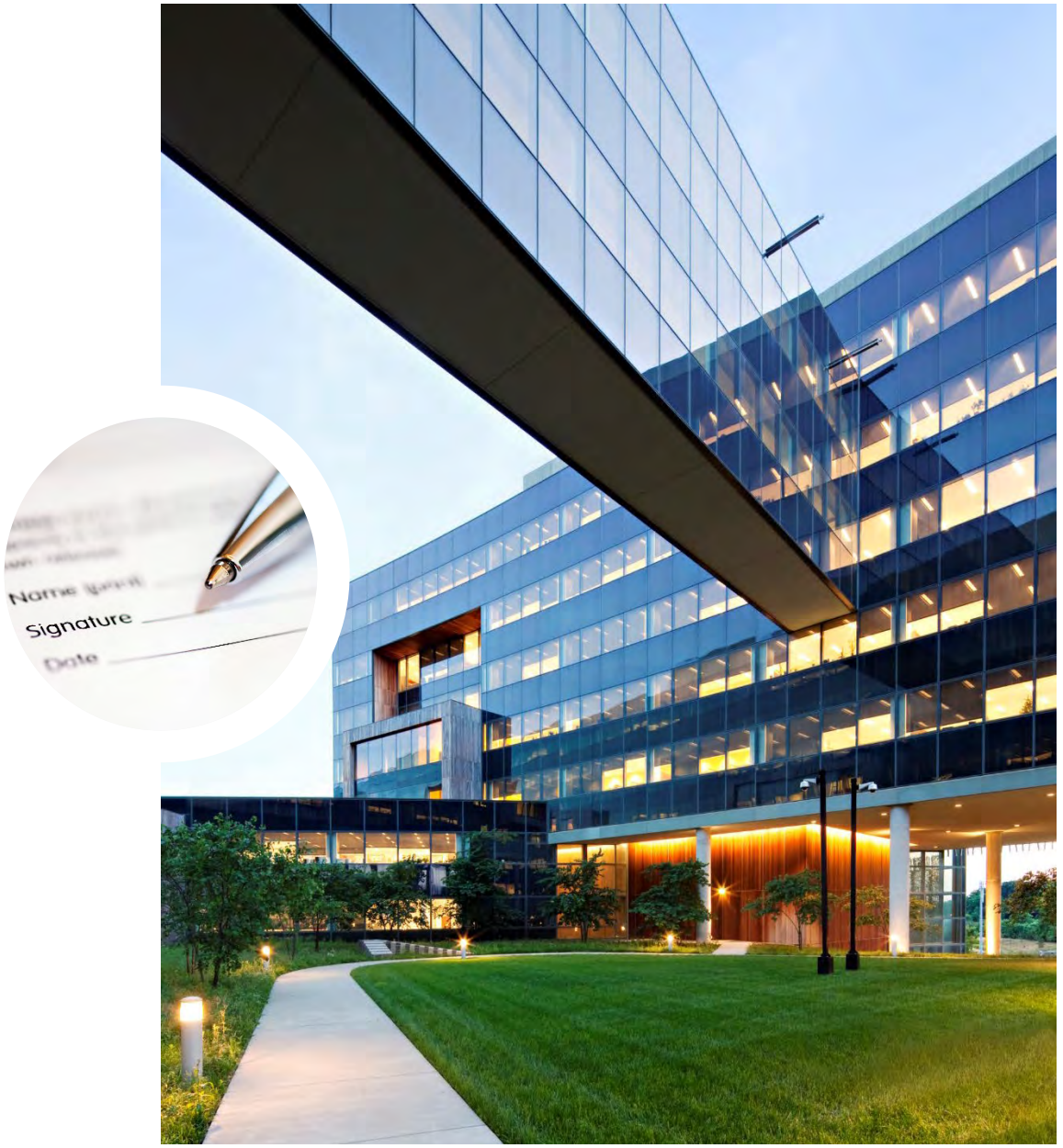
Digitization and Capture as a Service

October 26, 2023

2023 Federal Committee on Statistical Methodology (FCSM) Research and Policy Conference
University of Maryland, College Park, MD.

Nevada Basdeo
Kevin Schweickhardt
Brandon Dubbs
U.S. Census Bureau



United States® Census Bureau

June 30, 2024

United States® Census Bureau

Why Transition to Electronic Records? Challenges of Non-Digitial Physical Records

- Limited accessibility and inefficient retrieval
- Integration challenges and lack of security measures
- Data analysis inefficiency and collaboration hurdles
- Vulnerability to damage and deterioration
- Space requirements and backup limitations
- Mobility and version control issues

United States® Census Bureau

Decennial Census Digitization and Linkage (DCDL)

- Final component of creating longitudinal datasets that covers most of the U.S. population since 1940
- Demonstrates the infrastructure and methodology required to digitize historical records and capture data

Methodology:

1. Scan microfilm reels from 1960-1990 Censuses and create digital images
2. Use Optical Character Recognition (OCR) and Optical Mark Recognition (OMR) to capture names and additional information
3. Link new information to already digitized data

United States® Census Bureau

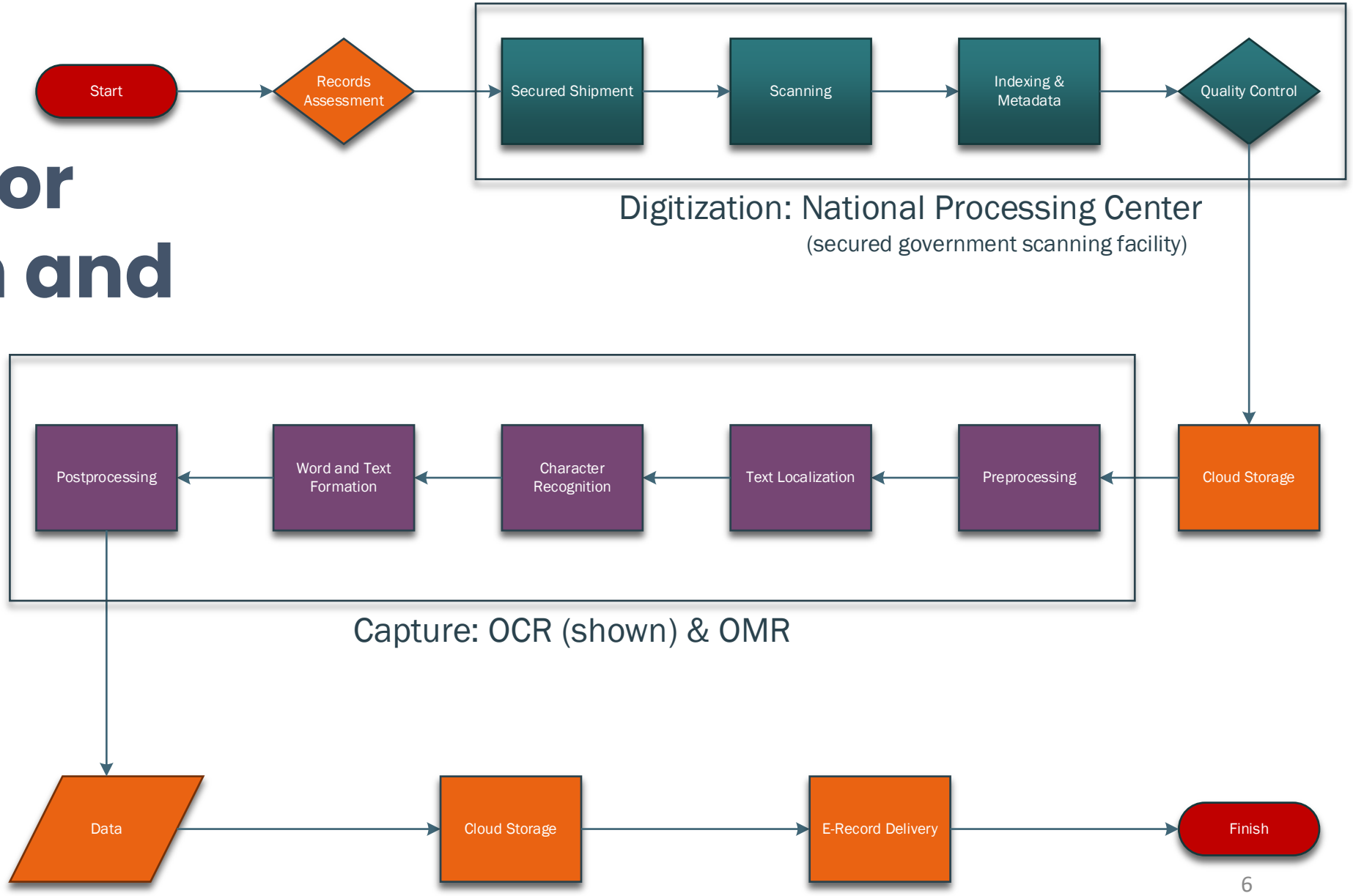
National Processing Center (NPC)

The Census Bureau's Secure Scanning Facility

- Conversion of documents, books, microfilm and microfiche into digital images.
- Secure shipment, handling, and storage
- Adherence to federal information security standards
- High volume production scanning equipment
- Skilled scanning operations staff, cleared and badged for secure records handling
- Production keying from image capabilities

United States® Census Bureau

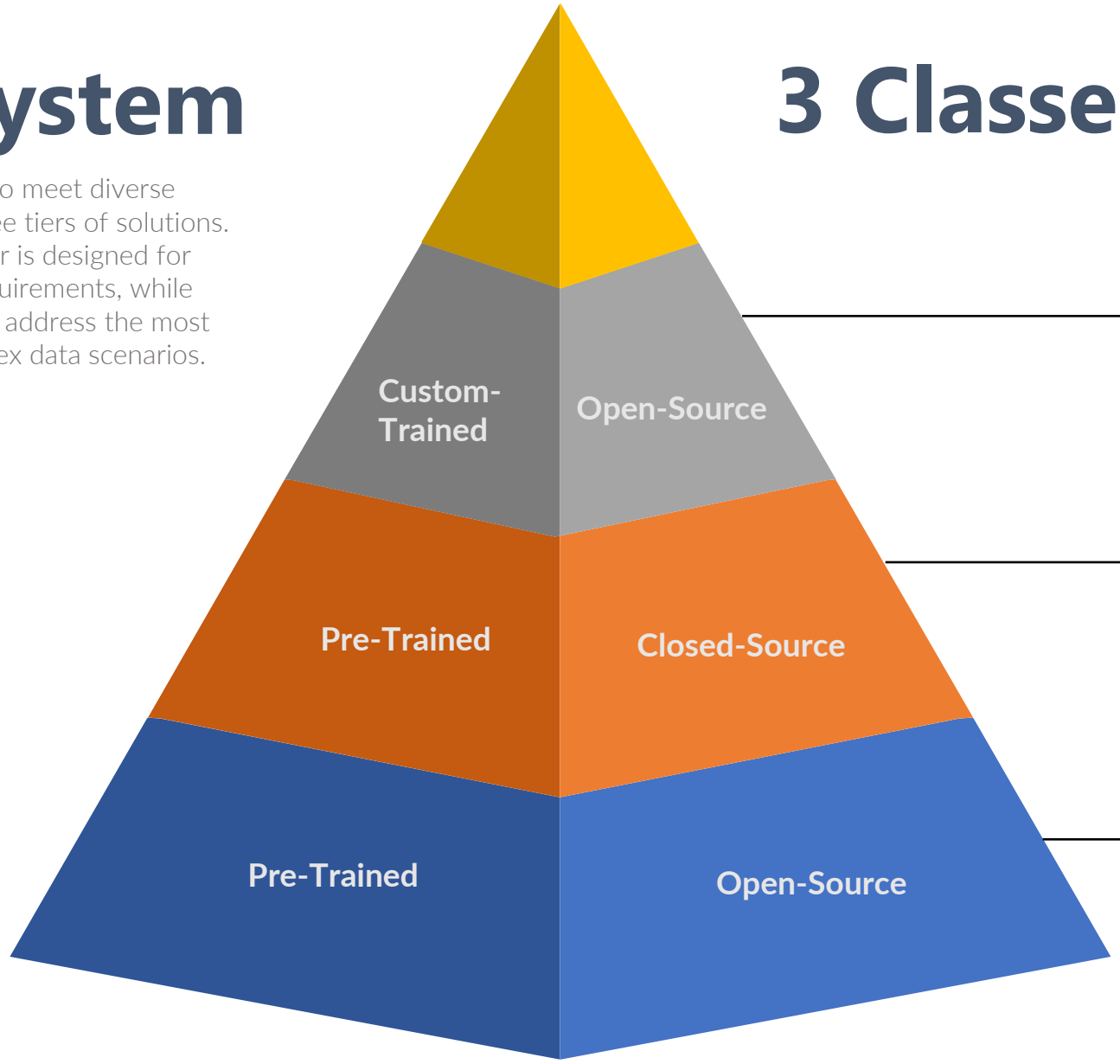
High Level Flowchart for Digitization and Capture



United States® Census Bureau

Tier System

Tailored approach to meet diverse needs, offering three tiers of solutions. The first or base tier is designed for straightforward requirements, while the third or top tier address the most intricate and complex data scenarios.



3 Classes of Models

Intricate and Complex Scenarios

- Ability to meet specific data capture requirements
- Can be extremely accurate
- Lengthy development and testing

Moderately Complex Scenarios

- Can be highly accurate on complicated forms
- Accelerated deployment, lots of features
- Ease of use and maintenance
- Method is a black box (proprietary)
- Can be costly (per page structure)

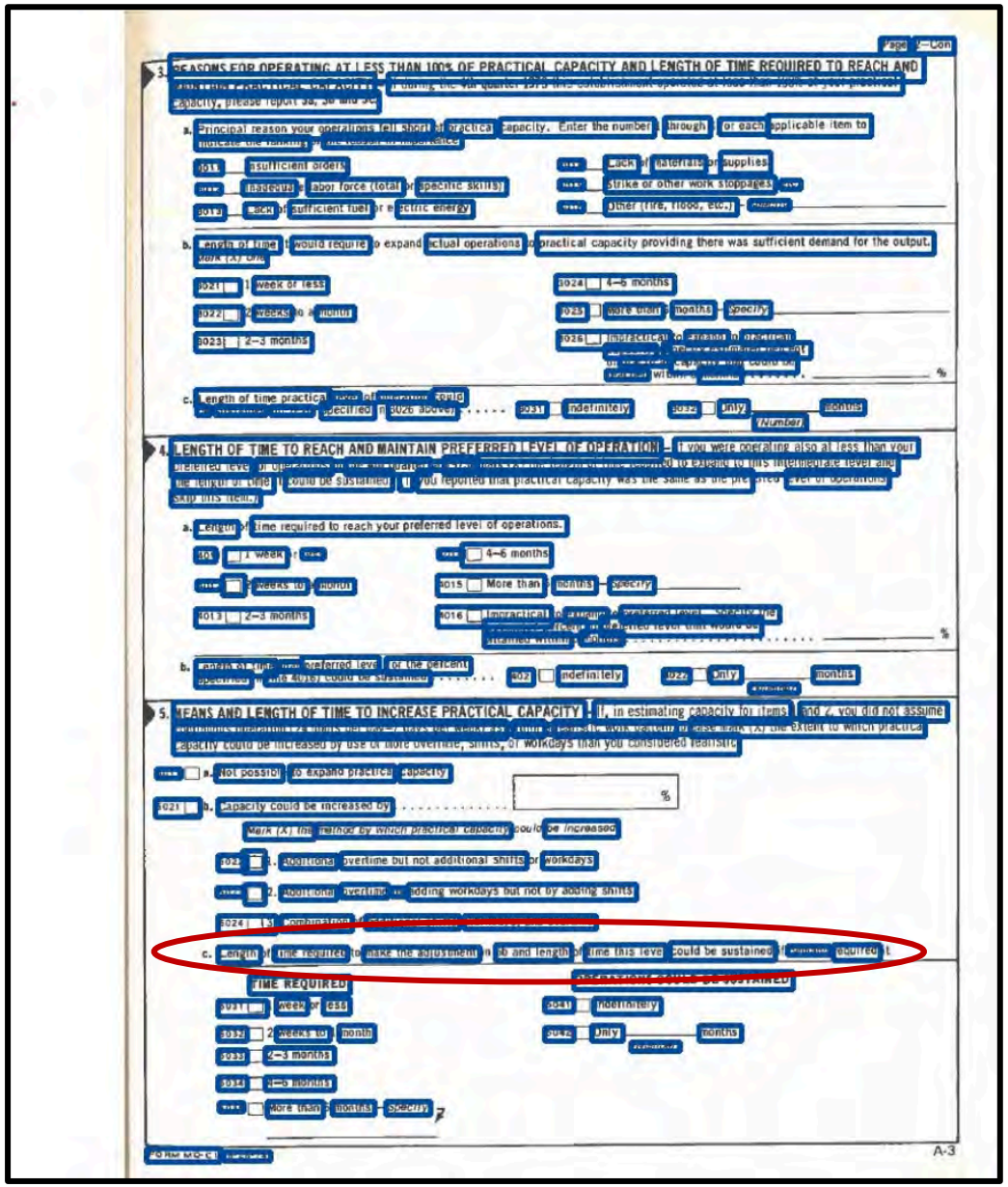
General use cases

- Can meet straightforward requirements
- Requires fine tuning; to accommodate specific needs
- Performs well for less complicated forms
- Quick to moderate development time
- Cost friendly

United States® Census Bureau

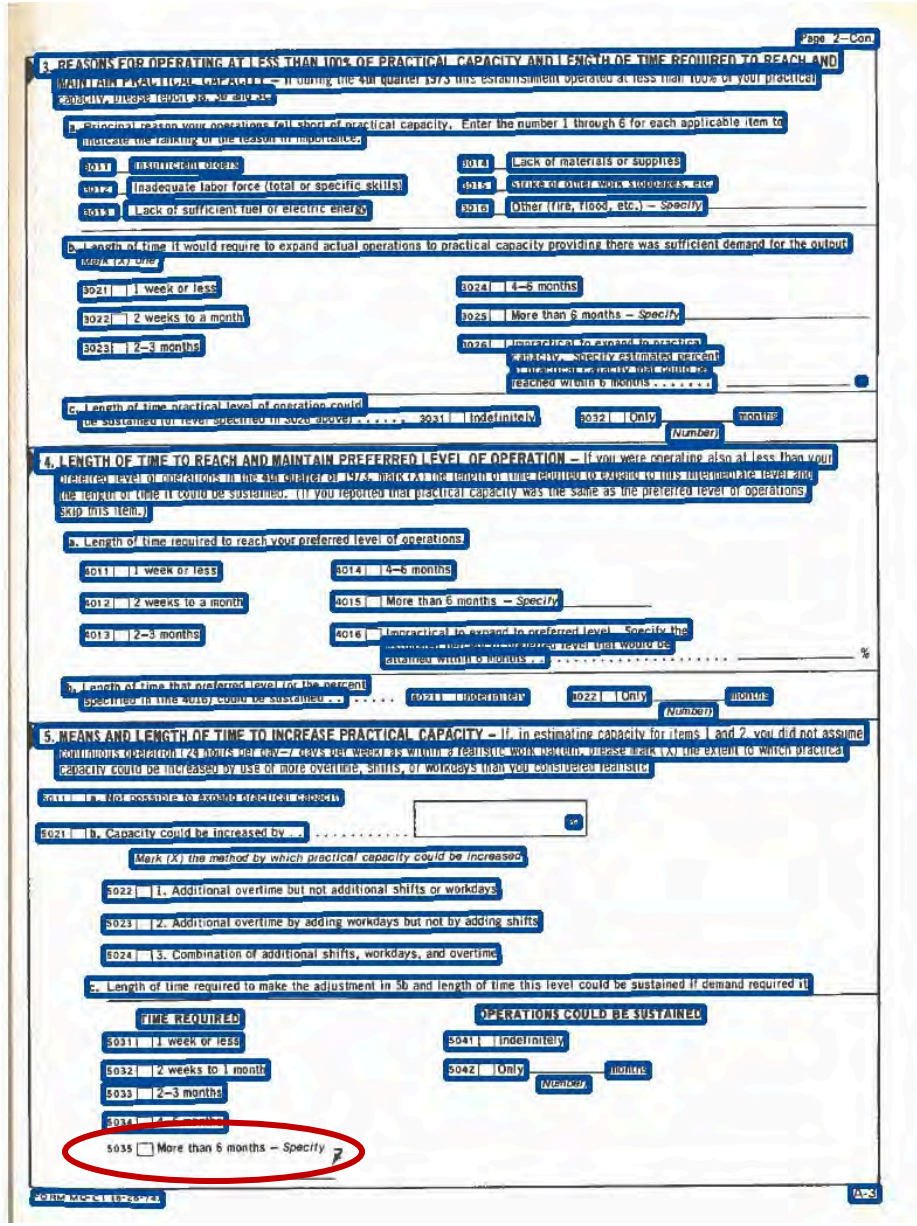
Open Source Pre-Trained (EasyOCR)

- Flexible
- Requires fine tuning
- Can accommodate specific needs
- Can be highly accurate on complicated forms
  - Segmented text when using “out-of-box”



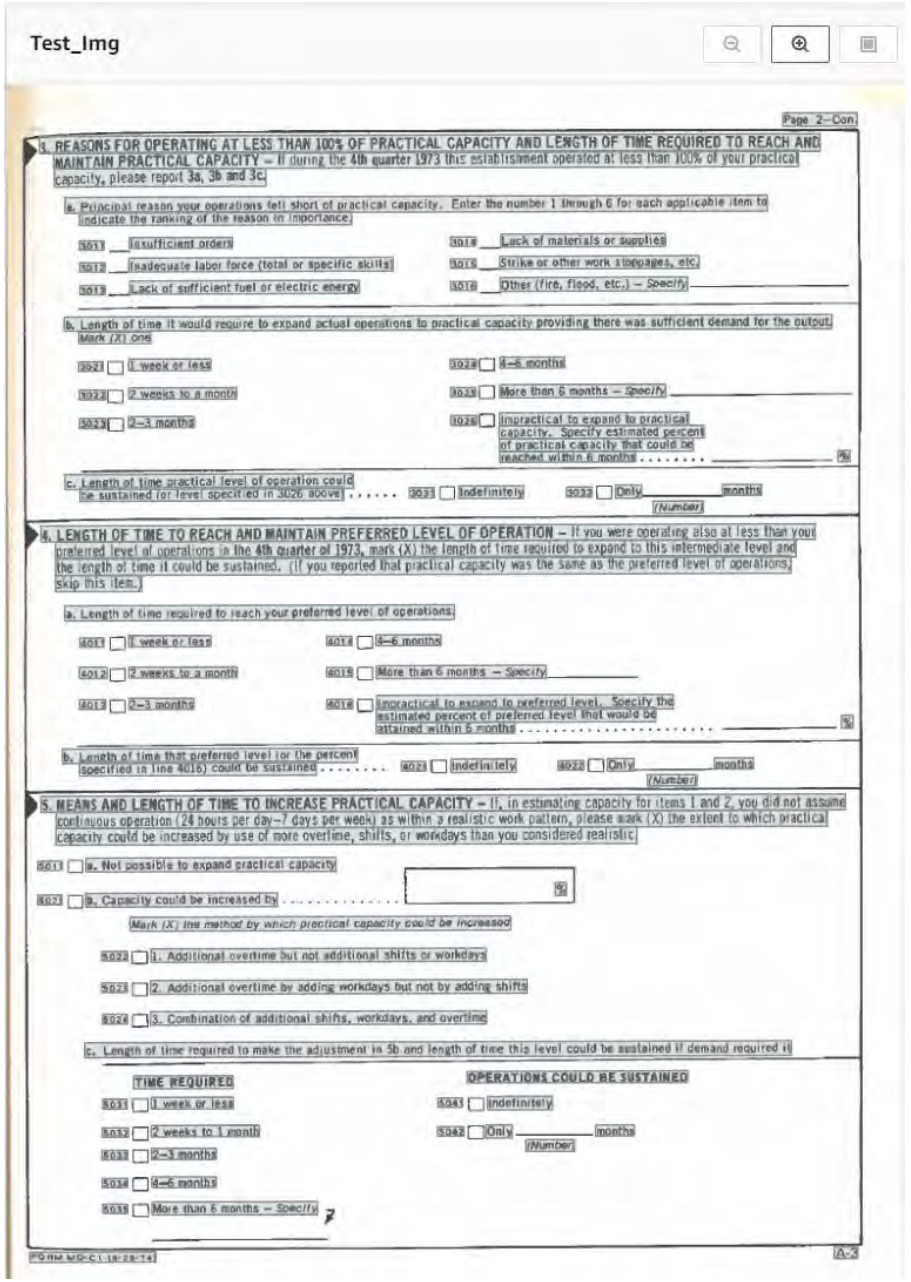
Open Source Pre-Trained (PaddleOCR)

- Flexible
- Requires fine tuning
- Can accommodate specific needs
- Can be highly accurate on complicated forms
  - But can miss some information when using “out-of-box”



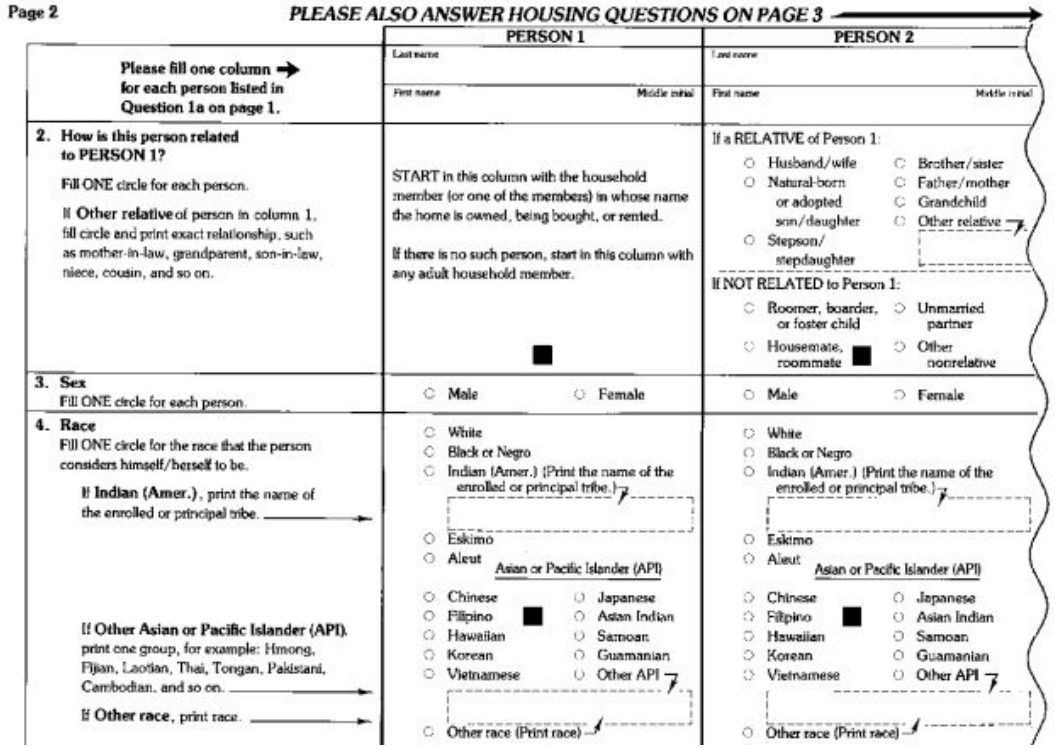
Closed Source Pre-Trained (Textract)

- Can be highly accurate on complicated forms
- Accelerated deployment, lots of features
- Ease of use and maintenance
- Method is a black box (proprietary)
- Can be costly (per page structure)



United States® Census Bureau

Custom-Trained OCR Model (DCDL)



Page 2 — PLEASE ALSO ANSWER HOUSING QUESTIONS ON PAGE 3

Please fill one column for each person listed in Question 1a on page 1.

PERSON 1 — Last name; First name; Middle initial

PERSON 2 — Last name; First name; Middle initial

2. How is this person related to PERSON 1?

Fill ONE circle for each person.

If Other relative of person in column 1, fill circle and print exact relationship, such as mother-in-law, grandparent, son-in-law, niece, cousin, and so on.

START in this column with the household member (or one of the members) in whose name the home is owned, being bought, or rented.

If there is no such person, start in this column with any adult household member.

If a RELATIVE of Person 1:
- Husband/wife
- Natural-born or adopted son/daughter
- Stepson/stepdaughter
- Brother/sister
- Father/mother
- Grandchild
- Other relative

If NOT RELATED to Person 1:
- Roomer, boarder, or foster child
- Housemate, roommate
- Unmarried partner
- Other nonrelative

3. Sex

Fill ONE circle for each person.

- Male
- Female

4. Race

Fill ONE circle for the race that the person considers himself/herself to be.

If Indian (Amer.), print the name of the enrolled or principal tribe.

If Other Asian or Pacific Islander (API), print one group, for example: Hmong, Fijian, Laotian, Thai, Tongan, Pakistani, Cambodian, and so on.

If Other race, print race.

- White
- Black or Negro
- Indian (Amer.) (Print the name of the enrolled or principal tribe.)
- Eskimo
- Aleut

Asian or Pacific Islander (API)
- Chinese
- Japanese
- Filipino
- Asian Indian
- Hawaiian
- Samoan
- Korean
- Guamanian
- Vietnamese
- Other API

Other race (Print race)

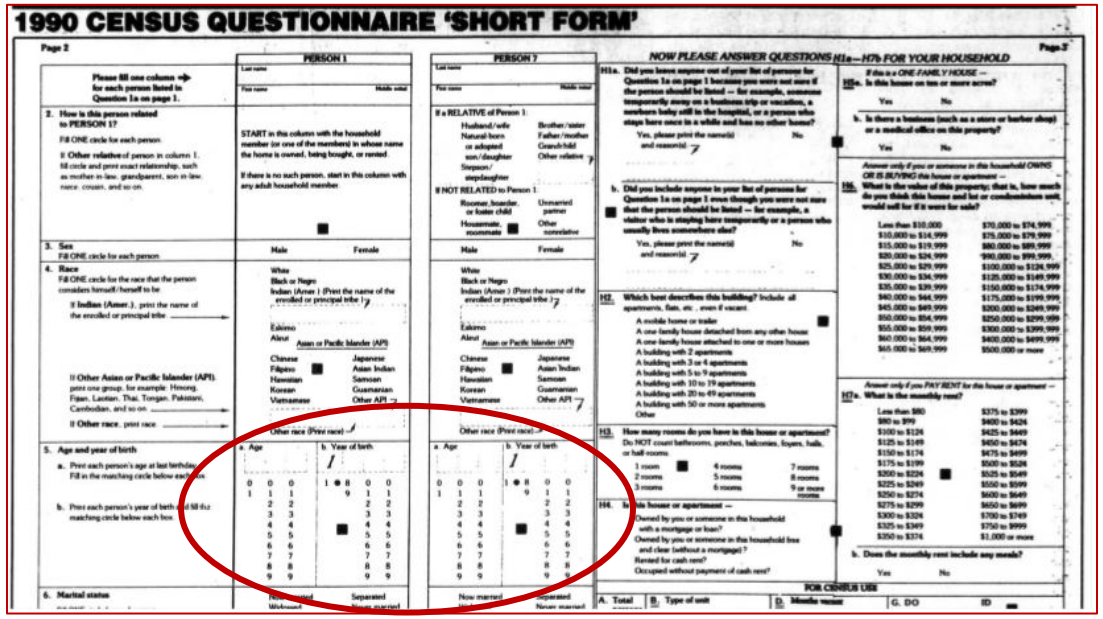
1990 Mean Exact Word Match (N = 9,700)

Name	Mean Match Rate (SE)
First	0.91 (0.28)
Last	0.89 (0.32)
Middle	0.88 (0.33)
Suffix	0.94 (0.24)

All results were approved for release by the U.S. Census Bureau, Data Management System number: DMS P-7532197 and approval number CBDRB-FY24-ERD006-001.

United States® Census Bureau

Custom-Trained OMR Model (DCDL)



OMR Accuracy

	OMR Marked Type 1*	OMR Marked Type 2*
Sample Size	541,000	533,000
Sample Pct	2.83%	0.15%
Percent Correct	98.66%	99.52%
Margin of Error	0.04%	0.02%
Lower CI	98.83%	99.50%
Upper CI	98.90%	99.55%

* Type 1 refers to empty bubbles and Type 2 refers to bubbles with numbers in it.

All results were approved for release by the U.S. Census Bureau, Data Management System number: DMS P-7532197 and approval number CBDRB-FY24-ERD006-001.

United States® Census Bureau

Digitization and Capture as a Service (DCaaS)

“Making Non-Digital, Digtial”

Securely transforming physical records into digital data, ensuring compliance with federal regulations, and enabling efficient use of valuable information.

Facilitating the Transition to Electronic Records for Federal Agencies

Decennial Census Digitization and Linkage (DCDL)

Methodology to digitize and capture microdata files from the 1960 – 1990 decennial census

OMB & NARA Memorandums

M-19-21 and M-23-07 specifying the federal transition to electronic records

DCaaS Digitization

Converting physical analog records (paper, microfilm, and microfiche) to digital records using federal digitization infrastructure capabilities

DCaaS Capture

Pre-trained and custom solutions to provide digital image data capture with a tiered and tailored approach to meet diverse needs

United States® Census Bureau

The Census Bureau DCaaS Team

Thank you for attending our presentation. Please feel free to reach out to the DCaaS team, we are happy to meet, to discuss digitization & capture challenges and solutions including your agency's electronic records, digitization and capture experience.

Nevada Basdeo
Program Director for Data Digitization
Business Development Staff
Economic Reimbursable Surveys Division
U.S. Census Bureau
nevada.basdeo@census.gov
301.763.7943

Kevin Schweickhardt
Data Digitization Program
Business Development Staff
Economic Reimbursable Surveys Division
U.S. Census Bureau - Contractor
kevin.l.schweickhardt@census.gov
301-763-5267

Brandon Dubbs
Data Digitization Program
Business Development Staff
Economic Reimbursable Surveys Division
U.S. Census Bureau - Contractor
brandon.m.dubbs@census.gov
301-763-0640

United States® Census Bureau