DEPARTMENT OF STATISTICS SINGAPORE

Empowering You with Trusted Data

Data Processing in Singapore's Census of Population

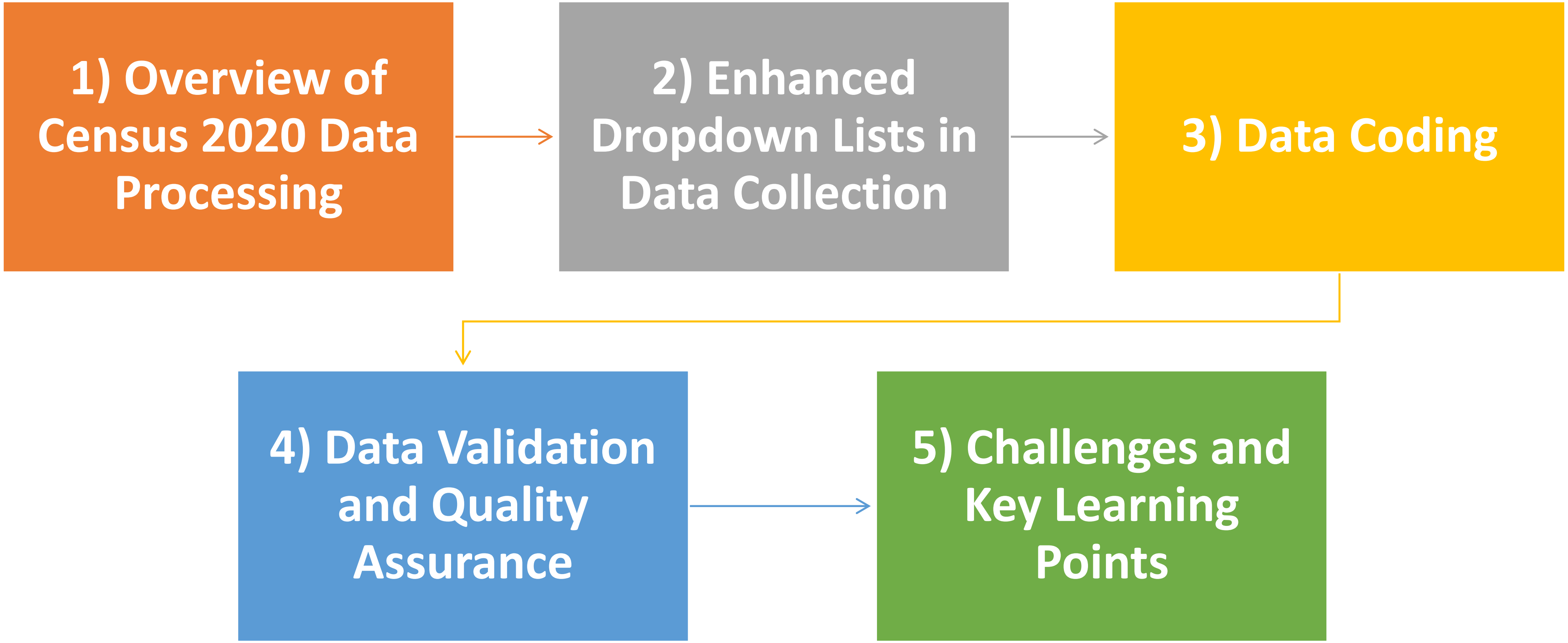
31st Population Census Conference

www.singstat.gov.sg

Agenda

CENSUS 2020

YOUR RESPONSE MATTERS.
TOGETHER, WE SHAPE OUR FUTURE.



1) Overview of Census 2020 Data Processing

2) Enhanced Dropdown Lists in Data Collection

3) Data Coding

4) Data Validation and Quality Assurance

5) Challenges and Key Learning Points

1. Overview of Census 2020 Data Processing

Introduction to Census of Population

Background

- Singapore Census is conducted once in 10 years, in years ending ‘0’
- For Census of Population 2020, the Census Reference Date is as at 30 June 2020, in line with mid-year reference of register-based data

Significance of Census

- Most comprehensive source of information providing a statistical profile of the population and households in Singapore.
- Collects information from the population and households and provides benchmark data for demographic and socio-economic statistics.
- Large sample size and coverage of the Census facilitate analyses on different population groups by fine disaggregation and by geographical area.
- Data from the Census provide key information used for public policy studies, private business decision making and for research and analysis.

Introduction to Census of Population

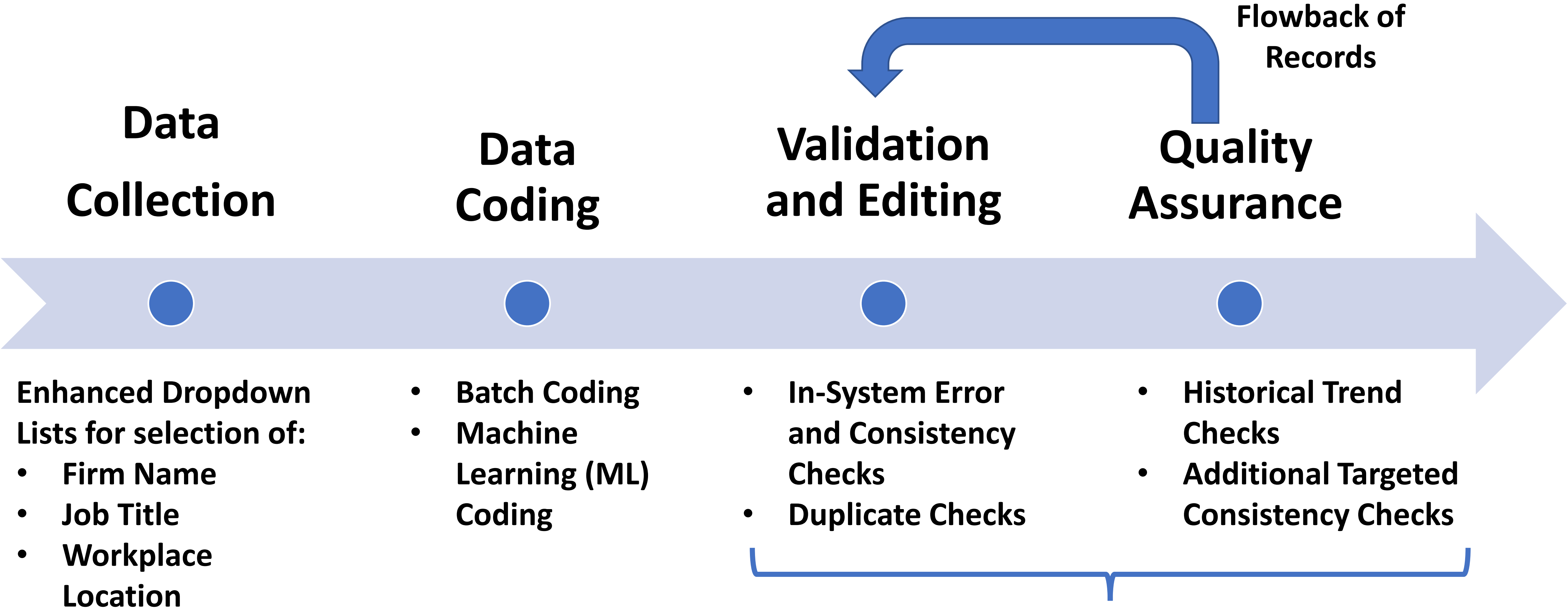
Data Items

- Total of **64** data items
  - Demographic and social characteristics
  - Household and housing characteristics
  - Economic and educational characteristics
  - Transport
  - Difficulty in performing basic activities

Register-based Approach using Admin Data from Different Sources, Supplemented by Survey

- Register-based Census adopted since 2000
  - Administrative records from multiple sources merged to provide basic demographic information such as age, sex and ethnic group for the whole population (full coverage)
  - Large-scale sample survey conducted to capture in-depth information on socio-economic and household characteristics (e.g. language, religion, transport, detailed household living arrangement, disability) not available from administrative sources
- Census 2020 survey covered some 150,000 households

Stages Involving Data Verification/Cleaning



Flowback of Records

Data Collection

Enhanced Dropdown Lists for selection of:

- Firm Name
- Job Title
- Workplace Location

Data Coding

- Batch Coding
- Machine Learning (ML) Coding

Validation and Editing

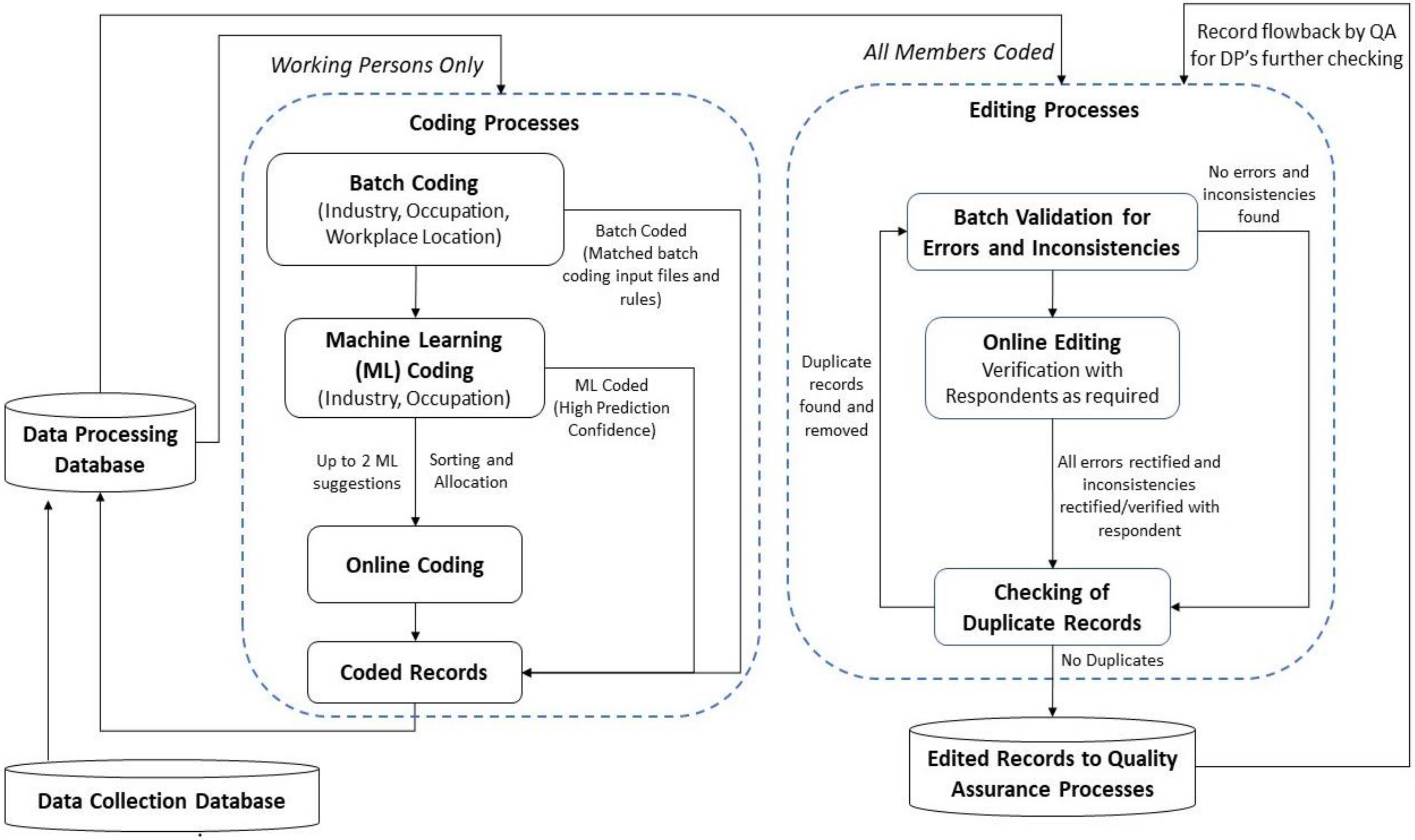
- In-System Error and Consistency Checks
- Duplicate Checks

Quality Assurance

- Historical Trend Checks
- Additional Targeted Consistency Checks

Validation with Independent External Data Sources

Overview of Data Processing in Census 2020



Working Persons Only
Coding Processes
Batch Coding
(Industry, Occupation, Workplace Location)
Batch Coded (Matched batch coding input files and rules)
Machine Learning (ML) Coding
(Industry, Occupation)
ML Coded (High Prediction Confidence)
Up to 2 ML suggestions
Sorting and Allocation
Online Coding
Coded Records
Data Processing Database
Data Collection Database
All Members Coded
Record flowback by QA for DP's further checking
Editing Processes
Batch Validation for Errors and Inconsistencies
No errors and inconsistencies found
Duplicate records found and removed
Online Editing
Verification with Respondents as required
All errors rectified and inconsistencies rectified/verified with respondent
Checking of Duplicate Records
No Duplicates
Edited Records to Quality Assurance Processes

2. Enhanced Dropdown Lists in Data Collection

Data Collection in Census 2020

Tri-Modal Data Collection Strategy

- Adopted since Census 2000:
  1. Online Submission via Self-Enumeration
  2. Phone Interview through Hotline using Computer-Assisted Telephone Interview (CATI)
  3. Face-to-face Interview with field interviewers using Tablets
- Cater to varied profile and needs of population while balancing resources considerations

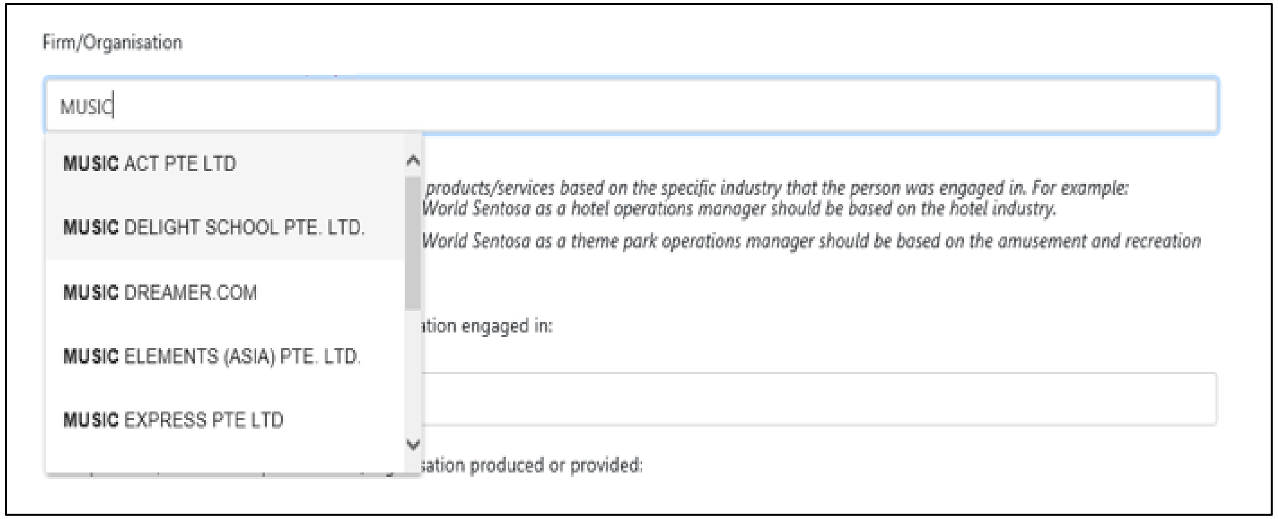
Preloaded Data

- Data available from admin sources not collected again
- Address details and list of persons staying at address presented for verification upon successful authentication where login details matched preloaded administrative data in database

Enhanced Dropdown Lists For Data Collection

Dropdown Lists for Data Collection (Industry and Occupation)

- Prior to Census 2020, company name and job title (for industry/occupation coding) were collected in free text.
- In Census 2020, respondents were able to select their firm name from a company dropdown list. Backend, each firm with exact match was tagged to its Unique Entity Number (UEN) and corresponding industry code(s)
- For firms with only 1 industry code, questions on main activity of firm were skipped
- On top of the names from the register, list was enhanced with addition of modifiers such as commonly known name of firm (in addition to official/formal name) to aid identification/selection



Firm/Organisation

MUSIC

MUSIC ACT PTE LTD

MUSIC DELIGHT SCHOOL PTE. LTD.

MUSIC DREAMER.COM

MUSIC ELEMENTS (ASIA) PTE. LTD.

MUSIC EXPRESS PTE LTD

products/services based on the specific industry that the person was engaged in. For example:
World Sentosa as a hotel operations manager should be based on the hotel industry.
World Sentosa as a theme park operations manager should be based on the amusement and recreation

ation engaged in:

sation produced or provided:

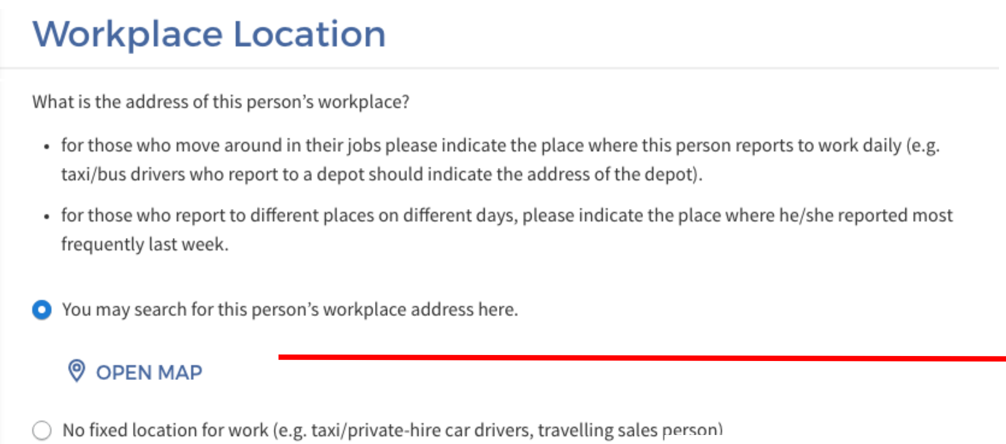
Company Register	Census Dropdown
Hanbaobao Pte. Ltd	**Mcdonalds** – Hanbaobao Pte. Ltd
Ministry of Defence	**MINDEF** – Ministry of Defence

Occupation Title	Census Dropdown
Pre-School Education Manager	Kindergarten Principal
	Childcare Centre Manager

- Respondents could also select their job title from a dropdown based on the occupational classification index (enhanced with common job titles and abbreviations)

Enhanced Dropdown Lists For Data Collection

Leveraging On Map Application For Reporting of Workplace Location



Workplace Location

What is the address of this person's workplace?

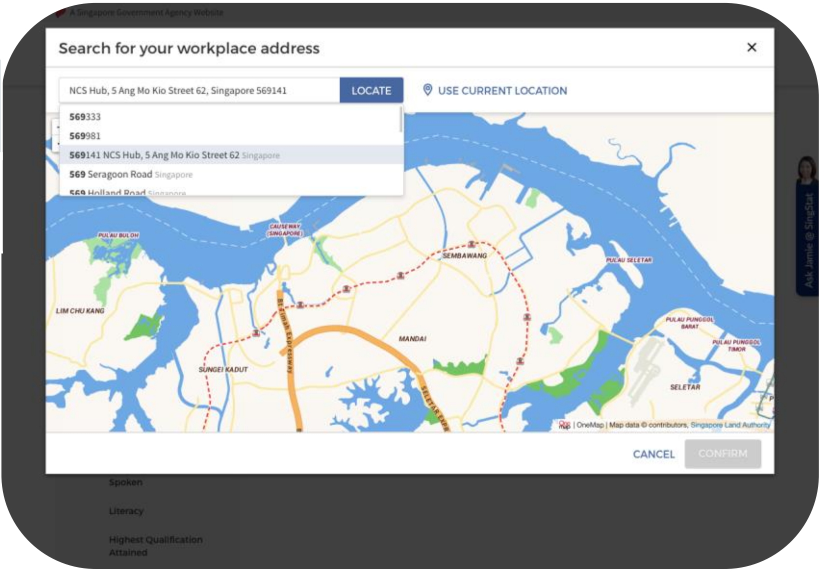
- for those who move around in their jobs please indicate the place where this person reports to work daily (e.g. taxi/bus drivers who report to a depot should indicate the address of the depot).
- for those who report to different places on different days, please indicate the place where he/she reported most frequently last week.

● You may search for this person's workplace address here.

OPEN MAP

○ No fixed location for work (e.g. taxi/private-hire car drivers, travelling sales person)

○ Work from home



A Singapore Government Agency Website

Search for your workplace address ×

NCS Hub, 5 Ang Mo Kio Street 62, Singapore 569141 | LOCATE | USE CURRENT LOCATION

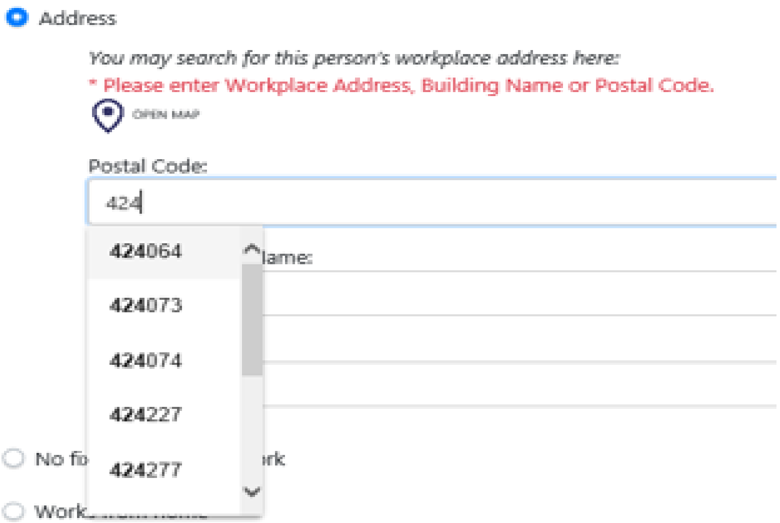
- 569333
- 569981
- 569141 NCS Hub, 5 Ang Mo Kio Street 62 Singapore
- 569 Seragoon Road Singapore
- 569 Holland Road Singapore

CANCEL | CONFIRM

Spoken

Literacy

Highest Qualification Attained



● Address

You may search for this person's workplace address here:

* Please enter Workplace Address, Building Name or Postal Code.

OPEN MAP

Postal Code:

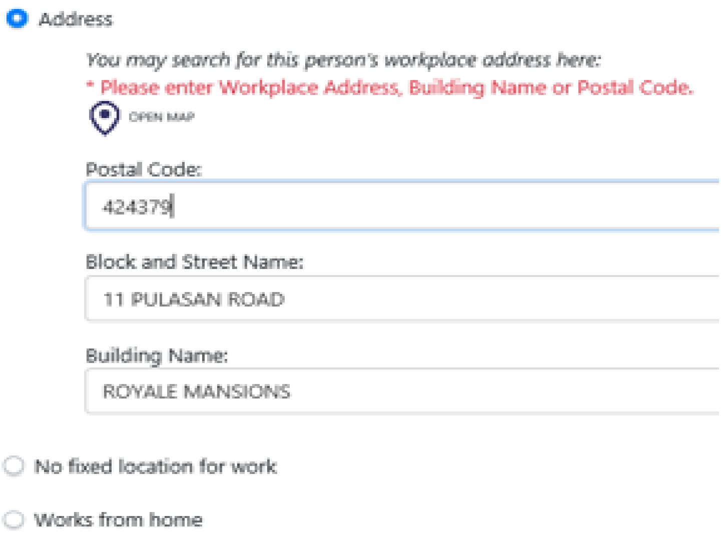
424

- 424064
- 424073
- 424074
- 424227
- 424277

lame:

○ No fi rk

○ Work



● Address

You may search for this person's workplace address here:

* Please enter Workplace Address, Building Name or Postal Code.

OPEN MAP

Postal Code:

424379

Block and Street Name:

11 PULASAN ROAD

Building Name:

ROYALE MANSIONS

○ No fixed location for work

○ Works from home

Alternative: Enhanced Dropdown List for selection using the Address Register

Enhanced Dropdown Lists For Data Collection

Dropdown Lists for Data Collection

- Use of Dropdown Lists helped to:
  - **Reduce respondent burden** as less inputs were required
    - Skipping of questions on main activity for single activity firms
    - Auto-completion of other address fields when unique address match found
  - **Reduce effort in data processing**
    - **More batch coding**
      - Industry for firms with only 1 industry
      - Workplace address based on “auto-completed” match found
      - Standardised respondent inputs (e.g. less typos) – improves keyword matching
    - **Facilitate online coding**
      - For firms with multiple industry codes, only the industry codes found in the business register were displayed

3. Data Coding Processes

Data Coding and Classification

Data Coding

- Data coding as 1st step in data processing
- Assignment of codes based on descriptive text information according to specific sets of classification codes
  - Industry: Singapore Standard Industrial Classification (SSIC) 2020 – *Based on International Standard Industrial Classification (ISIC) Rev 4*
  - Occupation: Singapore Standard Occupational Classification (SSOC) 2020 – *Based on International Standard Classification of Occupations 2008 (ISCO-08)*
  - Workplace Location: 6-Digit Postal Code – *code assigned to every house and building in Singapore, made up of sector code and delivery point. Mapped into planning areas*

- Batch coding (rule based, automated backend) used first
- Records not batch coded passed through Machine Learning Coding for automated coding
- Remaining for online coding, i.e. human intervention

Batch Coding Processes

Batch Coding (Rule Based)

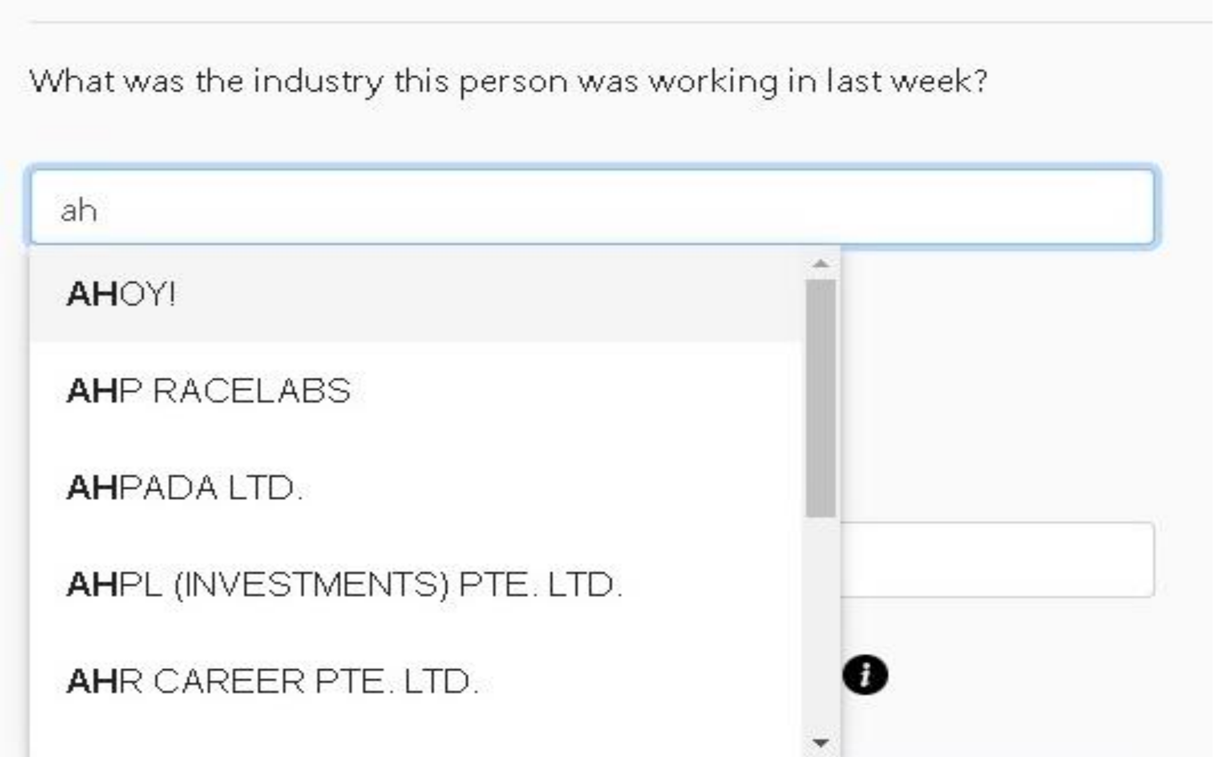
- Broadly 3 types of batch coding rules/processes:
  - Selection in data collection
  - Matching by keywords
  - Matching with administrative records

Data Item	Batch Coding Rate
Industry (SSIC)	~75%
Occupation (SSOC)	~30%
Workplace Location (Postal Code)	~90%

Batch Coding Processes

Selection from Dropdown Lists

- Selection of single activity firm name or occupational title
- “Exception” rules to flag for potential errors: List based on past surveys, refined with C2020 data
- Such records could be re-coded or flowed to online coding
- Highest share among batch coded records



What was the industry this person was working in last week?

ah

- AHOY!
- AHP RACELABS
- AHPADA LTD.
- AHPL (INVESTMENTS) PTE. LTD.
- AHR CAREER PTE. LTD.

DEPARTMENT OF STATISTICS SINGAPORE
Empowering You with Trusted Data

Batch Coding Processes

Matching by Keywords

- Match respondent reported firm name/job title together with main business/main tasks against a set of keywords. Other criterion such as industry/income added for more accurate matching

Collected Data	Coded Data
Firm Name: Ministry of Education Job Title: Teacher Main Task: Teaching Secondary 3 English Workplace Location: ABC Secondary School	Industry: Secondary Schools Occupation: Secondary school teacher

Matching with Admin Records

- ➢ Respondent’s reported firm name (captured in free-text) was compared to admin matched firm name
- ➢ If similar (measured by Jaro-Winkler Distance) and reported workplace location tallied with admin matched firm’s registered address, record was coded
- ➢ Similar approach for workplace location

Machine Learning (ML) and Online Coding

Machine Learning (ML)

- Supervised learning algorithms used to predict a likely SSIC/SSOC code based on free-text responses collected for industry and occupation.

Data Item	Workflow for ML
Occupation	**High** prediction confidence → Record was automatically coded by ML model **Moderate** prediction confidence → up to 2 codes at a broader level of classification offered as suggestions to human coders for further assessment/to facilitate online coding
Industry	Only ML suggestions were displayed for online coding

Online Coding

- For records that cannot be batch/ML coded, manual coding was done.
- Coders assigned codes to records referring to input files containing coding indices, as well as business/ address registers

4. Data Validation and Quality Assurance

Validation and Editing

Data Validation Processes

- After coding for all working members, house record underwent batch validation, comprising:
  1. **Imputation of skipped data items/removal of invalid data** – Due to survey branching rules at data collection
  2. **Checks against a series of “error” and “consistency” rules**

- Records which failed error/consistency rules would flow to **online editing** for data processing editors to correct/verify inconsistencies highlighted respondents contacted for clarifications where necessary

Validation and Editing

Error and Consistency Rules

- Error and consistency rules compiled based on past survey experiences
- **Error:** Missing Data/Invalid Codes or Entries for two or more data items which were **logically impossible**
- **Consistency:** Outlier scenarios which were **unlikely to occur** but **could still be valid**

Duplicate Checks

- Sampling unit in C2020 is an address, and data collection was carried out over a period of approximately 8 months → Possible for individuals/households to be enumerated more than once at multiple addresses
- Towards the end of data processing, checks conducted to retrieve these duplicate records
- Duplicated records studied and removed to prevent double counting.

Data Quality Assurance

Motivation and Importance

- To ensure data quality, independent backend checks were regularly done on edited records to identify issues or outliers not captured by the online error and consistency checks.

- Quality of Online Self-Enumeration responses (> 60% of all responses) were observed to be poorer due to a lack of understanding of questions (without guide from trained interviewers)

Identification of Potentially Erroneous Records

- Comparison of trends against historical data
- Systematic checks & flowback for coding for firms/industries/occupations with high error rate
- Cross-variable consistency checks (Extension of system checks)
- Identification of outliers
- Validation with independent external sources
  - Online Sources (E.g.: Google Maps/OneMap)
  - Alternative administrative data with proxy indicators

Validation with Independent External Sources

Admin Data as First Cut of Checks

- First cut of checks against administrative data done to reduce post-survey clarifications with respondents
  - If administrative records corroborated with reported information, no clarification with respondents would be required
  - Streamline clarification efforts and reduce respondent burden

Data Integration of Admin Data and Online Sources

- For some checks, admin data was combined with other sources of data
- E.g.: Verification of transport time to school
  - Use of admin school enrolment records to identify school address
  - Use of Google Maps/Onemap to cross check reported transport mode(s)/duration to school to identify records with large discrepancies for clarification

Using Admin Data in Data Quality Assurance

Impact of COVID-19 on Employment Items

- In response to COVID-19 situation in Singapore, the Government implemented a nation-wide Circuit Breaker from 7 April to 31 May 2020. Most economic activities were suspended during the period. Even after the Circuit Breaker, Work from Home was largely the default up till end 2020

Verification of Current Activity Status

- Arising from temporary work stoppage/leave, some respondents reported themselves as not working even though they were still employed and receiving wages.
- Checks against admin records to flag for such records were done for verification with the respondents

Verification of Workplace Location

- Many reported to be working from home, likely due to the temporary COVID arrangements, which was not reflective of their usual workplace location
- To streamline post-survey clarifications, respondent's admin company registered address was compared to the home address. If matched, no further clarification required for the usual workplace

Using Admin Data in Data Quality Assurance

Verification of Company Name

- Dropdown selection was long (>500K options) with many similar names, which may be prone to wrong selection.
- Companies of similar names could span across multiple industry codes and differ even at the broad industry level.

- A similar approach as coding of industry was used, where reported company name was compared to the administrative company name and cross checked with reported place of work and occupation.
- Potential erroneous records were flagged out for verification and/or rectification

5. Challenges and Key Learning Points

Challenges and Mitigating Factors

Increasing Resistance from Respondents to Provide Full UIN

- Census 2020 adopted a **deterministic record linkage** approach when using admin data
  - Based on individual identifiers (Unique Identification Number, UIN) that matched the Census Sample Survey and Admin Data Sources where possible
- UIN is important for identification of duplicate records as well as matching of admin data for checks and further statistical compilation

Mitigating Factors

- Preloading of data: Prior to the start of Census, each house record was preloaded with individuals who were registered in the sampled address. UIN asked only for newly added members.
- Further processing for non-preloaded records with incomplete UIN: Deterministic matching was done against admin registers to obtain full UIN where possible before matching with other admin records for checks.

Challenges and Mitigating Factors

Increasing Self-Enumeration

- Online Self-Enumeration was main mode of response (> 60%)
  - While convenient and handy especially in midst of the COVID situation, quality of these responses were observed to be poorer due to a lack of good understanding of questions (without guide from trained interviewers)

Mitigating Factors

- Tool-tips on definitions/scope of data items in online submission screens
- Increased use of admin data to identify records for clarifications and streamline clarification efforts

Challenges and Mitigating Factors

Impact of COVID-19 on Economic/Labour Activity

- COVID-19 changed the ‘norms’ of economic activity
  - Temporary work stoppages – affected data on labour force status
  - Working from home – affected data on workplace location
  - New jobs: Swabbing Personnel, Safe Distancing Ambassadors – affected occupation coding

Mitigating Factors

- Worked closely with Ministry of Manpower (MOM) on coverage & concepts of labour statistics, based on International Labour Organisation (ILO)’s updated guidelines
- Coordinated with MOM for consistent coding of new jobs based on online job descriptions

Key Learning Points

Implementing Quality Controls and Assessment

- Relatively high online submission rate take-up rate allowed for majority of the responses to continue to flow in despite the scaled down operations. However, quality of responses for online self-enumeration is observed to be poorer due to a lack of understanding of questions (without guide from trained interviewers) or missing details (for descriptive fields).

- Important to implement the below to ensure quality of responses:
  - Online help and prompts added to the form design
  - Completeness and validation checks put in upfront
  - More intense consistency checks performed backend, with common errors identified and call-backs to follow up

Key Learning Points

Leveraging Admin Register and Technology

- With basic data on population estimates compiled from population registered, top-line population data were produced on schedule though there was slight impact on release data for the detailed statistical releases (released in Jun 2021).

- Alleviate manpower constraints in data processing via:
  - Use of machine learning in data coding
  - Leverage on administrative data for:
    - Batch data coding
    - Streamlining of data consistency checks

DEPARTMENT OF STATISTICS SINGAPORE

Empowering You with Trusted Data

Thank You

Our Vision

National Statistical Service of Quality, Integrity and Expertise

Our Mission

We Deliver Insightful Statistics and Trusted Statistical Services that Empower Decision Making

www.singstat.gov.sg

Q & A