Managing Confidentiality and Provenance across Mixed Private and Publicly-Accessed Data and Metadata

Lars Vilhuber\textsuperscript{1}  John M. Abowd\textsuperscript{1}  William Block\textsuperscript{2}  
Carl Lagoze\textsuperscript{3}  Jeremy Williams\textsuperscript{2}

\textsuperscript{1}Labor Dynamics Institute, ILR,  
\textsuperscript{2}Cornell Institute for Social and Economic Research,  
\textsuperscript{3}University of Michigan

November 2013, FCSM 2013
NCRN

- This work is part of the NSF Census Research Network (NCRN) - Cornell Node ("Integrated Research Support, Training and Data Documentation")
- Funded by NSF Grant #1131848.
- For more information, see www.ncrn.cornell.edu.
Introduction

Overview of work

- Basic program outlined in Abowd, Vilhuber, and Block (PSD 2012) [3] and Lagoze, Block, Williams, Abowd, and Vilhuber, (IDCC 2013) [8]

- PROV extension described in more detail in Lagoze, Williams, Vilhuber (Metadata and Semantics Research Conference, November 2013) and Lagoze et al (European DDI User Conference, December 2013) [9]
Introduction

Some facts that motivated us

Stating the problem in the U.S. case

CED²AR: A proposed solution
  What is DDI
  DDI extension for confidentiality protection
  DDI extension for provenance tracing
Replication of research results

Critical element of science

- Replication of methods, data inputs, computational environment is a critical element of the scientific approach
- Journals, funding agencies (in the U.S.) have been moving to making archiving of inputs to scientific results more robust, even mandatory
Econometrica

“In its first issue, the editor of Econometrica (1933), Ragnar Frisch, noted the importance of publishing data such that readers could fully explore empirical results. Publication of data, however, was discontinued early in the journal’s history. [...] The journal arrived full-circle in late 2004 when Econometrica adopted one of the more stringent policies on availability of data and programs.

Problem will become worse

Increased use of restricted-access data

- Today’s young scholars pursue research programs that mandate inherently identifiable data
  - Geospatial relations,
  - Exact genome data,
  - Networks of all sorts,
  - Linked administrative records

- These researchers acquire authorized, generally unfettered, restricted access to the confidential, identifiable data and perform their analyses in secure environments.

- Archiving (curation) of input data is complicated
- Knowledge discovery is complicated
Decline in the use of classic public-use data

Use of Pre-Existing Survey Data in Publications in Leading Journals, 1980-2010

Note: "Pre-existing survey" data does not refer to surveys such as the CPS or BLS and do not include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.

Vilhuber, Abowd, Block, Lagoze, Williams
Increase in the use of administrative data in economics

Use of Administrative Data in Publications in Leading Journals, 1980-2010

Note: Administrative data refers to any dataset that was collected without directly surveying individuals (e.g., scanner data, stock prices, school district records, social security records). Sample excludes studies whose primary data source is from developing countries.

Vilhuber, Abowd, Block, Lagoze, Williams
Nature, 2012

“Many of the emerging ‘big data’ applications come from private sources that are inaccessible to other researchers. The data source may be hidden, compounding problems of verification, as well as concerns about the generality of the results.”

(Huberman, Nature 482, 308 (16 February 2012) doi:10.1038/482308d)

Other domains

- Biology (genetics data, chemical compounds)
- Computer science (search records, single-firm examples)
Introduction

Some facts that motivated us

Stating the problem in the U.S. case

CED²AR: A proposed solution
  What is DDI
  DDI extension for confidentiality protection
  DDI extension for provenance tracing
Why we think there is a problem

Core issues

a. Insufficient curation (starting with archiving)
b. No way to reference data (unique identifiers)
c. No consistent way to learn about the data (metadata dissemination)
d. Weak or non-existent provenance tracing
Generalized problem

Multiple data sources in the US

- U.S. Census Bureau (RDC) ▶ more
- Internal Revenue Service (confidential, public-use) ▶ more
- Bureau of Labor Statistics (confidential, public-use data) ▶ more

Present elsewhere?

- Canada:
  - Centre for Data Development and Economic Research (CDER: RDC-like for business data) ▶ more
  - better: Canadian RDC network ▶ more
- France: Réseau Quetelet ▶ more, Centre d’accés sécurisé distant aux données (CASD)
- Germany: IAB
Introduction

Some facts that motivated us

Stating the problem in the U.S. case

CED$^2$AR: A proposed solution

What is DDI

DDI extension for confidentiality protection

DDI extension for provenance tracing
Comprehensive Extensible Data Documentation and Access (CED²AR)

Core
We develop the core of a method for solving the data archive and curation problem that confronts the custodians of restricted-access research data and the scientific users of such data. Our solution recognizes the dual protections afforded by physical security and access limitation protocols, and allows for much improved provenance tracing.
Proposed solution

Extensible framework

- Based on existing standards (Data Documentation Initiative, DDI) with extension to accommodate disclosure protection mechanisms and provenance tracing
Proposed solution

Extensible framework

- Based on existing standards (Data Documentation Initiative, DDI) with extension to accommodate disclosure protection mechanisms and provenance tracing
- Connectors (import/export) to other sources and standards
Proposed solution

Extensible framework

- Based on existing standards (Data Documentation Initiative, DDI) with extension to accommodate disclosure protection mechanisms and provenance tracing
- Connectors (import/export) to other sources and standards
- To be filled by multiple sources of metadata (some the curators/owners, others “crowd-sourced”)

Vilhuber, Abowd, Block, Lagoze, Williams
Proposed solution

Extensible framework

- Based on existing standards (Data Documentation Initiative, DDI) with extension to accommodate disclosure protection mechanisms and provenance tracing
- Connectors (import/export) to other sources and standards
- To be filled by multiple sources of metadata (some the curators/owners, others “crowd-sourced”) 
- Interim solution for those datasets without unique identifiers (Digital Object Identifier, DOI)
Extensible framework

- Based on existing standards (Data Documentation Initiative, DDI) with extension to accommodate disclosure protection mechanisms and provenance tracing
- Connectors (import/export) to other sources and standards
- To be filled by multiple sources of metadata (some the curators/owners, others “crowd-sourced”)
- Interim solution for those datasets without unique identifiers (Digital Object Identifier, DOI)
What is DDI?
Example of DDI

<?xml version="1.0" encoding="UTF-8"?>
<codeBook xmlns="ddi:codebook:2_5">
<docDscr>
  <citation>
    <titlStmt>
      <titl>SIPP_Synthetic_Beta</titl>
      <altTitl>SSB</altTitl>
      <IDNo agency="DOI">TBD</IDNo>
    </titlStmt>
    <rspStmt>
      <AuthEnty affiliation="Cornell University">
        Virtual RDC
      </AuthEnty>
    </rspStmt>
  </citation>
</docDscr>
</codeBook>
CED²AR
The Comprehensive Extensible Data Documentation and Access Repository

SIPP Synthetic Beta

Browse Codebook (102 variables)
Prepared by: Cornell NSF-Census Research Network
Version: August 1, 2013

► Abstract

Citation Request
Please cite NSF Grant #1131848 when publishing work that makes use of this dataset.

► Confidentiality Declaration

Disclaimer

The data synthesis process employed by Census to protect the linked data from the risk of disclosing the identity of individuals is relatively new and substantially changes both the survey and administrative data. The intent of the modeling done as part of the synthesis is to preserve relationships among variables that are of interest to researchers while ensuring that personally identifiable information is not revealed to the data user. It has not been feasible to
Example DDI: ICPSR

Survey of Income and Program Participation (SIPP) 2004 Panel (ICPSR 4517)

Principal Investigator(s): United States Department of Commerce. Bureau of the Census

Summary:
This data collection is part of a longitudinal survey designed to provide detailed information on the economic situation of households and persons in the United States. These data examine the distribution of income, wealth, and poverty in American society and gauge the effects of federal and state programs on the well-being of families and individuals. There are three basic elements contained in the survey. The first is a control card that records basic social and demographic characteristics for... (more info)

Series: Survey of Income and Program Participation (SIPP) Series

Access Notes
- These data are freely available.

Dataset(s)

WARNING: Because this study has many datasets, the download all files option has been suppressed, and you will need to download one dataset at a time.

WARNING: This study is over 150MB in size and may take several minutes to download on a typical internet connection.
Example DDI: UK data archive

Vilhuber, Abowd, Block, Lagoze, Williams

Data Management of Confidential Data
Expanded DDI attributes

Standard DDI

Fragment of variable description*

```xml
<var ID="V1" dcml="0" files="F1" intrvl="discrete"
name="cur_end mar_flag">
  <location width="12"/>
  <labl>Flag: Linked marriage ended</labl>
  <valrng>
    <range UNITS="REAL" max="2" min="0"/>
  </valrng>
  <sumStat type="vald"> 123 </sumStat>
  <sumStat type="invd"> 456 </sumStat>
  <catgry>
    <catValu> 1 </catValu>
    <catStat type="freq"> 234 </catStat>
  </catgry>
</var>
```

* All values are fake
Expanded DDI attributes

Standard DDI

Fragment of variable description*

```xml
<!--var ID="V1" dcml="0" files="F1" intrvl="discrete" name="cur_end mar_flag">
  <location width="12"/>
  <labl>Flag: Linked marriage ended</labl -->
  <valrng>
    <range UNITS="REAL" max="2" min="0"/>
  </valrng>
  <!-- sumStat type="vald"> 123 </sumStat>
  <sumStat type="invd"> 456 </sumStat>
  <catgry>
    <catValu> 1 </catValu>
    <catStat type="freq"> 234 </catStat>
  </catgry -->
</!-->
```

* All values are fake
Expanded DDI attributes

Standard DDI
Fragment of variable description*

```xml
<!--var ID="V1" dcml="0" files="F1" intrvl="discrete"
name="cur_end mar_flag">
  <location width="12"/>
  <labl>Flag: Linked marriage ended</labl>
  <valrng>
    <range UNITS="REAL" max="2" min="0"/>
  </valrng -->
  <sumStat type="vald"> 123 </sumStat>
  <sumStat type="invd"> 456 </sumStat>
</catgry>
  <catValu> 1 </catValu>
  <catStat type="freq"> 234 </catStat>
</catgry -->
```

* All values are fake
**Expanded DDI attributes**

**Enhanced DDI**
Re-using existing attribute, but expanding scope.*

```xml
<var ID="V1" dcml="0" files="F1" intrvl="discrete"
  name="cur_end mar_flag">
  <location width="12"/>
  <labl>Flag: Linked marriage ended</labl>
  <valrng access="release">
    <range UNITS="REAL" max="2" min="0"/>
  </valrng>
  <sumStat access="restricted" type="vald">123</sumStat>
  <sumStat access="restricted" type="invd">456</sumStat>
</var>
```

* All values are fake

Vilhuber, Abowd, Block, Lagoze, Williams

Data Management of Confidential Data
**Expanded DDI attributes**

**Enhanced DDI**

Allows for verifiable filtering*

```xml
<var ID="V1" dcml="0" files="F1" intrvl="discrete" name="cur_end mar_flag">
  <location width="12"/>
  <labl>Flag: Linked marriage ended</labl>
  <valrng access="release">
    <range UNITS="REAL" max="2" min="0"/>
  </valrng>
<!-- sumStat suppressed -->
<!-- sumStat suppressed -->
  <catgry access="release">
    <catValu access="release">1</catValu>
    <catStat type="freq" access="restricted">
      [suppressed]
    </catStat>
  </catgry>
</var>
```

* All values are fake

Vilhuber, Abowd, Block, Lagoze, Williams
Application to confidentiality protection

Browse all Variables

Searching Synthetic Longitudinal Business Database

Uncheck any variables to be released, then press [SAVE].

Show 10 variables

<table>
<thead>
<tr>
<th>Confidential</th>
<th>Variable Name</th>
<th>Label</th>
<th>Codebook</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>act</td>
<td>dropped Activity Code</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>bestnaics</td>
<td>dropped Best NAICS code</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>bestsic</td>
<td>dropped Best SIC code</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>cbp</td>
<td>dropped</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>cfnn</td>
<td>dropped Census File Number</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>country</td>
<td>masked County FIPS codes</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>emp</td>
<td>synthetic March 12 Employment</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>firstflag</td>
<td>dropped First Link Flag</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>firstyear</td>
<td>synthetic First Year Establishment is Observed</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
</tbody>
</table>
Options

- Variable is suppressed, including all subordinate elements
- Variable description is released, but all subordinate statistical elements are suppressed (attribute of `<var> set to ”released””) [default]
- Expand all existing attributes, individually select subordinate elements to suppress (attribute of sub-element is set to ”suppressed”, content suppressed)
### Application to confidentiality protection

#### Browse all Variables

Searching Synthetic Longitudinal Business Database  
Uncheck any variables to be released, then press [SAVE].

Show ▼ 10 ▼ variables

<table>
<thead>
<tr>
<th>Confidential</th>
<th>Variable Name</th>
<th>Label</th>
<th>Codebook</th>
</tr>
</thead>
<tbody>
<tr>
<td>🗒️</td>
<td>act</td>
<td>dropped Activity Code</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td>🗒️</td>
<td>bestnaics</td>
<td>dropped Best NAICS code</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td>🗒️</td>
<td>bestsic</td>
<td>dropped Best SIC code</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td>🗒️</td>
<td>cbp</td>
<td>dropped</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td>🗒️</td>
<td>cfn</td>
<td>dropped Census File Number</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td>🗒️</td>
<td>county</td>
<td>masked County FIPS codes</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td>🗒️</td>
<td>emp</td>
<td>synthetic March 12 Employment</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td>🗒️</td>
<td>firstflag</td>
<td>dropped First Link Flag</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td>🗒️</td>
<td>firstyear</td>
<td>synthetic First Year Establishment is Observed</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
</tbody>
</table>
## Application to confidentiality protection

### Browse all Variables

Searching Synthetic Longitudinal Business Database  Uncheck any variables to be released, then press [SAVE].

Show ▼ variables

<table>
<thead>
<tr>
<th>Confidential</th>
<th>Variable Name</th>
<th>Label</th>
<th>Codebook</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>act</td>
<td>dropped Activity Code</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>bestnaics</td>
<td>dropped Best NAICS code</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>bestsic</td>
<td>dropped Best SIC code</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>cbp</td>
<td>dropped</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>cfn</td>
<td>dropped Census File Number</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>county</td>
<td>masked County FIPS codes</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>emp</td>
<td>synthetic March 12 Employment</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>firstflag</td>
<td>dropped First Link Flag</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
<tr>
<td></td>
<td>firstyear</td>
<td>synthetic First Year Establishment is Observed</td>
<td>Synthetic Longitudinal Business Database</td>
</tr>
</tbody>
</table>
Definitions

- First draft of specification in test use by our team
Definitions

- First draft of specification in test use by our team
- Full enhanced specification (based on DDI-Codebook 2.5) published on CED²AR
Definitions

- First draft of specification in test use by our team
- Full enhanced specification (based on DDI-Codebook 2.5) published on CED²AR
- Enhanced specification proposed to DDI Alliance
Definitions

- First draft of specification in test use by our team
- Full enhanced specification (based on DDI-Codebook 2.5) published on CED^2AR
- Enhanced specification proposed to DDI Alliance
- Expand to DDI-Lifecycle
The provenance problem
“data provenance, one kind of metadata, pertains to the derivation history of a data product starting from its original sources” [...] “from it, one can ascertain the quality of the data base and its ancestral data and derivations, track back sources of errors, allow automated reenactment of derivations to update the data, and provide attribution of data sources”

Support in DDI

Provenance and Metadata
Not (currently) a “native” component of DDI, closest thing is:

```xml
<xs:complexType name="othrStdyMatType">
  <xs:complexContent>
    <xs:extension base="baseElementType">
      <xs:sequence>
        <xs:element ref="relMat" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element ref="relStdy" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element ref="relPubl" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element ref="othRefs" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:extension>
  </xs:complexContent>
</xs:complexType>
```

Downside
No structure. Mostly verbose entries.
Introduction  Motivation  Problem  Solution  Additional info

Only a verbose description

Vilhuber, Abowd, Block, Lagoze, Williams

Data Management of Confidential Data
Introduction

Motivation

Problem

Solution

Additional info

---

UK Data Archive example

---

**SERIES ABSTRACT**

The Family Expenditure Survey (FES), which ran from 1961-2001, was a continuous annual survey that provided information on household and personal incomes, certain payments that recurred regularly (e.g. rent, gas and electricity bills, telephone accounts, insurances, season tickets and hire purchase payments), and included a detailed 14-day expenditure record. From 2001, the both the FES and the National Food Survey (NFS) were replaced by a new survey, the Expenditure and Food Survey (EFS), which subsequently became the Living Costs and Food Survey (LCF) from 2008.

---

**DATA ACCESS**

GN 33057 | FAMILY EXPENDITURE SURVEY, 1961-2001

---

**RELATED RESOURCES**

Related studies:

Family Resources Survey, 1979 (SN 1930)

---

Vilhuber, Abowd, Block, Lagoze, Williams

---

Data Management of Confidential Data
PROV model
W3C PROV Model based in the notions of

1. **entities** that are physical, digital, and conceptual things in the world;

2. **activities** that are dynamic aspects of the world that change and create entities; and

3. **agents** that are responsible for activities.

4. a set of **relationships** that can exist between them that express attribution, delegation, derivation, etc.
Incorporating PROV (LBD)

LBD Provenance

- Business Register (SSEL)
- Economic Census (quinquennial)
- Longitudinal Business Database (LBD)
- County Business Patterns (CBP)
- Annual Survey of Manufactures (ASM)
- Synthetic LBD (SynLBD)
- Business Dynamics Statistics (DDS)

Some exclusions
Enumeration frame
New establishments
Sample frame
Possible linkage
Possible linkage

Jarmin and Miranda (2002)

Vilhuber, Abowd, Block, Lagoze, Williams
Incorporating PROV (LBD)

LBD Provenance

Vilhuber, Abowd, Block, Lagoze, Williams
entity(cdr:LBD, [prov:type='cdr:dataset', prov:label="Longitudinal Business Data'])
entity(cdr:synthLBD, [prov:type='cdr:dataset', prov:label="Synthetic LBD"])
entity(cdr:BDS, [prov:type='cdr:dataset', prov:label="Business Dynamics Statistics"])
entity(cdr:BR, [prov:type='cdr:dataset', prov:label="Business Register"])
entity(cdr:cbp, [prov:type='cdr:dataset', prov:label="County Business Patterns"])
entity(cdr:asm, [prov:type='cdr:dataset', prov:label="Annual Survey of Manufacturers"])
entity(cdr:ec, [prov:type='cdr:dataset', prov:label="Economic Census"])
entity(cdr:jm, [prov:type='prov:Plan', prov:label="Jarmin Miranda 2002"])
entity(cdr:synthPlan, [prov:type='prov:Plan', prov:label="synthetic plan"])
entity(cdr:tax, [prov:type='cdr:dataset', prov:label="IRS Tax Records"])

agent(cdr:USCB, [prov:type='prov:Organization', prov:label="US Census Bureau"])
agent(cdr:CES, [prov:type='prov:Organization', prov:label="Center for Economic Studies"])
agent(cdr:IRS, [prov:type='prov:Organization', prov:label="Internal Revenue Service"])
agent(cdr:autoMatch, [prov:type='prov:SoftwareAgent'])
agent(cdr:SAS, [prov:type='prov:SoftwareAgent'])
agent(cdr:ESMPD, [prov:type='prov:SoftwareAgent',
  prov:label="Economic Statistical Methods and Programming Division"])

activity(cdr:synth, [prov:label="anonymize"])
activity(cdr:aggr, [prov:label="aggregate"])
activity(cdr:procLBD, [prov:label="process LBD"])
activity(cdr:ad, [prov:label="aggregation/disclosure protection"])
activity(cdr:asmSurvey, [prov:label="ASM Survey"])
activity(cdr:ecs, [prov:label="economic census survey"])

Vilhuber, Abowd, Block, Lagoze, Williams

Data Management of Confidential Data
The key PROV element embedded as DDI/XML

```xml
<stdyDscr> <!-- Standard DDI 2.5 -->
<othrStdyMat> <!-- Standard DDI 2.5 -->
<relStdy> <!-- Standard DDI 2.5 -->
  <!-- From here, PROV additions -->
  <prov:wasDerivedFrom>
    <prov:generatedEntity prov:ref="cdr:LBD"/>
    <prov:usedEntity prov:ref="cdr:BR"/>
  </prov:wasDerivedFrom>
  <prov:wasAssociatedWith>
    <prov:activity prov:ref="cdr:procLBD"/>
    <prov:agent prov:ref="cdr:CES"/>
    <prov:plan prov:ref="cdr:procLBDPlan"/>
  </prov:wasAssociatedWith>
</relStdy> <!-- Standard DDI 2.5 -->
</othrStdyMat> <!-- Standard DDI 2.5 -->
</stdyDscr> <!-- Standard DDI 2.5 -->
```
Additional PROV elements
These could be derived from existing DDI elements (still being developed)

```xml
<!-- Entities -->
<prov:entity prov:id="cdr:BR">
  <dct:title>Business Register</dct:title>
</prov:entity>
<!-- Plans = Methodology -->
<prov:plan prov:id="cdr:procLBDPlan">
  <prov:location xsi:type="xsd:anyURI">
    http://ideas.repec.org/p/cen/wpaper/02-17.html
  </prov:location>
  <prov:type>prov:Plan</prov:type>
</prov:plan>
```
More details forthcoming


- Lagoze, Williams, Vilhuber, Block “Encoding Provenance of Social Science Data: Integrating PROV with DDI”, accepted for 5th Annual European DDI User Conference (December 2013)
Usage scenario

CED²AR
The Comprehensive Extensible Data Documentation and Access Repository

Search

Searching Longitudinal Business Database, National QWI, Synthetic Longitudinal Business Database

Show 10 variables

Vilhuber, Abowd, Block, Lagoze, Williams
Usage scenario

Vilhuber, Abowd, Block, Lagoze, Williams
Usage scenario
**Usage scenario**

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>Description</th>
<th>Concept</th>
<th>Codebook</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp</td>
<td>March 12</td>
<td>Employment</td>
<td>Longitudinal Business Database</td>
<td></td>
</tr>
<tr>
<td>qwi_f</td>
<td>QWI FQ</td>
<td>employment</td>
<td>National QWI</td>
<td></td>
</tr>
<tr>
<td>emp</td>
<td>(synthetic)</td>
<td>Paid employment consists of full and part-time employees, including salaried</td>
<td>Synthetic Longitudinal Business</td>
<td></td>
</tr>
<tr>
<td></td>
<td>March 12</td>
<td>officers and executives of corporations, who we... [more]</td>
<td>Database</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Employment</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Highlighting provenance

### CED²AR

The Comprehensive Extensible Data Documentation and Access Repository

### Table

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>Description</th>
<th>Concept</th>
<th>Codebook</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp</td>
<td>March 12 Employment</td>
<td>Paid employment consists of full and part-time employees, including salaried officers and executives of corporations, who we... more</td>
<td>Synthetic Longitudinal Business Database</td>
<td>LBD National QWI</td>
</tr>
<tr>
<td>qwi_f</td>
<td>QWE FQ employment</td>
<td></td>
<td>National QWI</td>
<td></td>
</tr>
<tr>
<td>emp</td>
<td>(synthetic) March 12 Employment</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CED^2AR next steps

- Formalize the DDI extension
CED$^2$AR next steps

- Formalize the DDI extension
- Provide implementation outside of Census Bureau
CED\textsuperscript{2}AR next steps

- Formalize the DDI extension
- Provide implementation outside of Census Bureau
- Test implementation within the Census RDC
Thank you

- [3] for more details
- Labor Dynamics Institute
- VirtualRDC @ Cornell
- NCRN Cornell website
Data Management of Confidential Data
Extra slides

Census Bureau
IRS
BLS
CDER
CRDC
France
Dataset usage in Census RDC

1,505 project-dataset pairs

Many projects use multiple datasets.
Economic (business) datasets

- 71% of datasets are business (economic) datasets
- Primarily establishment-based records from the Economic Censuses and Surveys, the Business Register, and the Longitudinal Business Database (LBD)
- They form the core of the modern industrial organization studies [5, 11] as well as modern gross job creation and destruction in macroeconomics [4, 6].
- But there are no public-use micro-data for these establishment-based products
- Exception: recently-released Synthetic LBD [2, 7]
- Currently no active curation (of derived datasets) [a], no way to reference [b], convoluted way to learn about the data structure [c*]
LEHD data

Linked employer-employee data

► Longitudinal and cross-sectional detail
► New confidentiality protection methodologies [1, 10] have unlocked large amounts of data for public-use: highly detailed local area tabulations exist based on the LEHD data
► But: no public-use micro-data exist for this longitudinal job frame or any of its derivative files.
► Confidential data are dynamic (quarterly changes)
► Currently some active curation (archiving, 10-yr!) [a*], no way to reference (publicly) [b*], convoluted way to learn about the data structure [c*]
Internal Revenue Service/ Social Security Administration

- New projects (Chetty et al, 2012; von Wachter and co-authors) have created and/or used linked longitudinal data at the IRS or the Social Security Administration.
- Neither agency has long-run experience at the statistical data curation function [a], (meta)data dissemination [b,c].
- Although both IRS and SSA have produced statistical tables for a long time.
Not unique to Census Bureau

Bureau of Labor Statistics

- Long history of making time-series available
- Limited access to microdata at the BLS
- Unknown curation [a]
- Even for public-use data, no way to reference specific releases [b]
- No well-established way to learn about microdata [c]
Data Sets

A number of business micro databases can be accessed at CDER. Key databases are listed below. For more documentation on each of the databases, or documentation on other databases, please contact CDER at cdrer@statcan.gc.ca.

<table>
<thead>
<tr>
<th>Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual Survey of Manufacturing</td>
</tr>
<tr>
<td>Annual Survey of Manufacturing – Export and Import Registry Database</td>
</tr>
<tr>
<td>Canada Border Service Agency Customs Database</td>
</tr>
<tr>
<td>Capital and Investment Program</td>
</tr>
<tr>
<td>Longitudinal Employment Analysis Program</td>
</tr>
<tr>
<td>Longitudinal Worker File</td>
</tr>
<tr>
<td>National Accounts Longitudinal Microdata File</td>
</tr>
<tr>
<td>T2-LEAP</td>
</tr>
<tr>
<td>T2-LEAP-Export and Import Registry Database</td>
</tr>
<tr>
<td>Survey of Financing of Small and Medium Enterprises</td>
</tr>
<tr>
<td>Survey of Innovation and Business Strategies</td>
</tr>
<tr>
<td>Workplace Employee Survey</td>
</tr>
</tbody>
</table>

Annual Survey of Manufactures (ASM)

The ASM is a survey that covers all manufacturing locations together with associated head offices, sales offices and auxiliary units which have been classified to the manufacturing industries. Details of the industries include turnover, annual sales, cost of materials and supplies used, ...
Surveys available in the RDCs

The following data sets are currently available at the RDCs. For additional sources of data please refer to Statistics Canada Products and Services.

To read a short description about a specific survey used at the RDCs, click on the survey details.

To access detailed documentation on a specific survey used at the RDCs, click on the appropriate cycle or year. Many of the surveys below have multiple cycles. The links below will take you to the most recent cycle or wave released. Please select "Other reference period" in the "Definitions, Data Sources and Methods Pages" for links to documentation for the earlier cycles.

<table>
<thead>
<tr>
<th>Record Number</th>
<th>Survey Name</th>
<th>Acronym</th>
</tr>
</thead>
<tbody>
<tr>
<td>5108</td>
<td>Aboriginal Children’s Survey</td>
<td>ACS</td>
</tr>
<tr>
<td>3250</td>
<td>Aboriginal Peoples Survey</td>
<td>APS</td>
</tr>
<tr>
<td>3879</td>
<td>Adult Education and Training Survey</td>
<td>AETS</td>
</tr>
<tr>
<td>3207</td>
<td>Canadian Cancer Registry</td>
<td>CCR</td>
</tr>
<tr>
<td>3226</td>
<td>Canadian Community Health Survey - Annual Component</td>
<td>CCHS</td>
</tr>
<tr>
<td>5015</td>
<td>Canadian Community Health Survey – Mental Health</td>
<td>CCHS</td>
</tr>
<tr>
<td>5049</td>
<td>Canadian Community Health Survey - Nutrition</td>
<td>CCHS</td>
</tr>
<tr>
<td>5146</td>
<td>Canadian Community Health Survey – Healthy Aging</td>
<td>CCHS</td>
</tr>
<tr>
<td>5071</td>
<td>Canadian Health Measures Survey</td>
<td>CHMS</td>
</tr>
<tr>
<td></td>
<td>Biobank</td>
<td></td>
</tr>
<tr>
<td>4440</td>
<td>Canadian Tobacco Use Monitoring Survey</td>
<td>CTUMS</td>
</tr>
<tr>
<td></td>
<td>Census of Population</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Additional documentation</td>
<td></td>
</tr>
<tr>
<td>4508</td>
<td>Ethnic Diversity Survey</td>
<td>EDS</td>
</tr>
<tr>
<td></td>
<td>- User Guide</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Codebook</td>
<td></td>
</tr>
<tr>
<td>3504</td>
<td>Survey of Family Expenditures - Income</td>
<td>EFIE</td>
</tr>
</tbody>
</table>
... but also not perfect

Attempt to access data information on General Social Survey

Access forbidden! / Accès interdit !

Access forbidden DLI!

This web module may only be accessed from the institutional networks of Canadian postsecondary institutions participating in the Data Liberation Initiative (DLI). If you are a student or a member of a participating institution and you are unable to access these pages through your institutional network, please inform the DLI contact at your institution.

Accès interdit IDD !

L'accès à ce module Web est restreint aux réseaux institutionnels des établissements postsecondaires canadiens membres de l'Initiative de démocratisation des données (IDD). Si vous êtes un étudiant ou personnel d'un établissement membre de l'IDD et vous ne réussissez pas à accéder à ce module par le biais de votre réseau institutionnel, veuillez informer la personne-ressource de l'IDD à votre établissement.
Extra slides

Réseau Quetelet

Vilhuber, Abowd, Block, Lagoze, Williams


