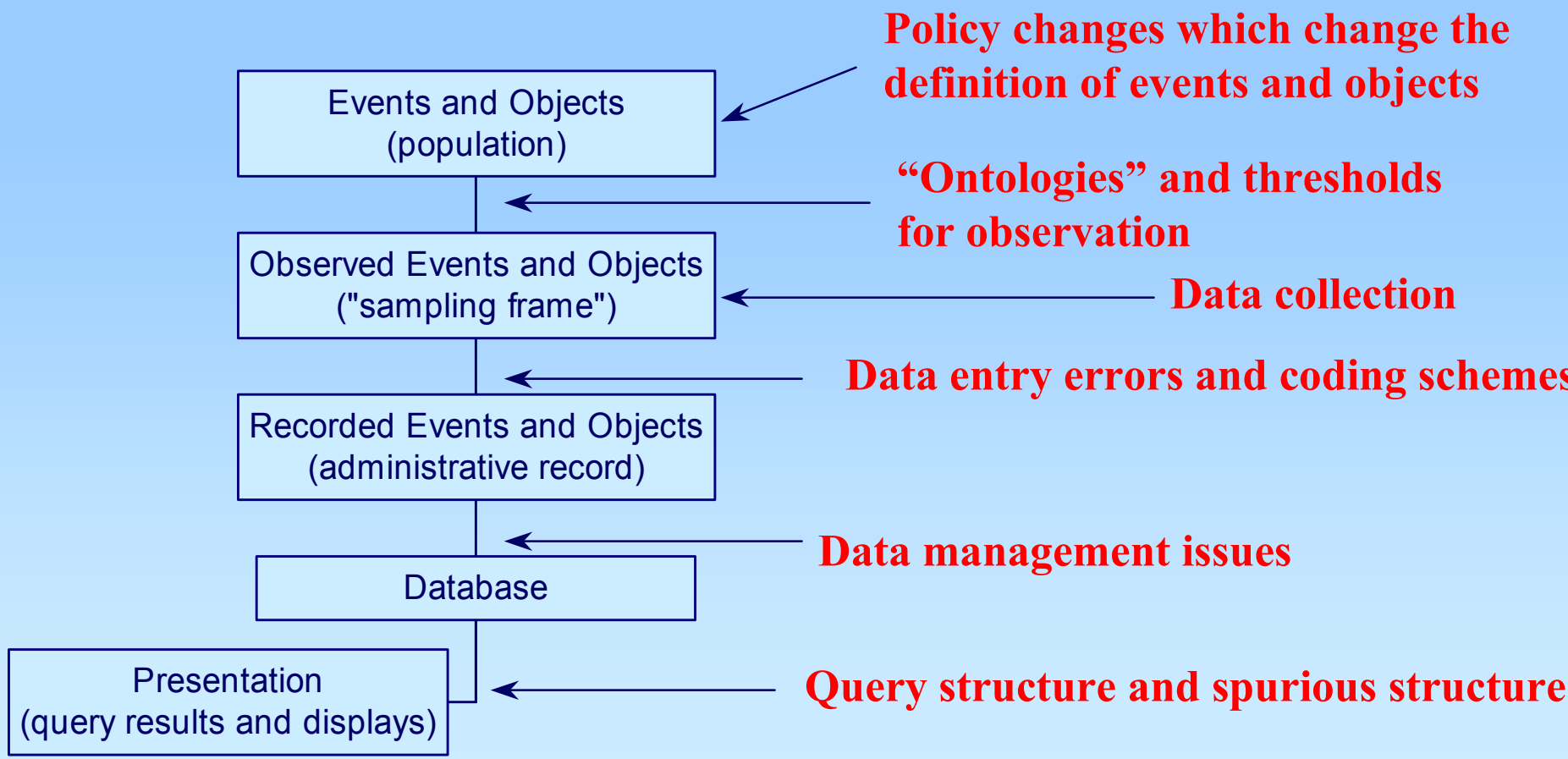
The Statistical Administrative Records System and Administrative Records Experiment 2000: System Design, Successes, and Challenges

Dean H. Judson

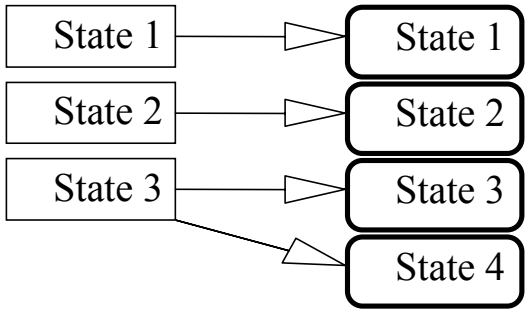
Planning, Research and Evaluation Division
U.S. Census Bureau

How Administrative Records Are Created and Used

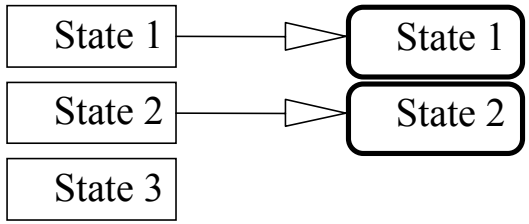


Ontologies and Data Quality

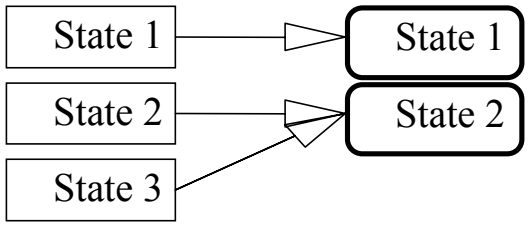
Proper Representation



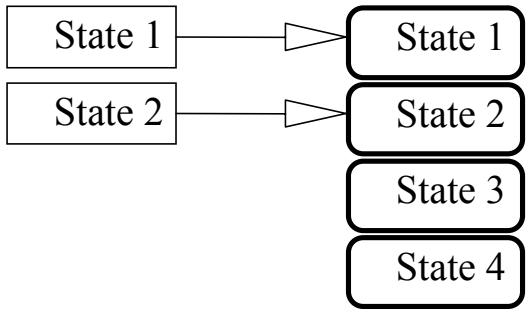
Incomplete Representation



Ambiguous Representation



Meaningless States

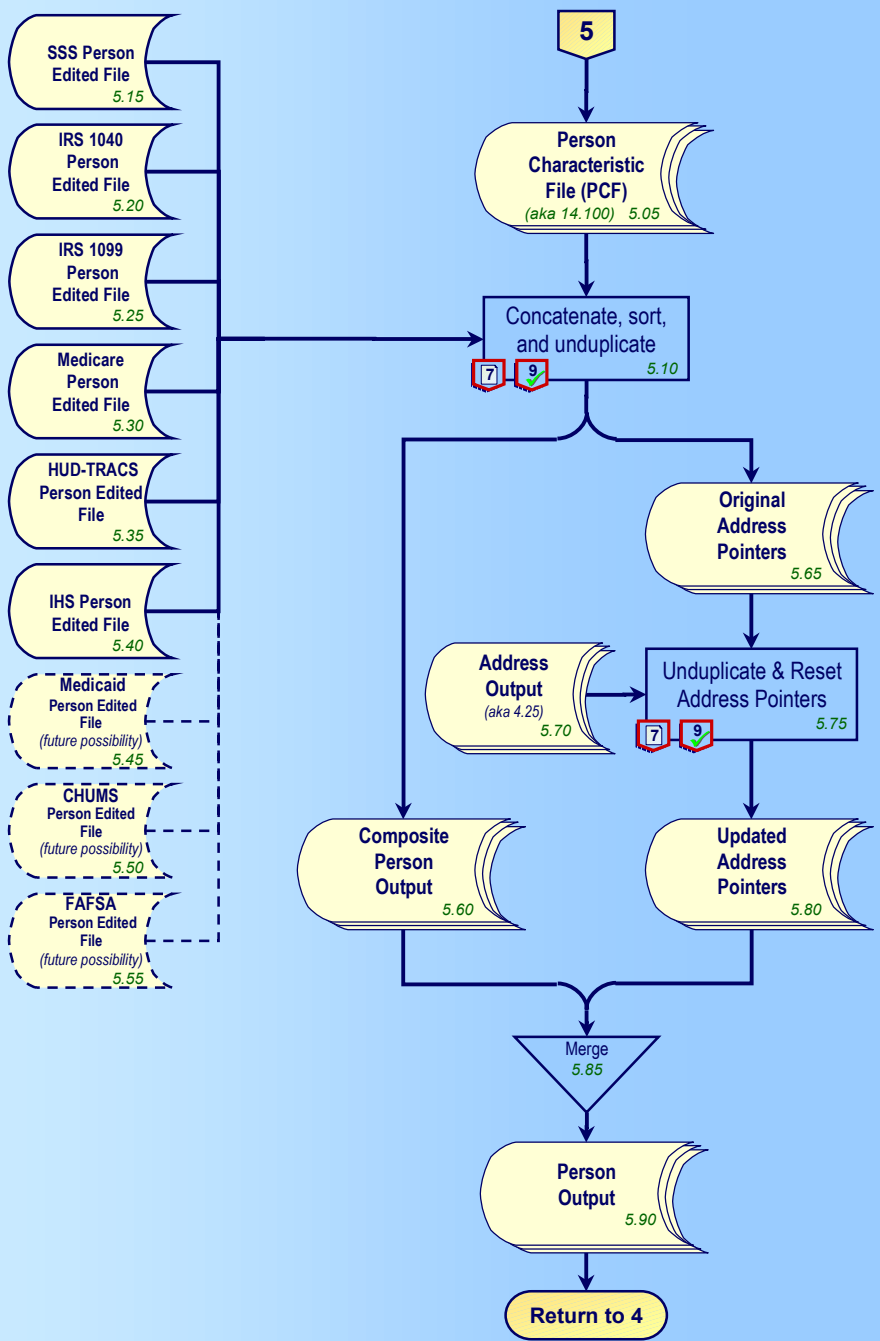


Source: Wand and Wang, 1996:90

Background and History

- Statistical Administrative Records System
  - Six large Federal input files: IRS 1040, IRS 1099, Selective Service, Medicare, Indian Health Service, HUD-TRACS
  - One lookup file: SSA/Census Numident
- AREX 2000
  - Attempt to use STARS data to simulate administrative records census

A Diagrammatic Depiction of Files Used to Create the Final StARS Database



Characteristics of Files Included in the STARS System

- IRS Individual Master 1040 File:
  - Tax year data; April, 2000 refers to "tax year" 1999
  - TY '99 file arrives October, 2000
  - Business entities, estates, other institutions included
  - 120 million records/year
  - Households below the filing threshold do not need to file
- Tax Filing Unit ≠ Housing Unit
  - Czajka, 2000: 10-20% of addresses are PO Boxes, business addresses, tax preparers
- Limited microdata content:
  - TY95+: SSN's of dependents requested, recorded
  - Czjaika, 2000: 1987 study: .5% of primary filer, 1.6% of secondary filer, 3.4% of dependents' SSN's in error
  - Age, race, sex hispanic origin microdata not available

Characteristics of Files Included in the STARS System, cont.

- IRS Information Returns (1099) File:
  - Tax year data; April, 2000 refers to "tax year" 1999
  - TY '99 file arrives October, 2000
  - Business entities, estates, other institutions included
  - 775 million records/year
  - Recipient address ≠ Housing Unit
  - Czajka, 2000: 10-20% of addresses are PO Boxes, business addresses, tax preparers
  - Limited microdata content: Age, race, sex hispanic origin microdata not available

Characteristics of Files Included in the STARS System, cont

- Selective Service File:
  - About 13 million records
  - Registration required in 1940, suspended in 1975, resumed in 1980
  - Presumably, males 18-25 are required to inform SSS when they move
  - Females, non-immigrant aliens, hospitalized, incarcerated, and institutionalized males, and members of the armed forces are exempt
  - Limited microdata content: Race, Hispanic origin microdata not available
  - Address information may not be current

Characteristics of Files Included in the STARS System, cont.

- Medicare Enrollment Database (EDB):
  - Current and historical Medicare enrollment
  - "Active" and "Inactive" cases
  - 35-40 million records at any one point in time; September '93: 77 million records (active + inactive)
  - Proxy recipients listed on the file (e.g., John Doe's benefits c/o Jane Doe; John Doe's benefits c/o nursing home)
  - A small portion of records at any point in time are probably deceased (Kim and Sater, 2000)
  - Used in population estimates system for 65+ household population estimates

Characteristics of Files Included in the STARS System, cont.

- Medicare EDB, cont.:
  - Recipient Address ≠ Housing Unit
    - Proxy recipients
  - Coverage is believed high (93-102%) but not perfect and unevenly distributed geographically
    - "Snowbird" states appear to have lower ratios of medicare to 65+ population than "non-snowbird" states

Characteristics of Files Included in the STARS System

- Indian Health Service patient file:
  - About 10 million patient/transaction records
  - Transaction record ≠ person record
  - Unduplication
    - about 10 million patient records, 2 million unduplicated SSN's
  - Many missing SSN's
    - about 20% missing SSN's

Characteristics of Files Included in the STARS System, cont.

- Housing and Urban Development Tenant Rental Assistance Certification System (HUD-TRACS):
  - HUD subsidy payments
  - Currently, about 3.3 million records
  - Short form data for all members of household (Race/Hispanic only for head of household)
  - Address information may represent project or landlord address

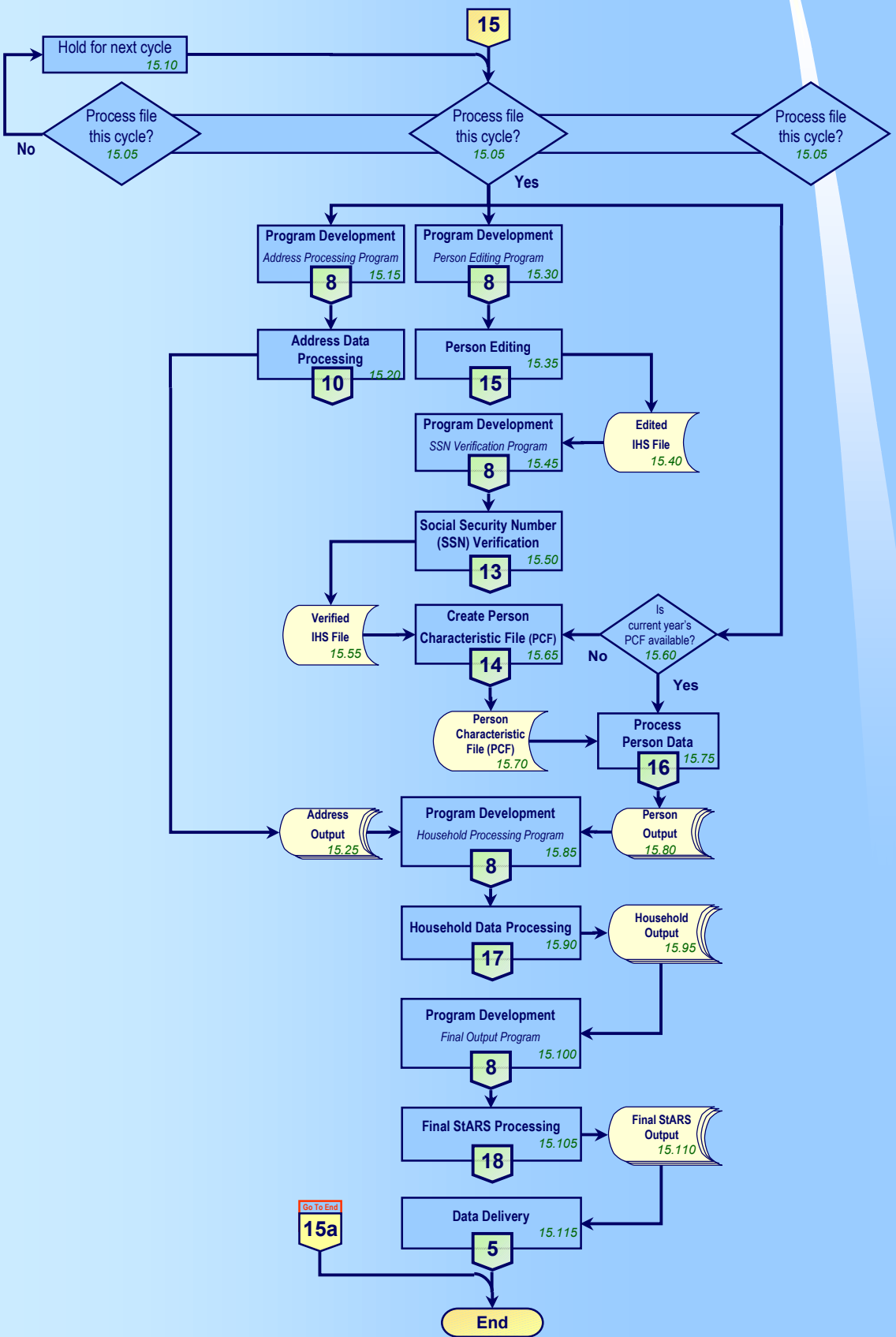
Characteristics of Files Included in the STARS System, cont.

- Census NUMIDENT File:
  - 750 million transaction records → 400 million individual SSN records
  - Post 1985: Enumeration at birth
  - For each SSN: Date of birth, gender, race, place of birth
- About 50-60 million persons on the file are deceased but not identified as such
- No current residence information on the file
- Taxpayer ID Numbers (TINs) not on the file
- About 35% of SSN's on file have alternate names (marriage, divorce, etc.)
- 6% missing gender
- Race coding has changed (prior to 1980, 3 races: White, Black, Other); 20% either "unknown" or "other"
- About 25% of SSN's have transactions with different race codes

STARS Processing Diagrams

- Two Goals:
  - For **person** data: One output record per person, assigned to an individual residence corresponding as closely as possible to Census residence definitions, in a household structure corresponding as closely as possible to Census household structure, containing microdata corresponding as closely as possible to Census short form microdata, and excluding persons which are not in the population of interest.
  - For **address** data: One output record per individual housing unit at a Basic Street Address, geocoded to Census TIGER geography, with address microdata and concepts corresponding as closely as possible to DMAF address fields and concepts, and excluding locations which are not in the population of interest.

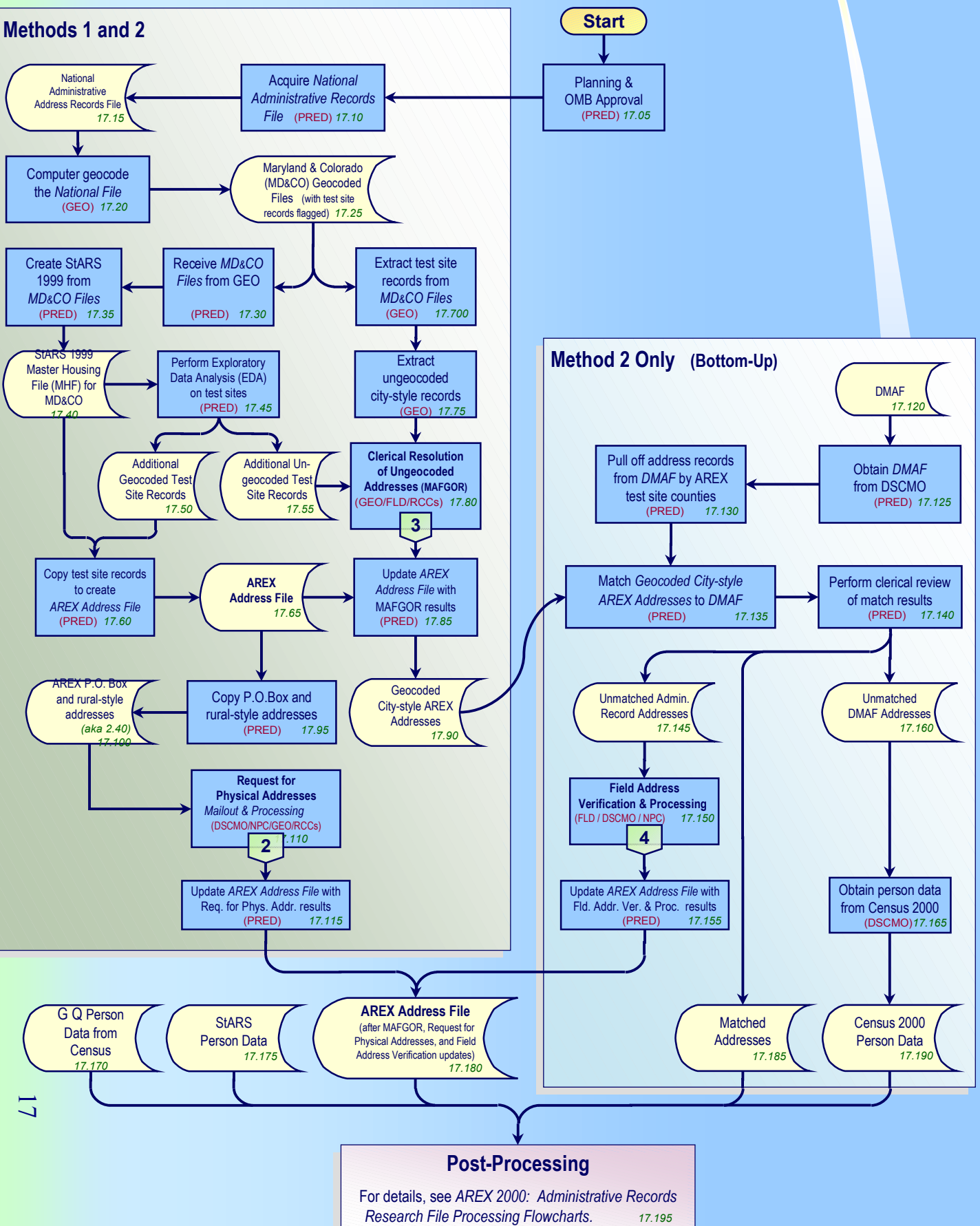
STARS Processing Overview



Administrative Records Experiment in 2000 (AREX 2000)

- Five selected sites in Maryland and Colorado
  - MD: Baltimore city, Baltimore county;
  - CO: El Paso county, Douglas county, Jefferson county
- Attempt to simulate an Administrative Records Census
- Not all aspects of an Administrative Records Census are simulated
  - Group Quarters survey
  - Coverage measurement survey
- Special operations not included in StARS
  - Request for physical address (PO boxes/RR's)
  - MAFGOR Geocoding
  - Field verification of addresses not matched to DMAF

AREX 2000 Overview Flowchart



AREX 2000 Evaluation Plans

- Evaluation 1: Comparison of both methods' site and block level counts of population by race, Hispanic origin, age groups and gender, with comparable decennial census counts
- Evaluation 2: Analyzing selected components of the AREX implementation processing
- Evaluation 3: Comparison of "bottom up" housing unit and household level information with comparable Census 2000 housing unit and household information
- Evaluation 4: Assessing the feasibility of using administrative records in lieu of a field interview to obtain data on nonresponding households

Major Analytic Issues with StARS Processing

- Ontologies
  - A delivery address suitable for receiving a payment check may not suffice for putting individuals at a street address
  - Difficult to distinguish individual units within the Basic Street Address
  - Race coding: Hispanic Origin is a separate race on NUMIDENT
  - Transaction data ≠ person data
  - How many names does a person have (and in what order)?
- Proxies – IRS & Medicare records
  - JOHN WILSON
  - C/O MARY SMITH
  - 1004 LAUREL LANE
  - ROCKMONT, MD 22345

  The address is for Mary Smith. John Wilson may or may not live there.

Major Analytic Issues with StARS Processing

- Addresses that are difficult to place on the ground
  - Huang and Kim, 2000: About 10 % of addresses are rural style
  - PO Boxes: 45% for IHS, 9.5% for Medicare, 7.5% for IRS 1040, 6.8% for SSS, 3.8% for IRS 1099, .4% for HUD-TRACS
  - Sater, 1995 IRS/CPS match: 86.5% of tax return cases had the same address as residence address, 94% coded to same county
    - John Smith
    - H&R BLOCK
    - P.O. BOX 12
    - GREENWAY, MD 29752
  - Addresses with both business and residential components
    - Dean H. Judson
    - JUDSON OLD GROWTH LOGGING & SPOTTED OWL EXTERMINATION SERVICES
    - 45850 BACKWOODS HIGHWAY
    - BOONDOCKS, OR 96432

Major Analytic Issues with StARS Processing, cont.

- Unduplication and matching
  - When addresses or personal characteristics are measured with substantial variation, it is often not obvious whether a particular pair of records represent a duplicate or not. Yet, with multiple files, unduplication decisions must be made.

CHUMS-enhanced IMH File						MAF				
A		Banana	St			1	Apple	St		
B	17	Banana	St			3	Apple	St	Apt	1
C	19	Banana	St	Apt	5	3	Apple	St	Apt	2
D	44	MLK, Jr.	Blvd			3	Apple	St	Apt	3
E	100	Route 4				3	Apple	St	Apt	4
F	7	Marie	Ln			7	Apple	St		
G		Wife Mrs. Smith				9	Apple	St		
H	5	Apple	St			#	Apple	St		
I	27	Apple	St			#	Martin Luther King, Jr.	Blvd		
J		Apple	St			#	Pennsylvania	Ave		
K	9999	Apple	St			7	Maria	Ln		
L	3	Apple	St	Apt	5					
M	1	Apple	St							
N	3	Apple	St	Apt	A					
O	3	Apple	St		ZZ					
P	3	Apple	St							
Q	3	Apple	St	Apt	1					

Major Analytic Issues with StARS Processing, cont.

Outcome of "CHUMS-enhanced IMH File" / MAF Match

MATCH					
Street	BSA	BSA+Unit		Possible Explanations	Example
NO	N/A	N/A	1	Street is not in MAF, either it was just missing or it's a new street	A,B,C
			2	Different, but valid representation of street name	D,E
			3	Misspelling of street name	F
			4	Erroneous street name	G
YES	NO	N/A	1	BSA is not in MAF, either it was just missing or it's a new BSA - There is a "hole" in MAF	H
			2	BSA is not in MAF, either it was just missing or it's a new BSA - A missing "street extension"	I
			3	Existing street with no incoming street number	J
			4	Erroneous street number	K
YES	YES	NO	1	Unit not in MAF, either it was just missing or it's a new unit	L
			2	Valid match - a BSA without separate units	M
			3	Different representation of a unit	N
			4	Erroneous unit information	O
			5	Missing unit information	P
YES	YES	YES	1	Valid match	Q

Major Analytic Issues with StARS Processing, cont.

- Variations in data from different sources
  - Huang and Kim, 2000: Of the 50% of SSN's found on multiple files,
    - about 1% have more than one gender recorded
    - about 32% have multiple addresses
    - about 2% have multiple races
- "Imputation" from the NUMIDENT
  - Many files have limited microdata. For those that are found on the NUMIDENT, we can "impute" microdata from the approximately equivalent NUMIDENT fields.

Major Analytic Issues with StARS Processing, cont.

- Changing information states
  - Distinct problem from "point in time" data collection
  - Information states change over time/over databases
    - Address information ages over time and varies over databases

SAM SMITH	SAM SMITH
BOX 2 RURAL ROUTE 37	486 MAIN STREET
WESTPORT, VA 32784	FAIRFIELD, VA 33412
(Dated 10/14/98 from Medicare)	(From TY97 IRS file, filed sometime in 1998)

    - Mortality information ages over time and varies over databases
    - One database provides information about the other, provided that matching can be performed
    - Data processing requires complex, and substantively important, decision logic at each step

References

- Bye, B (1998). Race and ethnicity modeling with SSA Numident Data: Interim report: File development and tabulations. Unpublished document available from the U.S. Bureau of the Census.
- Bryant, C. (1995). Comparing the LUCA address list to "local records." Paper presented at the 1995 State Data Center Meeting, San Francisco, CA, April 4, 1995.
- Czajka, J. (1999). Can we count on administrative records in future U.S. Censuses? Presentation at the Bureau of the Census, December 15, 1999.
- Huang, E., and Kim, J. (2000). One Percent Sample Study Report (SRD-DRAFT). Unpublished document available from the U.S. Bureau of the Census, February 10, 2000.
- Judson, D.H., and Popoff, C.L. (2000). Research Use of Administrative Records. Unpublished document.
- Judson, Dean H. (2000). The Statistical Administrative Records System: System Design, Successes, and Challenges. Unpublished document.
- Kim, Myoung Ouk, and Sater, Douglas (2000). Defining the Medicare Data Universe for the U.S. Census Bureau's Population Estimates Program. Paper presented at the Southern Demographic Association meetings, New Orleans, LA, August 29, 2000.
- Sater, D. (1995). Differences in Location of Households and Tax Filing Units. Paper presented at the 1995 meeting of the Population Association of America, San Francisco, CA, April 6, 1995.
- Wand, Yair, and Wang, Richard Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39: 86-95.
- Zanutto, E. (1996). Estimating a population roster from an incomplete census using mailback questionnaires, administrative records, and sampled nonresponse followup. Presentation to the U.S. Bureau of the Census, August 6, 1996.
- Zanutto, E., and Zaslavsky, A. (1999). Using Administrative Records to Impute for Nonresponse. Paper presented at the International Conference on Survey Nonresponse, Portland, OR., October 29, 1999.