Secure Data Query Service

2024 Consultants Panel
July 26, 2024

Amy O'Hara
Georgetown University

Secure Query System

- System linking clients, a data intermediary, and SOI, featuring
  - Data validation on client side
  - Administrative functions handled by intermediary
  - Automated matching process within SOI by SOI employees
  - Tabulation of pre-defined statistics
  - Automated disclosure avoidance review

Expected Clients

- Federal agencies, state agencies, local governments,and nonprofit organizations
  - Agree to terms
  - Agree to outputs
  - Prepare own data

Outreach

- State or Local Agencies
  - Education (K-12, Post-Secondary Education Institutions or Systems)
  - Workforce (Training and Employment Programs)
  - Health and Human Services (Crossover Youth, SNAP, TANF, Child Welfare)
  - Justice (Training, Reentry)
  - Housing (Housing Subsidy Recipients, Homeless/Transitional Housing Services)
- Non-Profit Service Providers
  - Training Programs
  - Randomized Controlled Trials
  - Intermediaries

User Requests

- Are their participants employed? What are they earning?
- Are they getting the EITC?
- What earnings are missing in state quarterly wages?
- How many out-of-state earners are missing in state quarterly wage data?
- Has LFP increased since intervention?
- Have wages increased since intervention?
- By how much have wages increased since intervention?
- Are trainees working in the industry they were trained in?
- What are career pathways? What are most valuable career pathways?
- Are they married now? Do they have children now?

Design Considerations

- Utility
- Privacy
- Efficiency

Utility: 1040 Match Output

Percent filed 1040 (1040 filers/total)	Percent with only Wage income
Filing status frequency (for 1040 filers)	Mean Total Wage Income on 1040
Percent claimed EITC (for 1040 filers)	25/50/75 Percentiles on Total Wage Income
Average EITC amount (for EITC>0)	Mean AGI (with standard deviation)
Percent claimed CTC (for 1040 filers)	25/50/75 Percentiles on AGI
Average CTC amount (for CTC>0)	Median AGI by Filestat
Percent with Schedule C	

Utility: Information Return Match Output

Percent with 1+ Form W-2 Percent with 1+ Form 1099-NEC Percent with both Form W-2 and 1099-NEC	Percent of matched persons with return address in same/different/unresolved origin state
Mean Total W-2	25/50/75 Percentiles on Total W-2
Mean Total 1099-NEC	25/50/75 Percentiles on Total 1099-NEC
Mean Total Earnings (W-2 + 1099-NEC)	25/50/75 Percentiles on Total Earnings (W-2 + 1099-NEC)

Privacy

- Cannot rerun same sample
- Disclosure Avoidance (DA)
  - Suppress statistics for small cells
  - Round percentages to two decimal places
  - Round dollars to nearest hundred
  - Present ranges instead of values
  - Noise injection if necessary
- Aim towards Privacy Preserving Record Linkage

Efficiency

- Administrative
  - Data validation on client side
  - Client TA and MOUs handled by intermediary
- Technical
  - CDW prep in advance
  - Modules written for current systems
    - Effective vs. efficient

Modules

- TIN Retrieval
- FTI Extraction
- Computation
- Tabulation
- Disclosure Avoidance

Data Validation

Identifier Columns

Social Security Number	Found: **SSN**	Missing or incomplete in **30** rows
Last Name	Found: **Last Name**	Missing or incomplete in **1** rows
First Name	Found: **First Name**	Missing or incomplete in **1** rows
Date of Birth		
Middle Name or Initial		
Suffix		
Street Number		
Street Name		
Unit Number		
PO Box		
City		
State		
ZIP Code		

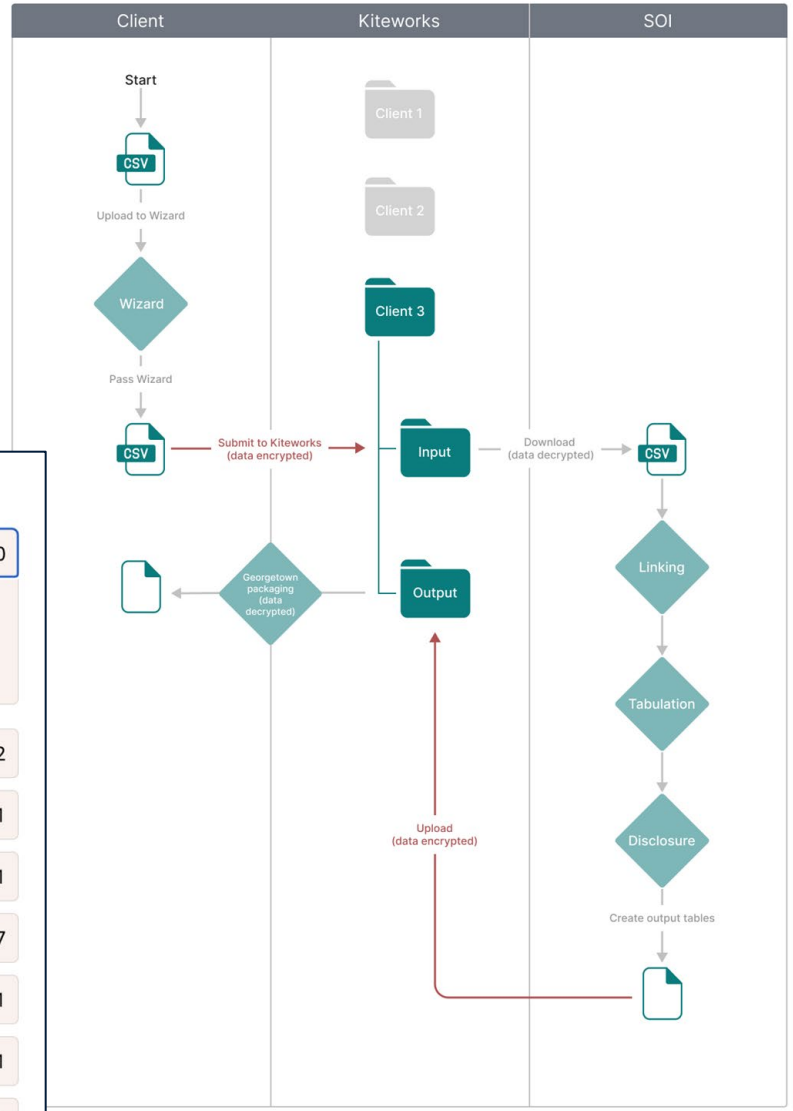
Population Grouping Columns

Sex	Fail: Smallest population is 10

Populations smallest to largest

ID: **<empty>**, Size: **10** | ID: **Female**, Size: **40** | ID: **Male**, Size: **50**

RACE	Fail: Smallest population is 2
LOCATION	Fail: Smallest population is 1
PROGRAM	Fail: Smallest population is 1
ETHNICITY	Fail: Smallest population is 27
MILITARY	Fail: Smallest population is 1
HIGHEST LEVEL ED	Fail: Smallest population is 1
NONCREDIT	Fail: Smallest population is 1



GEORGETOWN UNIVERSITY
McCourt School of Public Policy
MASSIVE DATA INSTITUTE

TIN Retrieval and FTI Extraction

- Match client identifiers to appropriate year of tax data
  - Exact on SSN
  - Probabilistic on Name and Address
- Extract 1040 or 1099/W2 data for TINs in client cohort

Computation and Tabulation

- Create flags and counts
  - E.g., filed EITC; sum of W-2s, sum of 1099-NECs, sum of both to get total earnings; count of W-2s and 1099-NECs
  - Entire cohort, and by group variables
- Produce output statistics
  - Retain the number of observations used to produce each statistic for DA

Client 1	1 In-State	2 Out of State	3 Discrepant
Total	*62%* 60-69%	*31%* 30-39%	*7%* 0-9%
Program Completed	*50%* 50-59%	*40%* 40-49%	*10%* 10-19%
Did Not Complete	*80%* 80-89%	*15%* 10-19%	*5%* 0-9%

	Mean AGI	Variance	AGI First Quartile	AGI Median	AGI Third Quartile
All	115,100	6,200	70,700	105,500	150,400
No Certificate	105,200	6,100	57,500	84,900	100,200
Certificate Completed	140,000	7,300	88,900	102,300	165,600

Next Steps

- Test minimum viable product for SQS-1040 and SQS-Info
- Improve matching and disclosure methods
- Consider additional SQS products (e.g., producing statistics before and after an intervention, lagged matches)
- Assess capacity building needs in state and local agencies

Questions and Comments?

Amy O’Hara
amy.ohara@georgetown.edu