My Desktop
Prepare & Submit Proposals
Proposal Status
Proposal Functions
Awards & Reporting
Notifications & Requests
Project Reports
Submit Images/Videos
Award Functions
Manage Financials
Program Income Reporting
Grantee Cash Management Section Contacts
Administration
Lookup NSF ID

Preview of Award 1131848 - Annual Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1131848
Project Title:	NCRN-MN: Cornell Census-NSF Research Node: Integrated Research Support, Training and Data Documentation
PD/PI Name:	Lars Vilhuber, Principal Investigator William C Block, Co-Principal Investigator
Recipient Organization:	Cornell University
Project/Grant Period:	10/01/2011 - 09/30/2018
Reporting Period:	10/01/2016 - 09/30/2017
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

Accomplishments

*** What are the major goals of the project?**

As part of the Cornell node's activities, we are building a Comprehensive Extensible Data Documentation and Access Repository (CED²AR) designed to improve the documentation and discoverability of both public and restricted data from the federal statistical system. The CED²AR will be based upon leading metadata standards such as the Data Documentation Initiative (DDI) and Statistical Data and Metadata eXchange (SDMX) and be flexibly designed to ingest documentation from a variety of source files.

We are also developing High Performance Logistic Regression Methods for Data Edits and Imputation for (a) multiple response variables (Census example: race/ethnicity coding) as well as (b) incompletely coded links (Census example: unit-to-worker imputation).

More recently, we have tackled the problem of efficient trade-offs between data quality and confidentiality (privacy loss) using techniques from economics, i.e., a formal production possibilities frontier (PPF). We consider situations where data quality will be inefficiently under-supplied. Results show that government data custodians should publish more accurate statistics with weaker privacy guarantees than would occur with purely private data publishing.

Finally, we are teaching a multi-site distance learning class on "Social and Economic Data" (INFO 7470). The course is designed to teach students basic and advanced techniques for acquiring and transforming raw information into social and economic data. The course is particularly aimed at American Ph.D. students from multiple fields (economics, political science, demography, sociology, etc.) who are interested in using confidential U.S. Census Bureau data, and the confidential data of other American statistical agencies that cooperate with the Census Bureau. We cover the legal, statistical, computing, and social science aspects of the data "production" process. More information is available at the course website http://www.vrdc.cornell.edu/info7470/.

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities:

CED2AR

Two versions of ced2ar v2 were released this past year.

1. ced2ar 2.8.2 was released on 4/5/2017.
1. New features:
1. New search functionality to Browse by Study. Codebooks are displayed in a tabbed horizontal layout. The tabs are the DDI complex types.
2. UI Navigation Customization. Administrators can set properties used to display/hide navigation tabs and the names of those tabs.
3. New Authentication option added. Setting accessMode to AdminOnly allows only users with the ROLE_ADMIN role to access the application. All others are prevented from accessing the pages.

- Resolved Issues:

1. 6 github issues were closed.
2. 7 jira issues were closed.

2. ced2ar 2.8.1.2 was release on 12/14/2016
1. New features:
1. Updated interface to expose the variable fields ‘Universe’ and ‘AnlysUnit’ for editing and viewing
2. Added fix to support PDF generation on Windows systems

V3

1. Developed new MVC application, similar to V2 but with concise templating
1. Implemented:

1. Views for codebooks and vars
2. Modular architecture extended with new modules:

1. Services-core provides intermediate functions between RDB module and the site module. These are exposed natively and as REST calls with JSON payloads.
2. Site: this is the “app” that ties everything together; notably it houses the MVC controllers which connect the views to the services
3. Two test modules for the rdb and services-core modules that can be used by the site module for integration tests (or used by an alternative site module developed by a third party)
4. Jsgen - the module that houses a Scala.JS SPA “app” (run ina browser), which supports interfacing with many of the Java and all JS APIs in a typesafe manner. Can be run standalone or integrated in the site module. See below for more detail.

- Partially implemented / in progress:

1. Schema map view (for controlling export of a codebook to an alternative schema)

- Developed an auxiliary library to search an XML schema or XML document for all XPaths.

- Initiated work on a SPA (single page application), which is designed to behave similarly to the MVC views (can browse by URL as well as in-app navigation) but is more responsive.

1. Implemented:

1. Largely converted existing MVC view pages mentioned above. As the MHTML views used here is very similar to JSP used in V2, provides a convenient porting route from V2 views as well.
2. Ability to run SPA locally and connect to a remote server

- Partially implemented / in progress:

1. Editing of codebook and var entries

Other activities with CED2AR

We collaborated with ICPSR on assessing deployment and usability issues. This improved the application. We also collaborated with the Census Bureau in assessing various options for use within the Bureau. We presented at the North American DDI conference (NADDI).

INFO7470

A new version of the course was started in August 2017. About 100 students are signed up, more components have been put on "flipped classroom" basis. The course is ongoing. We will assess how maintainable the course is in this technological form.

Privacy research

We supported organizing three workshops on privacy, with primary funding from NSF Grant 1012593 and a Sloan grant. A publication is under review by the AER. A separate publication is forthcoming in the ASA's "Chance" magazine on replication when data are private.

Specific Objectives:

Significant Results:

Key outcomes or Other achievements:

*** What opportunities for training and professional development has the project provided?**

A graduate student (Herbert) is working on research with confidential data.

*** How have the results been disseminated to communities of interest?**

The three workshops on privacy, the NADDI conference, and numerous working papers. Multiple papers have been submitted to academic journals. The CED2AR software is available for download as binary software for both servers and desktops. Source code is posted on Github. Publications are listed elsewhere in this report. Several presentations of the work at scientific conferences have been given (see attachment). INFO7470 materials were broadcast using a combination of EdX online tools and web posting. The recorded sessions will be made available in the future on Youtube. All papers and INFO7470 materials (presentations, videos), as well as other presentations, are made available on properly curated document archives at http://ecommons.cornell.edu.

*** What do you plan to do during the next reporting period to accomplish the goals?**

We are working with ICPSR to consider long-term impacts of CED2AR technology. Publications are pending. Software is being cleaned up for the ability of open-source groups to pick it up. CED2AR itself is being incorporated into Cornell infrastructure.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
NCRN-Cornell-Presentations-2016-2017.part1.pdf	This document lists all presentations made by members of the NCRN Cornell node between October 1, 2016 and September 30, 2017. All presentation and events are listed in a searchable form at https://www.ncrn.cornell.edu/events-and-activities/ . (this is part1 - October 2016-May2017)	Lars Vilhuber	10/02/2017
Events and Activities-May2017-August2017.pdf	This document lists all presentations made by members of the NCRN Cornell node between October 1, 2016 and September 30, 2017. All presentation and events are listed in a searchable form at https://www.ncrn.cornell.edu/events-and-activities/ . (this is part2 - May2017-August 2017)	Lars Vilhuber	10/02/2017
Events and Activities-May2017-August2017.pdf	This document lists all presentations made by members of the NCRN Cornell node between October 1, 2016 and September 30, 2017. All presentation and events are listed in a searchable form at https://www.ncrn.cornell.edu/events-and-activities/ . (this is part3 - August 2017-Sept 2017)	Lars Vilhuber	10/02/2017

Products

Books

Book Chapters

Inventions

Journals or Juried Conference Papers

Abowd, John M. and McKinney, Kevin L. (2016). Noise infusion as a confidentiality protection measure for graph-based statistics. *Statistical Journal of the International Association for Official Statistics*. 32 127-135. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.3233/SJI-160958

John M. Abowd and Francis Kramarz and Sebastien Perez-Duarte and Ian M. Schmutte (2018). Sorting Between and Within Industries: A Testable Model of Assortative Matching. *Annals of Economics and Statistics*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 21154430, 19683863

John M. Abowd and Kevin L. Mckinney and Nellie Zhao (2018). Earnings Inequality and Mobility Trends in the United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data. *Journal of Labor Economics*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Miranda, Javier and Vilhuber, Lars (2016). Using partially synthetic microdata to protect sensitive cells in business statistics. *Statistical Journal of the International Association for Official Statistics*. 32 69-80. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.3233/SJI-160963

Samuel Haney and Ashwin Machanavajjhala and John M. Abowd and Matthew Graham and Mark Kutzbach (2017). Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics. *Proceedings of the 2017 ACM International Conference on Management of Data*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3035918.3035940

Schmutte, Ian M. (2016). Differentially private publication of data on wages and job mobility. *Statistical Journal of the International Association for Official Statistics*. 32 81-92. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.3233/SJI-160962

Vilhuber, Lars and Abowd, John M. and Reiter, Jerome P. (2016). Synthetic establishment microdata around the world. *Statistical Journal of the International Association for Official Statistics*. 32 65-68. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.3233/SJI-160964

Vilhuber, Lars and Lagoze, Carl (2017). Making Confidential Data Part of Reproducible Research. *Chance*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://chance.amstat.org/2017/09/reproducible-research/

Licenses

Other Conference Presentations / Papers

Miranda, Javier and Vilhuber, Lars (2015). *Assessing the Data Quality of Public Use Tabulations Produced from Synthetic Data: Synthetic Business Dynamics Statistics (Presentation)*. Joint Statistical Meetings (JSM). . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Ping Li and John Abowd (2014). *Boosting Algorithms for Edit and Imputation of Multiple-response Variables (Presentation only)*. Federal Committee on Statistical Methodology Research Conference. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Drechsler, Jörg and Vilhuber, Lars (2015). *Synthetic Longitudinal Business Databases for International Comparisons (Presentation)*. Joint Statistical Meetings (JSM). . Status = PUBLISHED; Acknowledgement of Federal Support = No

Other Products

Other Publications

Barker, Brandon, Brumsted, Kyle, Simmer, Charles, & Vilhuber, Lars. (2016). *CED²AR V2.8.1.2.*. Software to edit and manage DDI metadata. http://doi.org/10.5281/zenodo.494998. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Brandon Elam Barker, Charles Simmer, Lars Vilhuber, Kyle Brumsted, & Ben Perry (2017). *CED²AR: 2.8.2.0*. Software for editing and management of DDI metadata. Zenodo. http://doi.org/10.5281/zenodo.495191. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Patents

Technologies or Techniques

Thesis/Dissertations
Shrivastava, Anshumali. *Probabilistic Hashing Techniques for Big Data*. (2015). Cornell University. Acknowledgement of Federal Support = Yes

Websites
Comprehensive Extensible Data Documentation and Access Repository (CED²AR)
https://github.com/ncrncornell/ced2ar/

Source code for CED2AR application.

NSF Census Research Network – Cornell Node
https://www.ncrn.cornell.edu/

website for dissemination of node news.

ecommons For Cornell University NCRN node
https://ecommons.cornell.edu/handle/1813/30503

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Vilhuber, Lars	PD/PI	2
Block, William	Co PD/PI	1
Lagoze, Carl	Faculty	1
Barker, Brandon	Other Professional	3
Brumsted, Kyle	Other Professional	6
Kambhampaty, Venkata	Other Professional	0
Perry, Benjamin	Other Professional	0
Brown, Warren	Staff Scientist (doctoral level)	2
Edwards, Anne	Staff Scientist (doctoral level)	0
Herbert, Sylverie	Graduate Student (research assistant)	6
Sexton, William	Graduate Student (research assistant)	1
Shrivastava, Anshumali	Graduate Student (research assistant)	0
Stanchi, Flavio	Graduate Student (research assistant)	0

Full details of individuals who have worked on the project:

Lars Vilhuber
Email: lars.vilhuber@cornell.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 2

Contribution to the Project: Lead PI, work on confidentiality, metadata, CED2AR, overall management.

Funding Support: This grant.

International Collaboration: No
International Travel: No

William C Block
Email: block@cornell.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Work on metadata.

Funding Support: This grant.

International Collaboration: No
International Travel: No

Carl Lagoze
Email: clagoze@umich.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 1

Contribution to the Project: Metadata, Provenance expertise

Funding Support: This grant.

International Collaboration: No
International Travel: No

Brandon Barker
Email: beb82@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 3

Contribution to the Project: Working on CED2AR software

Funding Support: This grant

International Collaboration: No
International Travel: No

Kyle Brumsted
Email: kjb245@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 6

Contribution to the Project: Software development

Funding Support: This grant.

International Collaboration: No
International Travel: No

Venkata Kambhampaty
Email: vkambhampaty@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 0

Contribution to the Project: Software development. Left in 2016.

Funding Support: This grant

International Collaboration: No
International Travel: No

Benjamin Perry
Email: bap63@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 0

Contribution to the Project: Programming. left in 2016.

Funding Support: This grant

International Collaboration: No
International Travel: No

Warren Brown
Email: warren.brown@cornell.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 2

Contribution to the Project: Expertise on ACS, INFO7470.

Funding Support: NSF (this grant)

International Collaboration: No
International Travel: No

Anne Michelle Edwards
Email: ame87@cornell.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: metadata expertise. Has left Cornell.

Funding Support: Cornell University

International Collaboration: No
International Travel: No

Sylverie Herbert
Email: sh2258@cornell.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 6

Contribution to the Project: Assistance in creating/editing/improving metadata based on available data outside the Census firewall, assistance in preparing INFO7470

Funding Support: This grant.

International Collaboration: No
International Travel: No

William Sexton
Email: wns32@cornell.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Assistance on confidentiality research

Funding Support: This grant.

International Collaboration: No
International Travel: No

Anshumali Shrivastava
Email: ansh@cs.cornell.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: He worked on developing a min search engine for metadata based on modern natural language processing techniques. He also worked on developing novel algorithms for graph data representations as well as large-scale statistics computations

Funding Support: NSF-EAGER 1249316 NSF-DMS 0808864 This grant.

International Collaboration: No
International Travel: No

Flavio Stanchi
Email: fs379@cornell.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Assistance in creating/editing/improving metadata based on available data outside the Census firewall

Funding Support: No other.

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
ICPSR	Other Nonprofits	Ann Arbor, MI
Roper Center	Academic Institution	Ithaca, NY

Name	Type of Partner Organization	Location
US Census Bureau	Other Organizations (foreign or domestic)	Washington, DC
University of Michigan	Academic Institution	Ann Arbor, Michigan

Full details of organizations that have been involved as partners:

ICPSR

Organization Type: Other Nonprofits
Organization Location: Ann Arbor, MI

Partner's Contribution to the Project:
In-Kind Support

More Detail on Partner and Contribution: We have had metadata contributions and discussions with ICPSR on the CED2AR project.

Roper Center

Organization Type: Academic Institution
Organization Location: Ithaca, NY

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Contribution to the development of metadata infrastructure/software.

US Census Bureau

Organization Type: Other Organizations (foreign or domestic)
Organization Location: Washington, DC

Partner's Contribution to the Project:
In-Kind Support
Facilities
Collaborative Research

More Detail on Partner and Contribution: Use of the Cornell Census Research Data implies a substantial Census Bureau participation since the Bureau pays substantially all of that RDC's operating expenses (unlike all the others, which bear these expenses themselves). The Census Bureau participated in the INFO7470 class, and we interact with the Census Bureau on the CED2AR project.

University of Michigan

Organization Type: Academic Institution
Organization Location: Ann Arbor, Michigan

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Training course provided by Michigan NCRN node, supported by this grant's CED²AR for the purpose of training new users of the SIPP Synthetic Beta.

What other collaborators or contacts have been involved?
Nothing to report

Impacts

What is the impact on the development of the principal discipline(s) of the project?

CED2AR has contributed by posing the problem of confidentiality of metadata, and providing a solution. It also has highlighted the feasibility of crowd sourcing such information, while maintaining control over the quality of the resulting documentation at the data curator level. Work on Privacy and Confidentiality has contributed by highlighting the need to think about privacy in the context of both data providers (who desire privacy) and data users (who desire accuracy), and to provide a framework to make optimal choices. INFO7470 has contributed to making future and current researchers aware of the source of the data they are using, of the constraints in constructing such data, including confidentiality constraints, and novel methods of accessing the data.

What is the impact on other disciplines?
Nothing to report.

What is the impact on the development of human resources?

The availability of improved metadata, and of better privacy protected public use data products, will enable more researchers to discover and use data, leading to new discoveries in the social sciences. INFO7470 trains new researchers in a variety of fields to use the resources of the statistical system effectively and appropriately.

What is the impact on physical resources that form infrastructure?
Nothing to report.

What is the impact on institutional resources that form infrastructure?

The availability of new metadata curation tools allows for institutions to adopt better, more transparent methods.

What is the impact on information resources that form infrastructure?
Nothing to report.

What is the impact on technology transfer?
Nothing to report.

What is the impact on society beyond science and technology?
Nothing to report.

Changes/Problems

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.