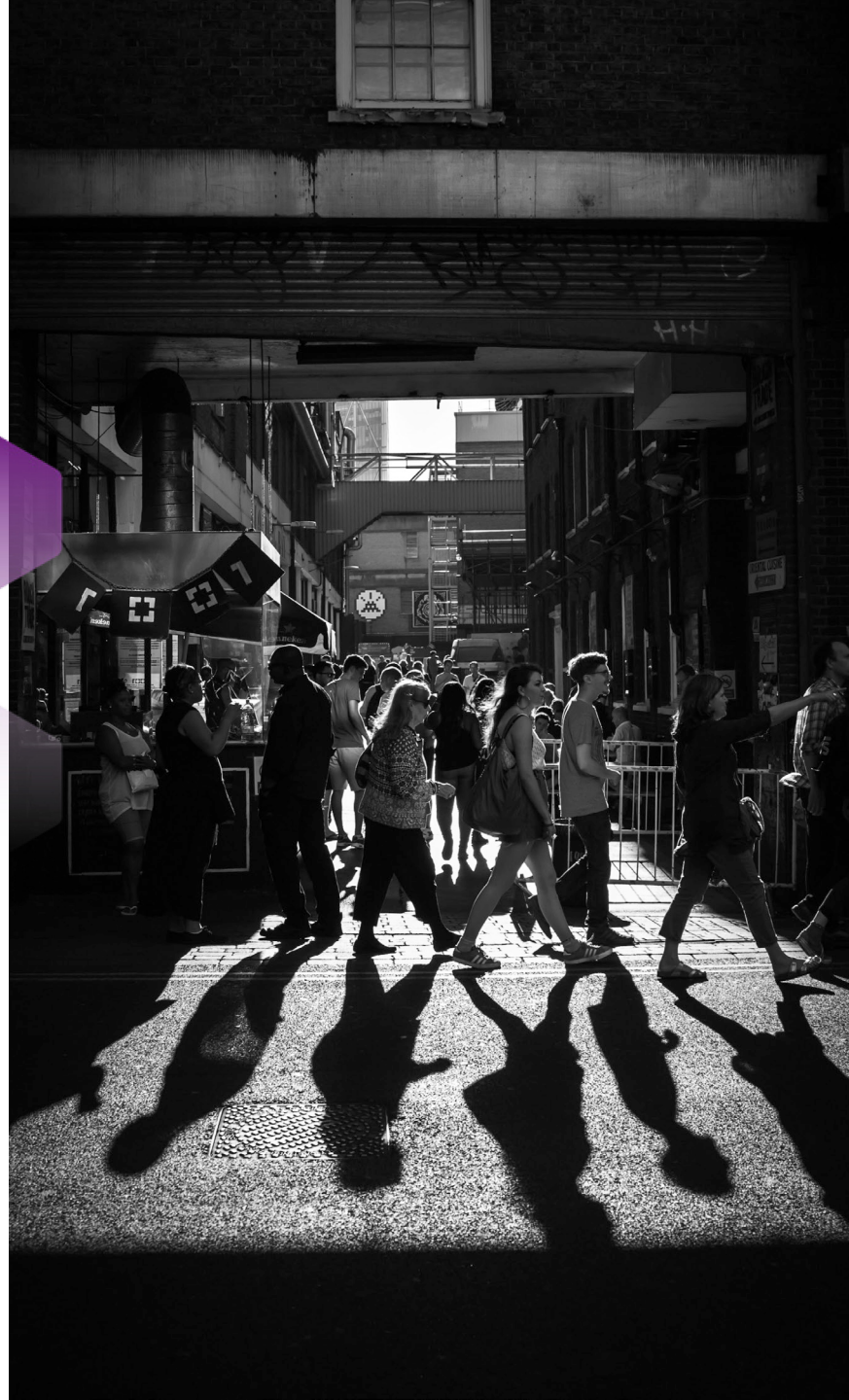
UK Data Service

International developments in policy: Tools and workflows for sensitive data management

Steve McEachern, Australian Data Archive (ADA)
Ryan Perry, Australian Data Archive (ADA)
Darren Bell, UK Data Archive (UKDA)
Deirdre Lungley, UK Data Archive (UKDA)

UKRI Economic and Social Research Council

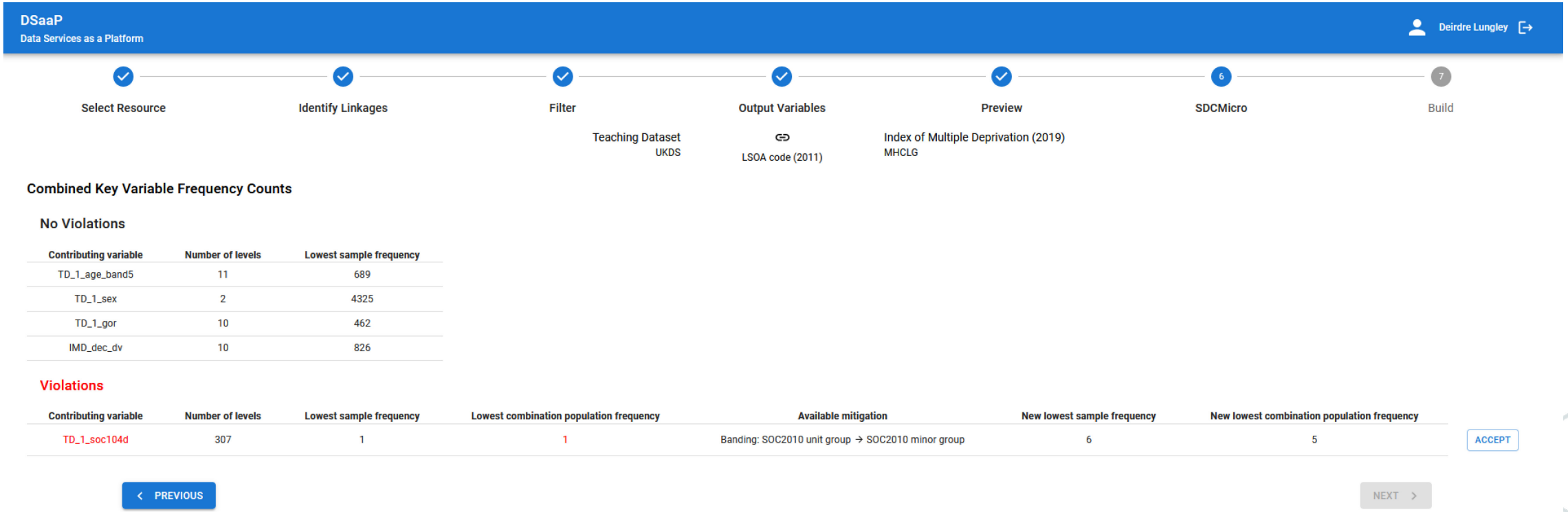


UK Data Service

Machine Learning for Privacy Metadata Annotations

Deirdre Lungley, UK Data Archive (UKDA)
Darren Bell, UK Data Archive (UKDA)

Goal: Automated Disclosure Risk Analysis (DRA)



Automated DRA Dependencies

- Identify key (potentially disclosive) variables.
- Fast computation of combination frequencies of these key variables.
- Stored population frequencies for these key variable combinations.
- Hierarchical conceptual ontology structures.

Improved Computational Performance

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

Our approach computes frequency counts using bitmasks* (strings of 1s and 0s) that are then combined for each row

We achieve a > 30x performance improvement** versus R[1]

Steps

1. Compute bitmasks for each key variable
2. Compute fk[2] for each row in the data frame

Note: -9 is treated as NaN (wildcard) and matches any value

* Bitmask operations are very simple and therefore very fast.
  Main code is written in Python with bitmask manipulation in C++

** Best performance is achieved if only fk is calculated, not Fk (using weights)

References:

1. Templ, Matthias. (2008). Statistical Disclosure Control for Microdata Using the R-Package sdcMicro. Transactions on Data Privacy. 1. 67-85.

2. Franconi, Luisa & Polettini, Silvia. (2004). Individual Risk Estimation in μ-Argus: A Review. 3050. 262-272. 10.1007/978-3-540-25955-8_20.

Improved Computational Performance (cont.)

	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

Friday 2nd June, Session 3, Powerful DDI-CDI Metadata – the How and the Why?, Tom Gilders, Darren Bell

Previous Automated Variable Annotation at the UK Data Archive

June 2012 - March 2013: SKOS-HASSET project (UKDS)

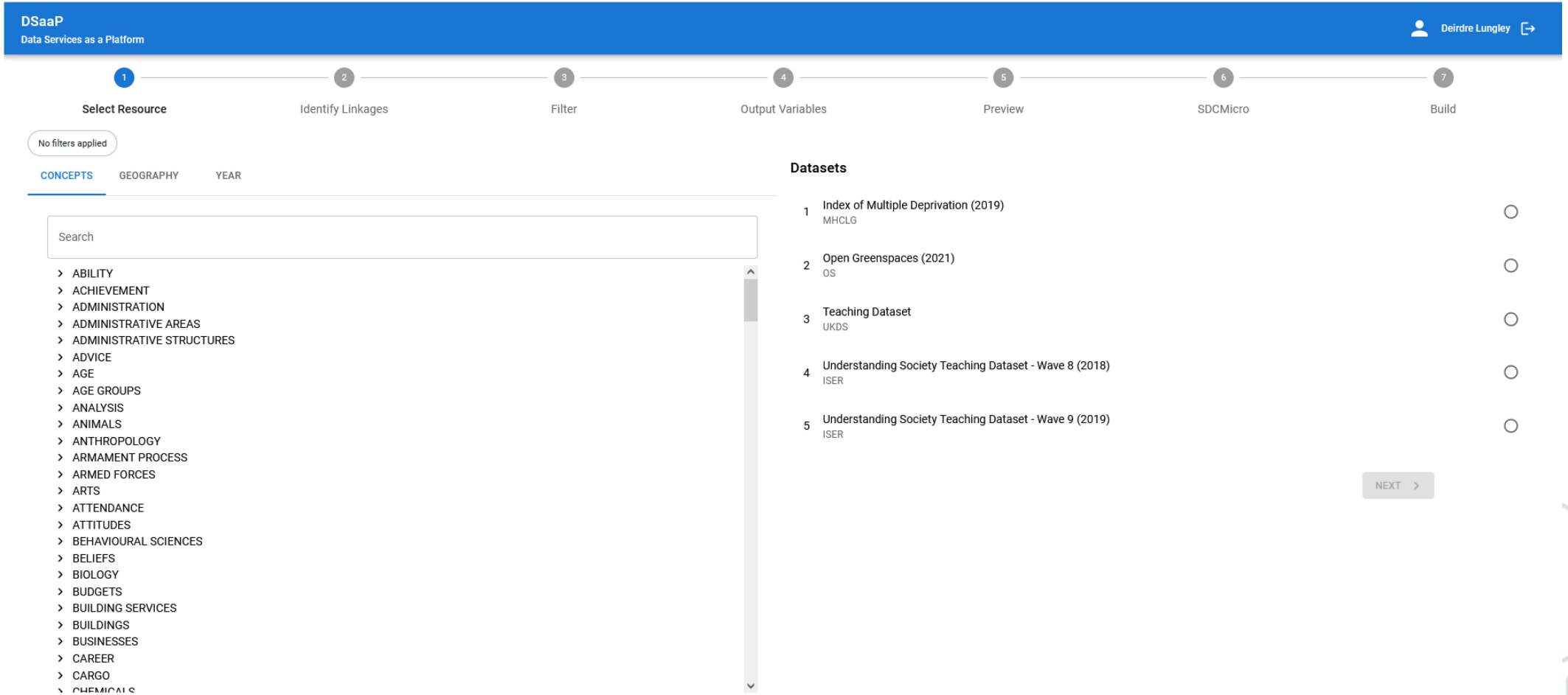
- Incorporated the use of the HASSET thesaurus to automatically index a wide range of the survey data resources of the UKDS.

UK Data Service

ML Variable Annotation objectives:

- Semantic tagging for resource discovery.
- Identify key variables (potentially disclosive) for our Disclosure Risk Analysis (DRA).

Semantic tagging for resource discovery



Identify key variables for DRA

- Key variables – potentially disclosive?
- Indirect identifiers can be any attributes or, more likely, combination of attributes that, are likely to be unique for at least some individuals in your dataset *and* in the population (UK Anonymisation Framework*).
- UKAnon example: A 16 year old widower living in rural Scotland.
- Disclosure Scenarios:
  - Nosy neighbour: an unsophisticated intruder who was trying to find a single specific individual.
  - Fishing attack (not phishing): finding unusual looking records in the dataset and attempting to find the corresponding individuals*.
- Key variable (demographic and socio-economic) public matching sources:
  - E.g. Electoral registers, land registry, estate agent listings.
  - Social media postings.
- Examples of key variables:
  - Address, age, sex, marital status, number of dependent children.

* Elliot, M., Mackey, E., & O'Hara, K. (2020). The Anonymisation Decision Making Framework: European Practitioners' Guide (2nd edition). UK Anonymisation Network. https://msrbcel.files.wordpress.com/2020/11/adf-2nd-edition-1.pdf

Context of our current ML variable annotation project: bounding the problem

- Objective – identify key variables. Initial subset:
  - Geographic (NUTS 3 to LSOA)
  - Sex
  - Age
  - Marital and civil partnership status
  - Household composition
  - SOC (1,2,3 and 4-digit codes)
  - SIC (1,2,3,4 and 5-digit codes)
  - Socio-economic Classification (NS-SeC)
  - Economic activity
  - Ethnic group
  - Country of birth
  - Language (main)
  - Religion
  - Highest qualification
  - Long term health problem or disability.
- Scope - Gold Standard Datasets.

Methodology:

- Apply model at variable group level, e.g. occupations:
  - UKDS gold standard datasets currently manually tagged with variable groups.
- Input features:
  - variable name
  - variable label
  - question text
  - variable group and subgroup.
- Multiple methods:
  - FastAI: language model based.
  - SVM: Support vector machines.
  - KNN: K-nearest neighbour.
- Multiple models:
  - Multi-class: Key variable related concepts.
  - Binary: Sensitive/Non-sensitive.
- Initially tune on one series – QLFS.
- Extend to other Gold Standard Datasets.
- Active learning – continually refining model.

Example Training Data: QLFS SOC Variable Group

	A	B	C	D	E
1	text	labels			
2	INDS07M Industry section in main job (1 charac) Standard Occupational Classification Main Job Industry	OCCUPATIONS|SIC2007|SECTION			
3	INDS07S Industry section in sec job (1 charac) Standard Occupational Classification Industry Occupation Second Job	OCCUPATIONS|SIC2007|SECTION			
4	INDD07M Industry division in main job (2 digits) Standard Occupational Classification Main Job Industry	OCCUPATIONS|SIC2007|DIVISION			
5	INDD07S Industry division in sec job (2 digits) Standard Occupational Classification Industry Occupation Second Job	OCCUPATIONS|SIC2007|DIVISION			
6	INDG07M Industry group in main job (3 digits) Standard Occupational Classification Main Job Industry	OCCUPATIONS|SIC2007|GROUP			
7	INDG07S Industry group in second job (3 digits) Standard Occupational Classification Industry Occupation Second Job	OCCUPATIONS|SIC2007|GROUP			
8	SC10MMJ SOC2010 Main Job Major Group Standard Occupational Classification	OCCUPATIONS|SOC2010|MAJORGROUP			
9	SC10SMJ SOC2010 Second Job Major Group Standard Occupational Classification	OCCUPATIONS|SOC2010|MAJORGROUP			
10	SC10MMN SOC2010 Main Job Minor Group Standard Occupational Classification	OCCUPATIONS|SOC2010|MINORGROUP			
11	SC10SMN SOC2010 Second Job Minor Group Standard Occupational Classification	OCCUPATIONS|SOC2010|MINORGROUP			

Current progress

- Currently covered key variable groups:
  - Standard Occupational Classification, e.g., SOC, SIC, NS-SEC, economic activity.
  - Respondent and household characteristics, e.g. sex, age, marital status.
  - Ethnicity, country of birth, language, religion.
- Consistent high accuracy within QLFS series as expected.
- Generalisation methodology:
  - Iteratively folding in additional gold standard datasets:
    - Binary classifier to identify variables for variable group, e.g. occupations.
    - Testing, updating training data, retesting.

Any questions?

UK Data Service

Thank you.

dmlung@essex.ac.uk