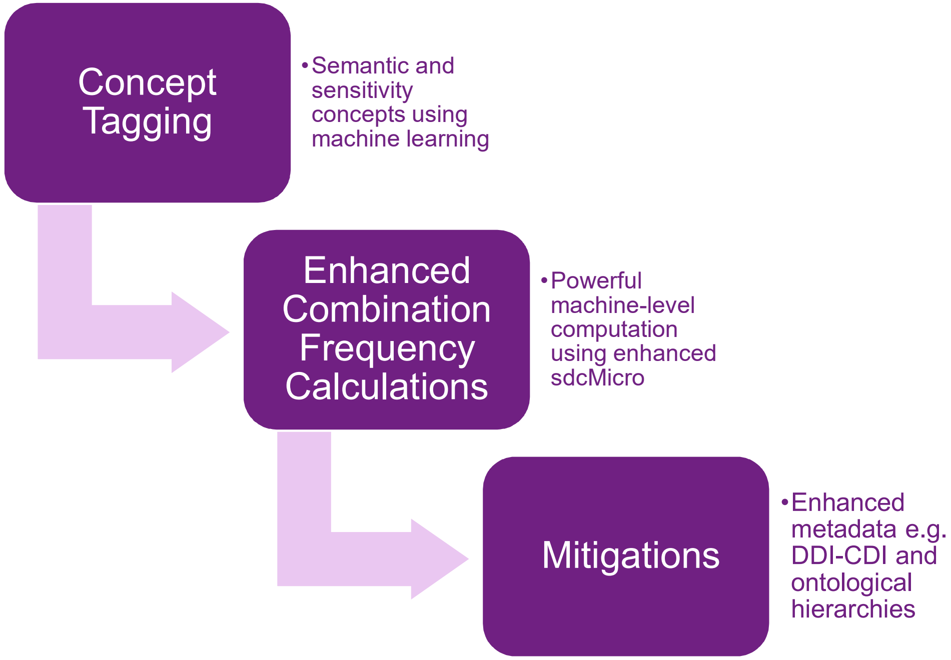


- Transparent disclosure risk methods

- Allows increased volume of data

- Reduce time to access data
- Reduce workload

How?



Concept Tagging

- Semantic and sensitivity concepts using machine learning

Enhanced Combination Frequency Calculations

- Powerful machine-level computation using enhanced sdcMicro

Mitigations

- Enhanced metadata e.g. DDI-CDI and ontological hierarchies

Concept Tagging

- Input features derived from:
  - Variable name
  - Variable label
  - Question text
  - Variable group and subgroup
- Multiple models:
  - Model per variable group, e.g., Standard Occupational Classification
  - Sensitivity model
- Machine Learning Methods:
  - FastAI: language based model
  - SVM: Support Vector Machines
  - KNN: K-Nearest Neighbour

UK Data Service

Enhanced Combination Frequency Calculations

- The existing tool we use is sdcMicro:
- GUI would not allow load of dataset as large as QLFS
- With scripting – possible but quite slow
- We have achieved 3-fold performance improvements by using C++ bitmask operations in place of the original R code
- Makes real-time disclosure analysis feasible

UK Data Service

Enhanced Combination Frequency Calculations II

	A	B	C	fk
0	22	77	44	**3**
1	33	77	66	**1**
2	22	77	-9	**4**
3	22	77	55	**2**
4	33	77	44	**2**
5	33	77	44	**2**
6	11	88	-9	**1**
7	22	-9	44	**3**

Tested with the Quarterly Labour Force Survey
~96,000 rows

10 key variables

2-, 3- and 4-way combinations = 375 permutations
36 million rows in total

~5s to compute the bitmasks*

~15s to compute the fk frequency counts for all combinations*
~240s to compute weighted Fk as well

* Intel core i7-12700H, 32 GB

Population-level Combination Frequencies

- Traditional Disclosure Risk Analysis (DRA) has relied on ‘data experts’
- We’re currently exploring the feasibility of using Census aggregate data to inform automated DRA
- Population-level combination frequencies can be checked when sample frequencies are low
- Demo to illustrate this

UK Data Service

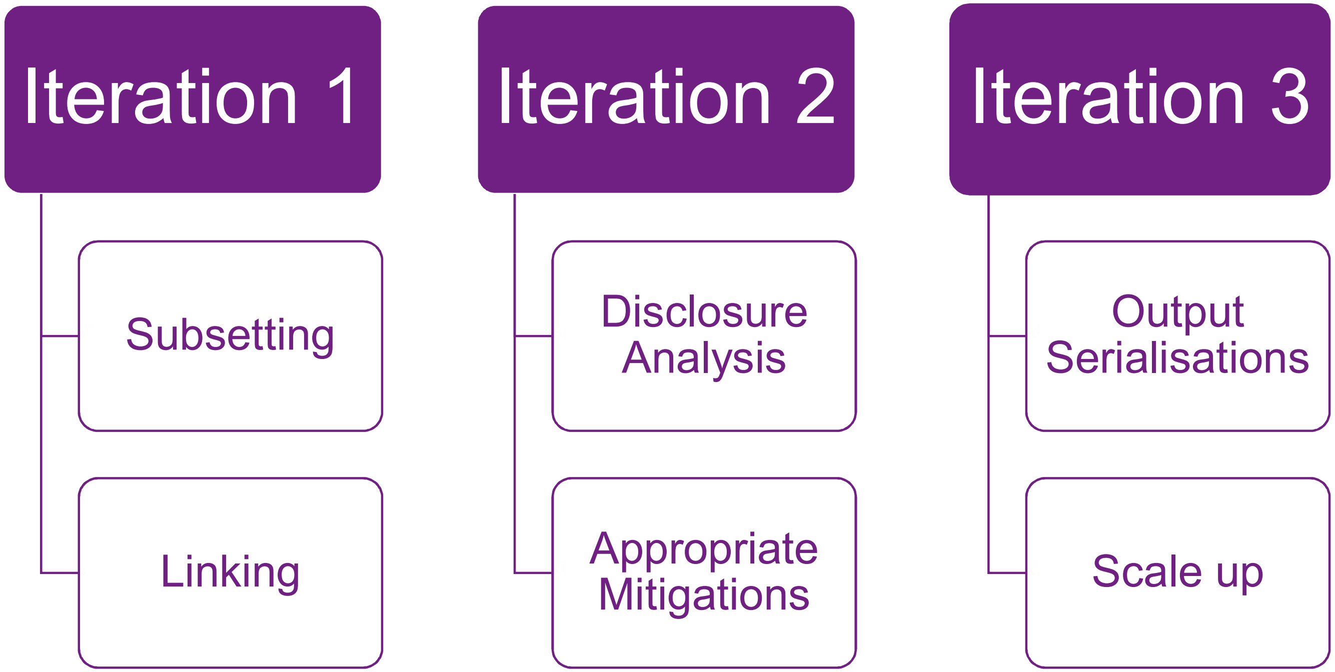
Mitigations

Our Data Product Builder aims to offer metadata driven mitigations:

- Top/Bottom coding examples:
    - Age top and bottom coding
    - Salary top-coding
- Rebanding examples:
    - Broader geographical concept
    - Broader occupation concept
    - Broader ethnicity concept

DEMO

UKDS Data Product Builder



Iteration 1

Subsetting

Linking

Iteration 2

Disclosure Analysis

Appropriate Mitigations

Iteration 3

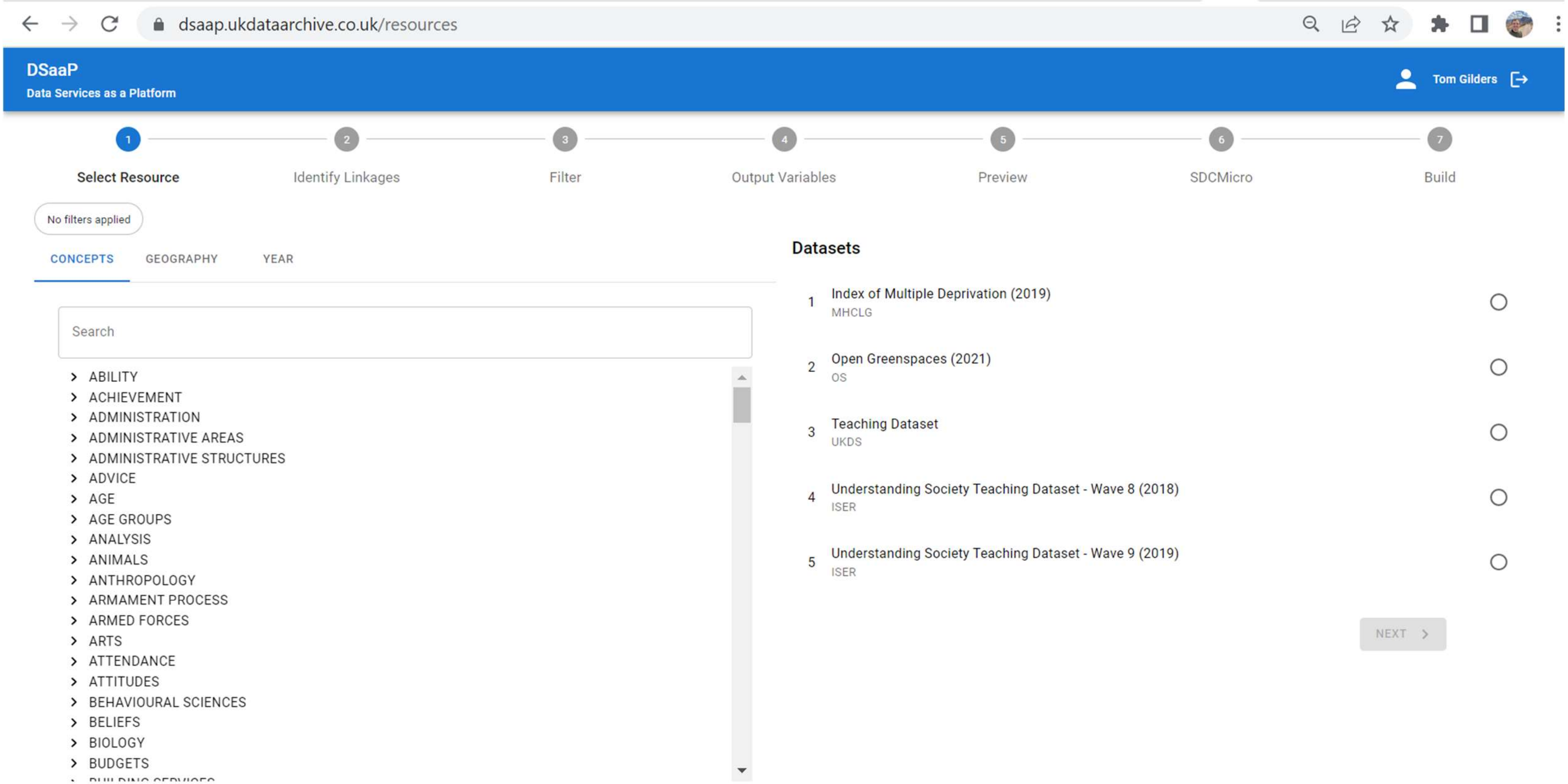
Output Serialisations

Scale up

Any questions?

UK Data Service

Thank you.



DSaaP
Data Services as a Platform

- ABILITY
- ACHIEVEMENT
- ADMINISTRATION
- ADMINISTRATIVE AREAS
- ADMINISTRATIVE STRUCTURES
- ADVICE
- AGE
- AGE GROUPS
- ANALYSIS
- ANIMALS
- ANTHROPOLOGY
- ARMAMENT PROCESS
- ARMED FORCES
- ARTS
- ATTENDANCE
- ATTITUDES
- BEHAVIOURAL SCIENCES
- BELIEFS
- BIOLOGY
- BUDGETS

Datasets

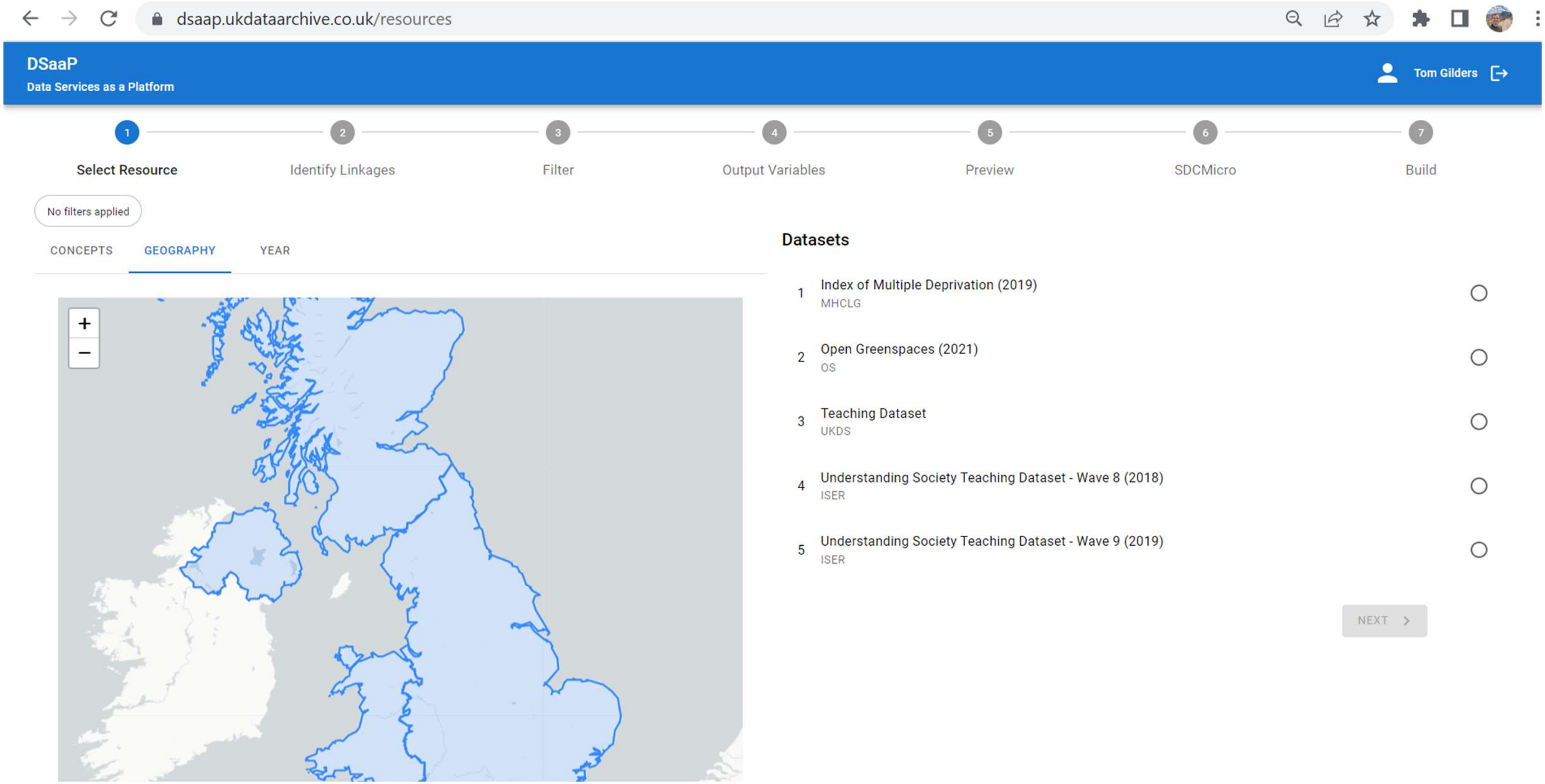
1 Index of Multiple Deprivation (2019) ○
MHCLG

2 Open Greenspaces (2021) ○
OS

3 Teaching Dataset ○
UKDS

4 Understanding Society Teaching Dataset - Wave 8 (2018) ○
ISER

5 Understanding Society Teaching Dataset - Wave 9 (2019) ○
ISER



dsaap.ukdataarchive.co.uk/resources

DSaaP

Data Services as a Platform

Tom Gilders

1 Select Resource — 2 Identify Linkages — 3 Filter — 4 Output Variables — 5 Preview — 6 SDCMicro — 7 Build

No filters applied

CONCEPTS | GEOGRAPHY | YEAR

Datasets

1 Index of Multiple Deprivation (2019) ○
MHCLG

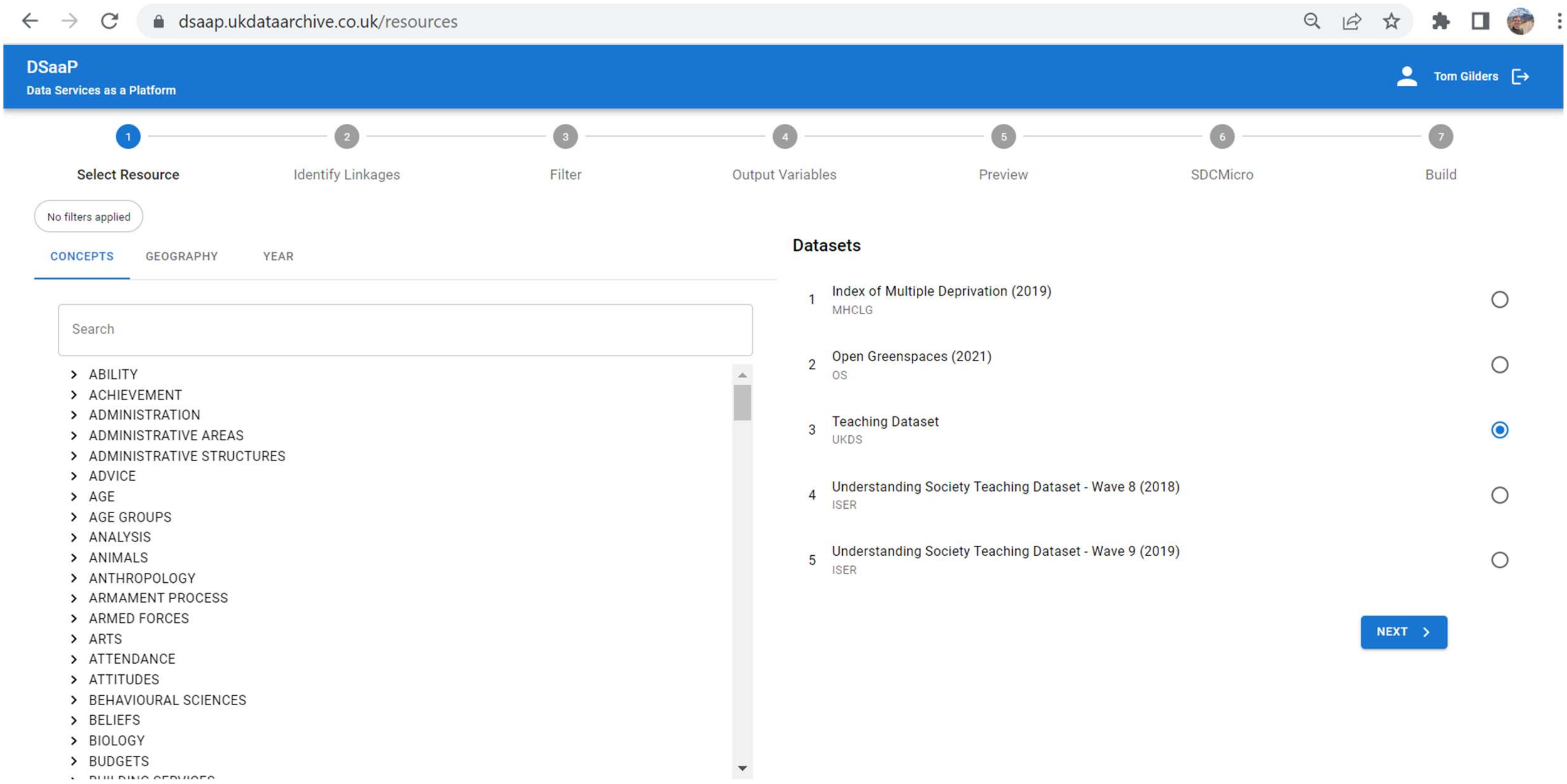
2 Open Greenspaces (2021) ○
OS

3 Teaching Dataset ○
UKDS

4 Understanding Society Teaching Dataset - Wave 8 (2018) ○
ISER

5 Understanding Society Teaching Dataset - Wave 9 (2019) ○
ISER

NEXT >



dsaap.ukdataarchive.co.uk/resources

DSaaP
Data Services as a Platform

Tom Gilders

1 Select Resource
2 Identify Linkages
3 Filter
4 Output Variables
5 Preview
6 SDCMicro
7 Build

No filters applied

CONCEPTS GEOGRAPHY YEAR

Search

- ABILITY
- ACHIEVEMENT
- ADMINISTRATION
- ADMINISTRATIVE AREAS
- ADMINISTRATIVE STRUCTURES
- ADVICE
- AGE
- AGE GROUPS
- ANALYSIS
- ANIMALS
- ANTHROPOLOGY
- ARMAMENT PROCESS
- ARMED FORCES
- ARTS
- ATTENDANCE
- ATTITUDES
- BEHAVIOURAL SCIENCES
- BELIEFS
- BIOLOGY
- BUDGETS
- BUILDING SERVICES

Datasets

1 Index of Multiple Deprivation (2019)
MHCLG ○

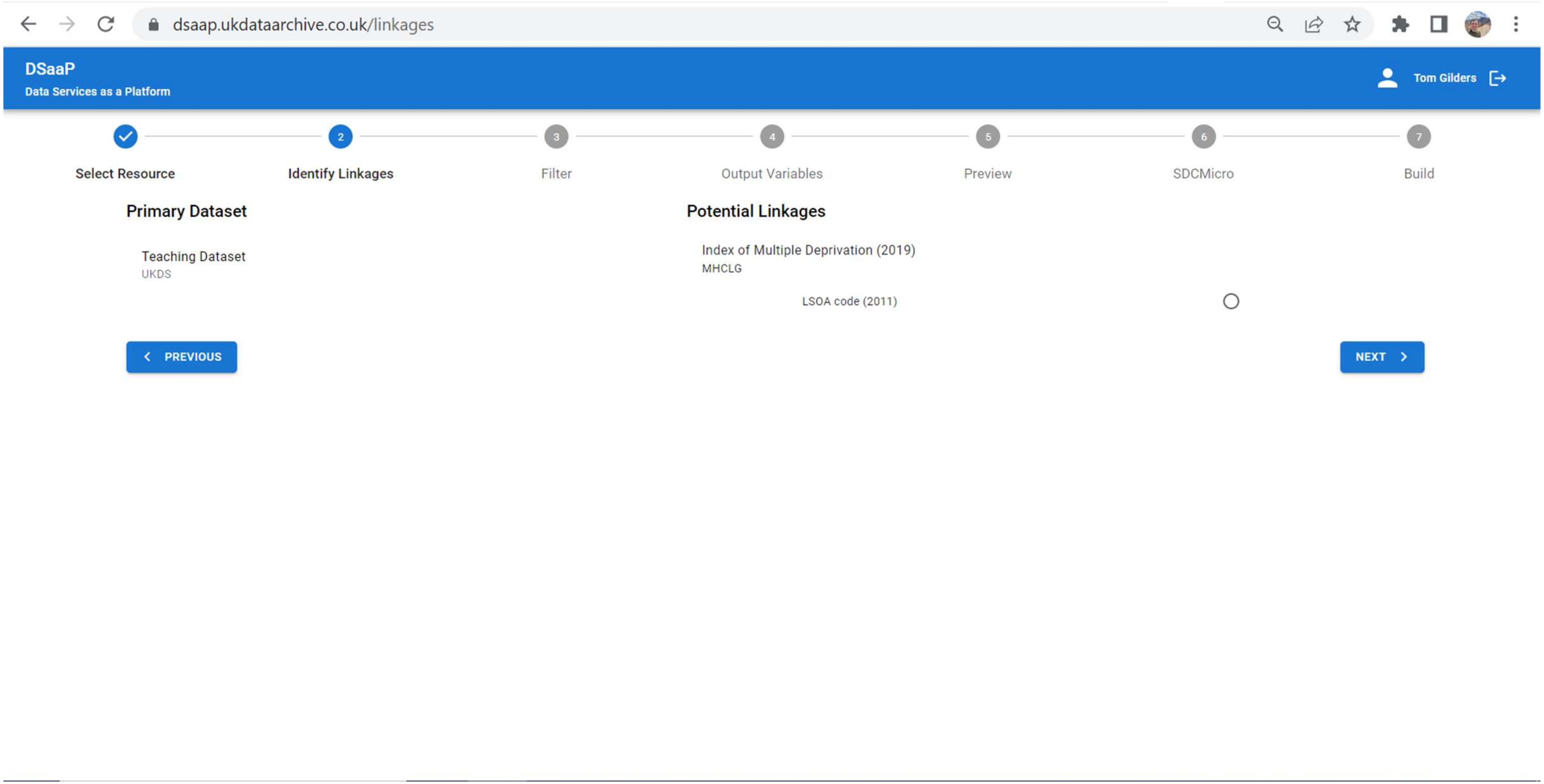
2 Open Greenspaces (2021)
OS ○

3 Teaching Dataset
UKDS ●

4 Understanding Society Teaching Dataset - Wave 8 (2018)
ISER ○

5 Understanding Society Teaching Dataset - Wave 9 (2019)
ISER ○

NEXT >



dsaap.ukdataarchive.co.uk/linkages

DSaaP
Data Services as a Platform

Tom Gilders

✓ Select Resource — 2 Identify Linkages — 3 Filter — 4 Output Variables — 5 Preview — 6 SDCMicro — 7 Build

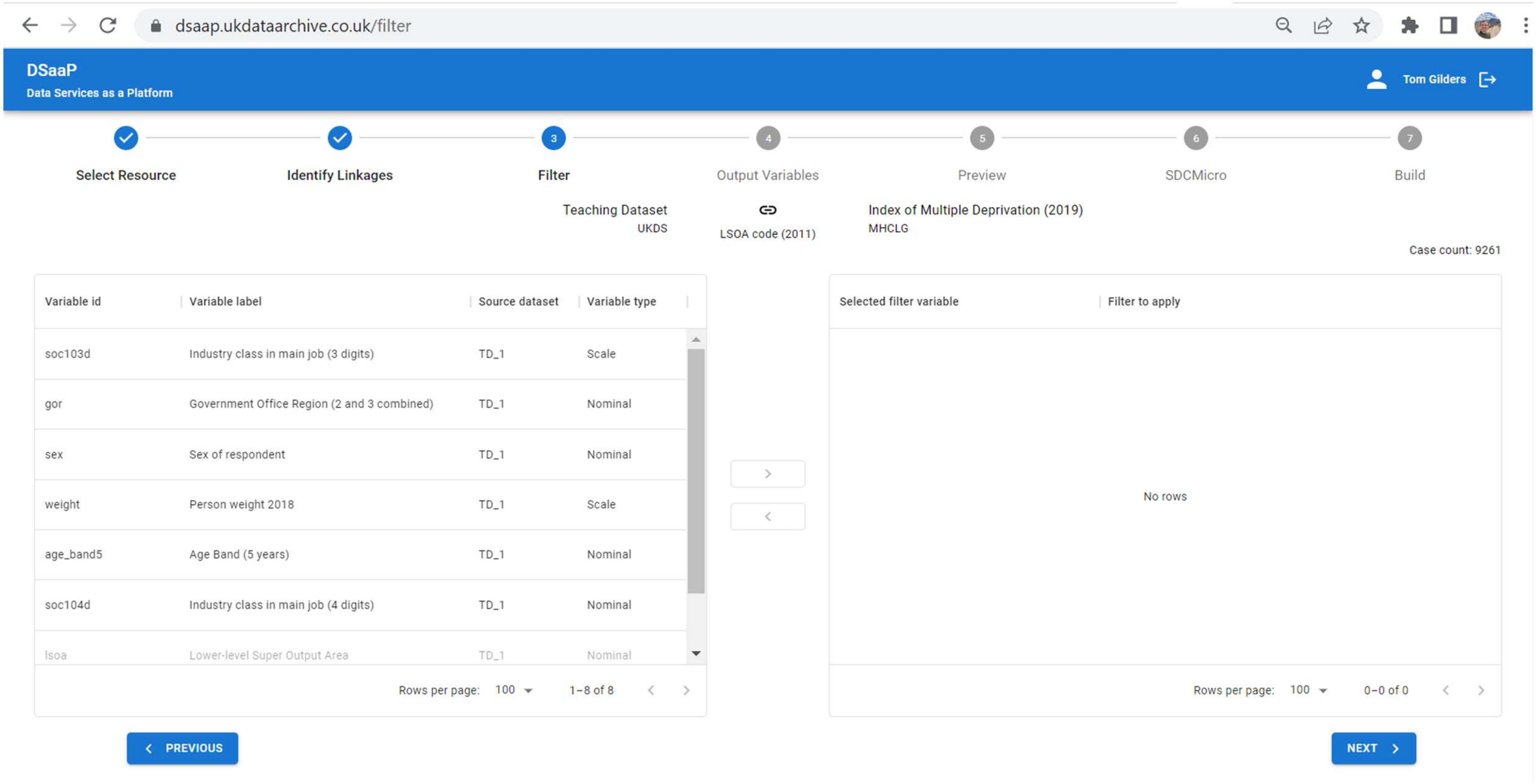
Primary Dataset

Teaching Dataset
UKDS

Potential Linkages

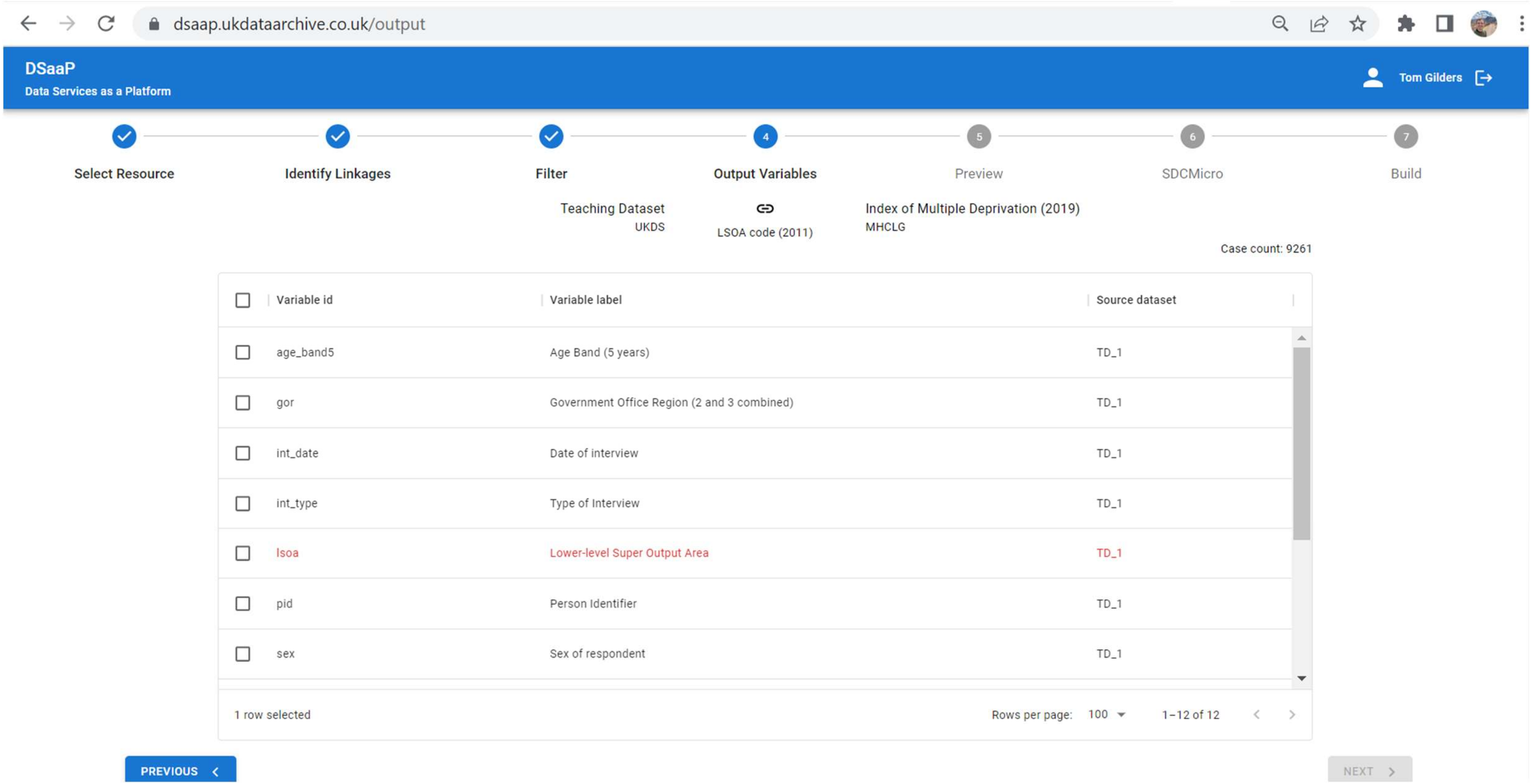
Index of Multiple Deprivation (2019)
MHCLG

LSOA code (2011) ○



Variable id	Variable label	Source dataset	Variable type
soc103d	Industry class in main job (3 digits)	TD_1	Scale
gor	Government Office Region (2 and 3 combined)	TD_1	Nominal
sex	Sex of respondent	TD_1	Nominal
weight	Person weight 2018	TD_1	Scale
age_band5	Age Band (5 years)	TD_1	Nominal
soc104d	Industry class in main job (4 digits)	TD_1	Nominal
lsoa	Lower-level Super Output Area	TD_1	Nominal

Selected filter variable	Filter to apply



Teaching Dataset
UKDS

LSOA code (2011)

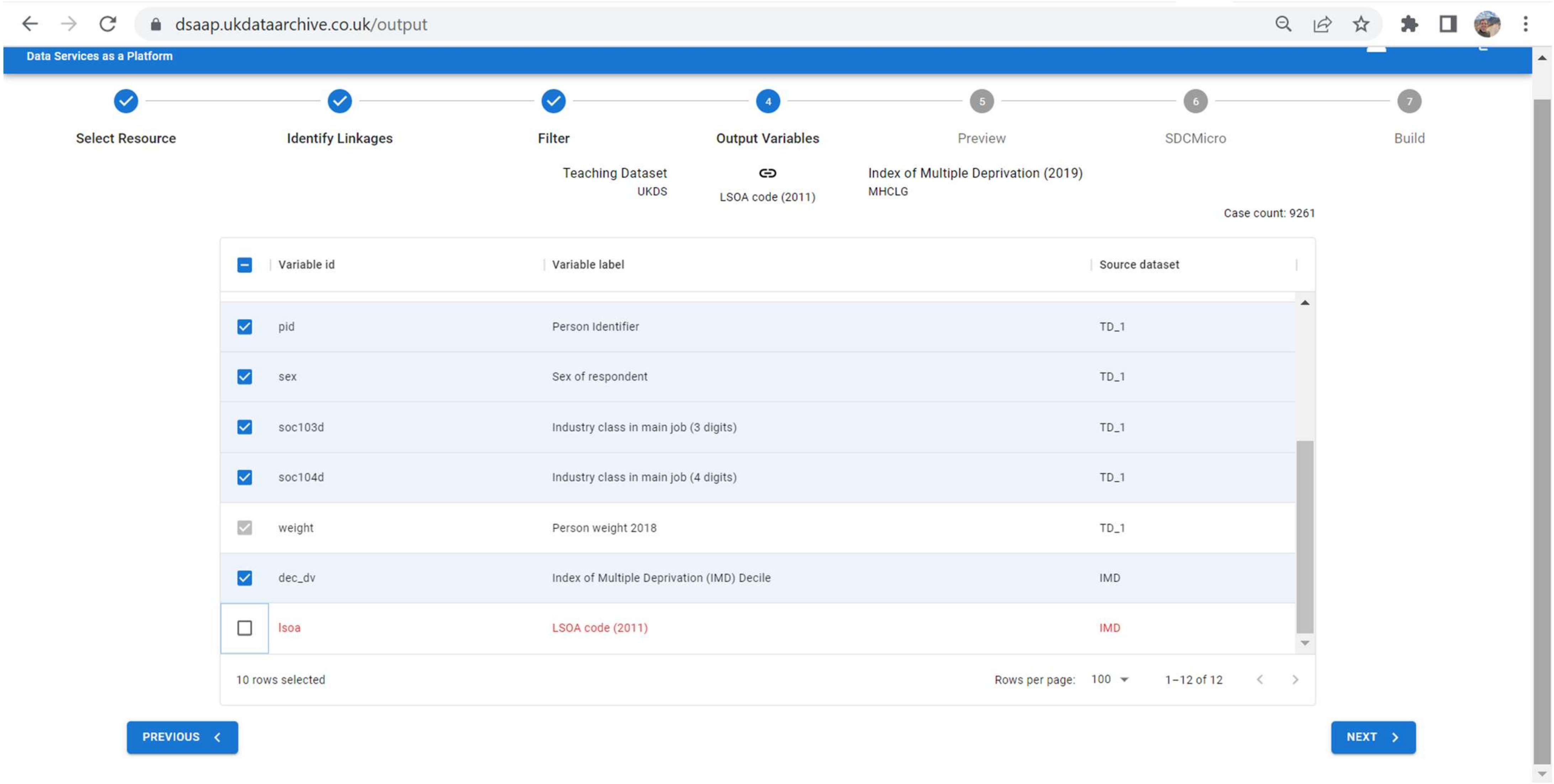
Index of Multiple Deprivation (2019)
MHCLG

Case count: 9261

□	Variable id	Variable label	Source dataset
□	age_band5	Age Band (5 years)	TD_1
□	gor	Government Office Region (2 and 3 combined)	TD_1
□	int_date	Date of interview	TD_1
□	int_type	Type of Interview	TD_1
□	lsoa	Lower-level Super Output Area	TD_1
□	pid	Person Identifier	TD_1
□	sex	Sex of respondent	TD_1

1 row selected

Rows per page: 100 1–12 of 12



dsaap.ukdataarchive.co.uk/output

Data Services as a Platform

Teaching Dataset
UKDS

LSOA code (2011)

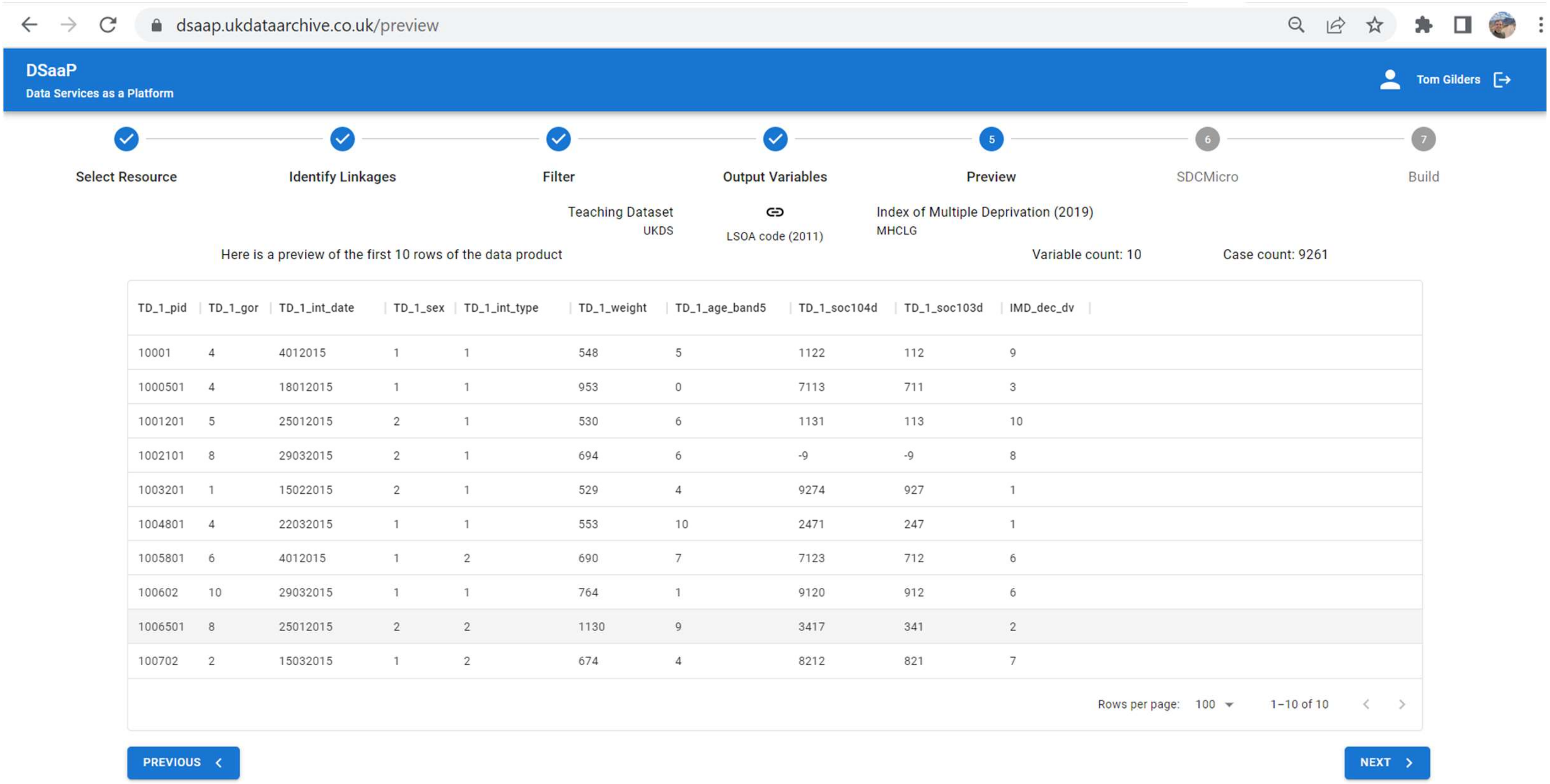
Index of Multiple Deprivation (2019)
MHCLG

Case count: 9261

☐	Variable id	Variable label	Source dataset
☑	pid	Person Identifier	TD_1
☑	sex	Sex of respondent	TD_1
☑	soc103d	Industry class in main job (3 digits)	TD_1
☑	soc104d	Industry class in main job (4 digits)	TD_1
☑	weight	Person weight 2018	TD_1
☑	dec_dv	Index of Multiple Deprivation (IMD) Decile	IMD
☐	lsoa	LSOA code (2011)	IMD

10 rows selected

Rows per page: 100 1–12 of 12



Teaching Dataset
UKDS

LSOA code (2011)

Index of Multiple Deprivation (2019)
MHCLG

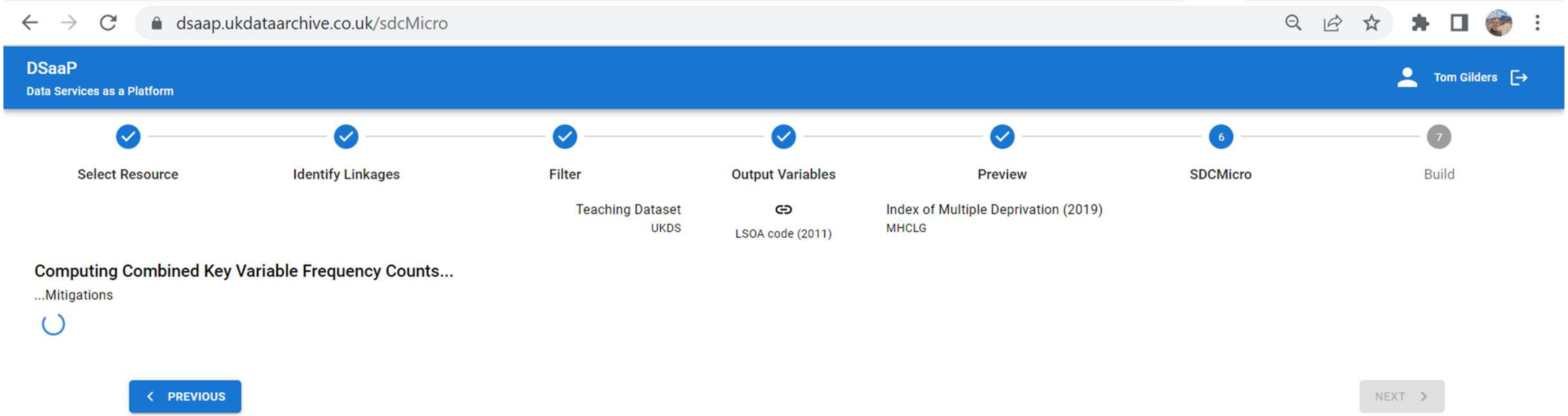
Here is a preview of the first 10 rows of the data product

Variable count: 10

Case count: 9261

TD_1_pid	TD_1_gor	TD_1_int_date	TD_1_sex	TD_1_int_type	TD_1_weight	TD_1_age_band5	TD_1_soc104d	TD_1_soc103d	IMD_dec_dv
10001	4	4012015	1	1	548	5	1122	112	9
1000501	4	18012015	1	1	953	0	7113	711	3
1001201	5	25012015	2	1	530	6	1131	113	10
1002101	8	29032015	2	1	694	6	-9	-9	8
1003201	1	15022015	2	1	529	4	9274	927	1
1004801	4	22032015	1	1	553	10	2471	247	1
1005801	6	4012015	1	2	690	7	7123	712	6
100602	10	29032015	1	1	764	1	9120	912	6
1006501	8	25012015	2	2	1130	9	3417	341	2
100702	2	15032015	1	2	674	4	8212	821	7

Rows per page: 100 ▾ 1–10 of 10



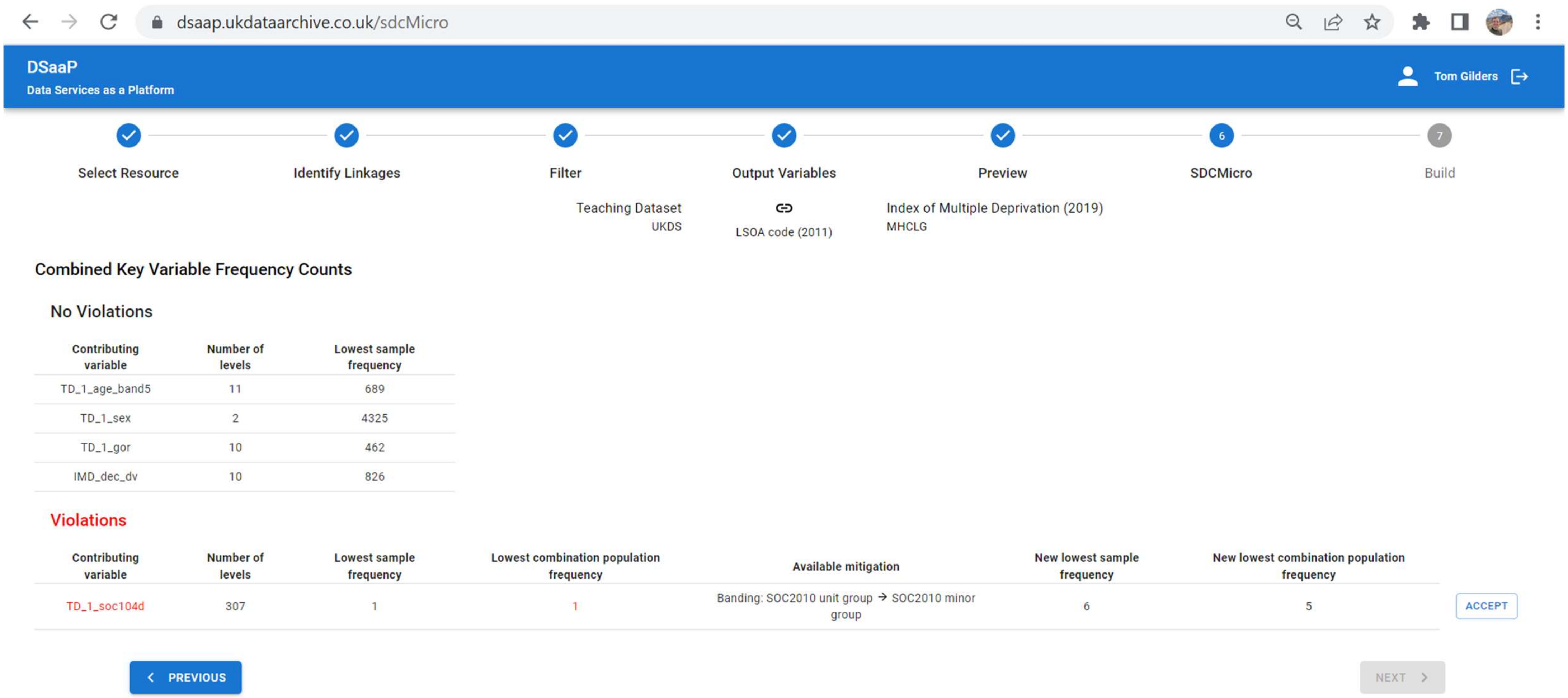
Teaching Dataset
UKDS

LSOA code (2011)

Index of Multiple Deprivation (2019)
MHCLG

Computing Combined Key Variable Frequency Counts...

...Mitigations



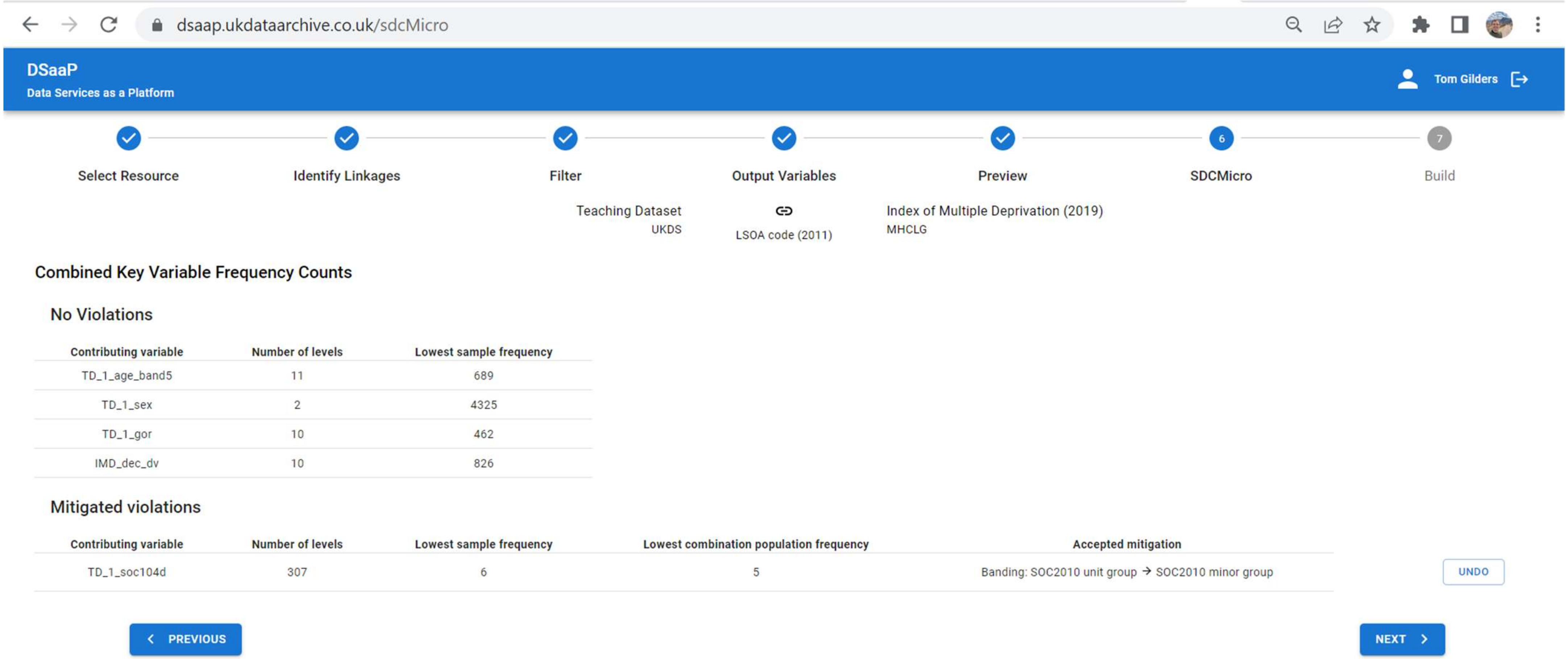
Combined Key Variable Frequency Counts

No Violations

Contributing variable	Number of levels	Lowest sample frequency
TD_1_age_band5	11	689
TD_1_sex	2	4325
TD_1_gor	10	462
IMD_dec_dv	10	826

Violations

Contributing variable	Number of levels	Lowest sample frequency	Lowest combination population frequency	Available mitigation	New lowest sample frequency	New lowest combination population frequency	
TD_1_soc104d	307	1	1	Banding: SOC2010 unit group → SOC2010 minor group	6	5	ACCEPT



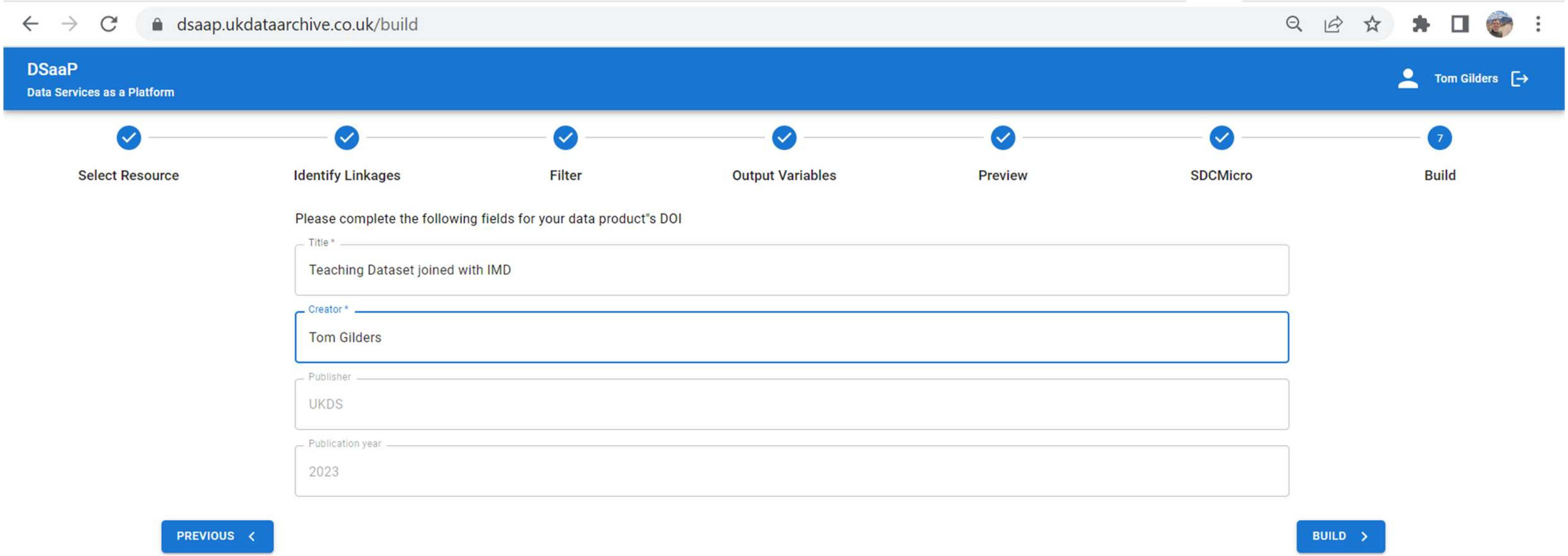
Combined Key Variable Frequency Counts

No Violations

Contributing variable	Number of levels	Lowest sample frequency
TD_1_age_band5	11	689
TD_1_sex	2	4325
TD_1_gor	10	462
IMD_dec_dv	10	826

Mitigated violations

Contributing variable	Number of levels	Lowest sample frequency	Lowest combination population frequency	Accepted mitigation	
TD_1_soc104d	307	6	5	Banding: SOC2010 unit group → SOC2010 minor group	UNDO



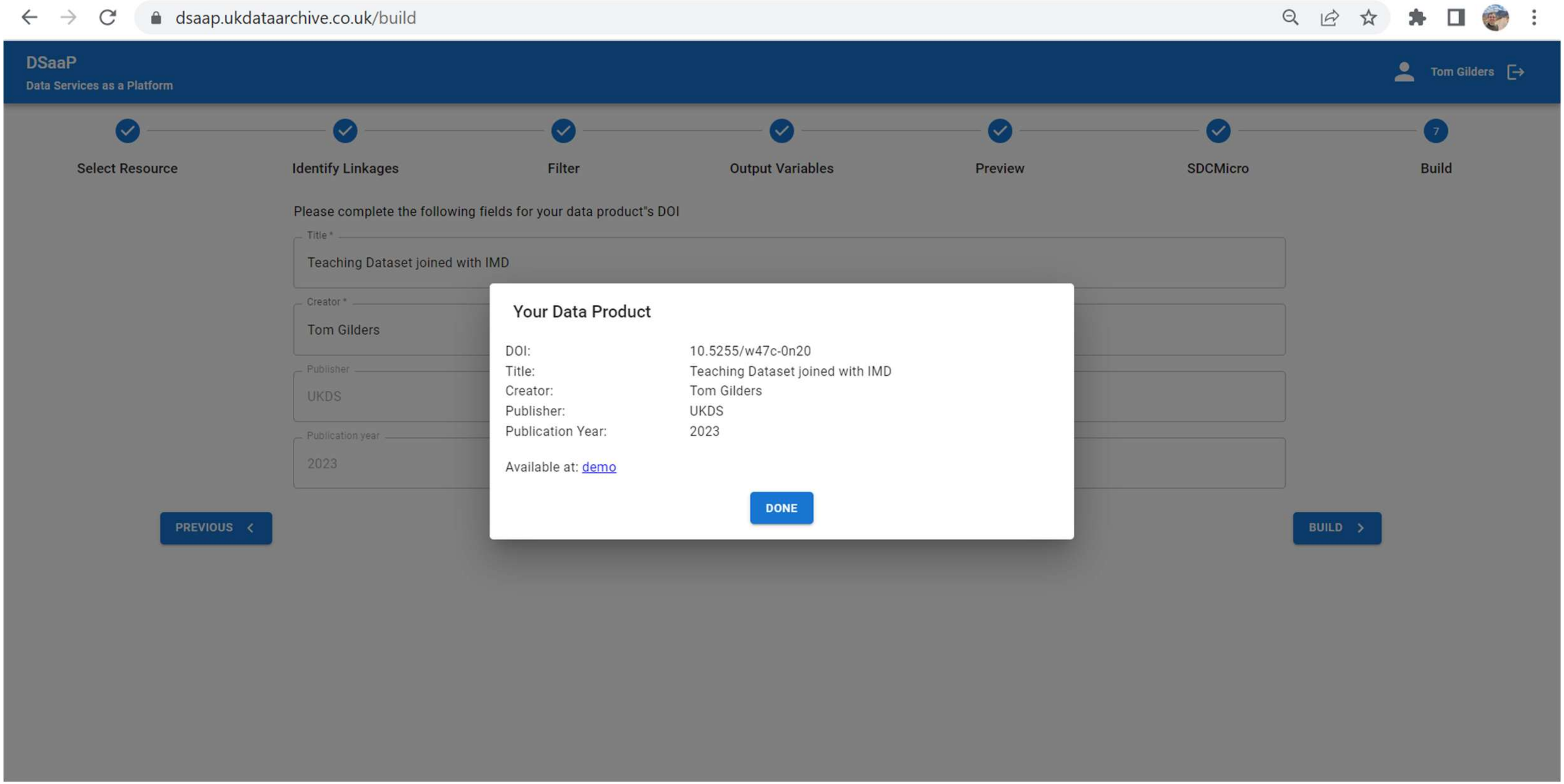
Please complete the following fields for your data product"s DOI

Title *: Teaching Dataset joined with IMD

Creator *: Tom Gilders

Publisher: UKDS

Publication year: 2023



Please complete the following fields for your data product"s DOI

Title *: Teaching Dataset joined with IMD

Creator *: Tom Gilders

Publisher: UKDS

Publication year: 2023

Your Data Product

DOI:	10.5255/w47c-0n20
Title:	Teaching Dataset joined with IMD
Creator:	Tom Gilders
Publisher:	UKDS
Publication Year:	2023

Available at: demo