# Portfolio Approach to Data Dissemination

Nick Greenia

IRS Statistics of Income Division

**The views expressed in this presentation are the presenter's and not necessarily those of the U.S. Internal Revenue Service.**

SOI Advisory Panel
November 6, 2009

# Presentation Outline

- Background – Legal requirements for confidentiality protection
- Underlying Tension in Data Dissemination – Factors driving demand
- Portfolio approach to access modalities
- Challenges & Tradeoffs
- Conclusion

# Confidentiality Protection is Legal Requirement

- SOI must maintain balance between dual roles: (1) statistical office and (2) data stewardship for administrative agency in our coordination (tax data extracts for external users) & advisory roles (Disclosure, Fed-State, Counsel, etc.)
- Voluntary compliance is backbone of tax system
- Protection of taxpayer confidentiality is a cornerstone of voluntary compliance
- Anything that threatens – **or is perceived to threaten** – taxpayer confidentiality is unacceptable and will not be tolerated

# Confidentiality (cont.)

- Unauthorized access (UNAX)
- Data Release Statute = IRC 6108(c) & 6103(j)(4)
  - Data must be anonymized before release so that data can't be associated, directly or indirectly, with a particular taxpayer (any taxpayer, not just those in a particular survey)
  - NOTE: Law says anonymous form is an absolute requirement, but methodology (how to do this!) is not mentioned

# Data Dissemination

- "Let not the motivation to be precise overtake the need to be relevant."
- Confidentiality protection is very important, but let's not forget that data dissemination is a core mandate of a statistical office.

# Data Dissemination (cont.)

- Different users need different accesses
- Resource constraints
- Confidentiality constraints
- New modalities require new techniques, technologies, attitudes – but not necessarily (or usually?) new legislation

# External Researcher Access Modalities re Need, Impact, Risk

- Disclosure-processed Macrodata (Tables)
  - Most of 6108a products (SOI's statistics mandate)
  - 6108b (Outside requests for special tables, studies)
- Disclosure-processed Microdata (PUF)
  - Confidentiality concerns re past methodologies, days numbered?
  - Synthetic data to rescue?
- Buffered remote access (black box)
  - Remote execution by user directly
  - Indirect remote access: User submits programs, provider runs
  - Sometimes "canned" datasets, programs
  - Disclosure review problem, scammable by savvy users, multiple runs – NCHS shut down

# Modalities (cont.)

- Virtual Data Enclave
  - Eric Bartelsman, disparate European datasets – **US Fed agency applications?**
  - Buffered remote access – big mkt for European researchers
- Gold Standard Microdata
  - Census Research Data Centers – review process per the IRS/Census Criteria Agreement (remote access?)
  - SOI's External Research Program (remote access some day?)
  - Data Enclave – menu of access protocols, smorgasbord of data, (virtual access option also)
    - NIST/NORC: underway, successful implementation
    - Proposed Federal Statistical Community of Practice (SCOP) – only in exploratory stage, far away from applications

# Other Agencies

- Outside requests (6108b): VA, NCHS (pending) for special tables, studies
- Partnering (6103n agents, 6108a-related), national economic imperative: FRB (SCF); BEA (Geo-coding); NSF (small business survey on innovation)

# Factors Driving More Data Dissemination (+ risk of unauthorized disclosure/access)

- **CIPSEA**: Proposed completion of Subtitle B (data sharing among BEA, BLS, Census requires tax law change) -- more agencies soon to follow?
- **Informed decisions** require more data to drive them
- Renewed **emphasis on transparency of decision-making**, [open]data.gov – get rid of the black boxes, provide more data access to citizenry

# Challenges & Tradeoffs

- Aggregate tables "easiest" to disclosure-protect, but least useful to many users
- Traditional PUF (top-coding, blurring) -- routinized methodology, but less useful for users, more liable to disclosure over time, days numbered?
- Virtual data enclave – untried in Fed system, agency skepticism, & besides, more focused on data sharing (especially tax data!!)
- Buffered remote access has promise, but which canned programs? Which canned data? Disclosure review?

# Challenges (cont.)

- Gold standard data most desirable by many users, but expensive to support RDC's -- and data enclaves -- below some access level, critical mass needed
- ----------
- Synthetic data seen by some as panacea – for both gold standard & PUF problems

# Some Concerns on Synthetic Data

- **Learning cost** for prospective researchers?
- "One stop shopping" file? or **will many synthetic data files be necessary** to accommodate many different research specialties?
  - Especially if the latter, opportunity cost for other access modalities?
  - If only partially synthetic, does disclosure risk increase?
- **Fidelity scores** (Jerry Reiter et al.?)
  - The better the utility score, the higher the release risk as synthetic data approach gold standard data?
  - Are the "best" synthetic data more appropriate for training?
- **Partially vs. Wholly synthetic** data
  - With fake data, why any disclosure concerns on the recent SIPP/SSA/IRS synthetic PUF? Only partially synthetic….
  - Does partially synthetic defeat the purpose of synthetic data's risk model attraction?

# On the other hand……

- Problems with all modalities
- Gold Standard Dilemma: Even with perfect safeguards & unlimited funding, access must be limited – credibility of confidentiality protection at risk; i.e., everyone cannot have access.
- Synthetic data – concerns, but promising for PUF, training, international data sharing?
- Portfolio approach: difficult & dynamic, requires innovation & constant re-tuning – but data dissemination is part of being a viable statistical office.