

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.

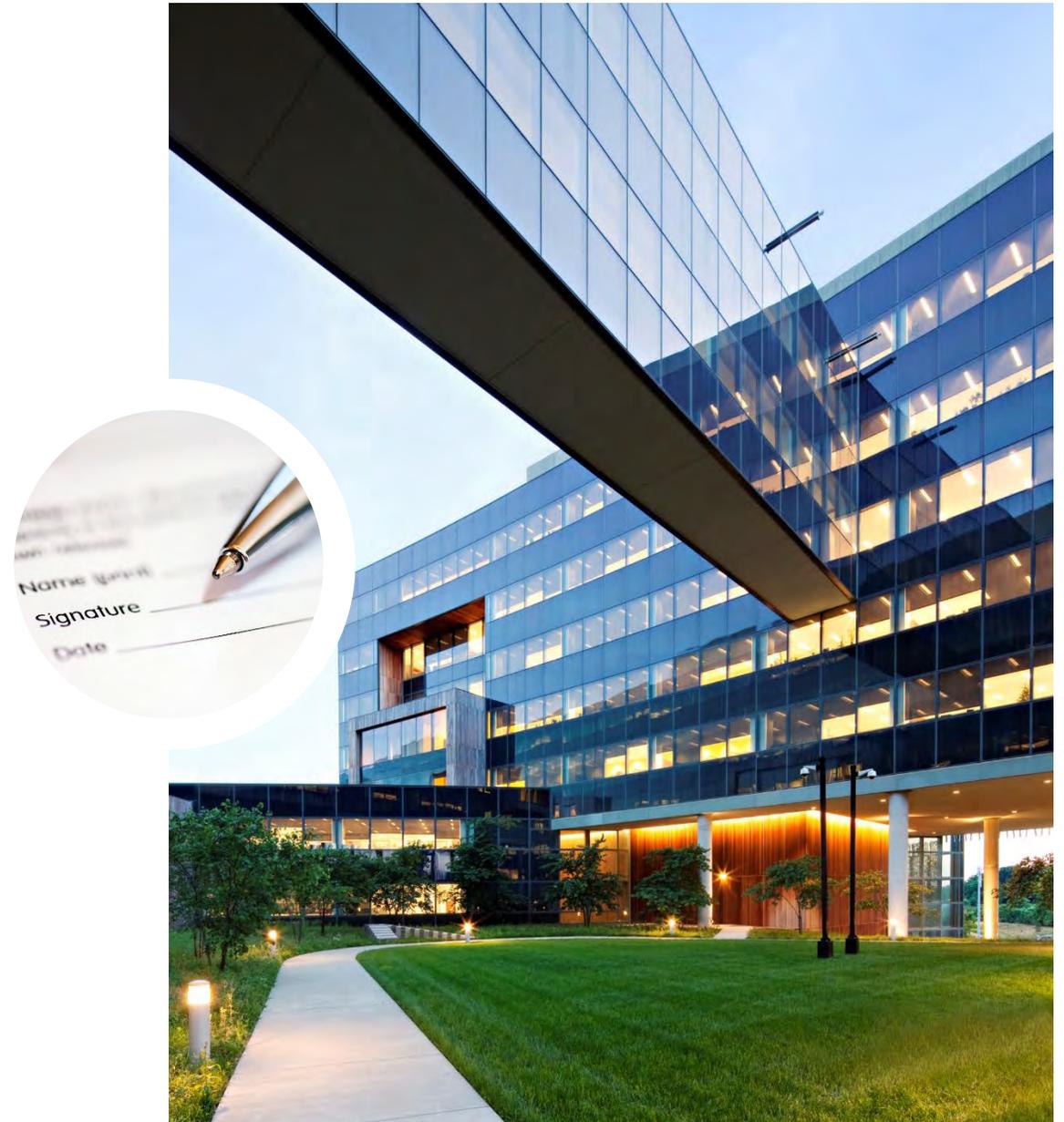
Nevada Basdeo

Kevin Schweickhardt

Brandon Dubbs

U.S. Census Bureau

Any conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. All results were approved for release by the Disclosure Review Board of the U.S. Census Bureau (Data Management System number: DMS P-7532197, Disclosure Review Board (DRB) approval numbers: CBDRB-FY24-ERD006-001.



June 30, 2024

Why Transition to Electronic Records? Challenges of Non-Digital 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

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

National Processing Center (NPC)

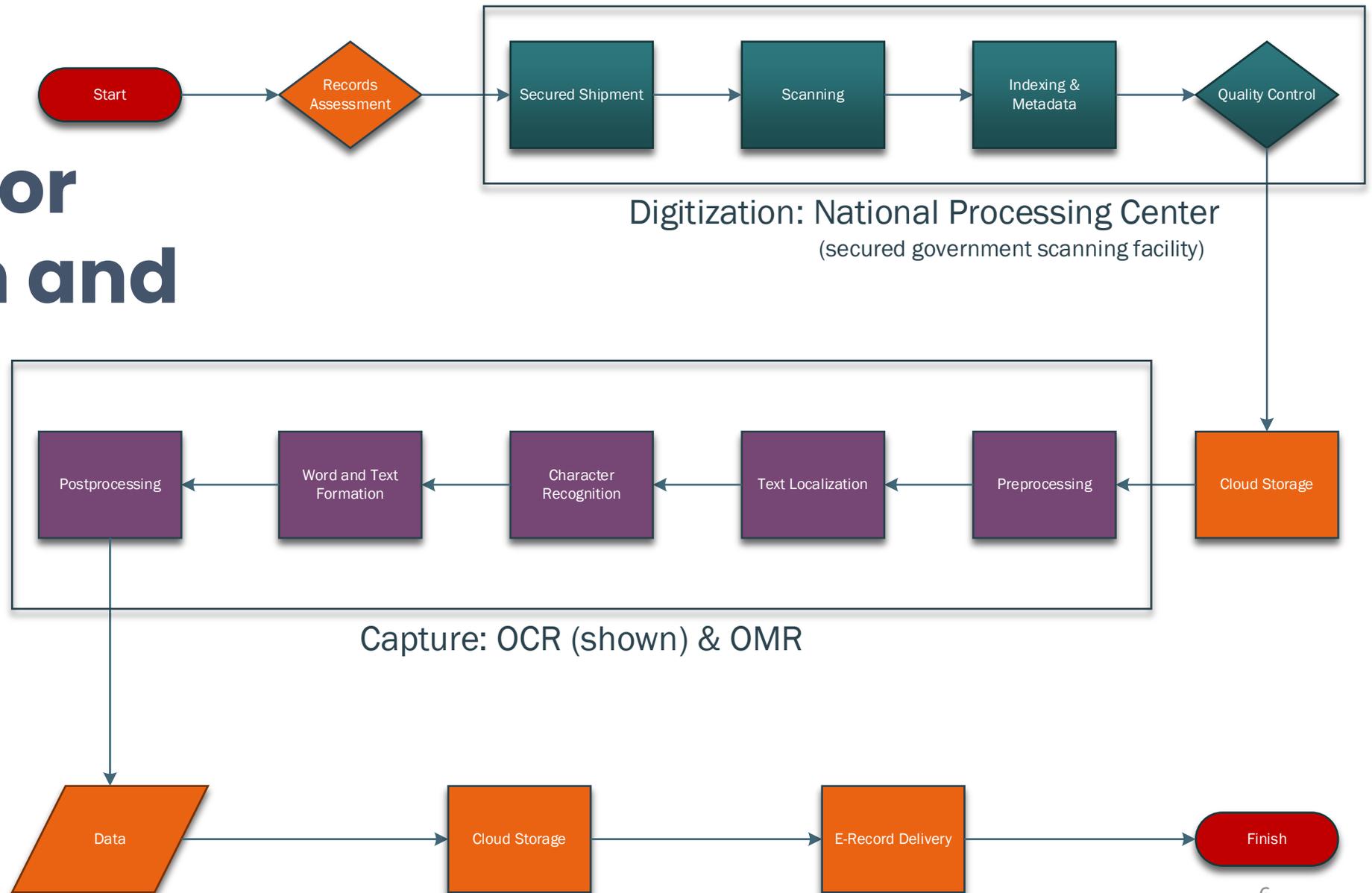
The Census Bureau's Secure Scanning Facility



- ✓ 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

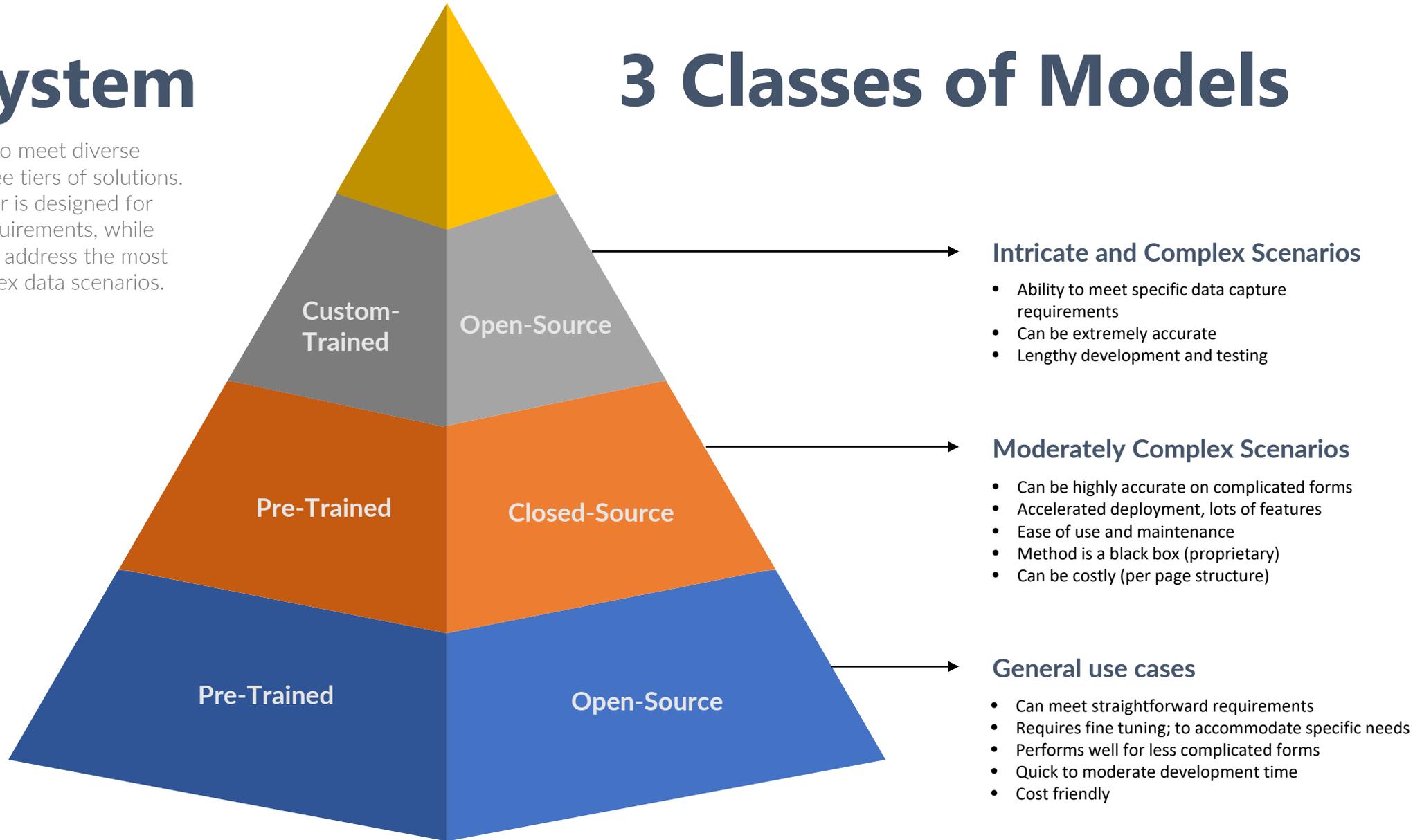
High Level Flowchart for Digitization and Capture



Tier System

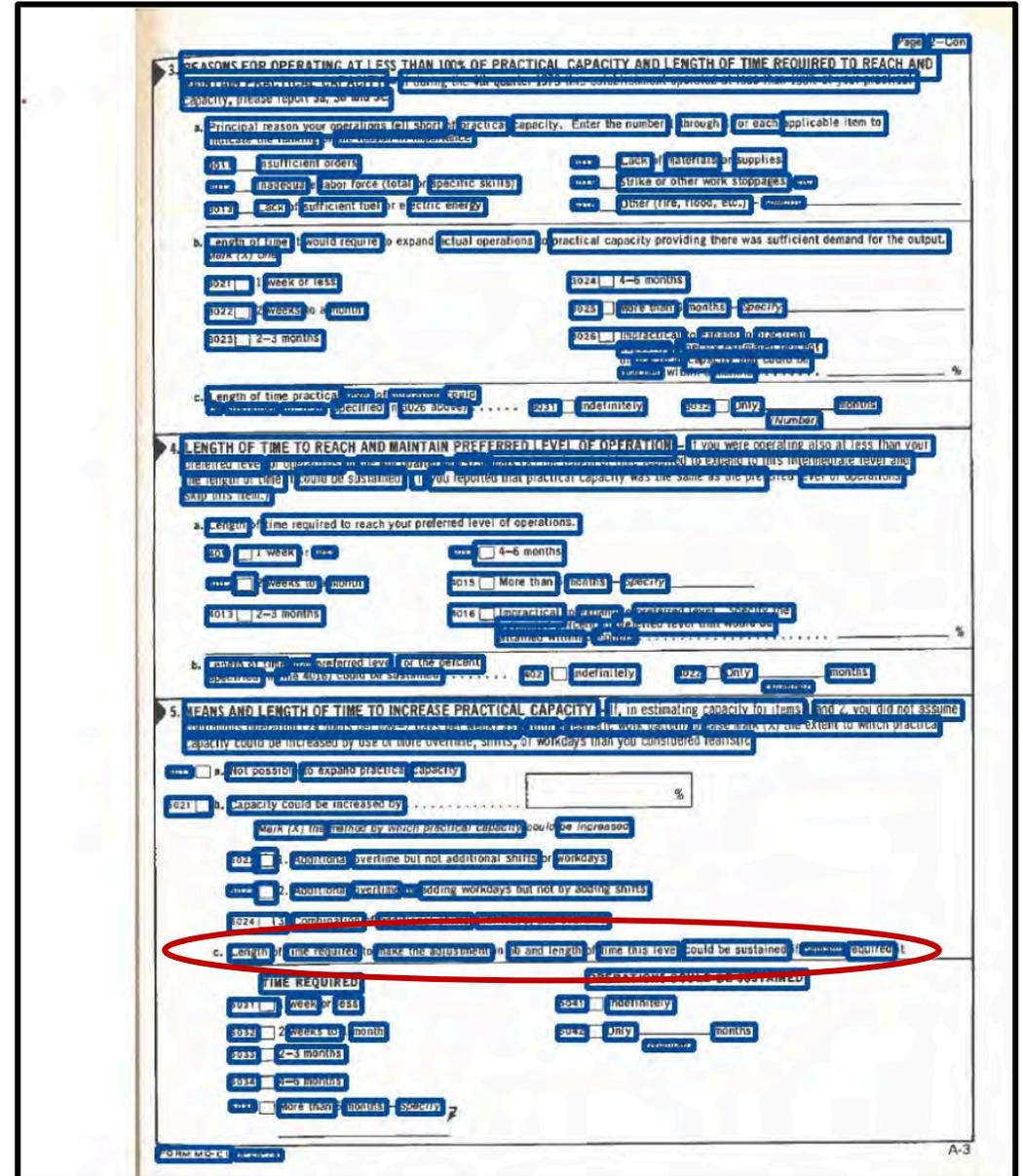
Tailored approach to meet diverse needs, offering three tiers of solutions. The first or base tier is designed for straightforward requirements, while the third or top tier address the most intricate and complex data scenarios.

3 Classes of Models



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”

Page 2 - Con

3. REASONS FOR OPERATING AT LESS THAN 100% OF PRACTICAL CAPACITY AND LENGTH OF TIME REQUIRED TO REACH AND MAINTAIN PREFERRED LEVEL OF OPERATION - If during the 4th quarter 1973 this establishment operated at less than 100% of your practical capacity, check (X) the reason or reasons below.

a. Principal reason your operations fell short of practical capacity. Enter the number 1 through 6 for each applicable item to indicate the ranking of the reason in importance.

3011 <input type="checkbox"/> Insufficient orders	3012 <input type="checkbox"/> Lack of materials or supplies
3014 <input type="checkbox"/> Inadequate labor force (total or specific skills)	3015 <input type="checkbox"/> Strike or other work stoppage
3013 <input type="checkbox"/> Lack of sufficient fuel or electric energy	3016 <input type="checkbox"/> Other (fire, flood, etc.) - Specify _____

b. Month or time it would require to expand actual operations to practical capacity providing there was sufficient demand for the output.

3021 <input type="checkbox"/> 1 week or less	3024 <input type="checkbox"/> 4-6 months
3022 <input type="checkbox"/> 2 weeks to a month	3025 <input type="checkbox"/> More than 6 months - Specify _____
3023 <input type="checkbox"/> 2-3 months	3026 <input type="checkbox"/> More than 6 months - Specify _____ If more than 6 months - Specify estimated percent reached within 6 months: _____

c. Length of time practical level of operation could be sustained to level specified in item 3021, 3022, 3023, 3024, 3025, 3026

3031 Indefinitely 3032 Only _____ month(s)

4. LENGTH OF TIME TO REACH AND MAINTAIN PREFERRED LEVEL OF OPERATION - If you were operating also at less than your preferred level of operations in the 4th quarter of 1973, check (X) the length of time required to expand to this level (practical level) and the length of time it could be sustained. (If you reported that practical capacity was the same as the preferred level of operations, skip this item.)

a. Length of time required to reach your preferred level of operations.

4011 <input type="checkbox"/> 1 week or less	4014 <input type="checkbox"/> 4-6 months
4012 <input type="checkbox"/> 2 weeks to a month	4015 <input type="checkbox"/> More than 6 months - Specify _____
4013 <input type="checkbox"/> 2-3 months	4016 <input type="checkbox"/> More than 6 months - Specify _____ If more than 6 months - Specify estimated percent reached within 6 months: _____

b. Length of time that preferred level (or the percent specified in item 4016) could be sustained.

4031 Indefinitely 4032 Only _____ month(s)

5. MEANS AND LENGTH OF TIME TO INCREASE PRACTICAL CAPACITY - If, in estimating capacity for items 1 and 2, you did not assume continuous operation (24 hours per day) 7 days per week as within a realistic work pattern, please mark (X) the extent to which practical capacity could be increased by use of more overtime, shifts, or workdays than you considered feasible.

5011 Not possible to increase practical capacity

5021 b. Capacity could be increased by _____

Mark (X) the method by which practical capacity could be increased.

5022 <input type="checkbox"/> 1. Additional overtime but not additional shifts or workdays
5023 <input type="checkbox"/> 2. Additional overtime by adding workdays but not by adding shifts
5024 <input type="checkbox"/> 3. Combination of additional shifts, workdays, and overtime

c. Length of time required to make the adjustment in 5b and length of time this level could be sustained if demand required it.

TIME REQUIRED	OPERATIONS COULD BE SUSTAINED
5031 <input type="checkbox"/> 1 week or less	5041 <input type="checkbox"/> Indefinitely
5032 <input type="checkbox"/> 2 weeks to 1 month	5042 <input type="checkbox"/> Only _____ month(s)
5033 <input type="checkbox"/> 2-3 months	5043 <input type="checkbox"/> _____
5034 <input type="checkbox"/> 3-6 months	5044 <input type="checkbox"/> _____
5035 <input type="checkbox"/> More than 6 months - Specify _____	5045 <input type="checkbox"/> _____

Closed Source Pre-Trained (Textextract)

- 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)

Test_Img

Page 2—Con

4. REASONS FOR OPERATING AT LESS THAN 100% OF PRACTICAL CAPACITY AND LENGTH OF TIME REQUIRED TO REACH AND MAINTAIN PRACTICAL CAPACITY — If during the 4th quarter 1973 this establishment operated at less than 100% of your practical capacity, please report 3a, 3b and 3c.

a. Principal reasons your operations fell short of practical capacity. Enter the number 1 through 6 for each applicable item to indicate the ranking of the reason in importance.

3011 <input type="checkbox"/> Insufficient orders	3016 <input type="checkbox"/> Lack of materials or supplies
3012 <input type="checkbox"/> Inadequate labor force (total or specific skills)	3017 <input type="checkbox"/> Strike or other work stoppages, etc.
3013 <input type="checkbox"/> Lack of sufficient fuel or electric energy	3018 <input type="checkbox"/> Other (fire, flood, etc.) — Specify _____

b. Length of time it would require to expand actual operations to practical capacity providing there was sufficient demand for the output. Mark (X) one.

3021 <input type="checkbox"/> 1 week or less	3026 <input type="checkbox"/> 4–6 months
3022 <input type="checkbox"/> 2 weeks to a month	3027 <input type="checkbox"/> More than 6 months — Specify _____
3023 <input type="checkbox"/> 2–3 months	3028 <input type="checkbox"/> Inoperational to expand to practical capacity. Specify estimated percent of practical capacity that could be reached within 6 months: _____ %

c. Length of time practical level of operation could be sustained (or level specified in 3026, 3028): 3029 Indefinitely 3030 Only _____ months (Number)

5. LENGTH OF TIME TO REACH AND MAINTAIN PREFERRED LEVEL OF OPERATION — If you were operating also at less than your preferred level of operations in the 4th quarter of 1973, mark (X) the length of time required to expand to this intermediate level and the length of time it could be sustained. (If you reported that practical capacity was the same as the preferred level of operations, skip this item.)

a. Length of time required to reach your preferred level of operations:

4011 <input type="checkbox"/> 1 week or less	4016 <input type="checkbox"/> 4–6 months
4012 <input type="checkbox"/> 2 weeks to a month	4017 <input type="checkbox"/> More than 6 months — Specify _____
4013 <input type="checkbox"/> 2–3 months	4018 <input type="checkbox"/> Inoperational to expand to preferred level. Specify the estimated percent of preferred level that would be attained within 6 months: _____ %

b. Length of time that preferred level (or the percent specified in line 4018) could be sustained: 4021 Indefinitely 4022 Only _____ months (Number)

5. MEANS AND LENGTH OF TIME TO INCREASE PRACTICAL CAPACITY — If, in estimating capacity for items 1 and 2, you did not assume continuous operation (24 hours per day—7 days per week) as within a realistic work pattern, please mark (X) the extent to which practical capacity could be increased by use of more overtime, shifts, or workdays than you considered realistic.

5011 a. Not possible to expand practical capacity

5012 b. Capacity could be increased by _____ %

Mark (X) the method by which practical capacity could be increased.

5021 <input type="checkbox"/> 1. Additional overtime but not additional shifts or workdays
5022 <input type="checkbox"/> 2. Additional overtime by adding workdays but not by adding shifts
5023 <input type="checkbox"/> 3. Combination of additional shifts, workdays, and overtime

c. Length of time required to make the adjustment in 5b and length of time this level could be sustained if demand required it

TIME REQUIRED	OPERATIONS COULD BE SUSTAINED
5031 <input type="checkbox"/> 1 week or less	5041 <input type="checkbox"/> Indefinitely
5032 <input type="checkbox"/> 2 weeks to 1 month	5042 <input type="checkbox"/> Only _____ months (Number)
5033 <input type="checkbox"/> 2–3 months	
5034 <input type="checkbox"/> 4–6 months	
5035 <input type="checkbox"/> More than 6 months — Specify _____	

FD-904 (5-6-73) (Rev. 2-28-74)

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		PERSON 2	
	Last name	First name Middle initial	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: <input type="radio"/> Husband/wife <input type="radio"/> Brother/sister <input type="radio"/> Natural-born or adopted son/daughter <input type="radio"/> Father/mother <input type="radio"/> Stepson/stepdaughter <input type="radio"/> Grandchild <input type="radio"/> Other relative →	
3. Sex Fill ONE circle for each person.	<input type="radio"/> Male <input type="radio"/> Female		<input type="radio"/> Male <input type="radio"/> 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. →	<input type="radio"/> White <input type="radio"/> Black or Negro <input type="radio"/> Indian (Amer.) (Print the name of the enrolled or principal tribe.) → <input type="radio"/> Eskimo <input type="radio"/> Aleut <input checked="" type="radio"/> Asian or Pacific Islander (API) <input type="radio"/> Chinese <input type="radio"/> Japanese <input type="radio"/> Filipino <input type="radio"/> Asian Indian <input type="radio"/> Hawaiian <input type="radio"/> Samoan <input type="radio"/> Korean <input type="radio"/> Guamanian <input type="radio"/> Vietnamese <input type="radio"/> Other API → <input type="radio"/> Other race (Print race) →		<input type="radio"/> White <input type="radio"/> Black or Negro <input type="radio"/> Indian (Amer.) (Print the name of the enrolled or principal tribe.) → <input type="radio"/> Eskimo <input type="radio"/> Aleut <input type="radio"/> Asian or Pacific Islander (API) <input type="radio"/> Chinese <input type="radio"/> Japanese <input type="radio"/> Filipino <input type="radio"/> Asian Indian <input type="radio"/> Hawaiian <input type="radio"/> Samoan <input type="radio"/> Korean <input type="radio"/> Guamanian <input type="radio"/> Vietnamese <input type="radio"/> Other API → <input type="radio"/> 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.

Custom-Trained OMR Model (DCDL)

1990 CENSUS QUESTIONNAIRE 'SHORT FORM'

Page 2

PERSON 1

Last name: _____ First name: _____ Middle name: _____

2. How is this person related to PERSON 1?

START in the column with the household member for one of the married in whose name the home is owned, being bought, or rented. If there is no such person, start in the column with any adult household member.

3. Sex

Male Female

4. Race

White Black or Negro Indian (Amer.) Other race (Print race) _____

5. Age and year of birth

a. Print each person's age at last birthday. Fill in the matching circle below each box.

b. Print each person's year of birth. Fill in the matching circle below each box.

6. Marital status

Never married Married Separated Widowed

PERSON 7

Last name: _____ First name: _____ Middle name: _____

8. How is this person related to PERSON 1?

9. Sex

Male Female

10. Race

White Black or Negro Indian (Amer.) Other race (Print race) _____

11. Age and year of birth

a. Age: 0 0 0 1 1 1 8 0 0 0

b. Year of birth: 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 6 6 7 7 7 8 8 8 9 9 9

12. Which best describes this building?

A. Total B. Type of unit C. Multiple units D. D.D. E. ID

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.

Digitization and Capture as a Service (DCaaS)

“Making Non-Digital, Digital”

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

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