Flexible dissemination system for Census 2021
Introduction

• We are proposing to develop an online flexible dissemination system where users can specify the data they need and define their own queries to build tables.

• This presentation shows you some features of a prototype built to assess the feasibility of a flexible dissemination system.

• First, we’ll explain the background to the flexible dissemination system.
The benefits of a flexible dissemination system

• The system would allow census data to be:

  ➢ flexible – users will be able to define their own queries with opportunities for greater detail in the tables

  ➢ timely – outputs will be available sooner than they were for the 2011 Census as they will be checked automatically online rather than manually by ONS

  ➢ accessible – the system will be hosted on the ONS website, meaning the majority of census data will be available from one location
Disclaimer

- In this presentation, images are for illustrative purposes only. They don’t reflect the final design of the dissemination system user interfaces or output table layout that might appear on the ONS website. Please read this disclaimer:

  - These slides have been produced to help you understand our proposals as outlined in the ‘Initial view on 2021 Census output content design consultation document.

  - The demonstration tool uses artificial test data.

  - The output categories in tables are examples based on 2011 Census data. We’ll be consulting with users closer to the 2021 Census to ensure output categories meet user needs and labels used accurately describe the population being counted.
Protecting confidentiality

- We have a legal obligation to protect the confidentiality of respondents and ensure detailed outputs don’t reveal that information.
- Statistical disclosure control ensures statistical outputs provide as much value and utility to users as possible, whilst maintaining the confidentiality of individuals, households and organisations.

For 2021, we’re proposing to protect the data by:
1. targeted record swapping
2. applying a small degree of perturbation to the data
3. applying a set of automated checks before releasing tables
What is data perturbation?

• Data perturbation means slightly changing estimates where otherwise confidentiality would still be compromised in a table after the record swapping.

• The effect of perturbation is not statistically significant and doesn’t damage the data quality, but the changes will result in a small amount of inconsistency between some related tables.

• We are currently investigating the minimum amount of perturbation required to protect confidentiality whilst ensuring accuracy.
Flexible dissemination system

- The diagram on the following slide illustrates how the future system might look.

- The green boxes at the top give a general picture of how we process the census data and the record swapping is shown in the top right.

- The blue boxes below set out the main processes involved in meeting a user request for data. These include applying the perturbation and the automated checks that apply the business rules to control access.
Overview of flexible dissemination system
Now we’re going to show you some features of the current approach:

1. We’ll show you a simplified version of the prototype we used to generate some tables.

2. We’ll explain how users might get more data for some geographic areas than for others.

3. We’ll show how the totals might differ due to perturbation.
The impact of automated checks

• We’ll use the demonstration version of our prototype to show you how the automated checks affect which data are provided.

• On the next slide is the starting page for a query.

• There are four selections to make:
  ➢ population (later versions would have households and communal establishments)
  ➢ geography
  ➢ variables (sometimes called topics)
  ➢ view or download the data

  In this demo, the population is selected by default.
Step 1: Choose your population

Choose your population

- People: Illustrative Demonstration

Your table

- Cell count: 0
- Population: People
- Geography: None selected
- Variables: None selected
Step 2: Choose your geography (Region)

- Our tool shows two geographies: London and the South West.
- We want to show data for an area with a diverse population so will choose London.
Step 3: Choose a lower level of geography (local authority)

• We’ll select the test local authority in London.
Step 4: Choose lower level of geography (Output Area)

- We’ll now select the Output Areas in our test local authority.
Step 5: Select variables

- We’ve selected the 19 category ethnic group.
- In this prototype, a description appears as the cursor hovers over the option.
Step 6: Select more variables

Number of areas for which data will be released
Selecting variables

• As we choose the variables, you’ll see a reading showing the number of areas for which data will be released for that selection.
• If we select too many variables, we won’t get any data for that area.
• We select the religion variable and see we get data for all areas. Users in urban areas are likely to find they can request more detail than was available in ONS tables from the 2011 Census, for most areas.
Step 7: Download data

- Lastly, we can choose whether to download the data or view on screen.
Now we’ll create another query, but for a different part of the country.

We’d like to show the effect of selecting an area with a more homogeneous population, so we choose the South West region.
Step 1: Choose your population

Choose your population

- People: Illustrative Demonstration

Your table

- Cell count: 0
- Population: People
- Geography: None selected
- Variables: None selected
Step 2: Choose your geography (Region)
Step 3: Choose your geography (local authority)

- We select our test rural local authority in the South West.
Step 4: Choose lower level of geography (Output Area)

- We then select all the test Output Areas in that local authority.
Step 5: Select variables

- As we choose the variables, you’ll see a reading showing the number of areas for which data will be released for that selection.
- We won’t get data for all the Output Areas.
Impact of automated checks

• These queries showed we’re getting data for fewer areas in the South West than in London. This is illustrated in the red and green circle.

• The automated checks in the flexible dissemination system are making data for some areas unavailable. This is to protect information about some people in those areas.

• Some queries may result in no data at all. There will be guidance on the ONS website about the protection and checks on the data. This will explain that census data for the less diverse areas of the country will be limited in detail to protect confidentiality.
Step 6: Download data

- We’ll download the table anyway, despite not getting all the data for that area.
We’ve just looked at an important feature of the flexible dissemination system. This is to make more data available than in previous censuses, but this won’t be possible for some areas.

Another feature of the flexible dissemination system is that the minor perturbation of data to protect confidentiality causes minor differences between tables.
Output table totals

• The next slides demonstrate the differences between output table totals.

• Repeating the same query for the same geographic area will always result in the same estimates.

• However, there will be some inconsistencies in the totals between different tables due to the perturbation applied by the cell-key method. This can occur by:
  1) geography
  2) variables
Comparing summed Output Area data with local authority tables

- The two tables on the next slide illustrate the small differences in table totals when data is aggregated by geography.

- The sum of estimates for small areas that constitute a larger area may not exactly match data in an identical table for the larger area, but the differences are likely to be small.
  
  ➢ For example - total for local authority may not equal the sum of output areas within that local authority.
### Geography

- Local authority total vs sum of aggregated output areas.

<table>
<thead>
<tr>
<th>LA=1</th>
<th>Occupation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnicity:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>3885</td>
<td>6014</td>
<td>4325</td>
<td>5726</td>
<td>5202</td>
<td>4172</td>
<td>4394</td>
<td>4282</td>
<td>6784</td>
<td>44784</td>
</tr>
<tr>
<td>Mixed</td>
<td>27</td>
<td>55</td>
<td>64</td>
<td>44</td>
<td>47</td>
<td>53</td>
<td>70</td>
<td>34</td>
<td>70</td>
<td>464</td>
</tr>
<tr>
<td>Asian</td>
<td>198</td>
<td>357</td>
<td>134</td>
<td>168</td>
<td>171</td>
<td>135</td>
<td>292</td>
<td>233</td>
<td>294</td>
<td>1982</td>
</tr>
<tr>
<td>Black</td>
<td>23</td>
<td>93</td>
<td>42</td>
<td>49</td>
<td>34</td>
<td>76</td>
<td>44</td>
<td>38</td>
<td>97</td>
<td>496</td>
</tr>
<tr>
<td>Chinese</td>
<td>23</td>
<td>40</td>
<td>11</td>
<td>6</td>
<td>25</td>
<td>15</td>
<td>23</td>
<td>15</td>
<td>33</td>
<td>191</td>
</tr>
<tr>
<td>subtotal</td>
<td>4156</td>
<td>6559</td>
<td>4576</td>
<td>5993</td>
<td>5479</td>
<td>4451</td>
<td>4823</td>
<td>4602</td>
<td>7278</td>
<td>47917</td>
</tr>
</tbody>
</table>

For White ethnic group category:
- total for row at local authority level is 44,784
- total for row in Output Area aggregation is 44,779
- Difference of 5 (0.015%)
Variables

• The next two tables illustrate the small differences in table totals that may occur when a common variable is cross-tabulated with another variable.

• In multivariate tables, which share a common variable, there may be small differences in the row totals when cross-tabulated with other variables, but the differences are likely to be small.
  ➢ For example - age by ethnic group vs age by marital status.
## Variables

### Age by marital status at local authority level

<table>
<thead>
<tr>
<th>LA=1</th>
<th>N.A.</th>
<th>Single</th>
<th>Married</th>
<th>Separated</th>
<th>Divorced</th>
<th>Widowed</th>
<th>Total:</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-15</td>
<td>89</td>
<td>12176</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12176</td>
</tr>
<tr>
<td>16-24</td>
<td>568</td>
<td>8313</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8552</td>
</tr>
<tr>
<td>25-34</td>
<td>23</td>
<td>5249</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8149</td>
</tr>
<tr>
<td>35-45</td>
<td>3</td>
<td>7810</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8716</td>
</tr>
<tr>
<td>45-54</td>
<td>3</td>
<td>8378</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8987</td>
</tr>
<tr>
<td>55-64</td>
<td>0</td>
<td>7518</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7831</td>
</tr>
<tr>
<td>65-74</td>
<td>0</td>
<td>5612</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5772</td>
</tr>
<tr>
<td>75+</td>
<td>0</td>
<td>4856</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4949</td>
</tr>
<tr>
<td>Total:</td>
<td>686</td>
<td>58161</td>
<td>986</td>
<td>4046</td>
<td>848</td>
<td>403</td>
<td>65130</td>
</tr>
</tbody>
</table>

### Age by ethnic group at local authority level

<table>
<thead>
<tr>
<th>LA=1</th>
<th>N.A.</th>
<th>White</th>
<th>Mixed</th>
<th>Asian</th>
<th>Black</th>
<th>Chinese</th>
<th>Total:</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-15</td>
<td>89</td>
<td>10146</td>
<td>420</td>
<td>1179</td>
<td>233</td>
<td>107</td>
<td>12174</td>
</tr>
<tr>
<td>16-24</td>
<td>568</td>
<td>6946</td>
<td>186</td>
<td>679</td>
<td>106</td>
<td>67</td>
<td>8552</td>
</tr>
<tr>
<td>25-34</td>
<td>23</td>
<td>6895</td>
<td>145</td>
<td>822</td>
<td>158</td>
<td>106</td>
<td>8149</td>
</tr>
<tr>
<td>35-45</td>
<td>3</td>
<td>7810</td>
<td>87</td>
<td>581</td>
<td>178</td>
<td>57</td>
<td>8716</td>
</tr>
<tr>
<td>45-54</td>
<td>3</td>
<td>8378</td>
<td>83</td>
<td>393</td>
<td>96</td>
<td>34</td>
<td>8987</td>
</tr>
<tr>
<td>55-64</td>
<td>0</td>
<td>7518</td>
<td>37</td>
<td>227</td>
<td>29</td>
<td>20</td>
<td>7831</td>
</tr>
<tr>
<td>65-74</td>
<td>0</td>
<td>5612</td>
<td>15</td>
<td>111</td>
<td>26</td>
<td>8</td>
<td>5772</td>
</tr>
<tr>
<td>75+</td>
<td>0</td>
<td>4856</td>
<td>13</td>
<td>54</td>
<td>22</td>
<td>4</td>
<td>4949</td>
</tr>
<tr>
<td>Total:</td>
<td>686</td>
<td>58161</td>
<td>986</td>
<td>4046</td>
<td>848</td>
<td>403</td>
<td>65130</td>
</tr>
</tbody>
</table>

Small difference in totals
Guidance for users

• ONS will publish guidance about the protection and checks on the data.

• Users will be advised that data won’t always match in some tables and be reassured that the effects are minimal.

• They will also see the guidance that mentions that any repeat of the same query always produces the exact same estimates.
Summary

- This presentation was intended to inform uses of the two main features associated with a flexible dissemination system for the 2021 Census.

1) There could be greater detail in tables, but not for all areas.
2) Estimates may differ between similar tables for the same areas but these differences are not statistically significant.

- If you would like to provide feedback, please contact us at census.outputs@ons.gov.uk.