PROCESSING OPTICAL MARK RECOGNITION DATA FROM THE 1990 DECENNIAL CENSUS

[US Census Bureau] | [Associate Directorate for Economic Programs]

Anna Capels, Data Analyst Fellow | *Colorado State University, Psychology and Data Science*

Keywords:

Data reformation, Longitudinal linkage, Optical Mark Recognition (OMR) pipeline

Summary:

Created a program within a OMR pipeline to reformat the output txt files from OMR to aid in linkage between multiple databases. Overall, the newly arranged data will be implemented in the Decennial Census Digitization and Linkage (DCDL) Project that aims to generate longitudinal datasets that span from 1940 to the latest decennial census (2020). Those datasets will then be used for novel research interested in behavior within the US spanning multiple generations.

PROCESSING OPTICAL MARK RECOGNITION DATA FROM THE 1990 DECENNIAL CENSUS

Associate Directorate for Economic Programs
U.S. Census Bureau
John Sullivan — Survey Statistician

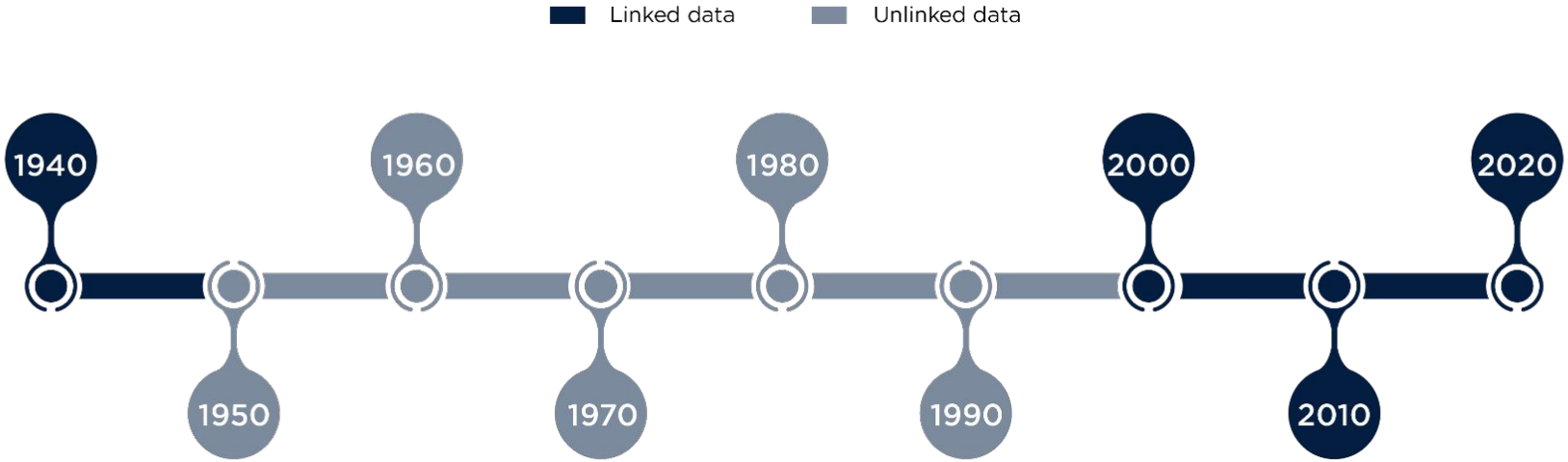
coding it forward >

United States™ Census Bureau

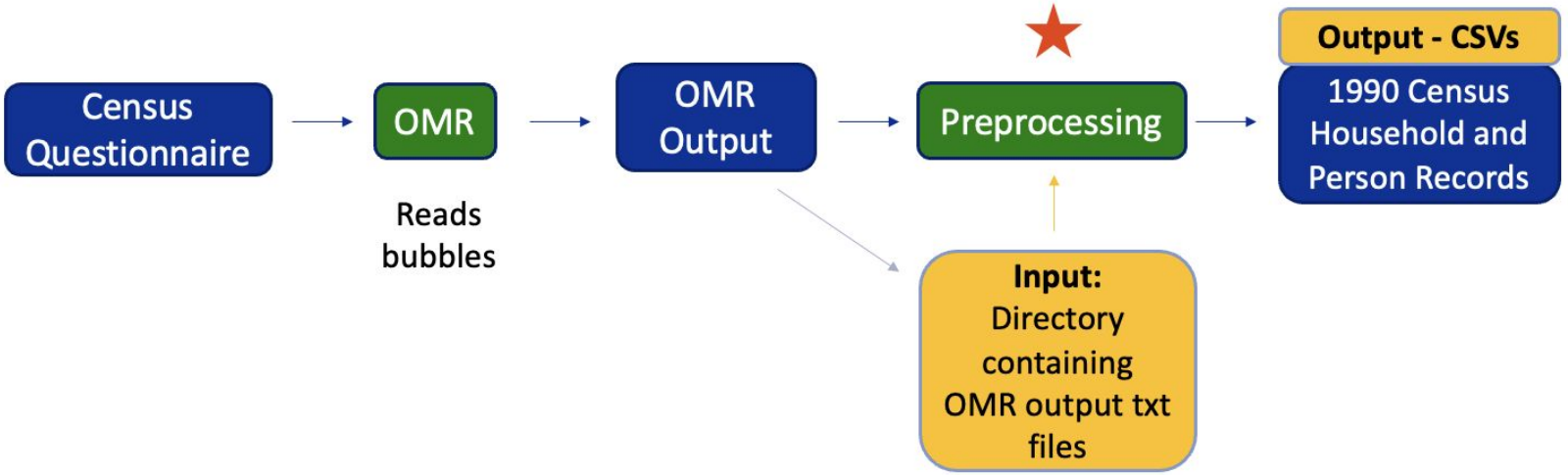
ANNA CAPELS
Colorado State University
Psychology & Data Science

BACKGROUND

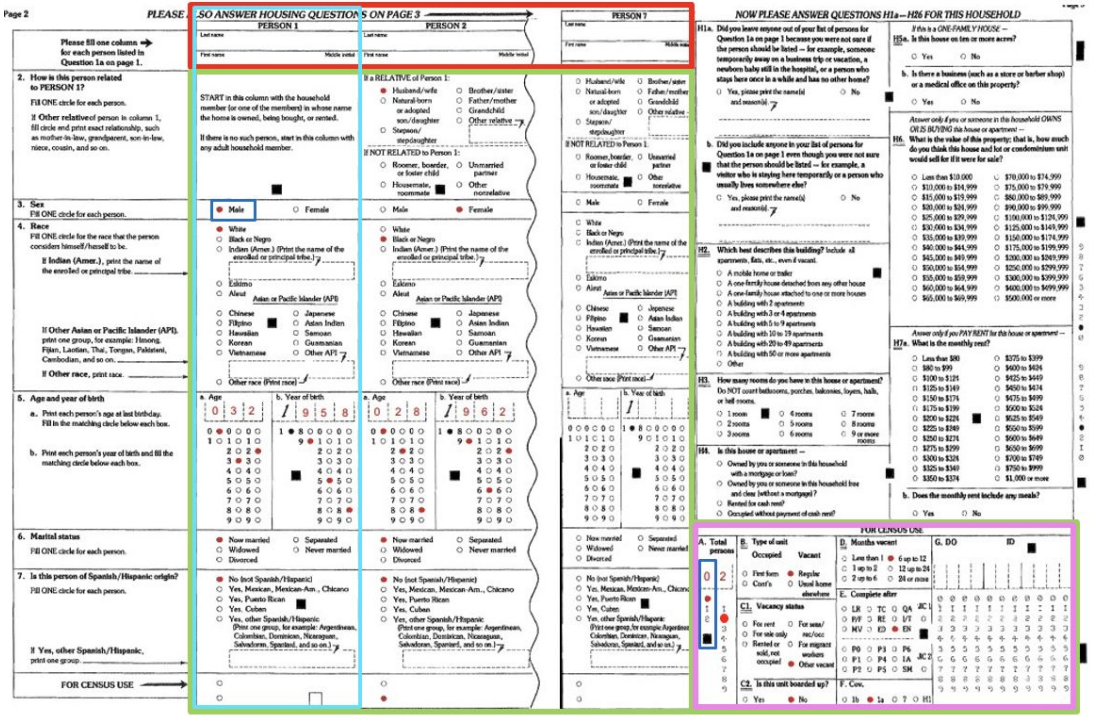
Decennial Census Digitization and Linkage (DCDL) Project: Linking microdata files from the 1950 - 1990 decennial censuses. Output will produce large longitudinal dataset to track behaviors across generations in the U.S. population from the 1940s to present-day.



OPTICAL MARK RECOGNITION (OMR) PIPELINE OVERVIEW



LAYOUT OF QUESTIONNAIRE & OMR OUTPUT



Names captured by OCR

OMR Capture

```
B:1990_72038972,01-Jan-24,07-16-2024 19:50:23
P:bw1990_72038972_0002.tiff,80,0
IBUB:19384653275
D:tot_per_d2_0,0
D:tot_per_d1_2,0
D:boarded_no,0
D:vacant_rented,0
D:unit_vac_reg,0
D:vacant_6to12,0
D:after_en,0
D:fcov_1a
P:bw1990_72038972_0002.tiff,80,60
IBUB:19384650198
D:male,1
D:race_white,1
D:age_d3_0,1
D:age_d2_3,1
D:age_d1_2,1
D:yob_d3_9,1
D:yob_d2_5,1
D:yob_d1_8,1
D:married,1
D:hisp_no,1
D:rel_spouse,2
D:female,2
D:race_black,2
D:age_d3_0,2
D:age_d2_2,2
D:age_d1_8,2
D:yob_d3_9,2
D:yob_d2_6,2
D:yob_d1_2,2
D:married,2
D:hisp_no,2
D:cancel_2,2
E:38
```

cif>

PREPROCESSING PT. 1 – REFORMATTING TO CSV OUTPUT

Before

	col_1	ID	col_3	RT	info_clean	b_key	p_key
0	B:1990_72038972	01-Jan-24	07-16-2024 19:50:23	B	1990_72038972	1990_72038972	0
1	P:bw1990_72038972_0002.tiff	80	0	P	bw1990_72038972_0002	1990_72038972	1990_72038972_0002
2	IBUB:19384653275	nan	nan	IB	19384653275	1990_72038972	1990_72038972_0002
3	D:tot_per_d2_0	0	nan	D	tot_per_d2_0	1990_72038972	1990_72038972_0002
4	D:tot_per_d1_2	0	nan	D	tot_per_d1_2	1990_72038972	1990_72038972_0002
5	D:boarded_no	0	nan	D	boarded_no	1990_72038972	1990_72038972_0002
6	D:vacant_rented	0	nan	D	vacant_rented	1990_72038972	1990_72038972_0002
7	D:unit_vac_reg	0	nan	D	unit_vac_reg	1990_72038972	1990_72038972_0002
8	D:vacant_6to12	0	nan	D	vacant_6to12	1990_72038972	1990_72038972_0002
9	D:after_en	0	nan	D	after_en	1990_72038972	1990_72038972_0002

1 row per person

IB_key = B_key + P_key

PER_ID = IB_key + ID

After

	PER_ID	ID	BUP	occupied_unit	...	SEX	RAC	MS	SOR	REL
1990_72038972_0002_193846501981	1			0	...	0	971	0	0	00
1990_72038972_0002_193846501982	2			0	...	1	972	0	0	01
1990_72038972_0002_193846532750	0		2	0	...					

AGE, YOB, NPU

Before

D:age_d3_0	1	nan	D	age_d3_0
D:age_d2_3	1	nan	D	age_d2_3
D:age_d1_2	1	nan	D	age_d1_2
D:yob_d3_9	1	nan	D	yob_d3_9
D:yob_d2_5	1	nan	D	yob_d2_5
D:yob_d1_8	1	nan	D	yob_d1_8

After

AGE	YOB
032	1958
028	1962

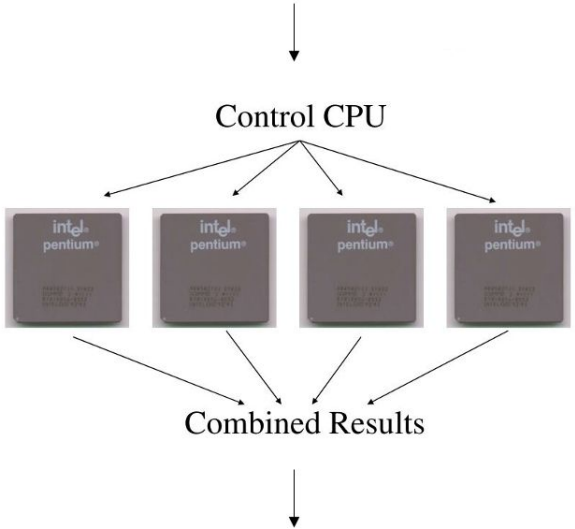
House and Person CSVs

Both	House	Person
Scan_date, B_key, P_key, UID, IB_Key, Flag for IB_key, PER_ID, ID	NPU, VS, BUP, MV, occupied_unit, vacant_unit, All after codes, All fcov codes	SEX, AGE, RAC, MS, SOR, REL, YOB, cancel_1, cancel_2

cif>

PREPROCESSING PT. 2 - PARALLEL PROCESSING

110,00+ files (300+ million data records)



Effect of Parallel Processing

LOOKING AHEAD

- **Link and compare newly transformed data to existing 1990 microdata**
  - Ex. Quality check state population level counts
  - Ex. Connect names
  - Ex. Connect geographical data (addresses)
- **Program to be a template for 1960-1980 OMR processing**
- **Linkable 1990 Decennial Census files**
  - Link to other records at the Census Bureau

cif>

THANK YOU!

Special thanks to:
John Sullivan, Sophie Schafer, and Katie Genadek!

cif>