CISER

Cornell Institute for Social and Economic Research
A LEADER IN SOCIAL SCIENCE DATA AND COMPUTING

The Cornell NSF-Census Research Node: Integrated Research Support, Training and Data

William Block, Co-PI
Warren Brown & Stefan Kramer, Senior Scientists
Florio Arguillas & Jeremy Williams, Project Staff

Cornell Institute for Social and Economic Research (CISER)

3rd Annual European DDI Users Group Meeting (EDDI11), Dec. 6, 2011

Cornell University

Overview

- Background on NCRN and the RDC Environment
- The Problem
- Proposed Solution
- Early Work

NSF 10-621: NSF-Census Research Network (NCRN) Program Solicitation

"The NSF-Census Research Network will provide support for a set of research nodes, each of which will be staffed by a team of scientists conducting interdisciplinary research and educational activities on methodological questions of interest and significance to the broader research community and to the Federal Statistical System, particularly the U.S. Census Bureau. The activities will be expected to advance both fundamental and applied knowledge as well as further the training of current and future generations of researchers in research skills of relevance to the measurement of economic units, households, and persons."

Total funding: $18,500,000

Expected Number of Awards: 8 - 12

NCRN Program Goals

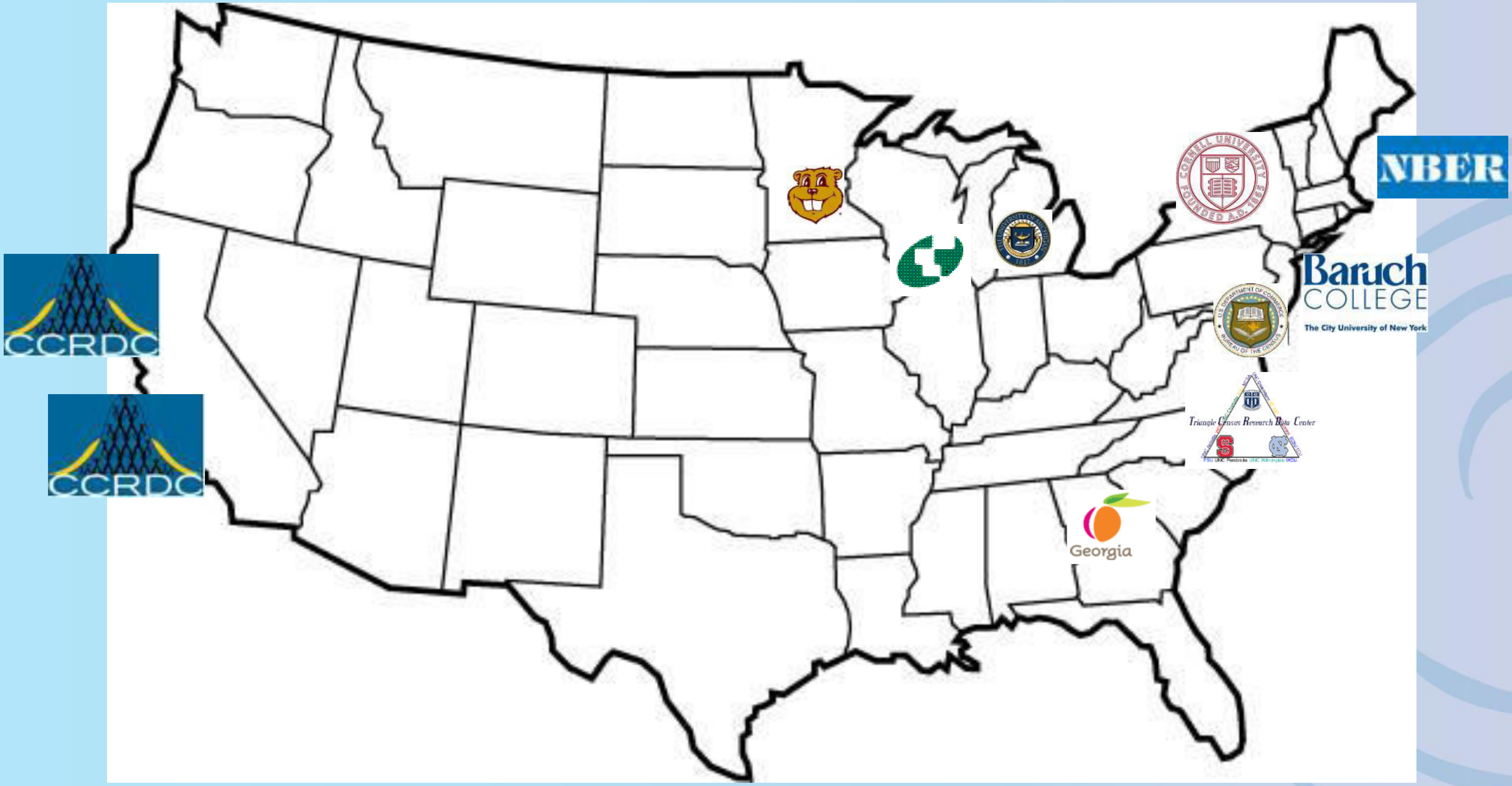
1. Establish a set of complementary research programs that advance the development of innovative methods and models for the collection, analysis, and dissemination of data in the social, behavioral, and economic sciences.

2. Relate fundamental advances in methods development to the problems of the Federal Statistical System, particularly the U.S. Census Bureau.

3. Facilitate the collaborative activities of scientists from across multiple disciplines, including the social, behavioral, and economic sciences, the statistical sciences, and the computer sciences.

4. Foster the development of the next generation of researchers in research skills of relevance to the measurement of economic units, households, and persons.

The RDC Network


CORNELL UNIVERSITY FOUNDED A.D. 1865
NBER
CCRDC
CCRDC
Baruch COLLEGE
The City University of New York
U.S. DEPARTMENT OF COMMERCE BUREAU OF THE CENSUS
Triangle Census Research Data Center
Georgia

Research Opportunities at the Cornell Census Research Data Center (RDC)

CISER, 391 Pine Tree Road

Data Available in the RDC

- Economic Data
  - Interview business establishments and firms
  - No public use versions
- Demographic Data
  - Interview individuals and households
  - Complete geography, no income topcoding
- Matched Employer-Employee Data
- Health Data: Partnerships with
  - NCHS
  - AHRQ

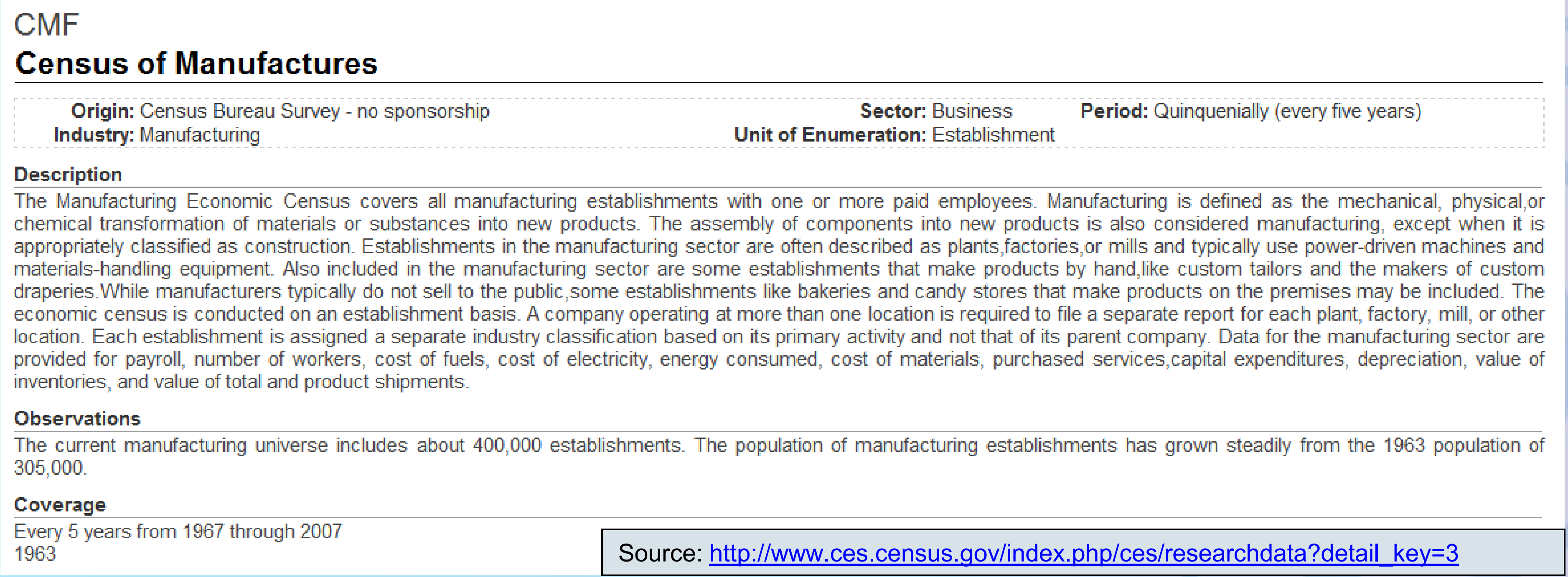
CISER

Current situation within Census RDCs

- Tension between confidentiality and user-friendliness
- Lack of consistent documentation at variable level
- Barriers to data discovery and use
- Long term curation precluded
- Scientific replication impossible
- Two examples: Census of Manufactures and American Community Survey

Present data documentation situation for RDC users

What can be learned online about the most widely used RDC dataset, the Census of Manufactures, ***outside*** of an RDC:



CMF

Census of Manufactures

Origin: Census Bureau Survey - no sponsorship **Sector:** Business **Period:** Quinquenially (every five years)
Industry: Manufacturing **Unit of Enumeration:** Establishment

Description

The Manufacturing Economic Census covers all manufacturing establishments with one or more paid employees. Manufacturing is defined as the mechanical, physical,or chemical transformation of materials or substances into new products. The assembly of components into new products is also considered manufacturing, except when it is appropriately classified as construction. Establishments in the manufacturing sector are often described as plants,factories,or mills and typically use power-driven machines and materials-handling equipment. Also included in the manufacturing sector are some establishments that make products by hand,like custom tailors and the makers of custom draperies.While manufacturers typically do not sell to the public,some establishments like bakeries and candy stores that make products on the premises may be included. The economic census is conducted on an establishment basis. A company operating at more than one location is required to file a separate report for each plant, factory, mill, or other location. Each establishment is assigned a separate industry classification based on its primary activity and not that of its parent company. Data for the manufacturing sector are provided for payroll, number of workers, cost of fuels, cost of electricity, energy consumed, cost of materials, purchased services,capital expenditures, depreciation, value of inventories, and value of total and product shipments.

Observations

The current manufacturing universe includes about 400,000 establishments. The population of manufacturing establishments has grown steadily from the 1963 population of 305,000.

Coverage

Every 5 years from 1967 through 2007
1963

Source: http://www.ces.census.gov/index.php/ces/researchdata?detail_key=3

No information provided here about which variables are contained in this dataset.

Present data documentation situation for RDC users cont.

What can be learned online about the most widely used RDC dataset, the Census of Manufactures, ***inside*** an RDC. One can browse folders, “read me” files, or

L*eave RDC*, use a computer with public Internet access!

Example from American Community Survey (ACS)

- Internal project experience
- Data and basic SAS program file for ACS “zero obs” files
- No internal documentation
- IPUMS documentation (public)

The Cornell NSF-Census Research Node: Integrated Research Support, Training, and Data

- The Comprehensive Census Bureau Metadata Repository (CCBMR; co-PI Block)
  - DDI-based metadata schema
  - synchronization between public and confidential instances of the repository
  - Disclosure Avoidance Review-compliant and friendly
  - Similar web interfaces for public (partial) and internal (full-information) documentation
  - Create and share a CCBMR toolkit
- Other elements of the Cornell NCRN Project
  - Integrated Doctoral Instruction on the Information Science of Restricted-access Data Analysis (PI: John Abowd; co-PI: Lars Vilhuber)
  - Adding Computational Statistics to the Administrative Record Toolkit (co-PI Ping Li)

Different metadata inside vs. outside RDCs

- Not all of the metadata about studies in RDCs that should be viewable inside an RDC can be made available on the public web
  - Example: certain value ranges, or the very existence, of a variable cannot be made known outside of an RDC
- Two different versions of a metadata: bad idea
- Better: complete internal version; derive subset for public

How DDI could help with this challenge

- Develop a metadata schema based on DDI, with modifications and additional fields/elements and/or attributes as warranted to capture and represent the necessary information about RDC datasets
- Information about specific confidentiality considerations and access conditions can already be expressed in the DDI 3.1 specification with elements such as <ConfidentialityStatement> and <AccessConditions>
- DDI could be extended through the addition of machine-actionable markup describing variables to control what information about them is revealed – example follows (assuming an XML-based implementation, which is not the only option)

How DDI could help with this challenge (cont.)

RDC Metadata (complete)	Derived Public Use Metadata (limited)

<d:VariableSet>
  <d: VariableItem>…:<d:/VariableItem>
  <d:Disclosability>
    <d:min disclosable=’’yes’’>0</d:min>
    <d:max disclosable=’’no’’>345678</d:max>
  </d:Disclosability>
</d:VariableSet>
	<d:VariableSet>
	<d: VariableItem>…:<d:/VariableItem>
	  <d:Disclosability>
	    <d:min>0</d:min>
	    <d:max>not disclosable</d:max>
	  </d:Disclosability>
	</d:VariableSet>

The above example denotes that one extreme value associated with this variable is not releasable, regardless of the public-use nature of other information on the variable. In this example, the maximum value would not be included in the publicly-released documentation.

Metadata platforms being considered for CCBMR

- XML database:
  - BaseX or
  - eXist
- or
- Relational database (RDB):
  - SQL server (MS Windows only) or
  - PostgreSQL or
  - MySQL or
  - Oracle (or possibly others)
- or
- Hybrid approach of using RDB and storing XML (such as variable-level DDI generated by Stat/Transfer) as objects inside it
- or
- Proprietary product:
  - Colectia Repository with SDK (with which we could create a customized web front-end)

Related to this topic: *Representing and Utilizing DDI in Relational Databases* (http://dx.doi.org/10.3886/DDIOtherTopics02)

Metadata specifications being considered for CCBMR

DDI

SDMX

DataCite

Approach: “as much as necessary, as little as possible” metadata for the project; *not* limited only to DDI

CISER

From draft functional requirements for CCBMR of Nov. 14, 2011

….

3. Searching and navigating

3.1. Nested Boolean searching (i.e.: (A OR B) AND (X OR Y OR Z))

3.2. Explicit truncation

3.3. Case-sensitivity option

3.4. Date-range searching

3.5. Choice which field(s) to search, e.g., title, description/abstract, variable-level elements

3.6. Browsing by … to-be-determined fields/elements

3.7. Searching by values from summary statistics in metadata?

4. Metadata export and transformation

4.1. Between RDC-internal and external version, generate elements/fields and their contents based on the (to be) developed suppression method (page 6 of proposal – "Disclosability" example)

4.2. In a format that will allow fairly easy creation of syntax/command files for statistical packages – DDI?

5. Metadata import and transformation

5.1. Ingest DDI 3.x metadata generated by tools such as Stat/Transfer from datasets housed in RDC that are selected for project

5.2. Transform ingested metadata into schema (to be) developed for project, e.g., rename and/or drop fields/elements

5.3. Import metadata from/for PUFs, e.g. from IPUMS – method TBD

6. Metadata editing

6.1. Web-based front end for adding and editing metadata elements

7. Persistent identifiers

7.1. Use DOIs? Needs further consideration.

CCBMR: planned features

- Nested Boolean searching, for instance:
  (Literacy OR Dropout rates) AND (Employee turnover OR Employee retention) AND (Number of employees)

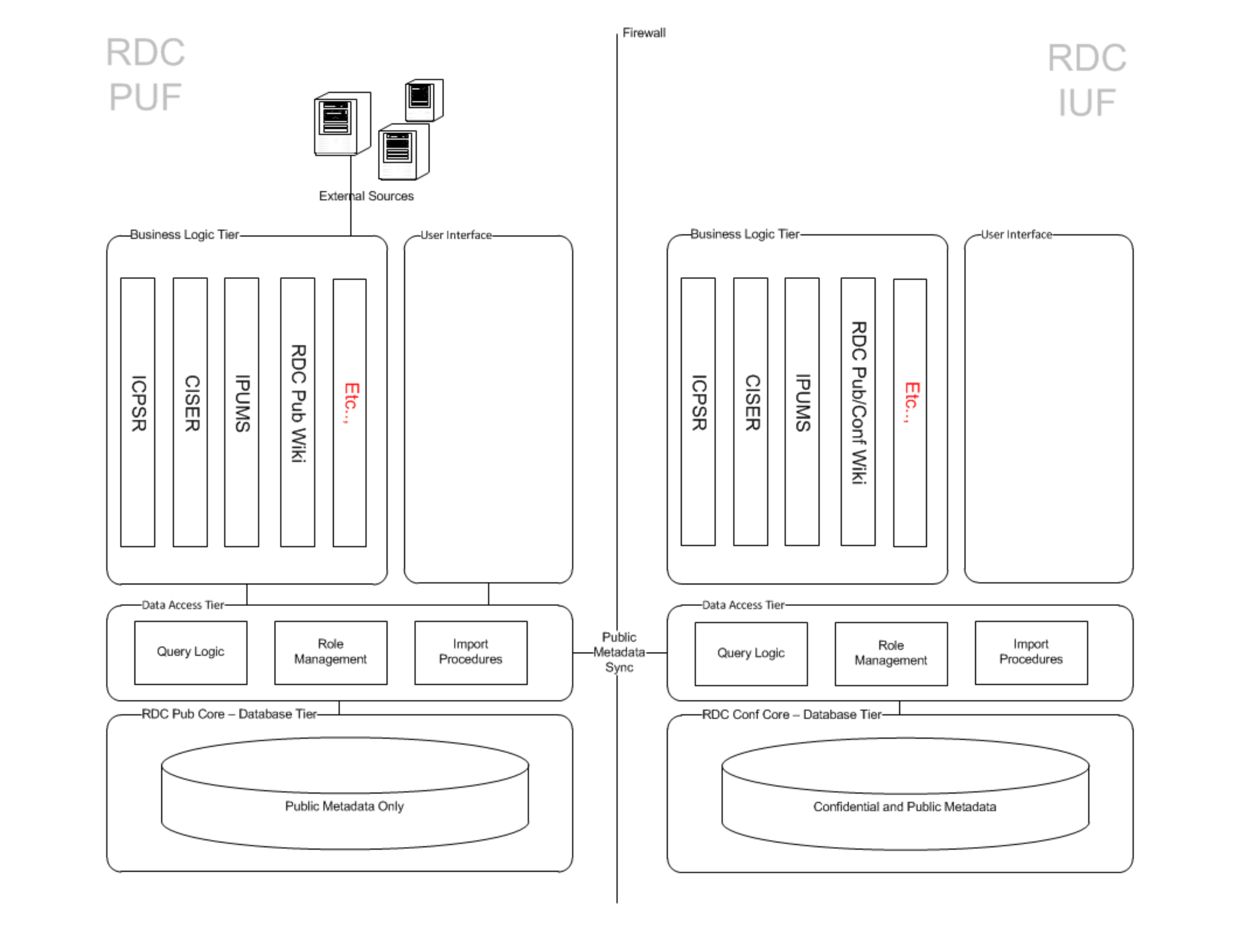
- Harnessing collective knowledge: incorporating social media (“data wikis”)

Initial CCBMR Datasets

- Longitudinal Business Database (LBD)
- American Community Survey (ACS)
- American Housing Survey (AHS)
- Longitudinal Employer-Household Dynamics (LEHD)

CISER

First draft of an enterprise application diagram



Project Collaborators

Pascal Heus – Metadata Technology

Jeremy Iverson – Colectica

Ingo Barkow – German Institute for Educational Research (DIPF)

David Schiller - Research Data Center (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)

Chuck Humphrey – University of Alberta; Canadian RDC Network

metadata technology

colectica

DIPF
Bildungsforschung und Bildungsinformation

Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)
FDZ

UNIVERSITY OF
ALBERTA

Thank you!

Questions? Comments?

William Block
block@cornell.edu