Statistical Disclosure Limitation at the U.S. Census Bureau and Beyond

Dr. Aref Dajani, Lead, Innovation and Review Group
Center for Enterprise Dissemination – Disclosure Avoidance
U.S. Census Bureau

Presented to the Transitioning from PAPI to CAPI Workshop
Sponsored by the International Population Center of the Population Division (POP/IPC)
September 23, 2022

Acknowledgments: Dr. Fabián Romero and Dr. Mitali Sen, POP/IPC

Outline of Presentation

- Values, ethics, and integrity
- What statistical disclosure limitation (“disclosure avoidance”) is and is not.
- How most national statistical offices protect data.
- The impact of computational efficiency and the advent of Big Data on both Computer Security and Disclosure Avoidance.
- The legal authority by which the Census Bureau disseminates data.
- The legal authority by which most of Europe disseminates data.
- Final comments moving forward.

Questions?

- At any point, feel free to direct questions to us in real time if there is anything you do not understand or you seek clarification.
- Please do not sit on your hands if you have questions.
  - Feel free to interrupt!! ☺

Values, Ethics, and Integrity (Slide 1 of 3)

- Values are agreed norms that allow teams to achieve a shared vision. Equivalent to Ground Rules.
  - *"The only inappropriate question is the one not asked."*
  - *"One person will speak at a time."*
  - *"Keep all electronic devices on mute unless you are speaking."*

Values, Ethics, and Integrity (Slide 2 of 3)

- Ethics are rules, regulations, and affirmations that allow individuals to work in accordance to established principles and guidelines.
- The following are documented ethics rules for the U.S. Census Bureau.
  - *"Employees shall put forth honest effort in the performance of their duties."*
  - *"Employees shall act impartially and not give preferential treatment to any private organization or individual."*
  - *"Employees shall not engage in outside employment or activities, including seeking or negotiating for employment, that conflict with official Government duties and responsibilities."*

Values, Ethics, and Integrity (Slide 3 of 3)

- Integrity is being compliant with established values and ethics within an organization.
  - Integrity is doing the right thing when nobody is looking.
  - Organizations that focus on success must not do so at the expense of integrity to attain or maintain public trust.

What are Data Ethics? (Slide 1 of 2)

- Data Ethics describe a code of behavior, specifically what is right and wrong, encompassing the following:
  - **Data Handling**: generation, recording, curation, processing, wrangling, dissemination, sharing, and use.
  - **Algorithms:** AI, artificial agents, machine learning, and robots.
  - **Corresponding Practices**: responsible innovation, programming, hacking, and professional codes.

What are Data Ethics? (Slide 2 of 2)

- Developing a code of behavior is not trivial and should be established very early in the product lifecycle.
- Like quality assurance and systems development, data ethics and data stewardship must not be an afterthought.
- Often, time, budget, and resources compete with the necessity to be compliant with data ethics.
- Not being thorough to meet project deadlines comes at a cost.

What is Data Stewardship?

- According to the Data Governance Institute:

  "Data Stewardship is concerned with taking care of data assets that do not belong to the stewards themselves. Data Stewards represent the concerns of others. Some may represent the needs of the entire organization."

Examples of Data Stewardship Activities

- Documenting rules and standards.
- Managing Data Quality issues.
- Executing operational Data Governance activities.
- Setting and managing guidelines around data.

Two Components to Data Stewardship at the U.S. Census Bureau

- Protecting data coming in from the field.
- Protecting data going out to our stakeholders.

Who Are Responsible for Disclosure Review?

- All of us, wherever we live, wherever we work.
- Everyone who releases information products outside their agency's firewall.
- We must protect sensitive and confidential data.
  - Must have a need to know to access sensitive and confidential data.
  - This is true even if we have access to "juicy data".
  - In the United States, no disclosure of sensitive or confidential data is allowed for life.

Email, Shared Network Drives, and Disclosure

- Email: sensitive or confidential data must be…
  - …encrypted as a password-protected file in an unencrypted email with the password sent separately in a subsequent email
  - …encrypted as an ordinary file in an encrypted email
  - *These are best practices implemented at the U.S. Census Bureau.*

- Shared Network Drives and the Internet: Upload carefully.
  - It is often just as easy to upload data to an external-facing servers as it is to an internally-facing server.
  - Uploading to an external server without authorization gets disseminators in a lot of trouble!!

What Disclosure Avoidance Is And Is Not!

- What is not Disclosure Avoidance?
  - Illegal breaches (“hacks”, intrusions) into computers
  - Data-free information products
    - Data can be quantitative or qualitative.
  - Release of an information product based purely on data that have already been released.

- What is Disclosure Avoidance?
  - Responsible release of quantitative and qualitative data

Where Do Computer Security and Disclosure Avoidance Intersect?

- With few exceptions, all internal data should be considered confidential until they have undergone disclosure review and have been approved for release.
- This includes reports, tables, graphs, and microdata.
- Use/set up secure network drives to minimize the need to email confidential data.
  - For most people, having to email confidential data (encrypted, of course) should be a rare occurrence.

The Evidence Act in the United States (Evidence-Based Policymaking Act of 2018)

- Federal agencies must have a "systematic plan for identifying and addressing policy questions relevant to the programs, policies, and regulations of the agency."
- They must have evaluation plans to identify and address these questions.
- They must establish a Statistical Official, an Information Officer, a Privacy Officer, Data Officers, and Evaluation Officers.
- Data must be open by default.
- Data must also be protected against inappropriate access and use.

Types of Information Products Released from National Statistical Agencies, including the U.S. Census Bureau (Slide 1 of 2)

- Tables and Other Aggregates of Direct Estimates
- Frequency Count Data
- Magnitude Data
- Public Use Microdata
- Graphs and Maps

Types of Information Products Released from National Statistical Agencies, including the U.S. Census Bureau (Slide 2 of 2)

- Model-based Output
- Reports
- Metadata
- Paradata
- *…and much, much more! (any information product derived from quantitative or qualitative data)*

This is how most national statistical offices protect data.

- **Information Reduction**
  - Suppressing or consolidating information that may be disclosive.
  - Examples are individual cell suppression or collapsing whole categories in tables.

- **Data Perturbation**
  - Altering values of information that may be disclosive.
  - Examples are data swapping of individual observations or injecting random noise in all observations.

What Are Microdata?

- At the Census Bureau, microdata refer to collected data that have been cleaned, edited, and sometimes imputed so that they can be used to produce statistical tabulations and analyses.
- These data are often released at the respondent record-level, as opposed to aggregate counts or magnitudes.
- Each record represents one respondent, such as a person, a household, or an establishment, and consists of values of characteristic variables for this respondent.

Impact of Computational Efficiency and the Advent of Big Data on Microdata

- Computational efficiency makes it easier to reconstruct protected microdata records from published tables generated from protected microdata.
  - Publishing too many tables is the problem. Better algorithms make database reconstruction easier. As a result, data users may have to get accustomed to receiving far fewer tables derived from internal use data.
- The advent of big data makes it easier to link public use microdata to external data, thereby re-identifying respondents whose identity we are trying to protect.
  - The Census Bureau acquires external data that are relatively easy and legal to obtain. With access to additional data, especially Big Data, it is even easier to link public use data to external files.

The Five Safes (Developed in the United Kingdom)

1. **Safe Projects**: Is the use of these data appropriate?
2. **Safe People**: Can the researchers be trusted to use the data in an appropriate manner?
3. **Safe Data**: Is there disclosure risk in the data themselves?
4. **Safe Settings**: Does the access facilitate limit unauthorized use?
5. **Safe Output**: Are the statistical results non-disclosive?

Information products at the U.S. Census Bureau fall under various legal authorities.

- Title 5 of the U.S. Legal Code mandates the dissemination of demographic information about Census Bureau employees and field representatives
- Title 13 mandates the dissemination of data collected by the Census Bureau from our respondents, whether people, households, or business establishments.
- Title 26 mandates the dissemination of tax data collected by the U.S. Internal Revenue Service
- Other laws mandate the dissemination of data collected from other government and non-government sources.
- Commingled information products have mixtures of multiple types of data. The laws of each type need to be followed with commingled data.

Information products fall under the General Data Protection Regulations (GDPR) through most of Europe and beyond.

- The GDPR passed the European Parliament in 2016.
- It was fully implemented in 2018.
- Applies to any personal data for people living in countries in the European Union (EU) or the European Economic Area (EEA).
- Their website provides all the information you will need: **gdpr.eu**

GDPR and non-GDPR nations in Europe.

- GDPR nations: Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, and the United Kingdom.
- Non-GDPR nations in Europe: Albania, Belarus, Bosnia and Herzegovina, Croatia, Kosovo, Moldova, Montenegro, North Macedonia, Russia, Serbia, Turkey, Ukraine, Georgia, and Armenia
- Non-GDPR nations that use data from GDPR nations are bound by all regulations in the GDPR when using those data.

Data Protection Principles in the GDPR

- Lawfulness, fairness, and transparency.
- Purpose limitation.
- Data minimization.
- Accuracy.
- Storage limitation.
- Integrity and confidentiality.
- Accountability.

People's Privacy Rights in the GDPR

1. The right to be informed
2. The right of access
3. The right to rectification
4. The right to erasure
5. The right to restrict processing
6. The right to data portability
7. The right to object
8. Rights in relation to automated decision making and profiling.

Final comments on the GDPR

- The GDPR mandates stronger data protection than regulations found in the United States.
- All of us should more about the GDPR, as their regulations are far reaching and impressive.

Final remarks

- It is not possible to responsibly disseminate information products from any survey or census with no disclosure risk.
  - This is true whether using legacy data protection methods or newly implemented data protection methods.
  - This is also true no matter how strong the regulations are.
  - We cannot completely avoid disclosure.
  - *We can only do the best we can.*

Questions? Comments?

გმადლობთ

շնորհակալություն

Дякую тобі

mulțumesc

Aref Dajani

aref.n.dajani@census.gov