Llywodraeth Cymru
Welsh Government

The Sensible
Code Company

aine@sensiblecode.io
duncan@sensiblecode.io

2021 Census Outputs and Dissemination Update

Suzie Dunsmith & Neil Townsend

ONS

We're committed to delivering 2021 Census results earlier, more flexibly and with greater accessibility

- Using innovative methods developed by our Statistical Disclosure Control experts we designed an approach to dissemination which meets these aims

- Last year we worked with Sensible Code Company who built these methods into a "proof of concept" prototype

- We are now developing methods, processes and specifications across several workstreams

- Output content – derived variables, classifications, geography etc
- Analysing table design to inform dissemination development
- Origin-destination outputs
- Microdata samples
- Admin data integration
- Metadata incl Welsh language requirements
- NS accreditation – OSR consultation
- Analysis and data visualisation
- UK data

Any questions or feedback please contact:

About the 2021 Census and/or other areas of ONS:

Respond to the Office for Statistics Regulation's: user consultation

AGENDA

Introductions SensibleCode/Welsh Government

Census 2021 ONS, UK Case Study

Interactive session/ Q&A

Our Challenge

"To learn more about the challenges being faced by professionals who are considering privacy issues on a regular basis; how they address these issues given the desire to open data and the fact that many more sources of data are being made available. What's being considered and the factors influencing these decisions"

The Sensible
Code Company

The Sensible Code Company

We make products that modernise the processing and dissemination of data

The Sensible Code Company

Problems

- disclosure control is a manual process

- surge of new data sources

- increasing capacity is a challenge

- pressure to publish more and sooner

- privacy preserving tech is new & landscape is foggy
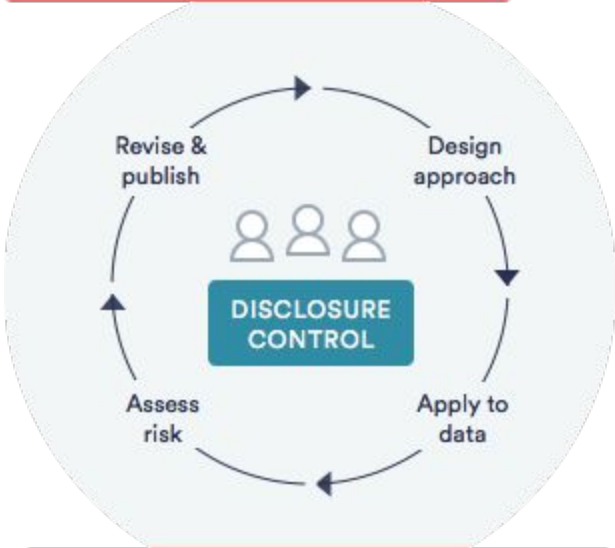
The Sensible
Code Company

- NSIs want to **modernise and automate SDC.**

- Disseminating **data closer to the collection date** increases their value to the economy.

- Users expect to be able to see **more granular data** for more diverse populations.

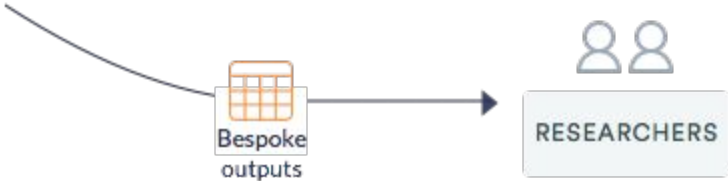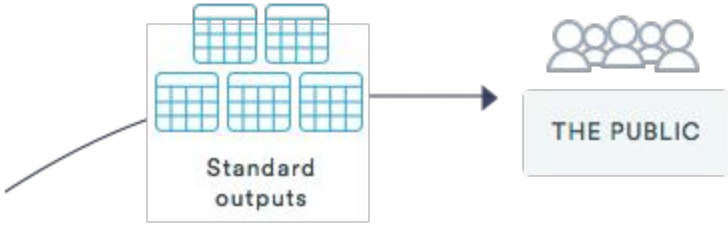- Users want to **query the data more flexibly.**

Flexible dissemination through real-time application of disclosure control techniques in response to user queries

The Sensible
Code Company

Manual creation and checking of outputs is very labour intensive

Data utility often limited due to a small range of standard outputs

Revise & publish

Design approach

DISCLOSURE CONTROL

Assess risk

Apply to data

Standard outputs

THE PUBLIC

Bespoke outputs

RESEARCHERS

Potential timelag of over a year between data collection and publication of outputs

Each release of data increases the chance of data leakage

The Sensible Code Company

Revise & publish

Design approach

DISCLOSURE CONTROL

Assess risk

Apply to data

Confidential dataset

TableBuilder

Detailed queries

Bespoke outputs

THE PUBLIC

RESEARCHERS

The Sensible Code Company
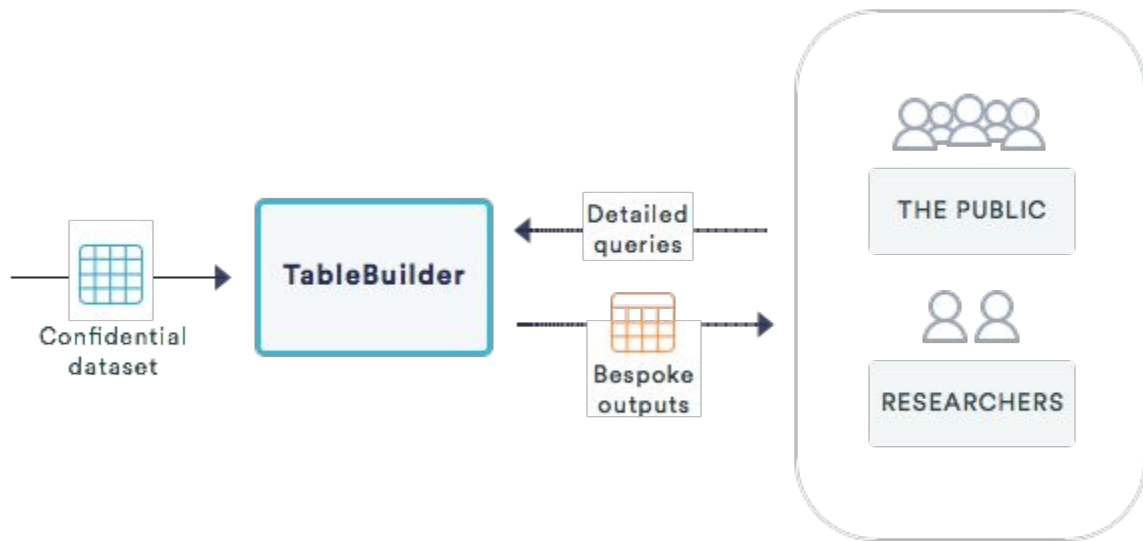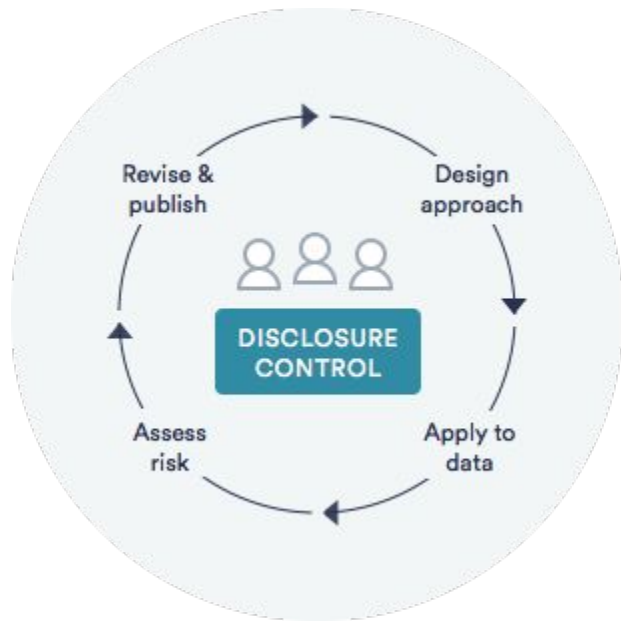
TableBuilder: what does it do?

- Best-in-class aggregation speed using an optimized data format

- Allow users to choose "**any**" output table within limits

  - dubbed "**Flexible Dissemination**"

- In **real-time**:

  - apply perturbative Statistical Disclosure Control (**SDC**)

  - use SDC **rules** post perturbation and redact data if necessary

The Sensible
Code Company

How it works

The Sensible
Code Company
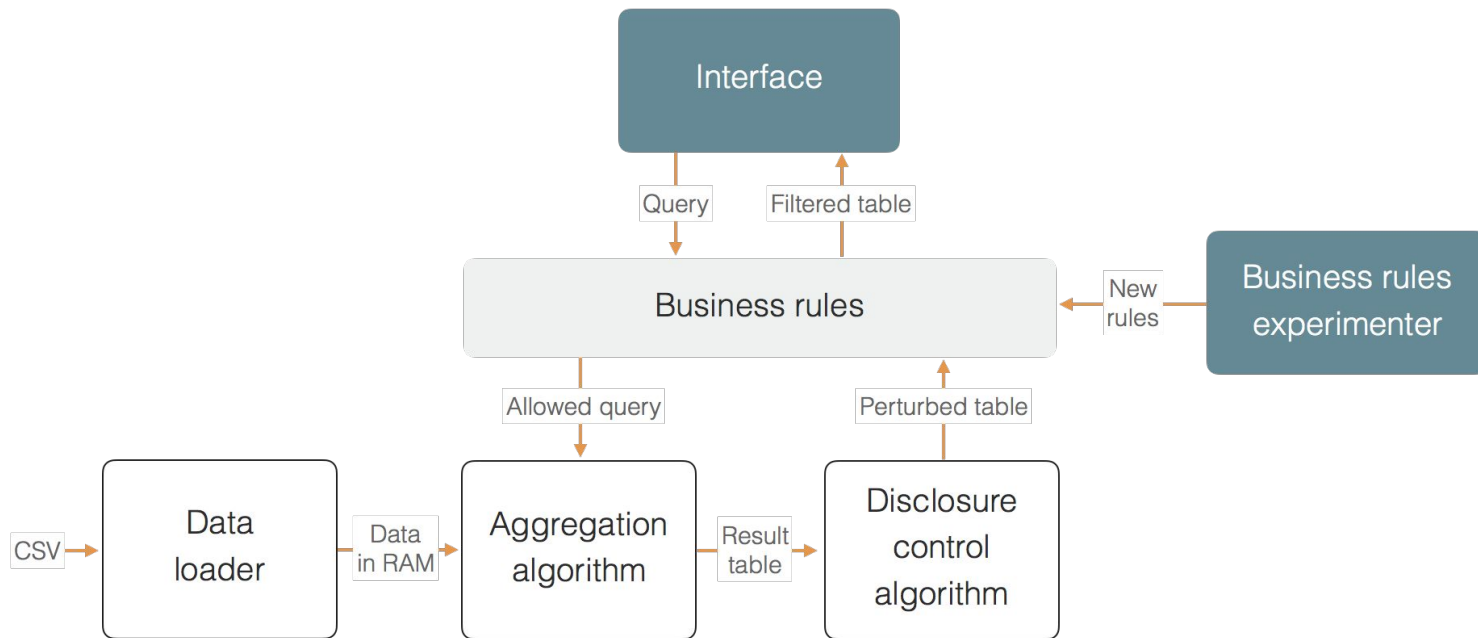
Census Data

Person dataset	
(data grid)	**57 million rows**
28 variables	
41 mappings	

Household dataset	
(data grid)	**22 million rows**
11 variables	
11 mappings	

Join both datasets to associate household variables and mappings with people

Geographical Data

2	**Countries**
10	**Regions**
350	**Local Authorities** (LA)
7,200	**Middle Layer** (MSOA)
35,000	**Lower Layer** (LSOA)
180,000	**Output Areas** (OA) *(average about **300** people)*

The Sensible
Code Company

Statistical Disclosure Control (SDC)

- TableBuilder does **perturbation** using the **cell-key method**

  - some modifications for ONS

  - **consistent** zero perturbation: always query whole data set

- Apply post-perturbation **rules**

  - a **publishable table** must pass all of the rules

SDC: Handling "Structural" Zeros

- **Naive** approach
  - Force zero the "**impossible**" combinations of categories
  - Problem: enumerating all the combinations

- TableBuilder: **automatic** preservation of structural zeros
  - Use zero count at **higher geographic level** as indicator
  - Sensitive to geographic variation

The Sensible Code Company

Edit disclosure rules for People with Households: England and Wales dataset

Effective rules

Apply to People with Households: England and Wales

```
MAX_VARS,3,REGION,7
MAX_CELLS,ROWCOUNT
RULE_1,77
RULE_2,95
RULE_3,90
RULE_4,DISABLED
RULE_5,99
RULE_6,50
```

Edit

Tips

- Enter your rules one per line, with the rule name and parameters separated by commas.

- The order of the rules has no bearing on the order in which they are executed.

- If a rule is not specified in your rules, then the rule will still be applied but with default parameters.

- To disable a rule, set its parameter to 'DISABLED'

The Sensible Code Company

SDC: Which tables can be published?

- Formalise SDC "**rules**"

  - Publishable tables must **pass all** of the rules

- Selective by **geography**

  - more tables are available in areas with **diverse** population

- Data controllers can **experiment** with rule parameters

The Sensible Code Company

Office for National Statistics

CENSUS DATA EXPERIMENTAL TABLEBUILDER

1 Population > 2 Geography > 3 Level > **4 Variables** > 5 Data

Choose your variables

🔍 Search available variables

☑ Age of individual   +

☐ Alternative household composition   +

☐ Approximated social grade

☐ Central heating   +

☐ Country of birth   +

☐ Country of birth of household reference person

☐ Distance travelled to work   +

☐ Economic activity   +

☐ English proficiency

Your table

Cell count: 120,600

Population: People with household variables: England and Wales

Geography: South East

Geographic level: MERGED_LA

Variables: Marital status (2 categories), Sex, General health, Age of individual

Data confidentiality

91% **61 out of 67 areas** pass confidentiality checks.

The Sensible Code Company

Demonstration

Q & A - Thank you

The Sensible
Code Company

aine@sensiblecode.io