**2021 Census Process flow – introductory paper for external review**

Cal Ghee, Steve Rogers

Census Statistical Design, Census Data Processing

February 2019

**The Panel are asked to note from this paper:**

The assumed first release strategy and assumptions behind the indicative timings

**At the panel meeting we will discuss:**

The Panel’s view from the ‘statistical’ angle of the risks and options

Balanced with

Their view from the ‘end user’ angle on the risks to the timeline, and options

**1 Introduction**

We have an ambition to release the first outputs from 2021 Census within a year of Census Day: before end March 2022. We have committed ourselves to this publicly, but not to what that first release would consist of. It is obviously important that subsequent outputs are consistent with the first release so the data need to be ‘stable’, and we want to leave ourselves with some contingency time, which means we have started with an initial assumption that we need to finish the processing stage by end December 2021. We have also committed to publishing all outputs within two years of Census Day, so any eating in to the contingency time and descoping what the first release is made of puts at risk getting the other products ready.

1. **Basic process flow**

Figure 1 below demonstrates the flow of processes between receiving the return data and completing a final outputs data set. The early processes deal with returns as they are streamed through: electronic returns come through directly, while for those on paper there is a lag as they are captured by the questionnaire management supplier. Once we get to the matching of census to census coverage survey (CCS) records, and to item edit and imputation, data need to progress in geographic batches, so these parallel processes can’t start until the data are all in. Note that we will ‘tune’ processes on early/incomplete cuts of data: for final runs that create the final outputs data set we will use all data.



Figure 1: Data Processing Flow

**3 Assumptions and timeline**

Figure 2 below demonstrates the timeline between receiving the data and getting those first outputs ready, given the following assumptions:

* There are no unexpected delays in the collection activities, including no extension to Census or CCS field work timings
* Electronic collection and the paper questionnaire processing contract run as expected
* The paper questionnaire processing supplier can meet the agreed contract dates regardless of how much paper they receive (estimated between 25 and 55 percent of all returns)
* No unexpected need for changes to any processing method
* No unexpected levels of census/CCS clerical matching
* No unexpected distributions or systematic errors in census responses
* Response levels are sufficient
* Matching can’t start until we receive all census returns submitted up to when the CCS started in the field, and all the CCS responses (allowing for a 4 week period of CCS self-response) – at the latest end July
* E&I can’t start until we have all the census returns, including those received late (on paper), and coding for all variables – end July or end Aug (tbc), based on:
  + late receipt of paper responses cut-off date 9th July based on 2011 patterns of response (electronic receipts earlier)
  + questionnaire management supplier will have captured and coded all paper responses by end July
  + questionnaire management supplier will have finished residual coding of all paper and electronic responses by end July (tbc - currently negotiating with supplier)
* to get the highest quality census/CCS matching possible, we need 60 matchers for 90 days (tbc – we will present alternative estimates)
* capture, cleaning and coding processes are streamed; matching, E&I, QA, coverage adjustment, post-coverage item imputation, assigning geographies, SDC record-swapping and QA are all batched by geography; coverage estimation is run at national level first

**4 Discussion at the Panel meeting March 2019**

In the presentation and discussion with the panel, we will discuss some of the risks to getting the response data, and within the processing stages, as well as what opportunities we have to speed up the processing stages.

**5 Next steps**

The Processing and outputs rehearsal (Feb-April 2020), and other large-scale simulations and tests will aim to test the operation of systems and validate assumptions where possible.

Note: these timings are for the FINAL runs of each process. TUNING runs (testing that interfaces work etc) start earlier



Matching initial estimate: 60 people, 90 days

Batched matching census/ccs

Batched E&I

Estimation

Batched adj’ment

Batched PCII

Batched Geog

Batched SDC

Batched QA

Coverage estimation preferred approach is national level first, so this needs all areas through Census/CCS matching and E&I before it can start

Contingency time

Outputs production

Returns streamed through capture, cleaning, coding