NSF–Census Research Network (NCRN)

Spring 2014 Meeting

May 22–23, 2014
Census Headquarters, Washington, DC

(the most up-to-date meeting agenda can always be found at http://www.ncrn.info/event/ncrn-meeting-spring-2014)

This version last updated 2014-05-06.

Thursday, May 22, 2014

9:00-9:30 Opening remarks by the Director of the Census Bureau, John Thompson [Conference Rooms 1-3]

9:30-12:00 Research Session I: **Data Documentation Initiative (DDI) Metadata within the Federal Statistical System: Implementation Challenges and Provenance Encoding** *(Organizer: William Block, Cornell)* [Conference Rooms 1-3] (each paper 30 minutes)

- 9:30-10:00 Jay Greenfield and Sophia Kuan (Booz Allen Hamilton): "Describing Adaptive/Responsive Protocols and Data Processing in DDI"
- 10:00-10:30 Tim Mulcahy and Michael Davern (NORC): DDI and Record Linkage
- 10:30-11:00 Bill Block and Jeremy Williams (Cornell University): Integrating PROV with DDI: Mechanisms of Data Discovery within the US Census Bureau
- 11:00-11:30 Abdul K. Rahim and Pascal Heus (Metadata Technologies North America) "Using Metadata Standards and Technology to Improve Accessibility, Opening and Reuse of Census Data"
- 11:30-12:00 Discussion and questions (30 minutes)

12:00- 1:00 Lunch (Census cafeteria, on your own)

1:00-2:00 Research Session II: **SWELL (Summer Work Group for Employer List Linkage)** *(Organizers: Matthew Shapiro, Michigan; Lars Vilhuber, Cornell)* [Conference Room 1-3]

- Nada Wasi (University of Michigan), "Employer List Linking: A collaboration project between the Census Bureau, University of Michigan, and Cornell University"
- Mark Kutzbach (US Census Bureau), "ACS-LEHD job linking: early results from employer list linkages"

2:00- 5:00 Individual meetings with Census personnel (self-organized) [Conference Room 4 or elsewhere]

3:00-4:00 *(by invitation only)* Meeting of the PIs and Co-PIs

4:00-5:00 *(by invitation only)* Meeting with Director of the Census Bureau [Room 8H008]

7:30-9:30 *(by invitation only)* Group dinner - Capitol City Brewing Company

Friday, May 23, 2014

9:00-12:00 Research Session III: **Designing a questionnaire for the 21st century: Adaptive design and other survey topics** *(Organizer: Lars Vilhuber, Cornell)*
[Conference Rooms 1-3]

- 9:00-9:30 Adaptive Design at the Census Bureau - Overview (Peter Miller, Michael Thieme, and Anup Mathur, U.S. Census Bureau)
- 9:30-10:00 Web Surveys, Online Panels, and Paradata: Automating Adaptive Design (Allan McCutcheon, University of Nebraska-Lincoln)
- 10:00-10:30 Imputation of multivariate continuous data with non ignorable missingness (Thais Paiva, Duke University)
- 10:30-11:00 Aiming at a More Cost-Efficient Census Via Online Data Collection: Privacy Trade-Offs of Geo-Location (Laura Brandimarte and Alessandro Acquisti)
- 11:00-12:00 Discussion and questions

Conference ends.

Abstracts

Research Session I

Jay Greenfield and Sophia Kuan (Booz Allen Hamilton): "Describing Adaptive/Responsive Protocols and Data Processing in DDI"

DDI 4 is developing a process model capable of both describing an adaptive protocol and providing a machine readable representation that can assist during protocol execution. The information objects that make up the process model come from a number of sources/standards including the General Statistical Information Model (GSIM), a semantic markup language for web services called OWL-S and several longstanding DDI study objects. The process model is in the early stages of development. Besides the protocol representation/adaptive protocol use case, it is also being shaped by a second use case. In this use case the process model supports the description and actual orchestration of data transformations like those that play a role in provenance chains.

Abdul K. Rahim and Pascal Heus (Metadata Technologies North America) "Using Metadata Standards and Technology to Improve Accessibility, Opening and Reuse of Census Data"

This presentation highlights the importance and benefit of industry standard information technology and global metadata specifications - such as the Data Documentation Initiative - for the management and publication of open data. Many organizations make statistical data publicly or securely available to end users using traditional formats, but these often carry limited metadata and do not leverage global standards and modern tools or practices that have emerged in the past decade, following tremendous progress made in information technology and global efforts by leading agencies to establish common practices. Using the US Census Bureau's PUMS as an example, we illustrate how leveraging DDI, our OpenDataForge tools suite, and some custom developments, we rapidly repackaged the 2000 Census data files into open formats, ready for reuse by researchers, and further automated their publication into business intelligence platform for easy access by end users. We hope through this light exercise to demonstrate the benefits of standards based data management, and hint to the potential impact of the adoption of such practices across the data life cycle, form conceptualization and production to analysis and delivery to decision makers.

Bill Block and Jeremy Williams (Cornell University): Integrating PROV with DDI: Mechanisms of Data Discovery within the US Census Bureau

Within the United States Census Bureau, datasets are often derived by complex methods that are not always well documented. This derivation process, or provenance, can be hard to understand for a researcher attempting to use or explore a given dataset. Without understanding the provenance of a dataset, it can be impossible establish whether it is appropriate to use for a given investigation, because its history remains a black box with no way to see inside. The infrastructure upon which the semantic web is built provides a means to label the relationships of social science datasets with logical meaning according to standardized ontologies and controlled vocabularies. This paper outlines the work of the Comprehensive Data Documentation and Access Repository (CED^{2}AR) to integrate provenance metadata encoded according to the W3C PROV ontology with a DDI-based repository with the aim of making US Census data more discoverable and accessible.

Research Session II

Nada Wasi (University of Michigan): Employer List Linking: A collaboration project between the Census Bureau, University of Michigan, and Cornell University

This project aims to develop tools that can be commonly used for linking self-reported employer from a person-level survey to employer information in administrative records files. The constituent subprojects are (1) linking American Community Survey (ACS) with the Longitudinal Employer-Household Dynamics (LEHD); (2) linking the Survey of Income and Program Participation (SIPP) with the Business Register (BR); and (3) linking the Health and Retirement Study (HRS) with the BR. We develop tools for pre-processing the data, linking employers using various string and proximity comparators, and clerical review machine-generated matched pairs. An advantage of the collaboration is that each set of tools can be tested and fine-tuned from different types of real-world datasets. Preliminary results and challenges will be discussed.

Research Session III

Peter Miller, Michael Thieme, and Anup Mathur: Adaptive Design at the Census Bureau - Overview

In this presentation we will discuss how the Census Bureau is researching adaptive design capabilities for censuses and surveys, and is employing an enterprise architectural approach to deliver Adaptive Survey Design capabilities. Research activities include developing data resources for guiding fieldwork and tests of adaptive design interventions in simulation and field research. Using the enterprise architectural approach, we are developing systems that provide survey lifecycle capabilities on shared platforms across surveys and censuses.

Allan McCutcheon (University of Nebraska-Lincoln): Web Surveys, Online Panels, and Paradata: Automating Adaptive Design

The rising cost of telephone survey data collection, declining telephone survey response rates, and the speed of online survey data collection are leading some researchers to explore the use of web surveys and online panels. While these new modes of data collection present their own set of challenges (e.g., assuring probability sampling), they also present a new set of opportunities for survey researchers. This presentation will focus on an ongoing, five-year research project that is part of the NSF/Census Research Network (NCRN) involving online, probability-based panels in multi-mode surveys. The research is exploring the use of data and paradata from the internet portion of the survey to develop a machine-learning, 'smart agent' to implement near real-time adaptive design for online panels and other web surveys in an effort to reduce survey breakoff and panel attrition and to improve data quality. The project draws upon contributions from a team of survey methodologists, statisticians, and computer science engineers, and the cooperation of a leading industry partner (Gallup).

Laura Brandimarte and Alessandro Acquisti (Carnegie Mellon University): Aiming at a More Cost-Efficient Census Via Online Data Collection: Privacy Trade-Offs of Geo-Location

In an effort to reduce costs associated with data collection, the director of the Census Bureau announced in 2010 that the 2020 Census of the United States would take place via two channels: the classical paper questionnaire, mailed to each physical address recorded in the Census database (the Master Address File), and the online questionnaire – an option that will be offered for the first time to the whole US population. Besides an adequate awareness campaign, the success of the initiative will depend on addressing potential security and privacy concerns and providing adequate incentives for a sizable portion of US citizens to transition from the offline to the online format. In this research in progress, we analyze a specific incentive that the Census is considering – namely, pre-populating location information in the form, so to reduce completion time and effort. We study how individuals may react to this initiative, given the privacy concerns that pre-populating may arise. We describe two online experiments of geo-location (one on-going and one yet to be started) which investigate the impact of awareness of being geo-tracked on willingness to provide, among other pieces of information, Census-related personal information.

Thais Paiva, Duke University: Imputation of multivariate continuous data with non ignorable missingness

Regular imputation methods have been used to deal with non-response in several types of survey data. However, in some of these studies, the assumption of missing at random is not valid since the probability of missing depends on the response variable. In our adaptive design motivating example, the response variable distribution can depend on the wave on which the data is collected. We propose an imputation method for multivariate data sets when there is non-ignorable missingness. A Dirichlet process mixture of multivariate normals is fit to the observed data under a Bayesian framework to provide flexibility. We provide some guidelines on how to alter the estimated distribution using the posterior samples of the mixture model and obtain imputed data under different scenarios. Lastly, we apply the method to a real data set.