Preview of Award 1131848 - Annual Project Report

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:	John M Abowd, Principal Investigator William C Block, Co-Principal Investigator Ping Li, Co-Principal Investigator Lars Vilhuber, Co-Principal Investigator
Recipient Organization:	Cornell University
Project/Grant Period:	10/01/2011 - 09/30/2016
Reporting Period:	10/01/2013 - 09/30/2014
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).

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:

Development of the CED2AR tool continues, incorporating feedback from diverse partners and users. Additional workflows (metadata creation within the Census Bureau, workflow documentation in other restricted-access environments such as Synthetic Data Server) are being incorporated. Work continues on obtaining the appropriate permissions to deploy CED²AR within the Census Bureau network.

Standards-enhancing work on incorporating provenance information (PROV) into DDI has advanced, and draft standards have been published on our website.

With additional support from Cornell University, INFO 7470 is being converted into an online "flipped classroom" style class.

Specific Objectives:

The INFO7470 class is aimed at American Ph.D. students 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. No equivalent class exists, and takeup is wide, and generally found to be useful. Converting the class into an online class will allow certain sections of the class to be made available more frequently, and permanently.

The CED2AR tool is the platform on which to achieve the overarching goal of liberating metadata on confidential data. Once stable, it will be used to host the data within the confidential confines (the Census Bureau) as well as the filtered metadata outside. In addition, the process of generating metadata on data and its provenance has proven useful in simple workflow documentation contexts, such as the release request workflow in the Census RDC.

Work into support for provenance in DDI-C will support more robust documentation of data, in support of better data discovery and replication exercises.

Significant Results:

CED2AR has been officially released on the non-confidential part, hosting unique metadata on data accessible through the Synthetic Data Server.

A draft DDI-C enhancement has been published.

Key outcomes or Other achievements:

CED2AR has been actively used in training of student and faculty users of the Synthetic Data Server (SIPP Synthetic Beta).

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

CED2AR has been actively used in training of student and faculty users of the Synthetic Data Server (SIPP Synthetic Beta).

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

CED²AR and related research has been presented at key conferences (iASSIST, Digital Libraries, EDDI, NADDI) in 2013-2014. CED²AR also has a production implementation, accessible on the open internet, and turning up in search results.

Work on boosting algorithms has been presented at several conferences.

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

The integration of PROV into DDI-C will be made robust, and will contribute to CED²AR. An implementation of CED²AR at the Census Bureau is expected to be implemented. A workflow tool incorporating the refinements implemented in CED²AR and relying on the integration of PROV will be tested on the Synthetic Data Server at Cornell and at the Census Bureau. Multiple additional features of CED²AR, including critically for an enhanced implementation at the Census Bureau: authentication, will be implemented.

Applying boosting algorithms to the Census-specific application remains a goal.

Products

Books

Book Chapters

Conference Papers and Presentations

Anshumali Shrivastava and Ping Li (2013). *A New Mathematical Space for Social Networks*. Neural Information Processing Systems (NIPS) Workshop. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Anshumali Shrivastava and Ping Li (2014). *A New Space for Comparing Graphs*. ASONAM. . Status = ACCEPTED; 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 = OTHER; Acknowledgement of Federal Support = Yes

Carl Lagoze and Lars Vilhuber and Jeremy Williams and Benjamin Perry and William C. Block (2014). *CED�AR: The Comprehensive Extensible Data Documentation and Access Repository*. ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014). London, United Kingdom. Status = ACCEPTED; Acknowledgement of Federal Support = Yes

Lagoze, Carl and Willliams, Jeremy and Vilhuber, Lars (2013). *Encoding Provenance Metadata for Social Science Datasets*. Metadata and Semantics Research. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Carl Lagoze and William C. Block and Jeremy Williams and Lars Vilhuber (2013). *Encoding Provenance of Social Science Data: Integrating PROV with DDI*. 5th Annual European DDI User Conference. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Ping Li and Cun-Hui Zhang (2013). *Exact Sparse Recovery with L0 Projections*. 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Anshumali Shrivastava and Ping Li (2014). *In Defense of MinHash Over SimHash*. Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS). Reykjavik, Iceland. Status = ACCEPTED; Acknowledgement of Federal Support = Yes

Ping Li and Anshumali Shrivastava and K\"onig, Arnd Christian (2013). *b-Bit Minwise Hashing in Practice*. Internetware'13. Changsha, China. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Inventions

Journals

Anshumali Shrivastava and Ping Li (2014). Graph Kernels via Functional Embedding. *CoRR*. abs/140 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://arxiv.org/abs/1404.5214

Carl Lagoze and William C. Block and Jeremy Williams and John M. Abowd and Lars Vilhuber (2013). Data

Management of Confidential Data. *International Journal of Digital Curation*. 8 (1), 265-278. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.2218/ijdc.v8i1.259

Licenses

Other Products

Other Publications

Collective (Cornell NSF-Census Research Network) (2013). *Comprehensive Extensible Data Documentation and Access Repository. Codebook for the NBER-CES Manufacturing Industry Database (2009) [Codebook file]. Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, 2013*. Codebook for NBER-CES productivity database, available at http://www2.ncrn.cornell.edu/ced2ar-web/codebooks/nber-ces-naics. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Collective (Cornell NSF-Census Research Network) (2013). *Comprehensive Extensible Data Documentation and Access Repository. Codebook for the SIPP Synthetic Beta 5.1 [Codebook file]. Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, 2013*. Codebook for SIPP Synthetic Beta, available at http://www2.ncrn.cornell.edu/ced2ar-web/codebooks/ssb. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Collective (NSF-Census Research Network - Cornell node) (2014). *Comprehensive Extensible Data Documentation and Access Repository. Codebook for the Synthetic LBD Version 2.0 [Codebook file]. Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, 2013*. Codebook for Synthetic LBD, available at http://www2.ncrn.cornell.edu/ced2ar-web/codebooks/synlbd. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Patents

Technologies or Techniques

Thesis/Dissertations

Websites

CED²AR production server
http://www2.ncrn.cornell.edu/ced2ar-web/

CED2AR is designed to improve the discoverability of both public and restricted data from the federal statistical system. The project is based upon leading metadata standards and is flexibly designed to ingest data from a variety of sources.

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Abowd, John	PD/PI	1
Block, William	Co PD/PI	1
Li, Ping	Co PD/PI	3
Vilhuber, Lars	Co PD/PI	3
Anderson, Margo	Faculty	0

Lagoze, Carl	Faculty	1
Nagle, Nicholas	Faculty	0
Slavkovic, Aleksandra	Faculty	0
Perry, Benjamin	Other Professional	12
Williams, Jeremy	Other Professional	3
Brown, Warren	Staff Scientist (doctoral level)	1
Shrivastava, Anshumali	Graduate Student (research assistant)	6
Stanchi, Flavio	Graduate Student (research assistant)	6

Full details of individuals who have worked on the project:

John M Abowd
Email: john.abowd@cornell.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Lead PI. Participating in all aspects of the research. Leading the effort to develop improved edit and imputation methods. Supporting the effort to convert the Ph.D. course on methods for using and analyzing large confidential data systems to an online course.

Funding Support: 1 month 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: Co-PI; provide day-to-day oversight and management of the CED²AR software development effort, in particular the UI related to DDI metadata creation and editing. Incorporate support for provenance metadata into DDI-C via the PROV metadata standard; outreach, presentations.

Funding Support: None.

International Collaboration: No
International Travel: Yes, France - 0 years, 0 months, 3 days; Canada - 0 years, 0 months, 5 days

Ping Li
Email: pingli@stat.rutgers.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 3

Contribution to the Project: Advise the Ph.D. student Anshumali Shrivastava on research problems on large-scale statistics computing, graph representations, and database search and indexing. Work on developing novel algorithms for predicting multiple responses.

Funding Support: NSF-EAGER 1249316 NSF-DMS 0808864

International Collaboration: No
International Travel: No

Lars Vilhuber
Email: lars.vilhuber@cornell.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 3

Contribution to the Project: Managing PI, contributed to CED²AR development, outreach, presentations

Funding Support: None

International Collaboration: No
International Travel: Yes, France - 0 years, 0 months, 3 days; - 0 years, 0 months, 0 days; - 0 years, 0 months, 0 days

---

Margo Anderson
Email: margo@uwm.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Margo Anderson kindly contributed a session to the INFO7470 class.

Funding Support: Census Fellowship/sabbatical

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

Nicholas Nagle
Email: nnagle@utk.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Contributed a session to INFO7470

Funding Support: none (possibly NCRN to another node)

International Collaboration: No
International Travel: No

Aleksandra B. Slavkovic
Email: sesa@psu.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Aleksandra Slavkovic contributed to a INFO7470 session.

Funding Support: Sabbatical

International Collaboration: No
International Travel: No

Benjamin Perry
Email: bap63@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 12

Contribution to the Project: Programming

Funding Support: This grant

International Collaboration: No
International Travel: No

Jeremy Williams
Email: jw568@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 3

Contribution to the Project: Programming and writing up the results of project research.

Funding Support: This grant and Cornell University

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: 1

Contribution to the Project: Expertise on ACS

Funding Support: NSF (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: 6

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

International Collaboration: No
International Travel: No

Flavio Stanchi
Email: fs379@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

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
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.

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.

Have other collaborators or contacts been involved? No

Impacts

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

We have published enhanced standards, and proposed them to the appropriate organization. We are lobbying for integration of the standards.

What is the impact on other disciplines?

The project is crossdisciplinary (information science, computer science, statistics, economics) and influences all four domains.

What is the impact on the development of human resources?

The metadataoriented outcomes of this project will facilitate knowledge discovery and creation by graduate students and other scientists in the social sciences. It has been successfully used in a course on the use of the SIPP Synthetic Beta, and was well received.

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

What is the impact on institutional resources that form infrastructure?

(practices) The improvements to metadata standards are being proposed to international standards bodies, and thus likely to be adopted by many institutions worldwide.

What is the impact on information resources that form infrastructure?

The metadataoriented outcomes of this project will facilitate knowledge discovery and creation of data, thus improving the overall set of information resources. Such resources do not need to, and won't, be housed within the project.

What is the impact on technology transfer?

(public use) The improvements to metadata standards are being proposed to international standards bodies, and thus likely to be adopted by many institutions worldwide.

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

Ping Li has moved to Rutgers University. He will continue to contribute from there, and the project continues to support his Cornell-based Ph.D. student. We do not expect future Cornell Ph.D. students in statistics to be supervised by Ping Li; they may still, however, be supervised by John Abowd (has appointment in Statistics).

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.