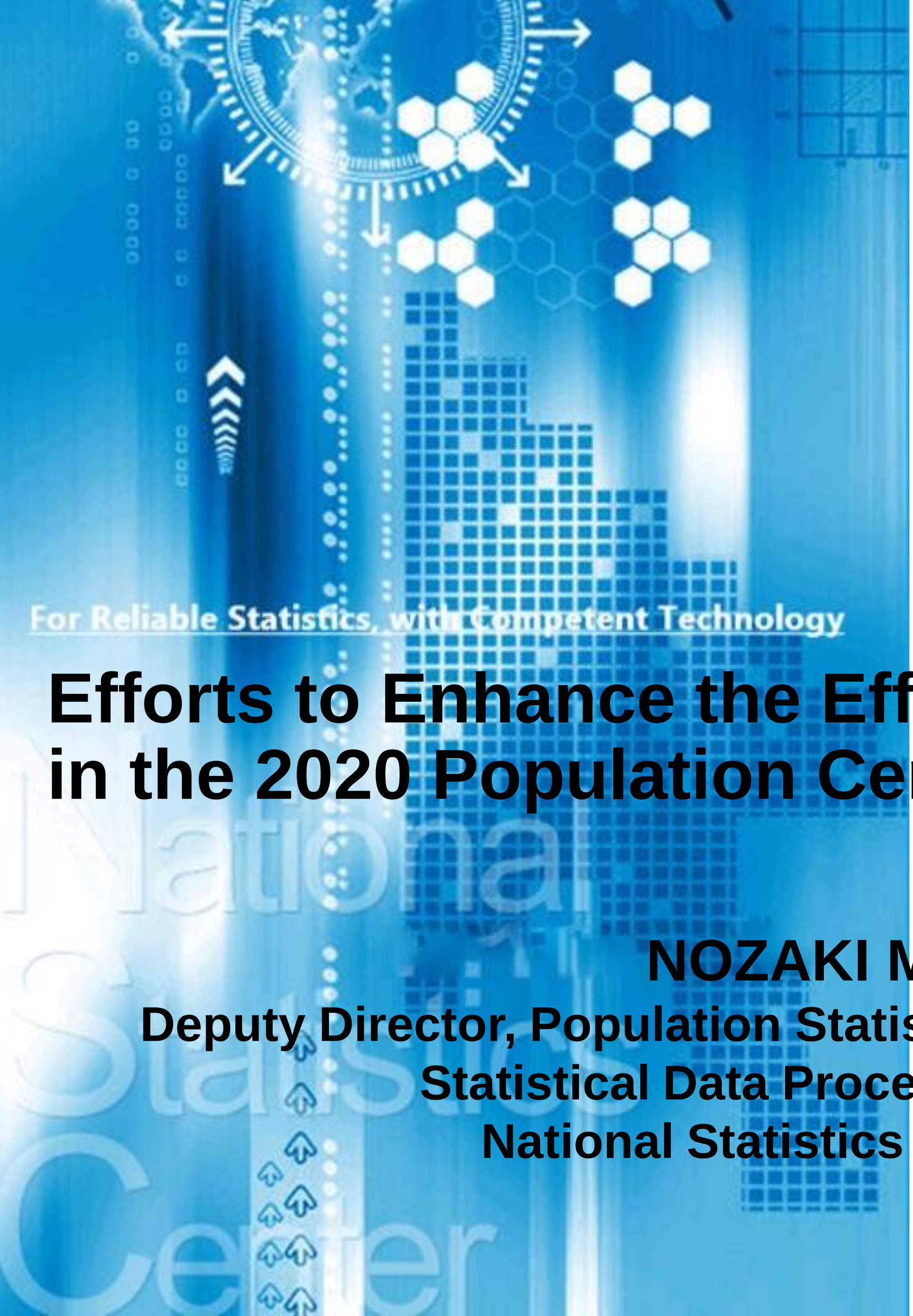
NSTAC National Statistics Center



For Reliable Statistics, with Competent Technology

Efforts to Enhance the Efficiency of Data Processing in the 2020 Population Census in Japan

NOZAKI Masashi

Deputy Director, Population Statistics Data Processing Division,
Statistical Data Processing Department,
National Statistics Center (NSTAC)

Table of Contents

1. Flow of the 2020 Census

2. Flow of Data Processing in the 2020 Census

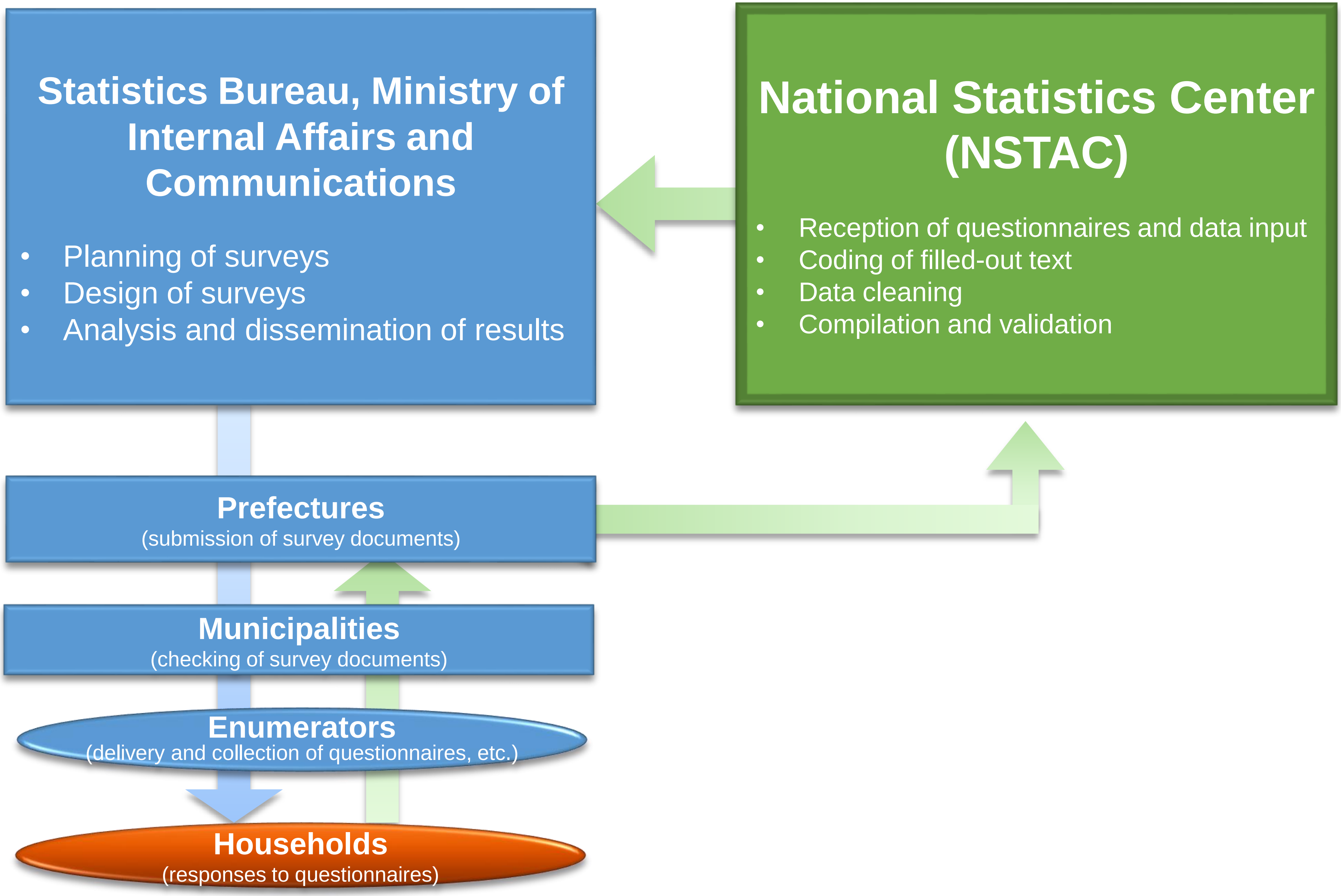
3. Issues with Data Processing in the 2020 Census

4. Measures to Resolve Issues

5. Effects of the Measures to Resolve Issues

6. Future Tasks for the Population Census

1. Flow of the 2020 Census



Statistics Bureau, Ministry of Internal Affairs and Communications

- Planning of surveys
- Design of surveys
- Analysis and dissemination of results

National Statistics Center (NSTAC)

- Reception of questionnaires and data input
- Coding of filled-out text
- Data cleaning
- Compilation and validation

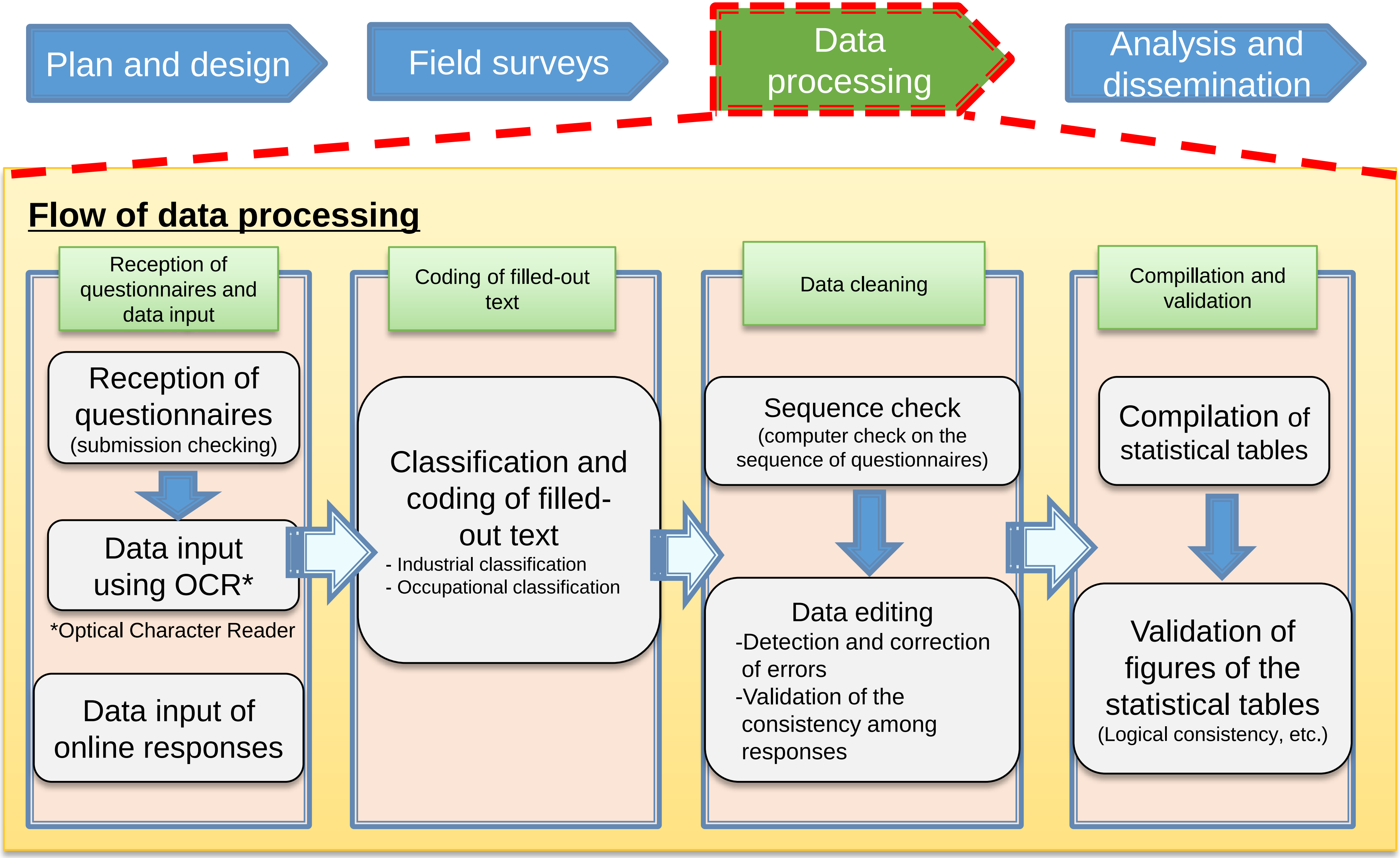
Prefectures
(submission of survey documents)

Municipalities
(checking of survey documents)

Enumerators
(delivery and collection of questionnaires, etc.)

Households
(responses to questionnaires)

2. Flow of Data Processing in the 2020 Census



Plan and design

Field surveys

Data processing

Analysis and dissemination

Flow of data processing

Reception of questionnaires and data input

Reception of questionnaires (submission checking)

Data input using OCR*

*Optical Character Reader

Data input of online responses

Coding of filled-out text

Classification and coding of filled-out text

- Industrial classification
- Occupational classification

Data cleaning

Sequence check (computer check on the sequence of questionnaires)

Data editing

-Detection and correction of errors

-Validation of the consistency among responses

Compillation and validation

Compilation of statistical tables

Validation of figures of the statistical tables (Logical consistency, etc.)

3. Issues with Data Processing in the 2020 Census

Background

・Increase in the number of target households

While the overall population is decreasing, the number of households **is on the rise**.

○ Population and number of households for the 2015 Census and the 2020 Census

	2020	2015	Change from 2015 (number)	Change from 2015 (rate)
Population (persons)	126,146,099	127,094,745	-948,646	-0.7%
Number of households (households)	55,830,154	53,448,685	2,381,469	4.5%

・Challenges to field work

Partly due to the spread of COVID-19, face-to-face contact had become difficult.

⇒ Negative influence on responses to the questionnaires (increase in incomplete entries)

Larger number of errors to be corrected in data editing.

・Reduction of resources

Due to reductions in staffing levels at NSTAC, data processing for the 2020 Census had to be conducted with fewer human resources.

Mission

To enhance efficiency by reviewing the entire process of data editing

4. Measures to Resolve Issues

Tasks for the 2020 Census

(i) Improvement of the auto-coding rate for industrial and occupational classifications by using Artificial Intelligence (AI)

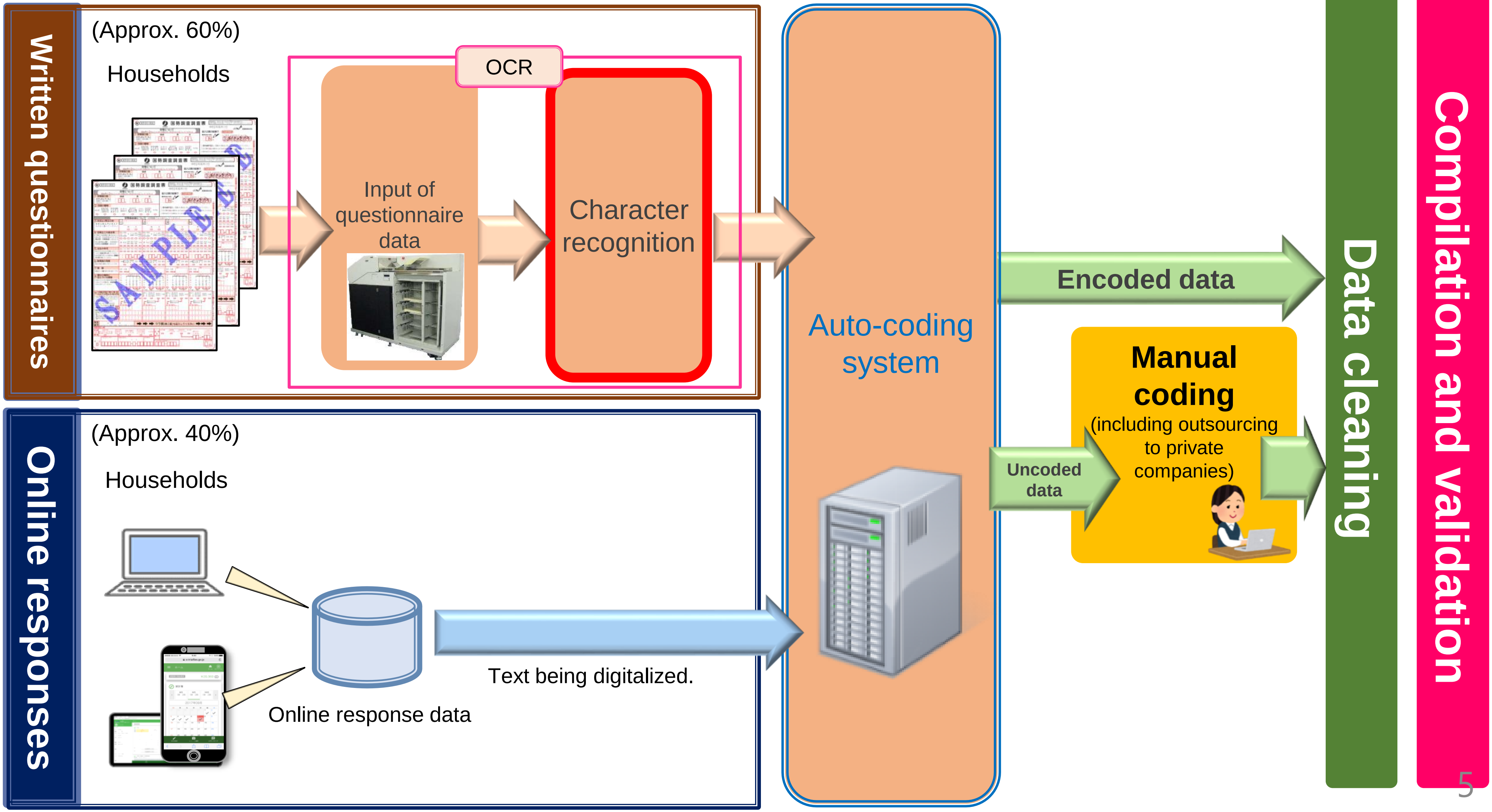
(ii) Expansion of mechanical correction through the review of data editing

(iii) Enhancement of the efficiency of the manual data-editing process

4. Measures to Resolve Issues

(i) Improvement of the auto-coding rates for industrial and occupational classifications by using AI

Flow of coding of industrial and occupational classifications



Written questionnaires

(Approx. 60%)

Households

OCR

Input of questionnaire data

Character recognition

Online responses

(Approx. 40%)

Households

Online response data

Text being digitalized.

Auto-coding system

Encoded data

Uncoded data

Manual coding

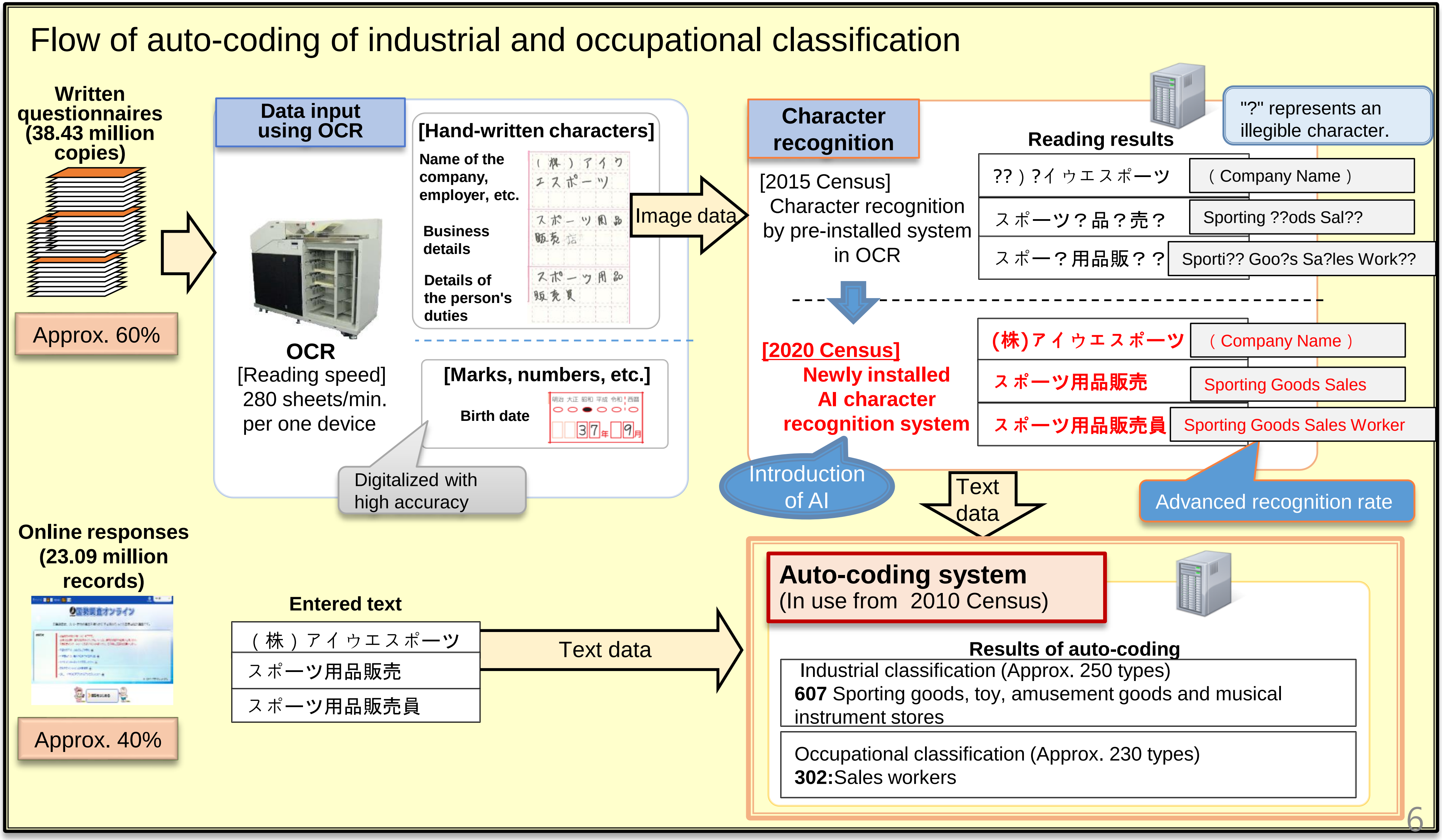
(including outsourcing to private companies)

Data cleaning

Compilation and validation

4. Measures to Resolve Issues

(i) Improvement of the auto-coding rate for industrial and occupational classification by using AI



Flow of auto-coding of industrial and occupational classification

Written questionnaires (38.43 million copies)

Approx. 60%

Data input using OCR

OCR

[Reading speed]
280 sheets/min.
per one device

[Hand-written characters]

Name of the company, employer, etc.

Business details

Details of the person's duties

[Marks, numbers, etc.]

Birth date

Digitalized with high accuracy

Image data

Character recognition

[2015 Census]
Character recognition by pre-installed system in OCR

Reading results

"?" represents an illegible character.

??）?イ ウエスポーツ	（Company Name）
スポーツ？品？売？	Sporting ??ods Sal??
スポー？用品販？？	Sporti?? Goo?s Sa?les Work??

[2020 Census]
Newly installed AI character recognition system

(株)アイウエスポーツ	（Company Name）
スポーツ用品販売	Sporting Goods Sales
スポーツ用品販売員	Sporting Goods Sales Worker

Introduction of AI

Advanced recognition rate

Text data

Online responses (23.09 million records)

Approx. 40%

Entered text

（株）アイウエスポーツ
スポーツ用品販売
スポーツ用品販売員

Text data

Auto-coding system
(In use from 2010 Census)

Results of auto-coding

Industrial classification (Approx. 250 types)
607 Sporting goods, toy, amusement goods and musical instrument stores

Occupational classification (Approx. 230 types)
302: Sales workers

4. Measures to Resolve Issues

(i) Improvement of the auto-coding rate for industrial and occupational classification by using AI

The AI character recognition system improved the accuracy of recognition of hand-written characters.

○ Percentage of legible characters

2015 Census: 76.2% ⇒ 2020 Census: 92.8%

OCR tried to recognize each character independently.

Even if a character cannot be recognized, AI-OCR predicts a sequence of characters as a word.

The auto-coding rates for industrial and occupational classifications were improved

Response method	Classification type	Auto-coding rate	
		2015	2020
Written questionnaires	Industrial classification	25.6%	**71.3%**
	Occupational classification	24.3%	**73.0%**
Online responses	Industrial classification	67.1%	75.2%
	Occupational classification	71.4%	78.5%

Number of manual coding was reduced from 58.6 million to 28.5 million.

4. Measures to Resolve Issues

(ii) Expansion of mechanical correction through the review of data editing

Analyzed the pattern of manual checking and correction for error data in the 2015 Census, thus expanding mechanical correction in the 2020 Census.

In the 2020 Census, while the number of errors increased from 26 million to 46 million, the number of manual data checking and correction decreased from 5.5 million to 4.3 million.

《2015 Census》
Generally, mechanical correction for mechanically detected errors. However, to maintain the accuracy of data cleaning, manual data correction was carried out for 21% of error data.

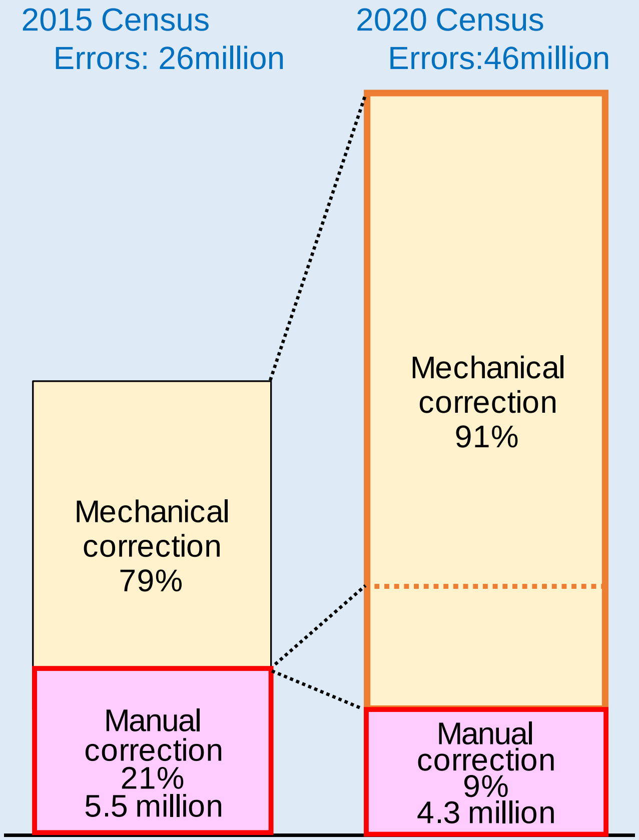
Review of Data Checking Process

Analyzed the reference and correction patterns for manual checking and correction in the 2015 Census, investigated mechanical correction.

Checked the impact of mechanical correction on statistical figures.

Determined the scope and methods of mechanical error correction.

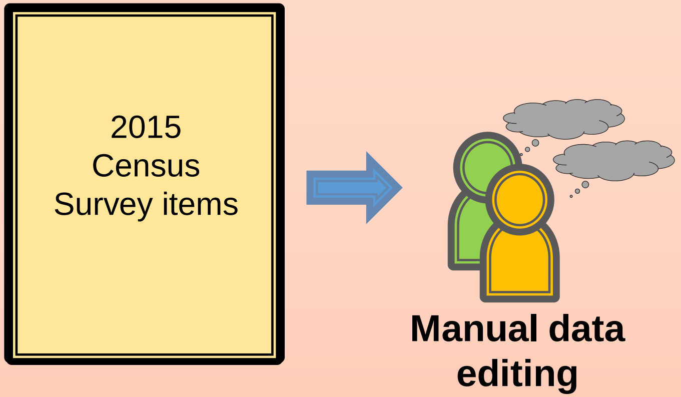
《2020 Census》
Reduced the ratio of manual correction of error data to 9%, while maintaining the accuracy of data cleaning.



4. Measures to Resolve Issues

(iii) Enhancement of the efficiency of the manual data-editing process

2015

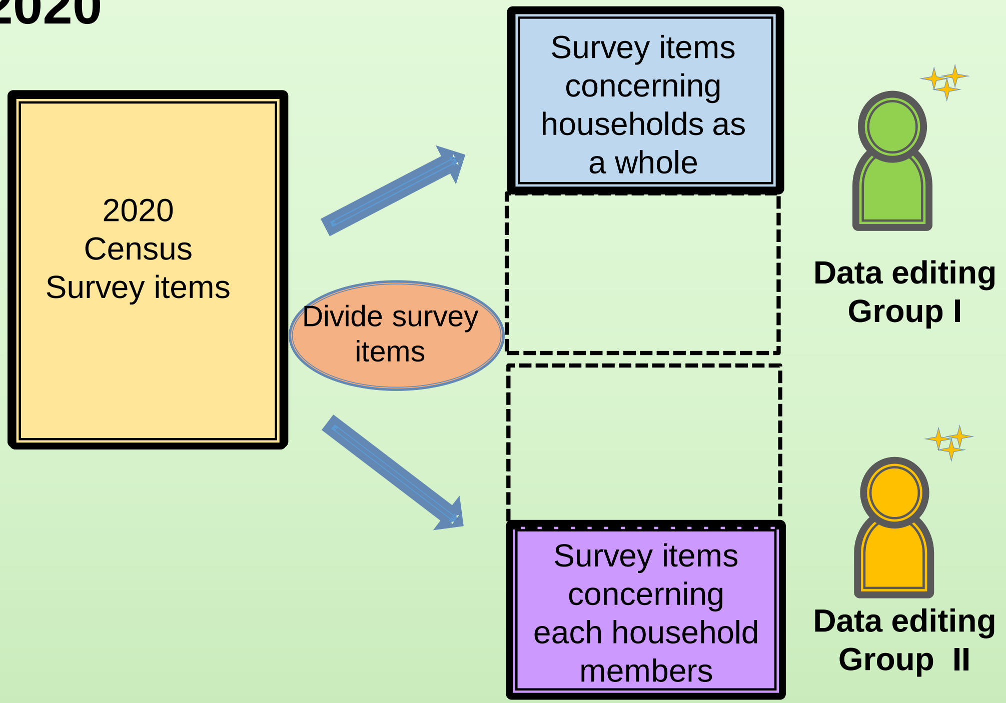

2015
Census
Survey items

Manual data editing

- Errors were categorized into approximately 240 patterns which require corresponding manual checking and correction procedures.
- Data editing staff were requested to understand proper methods covering all types of errors.

Not easy to secure skilled staff and enhance the efficiency of data editing.

2020


2020
Census
Survey items

Divide survey items

Survey items concerning households as a whole

Data editing Group I

Survey items concerning each household members

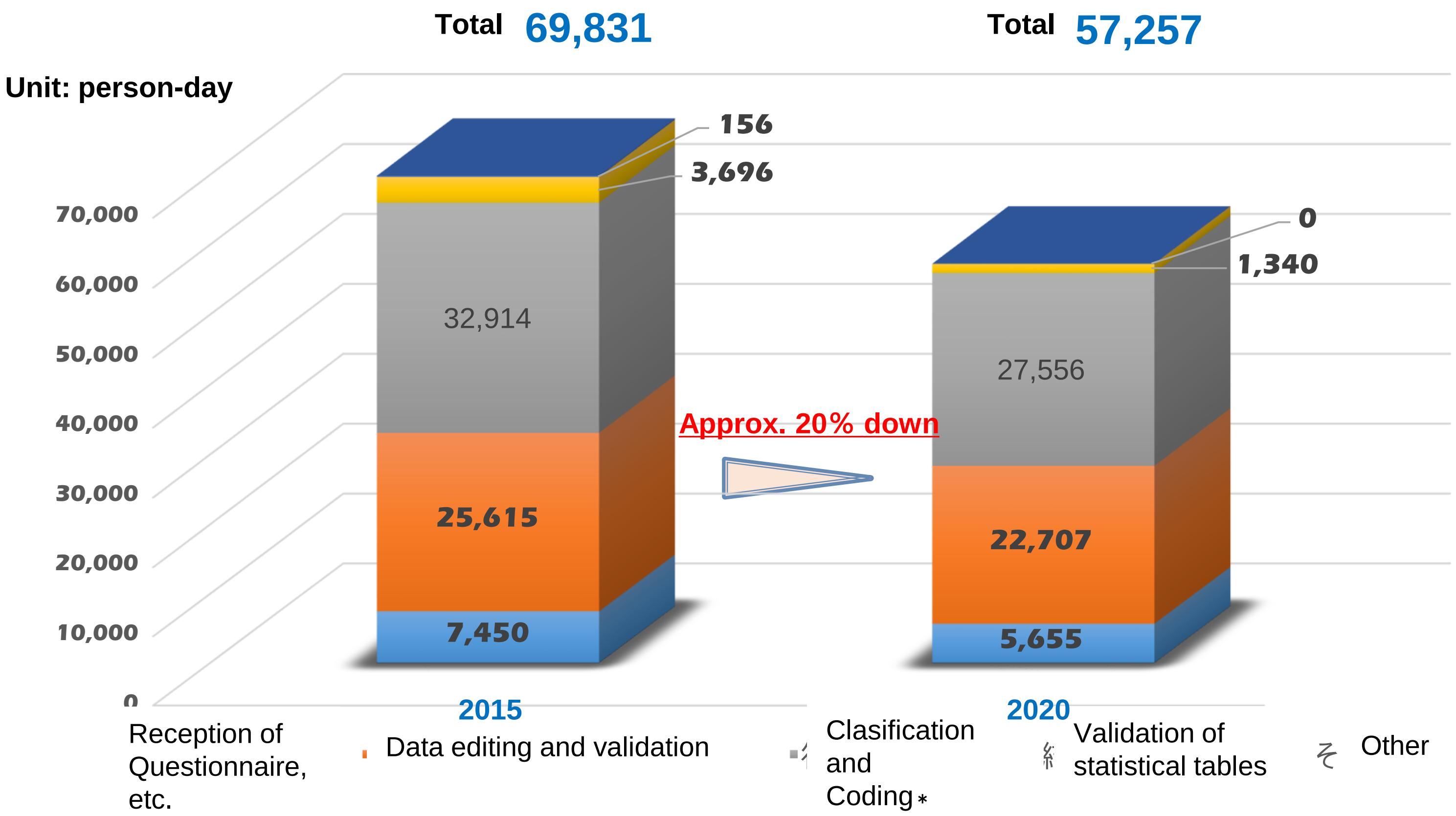
Data editing Group II

- Survey items in the 2020 Census were divided into two categories.
- Skilled staff were allocated separately to “Group I” for items concerning households as a whole and “Group II” for items concerning household members.

With skilled staff who learned specific checking and correction methods, high efficiency in data editing was expected.

5. Effects of the Measures to Resolve Issues

Effects on the number of required data processing staff (person-day basis)



* Excluding coding jobs outsourced to private companies
(NSTAC staff were responsible for supervising and manual coding in difficult cases.)

The number of data processing staff was reduced by approx. 20%. (person-day basis)

6. Future Tasks for the Population Census

NSTAC

○ Study of web-scraping technology to obtain information for industrial and occupational classifications from the websites of companies

○ Study of auto-coding methods using machine learning to further heighten auto-coding rate
(presently : rule-based expert system)

○ Imputation of unknown data with newly developed statistical methodologies

Thank you for your attention.