

Administrative Records and Third Party Data Use in the 2020 Census Working Group

SUMMARY: The purpose of this Working Group was to review the Census Bureau’s research program and plans involving administrative records and third party data use in the 2020 Census. The Working Group reviewed interim findings from 2020 Research and Testing projects using administrative records and commercial data, focusing on person and housing unit coverage by demographic characteristics. The Working Group discussed privacy, confidentiality, and consent issues, and provided feedback on research and plans to explore public attitudes on administrative records and commercial data uses in this document and through NAC presentations. We identified topics for further research.

This report contains the following sections:

1. Issue
2. Process
3. Sources And Materials
4. Recommendations And Findings

1. ISSUE

The Census Bureau is committed to designing and conducting a 2020 Census that costs less per housing unit than the 2010 Census, while maintaining high quality results. A major cost driver for the 2010 Census involved collecting information from housing units that did not respond to enumeration attempts. The Non-Response Follow-Up (NRFU) operations sent enumerators to knock on doors up to six times. To reduce costs for the 2020 Census, the Census Bureau is investigating the strategic reuse of federal, state, and private data sources.

From Working Group Charter (1/23/2013)

To reduce costs for the 2020 Census while maintaining quality, the Census Bureau is investigating the strategic reuse of administrative records and private data sources. Administrative data refers to any information collected by Federal or state agencies for the purpose of administering programs or providing services. Private, or commercial, data refer to information collected by third parties, which were acquired by the Census Bureau. The purpose of this subcommittee of the National Advisory Committee is to explore privacy, confidentiality, and consent issues, as well as provide feedback on research and plans to explore public attitudes on Administrative Records and Third Party Data (ARTPD) uses.

Administrative Records and Third Party Data (ARTPD) can be used in different parts of the census/survey lifecycle. This document describes the spectrum of uses from frame and contact development, through data collection and processing, and post-collection analysis. Examples of current and possible ARTPD sources are included in this document.

Frame and Contact. ARTPD can be used with the Master Address File (MAF),¹ or can be used in lieu of the MAF, in development of the survey frame.² ARTPD addresses can be compared to the MAF to add units, update address information, or to validate information in the MAF.

- Administrative records could be used to validate and improve address ranges.
- ARTPD can be used for sampling to augment Title 13-based samples or provide non-Title 13 addresses
- ARTPD content can be appended to the MAF (or commercial data) to target units when specific samples are needed.
- ARTPD contact information, such as telephone numbers and email addresses, can form a non-address frame or can be appended to an address frame.

Data Collection and Processing

- During data collection, record linkage could provide real-time association to improve dependent interviewing or target coverage/edit probes.
- After data collection, ARTPD can be used in edit and imputation. This may include replacing misreported or missing items directly or through modeling.
- With sufficient ARTPD coverage, ARTPD could be used to replace data collection allowing items to be removed from the form.

Post-Collection. ARTPD can be used to investigate bias resulting from sampling or non-responding units through comparison to unit level or area level characteristics observed across files.

From Census ARTPD WG Summary (for outside experts)

2. PROCESS

Working Group Focus

Examine Research Plans and Results - Assess completed research and research plans to use administrative records and third party data for the 2020 Census; identify omissions and propose topics for future analysis. Consider criteria that will be used to select administrative records and third party data sources, and the impact of administrative records use on demographic and geographic disparities.

Awareness of Proposed Administrative Records Uses – Consider attitudes and actions that may result from administrative records and third party data use; consider impact of public opinion on strategies and implementation. Assess privacy concerns, including the public's ability or willingness to provide informed consent. Discuss costs and benefits of using federal, state and commercial lists.

¹ The Master Address File (MAF) contains an accurate, up to date inventory of all known living quarters in the United States, Puerto Rico and island areas. The MAF is used to support most of the census and survey operations that the Census Bureau conducts, including the decennial census, American Community Survey, and ongoing demographic surveys. The content of the MAF includes address information, Census geographic location codes, as well as source and history data.

² A survey frame refers to the list of units comprising the universe from which a sample is drawn. An example of a survey frame is the list of all addresses in the U.S., which is then used for census form delivery.

Acceptance of Administrative Records Uses – Assist the Census Bureau in developing strategies and options to educate the public about the use of administrative records and third party data to increase public trust.

Working Group Members

NAC Members	Census Bureau
Barry Steinhardt. Friends of Privacy USA.	Amy O’Hara, Center for Administrative Records Research & Applications (CARRA)
Eric Hamako, Program Coordinator for Institutional Diversity & Equity, Smith College	Catherine Massey, CARRA
Mary McGehee. Section Chief, Survey Unit Health Statistics Branch, Center for Public Health Practice Arkansas Department of Health Little Rock, AR	Howard R. Hogan, Chief Demographer, Office of Director
Ditas Katague. Chief of Staff California Public Utilities Commission	Jennifer Hunter Childs, Center for Survey Measurement
Kirsten Martin. Assistant Professor School of Business Administration The George Washington University	Nancy Bates, Senior Researcher for Survey Methodology
Wei Li. Professor Asian Pacific American Studies Arizona State University	Sonya Rastogi, Census Applications, CARRA
Randall Akee. Assistant Professor Luskin School of Public Affairs University of California Los Angeles	Kimberly Collier, Office of External Engagement
	Jeri Green, Office of External Engagement
	Tom Loo, Office of External Engagement

3. SOURCES AND MATERIALS**a. Reports within Census and Working Group**

Title	Author/Date	Source	Summary
2020 Census Privacy and Confidentiality Study Plan	Jennifer Hunter Childs and Nancy Bates (2012)	Census Bureau. Email 1/31/2013	
2010 Census Match Study Report	Sonya Rastogi and Amy O'Hara	Census Bureau. Email 1/31/2013	This study evaluates the administrative data and the 2010 Census at different levels of geography and by factors such as Hispanic origin, race, and mode of data collection. This report also evaluates the quality and coverage of Hispanic origin, race, sex, and age response data in administrative records relative to the 2010 Census.
Gallup Poll:	Nancy Bates and Jennifer Childs	Census Bureau Call. Email 2/19/2013.	
Development of the Federal Statistical System Public Opinion Survey	Childs, J., Wilson, S., Martinez, S.W., Rasmussen, L. and Wroblewski, M. (2012).	Census Bureau. Email 1/31/2013	
Development of the Federal Statistical System Public Opinion Survey	Jennifer Hunter Childs, Steaphanie Willson, Shelly Wilkie Martinez, Laura Rasmussen, Monica Wroblewski	Census Bureau, National Center for Health Statistics, Office of Management and Budget, Internal Revenue Service	This study looks at trust in the federal statistical system, the credibility of federal statistics, and attitudes toward and knowledge of statistical uses of administrative records.
Quality Criteria Checklist (QCC)	The Quality Criteria Group (Berning, Brown, Konicki, O'Hara, Sheppard, Noon)	Census Bureau. Email 3/25/2013	Can be used for any administrative records or third party data (ARTPD) file, for any specified statistical use.
NAC ARTPD WG – List of Administrative Records and Third Party Data (May 2013)	Amy O'Hara	Census Bureau	List of administrative records being used by the Census Bureau.
Presentation. Update on the Use of Administrative Records During Non-Response Followup in the	Kevin Deardorff, Decennial Management Division	Census Bureau. Email 5/17/2013	Presented to the National Academy of Sciences Panel to Review the 2010 Census

2020 Census			
NAC Meeting Presentation	ARTPD WG	NAC March 2013 Meeting	
NAC Meeting Presentation	ARTPD WG	NAC December 2013 Meeting	
Analyzing Data Sets: The ethics of using Big Data	Kirsten Martin, <i>George Washington University</i>	ARTPD WG. May 2013 Email	Discusses issues surrounding the use of Big Data and the concerns that should be addressed when considering their use.
FTC Warns data brokers on privacy rules		Craig Timberg <u>Washington Post 5/7/2013</u>	
Census ARTPD WG Background	Amy O'Hara	Census Bureau	Description of uses of administrative records and list of administrative records acquired and data the Census would like to obtain.
Attitudes Towards the Use of Administrative Records	Ryan King, Jennifer Hunter Childs, Monica Wroblewski, Darby Miller Steiger	Census Bureau, Westat Email 10/22/2013 Call: 11/8/2013	Summarizes Gallop polling specifically aimed towards assessing public opinion of the use of administrative records for statistical purposes.
2020 Research Report and Call	Tom Mule	Census Bureau Email Call: 2/5/2014	Discussed Decennial Nonresponse Followup (NRFU) Operations and the results from an initial application of administrative records usage for NRFU.
Race and Ethnicity and ARTPD	Sonya Rastogi	Call 1/15/2014	Comparison of race responses in the 2010 Census and administrative records show that they are largely in agreement for most single races (white, black, Asian), but are less successful for AIAN, NHPI, and multiple races.

b. Interviews/Meetings with outside experts

Call/Meeting	Attendees / Background	Summary
Data Quality Subgroup (11-19-13)	<ol style="list-style-type: none"> David A Swanson, PhD – Expert, Professor of Sociology, University of California, Riverside Ron Prevost, PhD – Expert, Former employee of U.S. Census Bureau Population Division 	Included in document Data_Quality_Expert_Call_11-19-13
Privacy Subgroup Calls	<p>David Vladeck -- the former head of the FTC's Consumer Protection Division</p> <p>Ed Mierzwinski – PIRG</p> <p>Bob Gellman – independent Privacy consultant</p>	
NAC Panel Meeting (12/2013)	Cavan Capps, Ph.D., Christa D Jones, Sonya Rastogi, Ph.D., Gina K. Walejko, Ph.D., Kevin Deardorff,	
2/14/2014 Meeting at Privacy Working Group, Washington DC.	February 14, 2014, ITIF in Washington DC. Privacy experts from academia, policy, regulatory, NGOs, etc.	

c. Books, papers, articles

Privacy References/Sources	Notes
Etzioni, 2012. "The Privacy Merchants. <i>Jnl of Constitutional Law</i>	
Ohm, P. 2012. The Fourth Amendment in a World Without Privacy. <i>Mississippi Law Journal</i> , 81(5), 1309.	
Slobogin, C. (2008). Government data mining and the fourth amendment. <i>The University of Chicago Law Review</i>	
Tene, O., & Polonetsky, J. (2012). Privacy in the age of big data: a time for big decisions. <i>Stanford Law Review Online</i> , 64, 63.	
Tene, O., & Polonetsky, J. (2012). Big Data for All: Privacy and User Control in the Age of Analytics. <i>Northwestern Journal of Technology and Intellectual Property</i> , Forthcoming.	
Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. <i>Information, Communication & Society</i> , 15(5), 662-679.	
Van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. <i>Ethics and Information Technology</i> , 6(2), 129-140.	
Ambrose, M. (2012). You are what Google says you are: The Right to be Forgotten and Information Stewardship. <i>International Review of Information Ethics</i> , 17.	
Nissenbaum, H. (2011). A contextual approach to privacy online. <i>Daedalus</i> , 140(4), 32-48.	
Nissenbaum, H. (2004). Privacy as contextual integrity. <i>Washington Law Review</i> , 79(1).	
Hartzog, W. (2012). Chain-Link Confidentiality. <i>Georgia Law Review</i> , 46.	
Mierzwinski, Ed, and Jeff Chester. "Selling Consumers Not Lists: The New World of Digital Decision-Making and the Role of the Fair Credit Reporting Act." <i>SUFFOLK UNIVERSITY LAW REVIEW</i> 46 (2012): 845.	
Data Quality References/Sources	
<p data-bbox="154 1165 170 1186">s</p> <p data-bbox="203 1165 544 1186">Sources provided by Swanson:</p> <ol data-bbox="203 1197 1153 1522" style="list-style-type: none"> <li data-bbox="203 1197 1153 1260">1. On Estimating a De Facto Population and Its Components in <i>Review of Economics & Finance</i>, June 30, 2011 <li data-bbox="203 1291 1153 1354">2. Link to school enrollment by gender, grade, race, ethnicity, and low SES at http://dq.cde.ca.gov/dataquest/dataquest.asp <li data-bbox="203 1386 1153 1522">3. Link to California Driver License Address Change (DLAP) which the state uses to estimate internal migration and migration into California from other states and out of California to other states (http://www.dof.ca.gov/Research/demographic/reports/estimates/e-1/view.php) 	<p data-bbox="1177 1165 1404 1543">Paper published by Dave Swanson and Jeff Tayman that describes methodology for estimating a population for which there is no readily accessible census data for size, composition, and distribution.</p> <p data-bbox="1177 1575 1404 1848">Examples of demographic data available from the California Department of Education California driver license data used to estimates migration.</p>
1.	

4. RECOMMENDATIONS AND FINDINGS

A. Definition of Terms

The Working Group uses a handful of terms throughout this report that require definition.

A. When we use the term “Administrative Records” we are referring to government records executive branch agencies or state and local authorities.

“Third-party data” is exactly that. It is data obtain from private sources usually commercial.

There are times, however, when data obtained from a governmental source is actually third-party data. This happens when an agency purchases commercial data and incorporates it, in whole or in part, in to a data set that it transfers to the CB.

B. The WG’s focus was on “Personally Identifiable Information (PII)”

The National Institute for Standards and Technology (NIST) defines PII as "any information about an individual... including

(1) any information that can be used to distinguish or trace an individual’s identity ... and

(2) any other information that is linked or linkable to an individual ...

From Special Publication 800-122

There are many cases of non-PII – usually aggregate statistical data –that are used by divisions of CB e.g. in the creation of the Economic Census.

The collection and use of that data has little to no privacy implications. Our findings and recommendations do not include non-PII data.

B. Key Findings**Summary Table –Of Findings Below**

- Data sets – AR & TPD – vary in how data is collected, maintained and used.
- Data stewardship practices of some third-party data sources are unknown.
- Some data sets will contain information that was surreptitiously gathered.
- Data quality of TPD is more problematic than Administrative Records.
- Both AR & TPD can be racially or ethnically skewed. For the most part these data sets tend to over count the white and economically advantaged populations.

The use of data sets or Big Data – the aggregation and analysis of large data sets in order to identify both trends and personally identifiable data – is becoming more popular in commercial and government sectors. However, not all data sources are equal. Data sets should be analyzed to determine not only if the data is accurate and if the knowledge is fit for use, but also if the data stewardship practices of the data source match that of the organization, and if the use of the data source has any harmful consequences. While the Bureau has been conscientiously analyzing how data is disseminated and used once gathered by the Bureau, the turn to analyzing how data is gathered by alternative sources is less established. The goal of this section is to outline guiding principles, questions and our findings for analyzing data sets – such as administrative records and third party data – for use in statistical analysis.

In our exploration of these issues we learned that:

- The currently available and/or tested government administrative records (AR) exacerbate racialized disparities in the quality of data available to the Census Bureau. Such racialized disparities may be attributed to both “coverage” issues and “response” issues in the AR databases’ quality. Such administrative records databases better “cover” the White population than racial minority populations and are also more likely to produce cross-database response agreement for the White population than for racial minority populations. Such problems are particularly pronounced for the American Indian and Alaska Native (AIAN), Native Hawaiian and Other Pacific Islander (NHPI), and Two Or More Races (TOMR) populations; for such relatively small populations, data quality issues may cause even more pronounced distortions than for larger populations.
- A particular problem is that there are low matching ratios for AIAN alone, NHPI alone, some other race, and two or more races using administrative record data. NB: matching ratios need to be explained.

- Unknown data stewardship practices of some third-party data sources. Just because the data is available in a data set, doesn't mean the information was knowingly disclosed by an individual. Data sets – both administrative and commercial – vary in how the data is collected, how it is maintained, who has access to the data, and how the data is analyzed. Individuals frequently do not knowingly disclose data online, and some data sets will contain information that was surreptitiously gathered.
- The information in ARTPD may not be meant for the Census Bureau. Individuals disclose information within a particular purpose or context with rules in mind at the time of disclosure. When interacting with a firm, a website, another individual, individuals reveal information with an understanding as to who can see that information, how it might be used, and the context in which it is revealed. Disclosure of information is not synonymous with information being public – disclosure is done within expectations of privacy.
- The Privacy Act gives the CB broad authority to request data from other government sources. Those agencies are not required to comply.

In fact, other statutes may prohibit disclosure. Federal agencies like the CB must publically disclose the use or transfer of ARTPD as required by the Privacy Act. In some manner the Bureau will likely be required to disclose the use of ARTPD data sets and that there will be more than just a public relations issue that might affect respondent cooperation. The use of these new data sets has the potential to light a firestorm from privacy groups and members of Congress concerned about privacy

**D. RECOMMENDATIONS VOTED ON BY NATIONAL ADVISORY COMMITTEE
March 2014 Meeting**

- 1. Data sources should be distinguished as either AR or TPD in CB polling, communications, reports, etc.**

WG Explanation: Data sources should be identified as either AR or TPD. Findings should be generalized to a specific type of record and the term 'administrative records' should not be used to mean third-party data.

- 2. Any ARTPD under consideration by CB for decennial census should assess privacy and inclusiveness issues.**

WG Explanation: If a particular data set does not significantly aid the Bureau to accomplish a statutory obligation, it should not be used and there is no need to even consider the privacy issues.

- 3. CB should consider public perception of ARTPD sources in making its decision about how to use ARTPD and develop public messaging strategies on the use of ARTPD**

- 4. TPD containing PII should not be used if it was obtained surreptitiously or by "scraping" the web and there should be a presumption against the use of TPD containing PII. That presumption can only be overcome by the CB Staff advocating for its use, when a rigorous analysis similar to that outlined in the attached "decision" tree, clearly and unambiguously shows that data cannot be obtained by some other means and when the benefits of use strongly outweigh the negative consequences.**

WG Explanation: The NAC asked Martin and Steinhardt to modify the recommendation with language that was mutually agreeable. This is the revised recommendation that was approved by the full WG.

- 5. Continue Gallup polling. But anticipate reaction of those with outsized influence**

WG Explanation I: The CB should continue Gallup polling of the general population. But must anticipate the reaction of those of great influence such as members of Congress, the press, and the organized privacy community. CB may find it useful to brief relevant groups on issues that has or could draw their attention.

WG Explanation II: The CB should continue using Gallop polling to identify general trust in using ARTPD. Better polling questions include very specific data sources, use of the data within Census, and specific alternatives to the use of the data source. For example, asking if the use of medical records from a health exchange for NRFU is better

than asking neighbors. This includes the specific files, the possible use within Census, and the alternative (asking neighbors).

6. ARTPD needs to be rigorously analyzed considering:

1. Fit of reputation and data stewardship practices aligned with those of the CB.
2. How data was collected
3. Costs – such as acquisition of ARTPD, data quality and coverage, loss of reputation, loss of public trust
4. Benefits – such as cost savings, improvement in quality and coverage

- See Decision Tree for Analyzing ARTPD Below

7. CB should seek to acquire ARTPD for HTR/HTC populations that better covers those groups.

WG Explanation: The Census Bureau should target HTR/HTC populations through datasets with better coverage of those groups – e.g., WIC and SNAP are possibilities among the many which should be considered. Time and money required to acquire such data sets would, we believe, offset the costs and inaccuracies of NRFU for these groups as well as populating the MAF. One of the best sources for this data might be states involved in the Federal-State Cooperative Program for Population Estimates that large numbers of the populations of interest

8. CB should look at ARTPD coverage of the HTC by area to determine if more targeted efforts are needed in certain parts of the country

9. CB should consider using statistical methods on ARTPD for data quality and coverage (e.g. imputation for missing data).

10. Continued outreach and discussions with external data quality and privacy stakeholders.

11. NAC should have a continuing advisory role as the CB continues to test and then integrate ARTPD for the 2020 Census.

E. MOVING FORWARD/ THE *REKPOLITK* OF ARTPD

§The CB needs to tread very carefully in its use of ARTPD.

§The CB could face a public relations issue that lowers respondent cooperation.

§The use of ARTPD – especially TPD it purchases – could ignite a firestorm from privacy groups and members of Congress

APPENDIX 1 – A DECISION TREE FOR ANALYZING DATA SETS

Questions around data sets can be viewed in three phases: concerns around the data set itself, questions around how to transition the data into the organization, and issues around how the data will be used and retained when in the target organization.

ABOUT ARTPD ORGANIZATION	
***What are the information stewardship practices of the data source? Are they consistent with the organization's?	***What groups are under represented or over represented by this data source?
***Does the data sources reputation complement the organizations? Is this a worthy partner?	***What information 'supply chain' is the Census Bureau joining?
ABOUT ARTPD DATA	
1. About Data Set	2. About Transitioning Information
Disclosure – How data is initial gathered...	What is the minimum amount of data necessary for knowledge needed?
Who gathered the data from the individual (if not the ARTPD or data source)?	What can be done to clean the data before acquisition?
How did the data source access the data? What practices were used to gather from the individual?	Is the level of analysis/detail consistent with the context and the expectations at disclosure?
**Was the individual notified of the policies? Was there consent? (or Fair Information Practices – FIP ³)	Is the data being combined with new data in surprising ways? Is new knowledge created?
What part of the initial disclosure by the individual was expected to be known broadly?	Does knowledge gained match expectations at disclosure?
What was the context of the initial disclosure? What was the purpose of sharing the data? Does it include the use proposed?	What are the alternative mechanisms to achieve this knowledge? What (vulnerable) groups will no longer be treated with this alternative?
After Disclosure	3. About Information Use
What biases exist in the access of the data? Does access favor or target particular groups? Is the targeting warranted?	What knowledge do you hope to gain? What is the immediate use for the data (MAF, NRFU, etc)?
How old is this data? Was the data expected to be retained this long?	What biases exist in the use of the data (vulnerable populations targeted by alternative tactics?)
Is the data source in a protected or sensitive context (e.g., medical, financial, educational)?	What are the reporting requirements for this data source or this data set? What information do we want to 'keep' when it is reportable to individuals?

³ Fair Information Practices (FIP) are seen by some to assuage privacy concerns online and include: 1) Notice/Awareness; (2) Choice/Consent; (3) Access/Participation; (4) Integrity/Security; and (5) Enforcement/Redress. More recently, see the White House report <http://www.whitehouse.gov/sites/default/files/privacy-final.pdf>.

Do individuals know that their information is in this data set?	What is the retention policy for this data?
---	---