Data Management of Confidential Data

Carl Lagoze
University of Michigan School of Information

William C. Block, Jeremy Williams
Cornell Institute Social and Economic Research, Cornell University

John Abowd, Lars Vilhuber
Labor Dynamics Institute, Cornell University

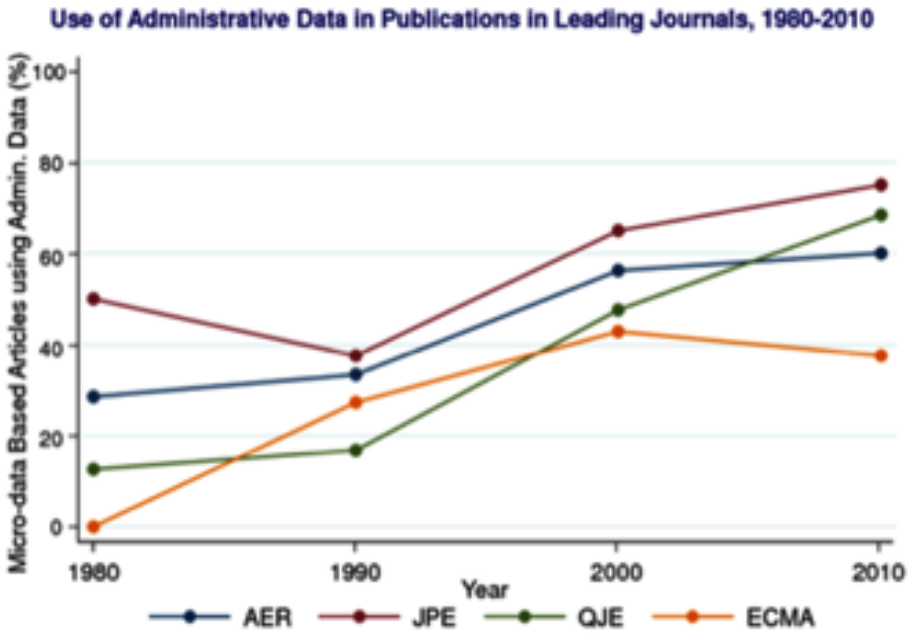
Cornell NSF Census Research Network (NCRN)

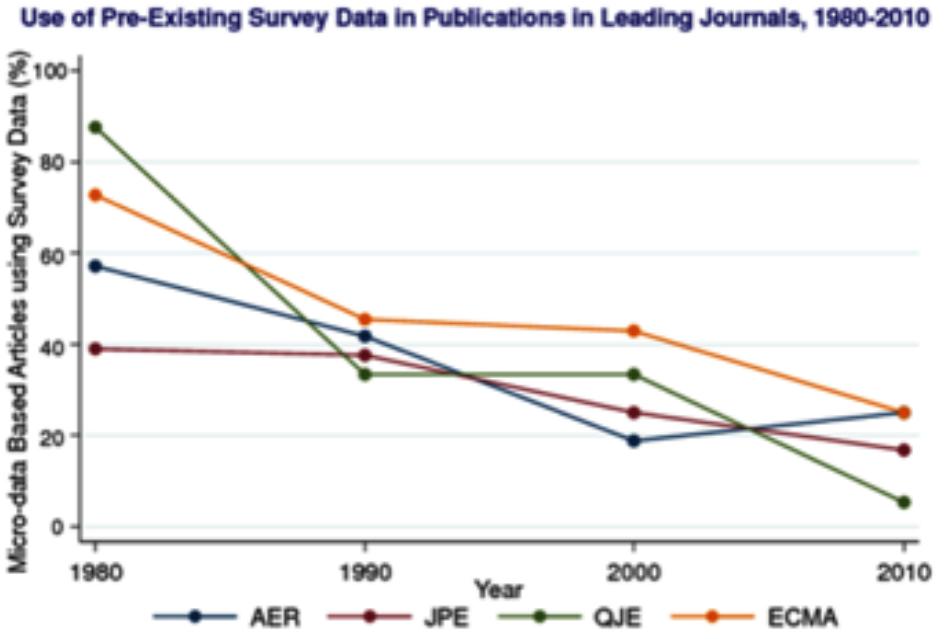
Motivation: Replicating of research results

- Replication of methods, data inputs, computational environment is a critical element of the scientific approach
- Various stakeholders in science including journals, funding agencies (in the US), learned societies have been moving to make archiving of inputs to scientific results more robust, even mandatory

Problem:
increased use of restricted-access data

Increasing number of scholars pursuing research programs that mandate inherently identifiable data (e.g., geospatial relations, exact genome data, etc.)





Carl Lagoze
IDCC – 16 January 2013

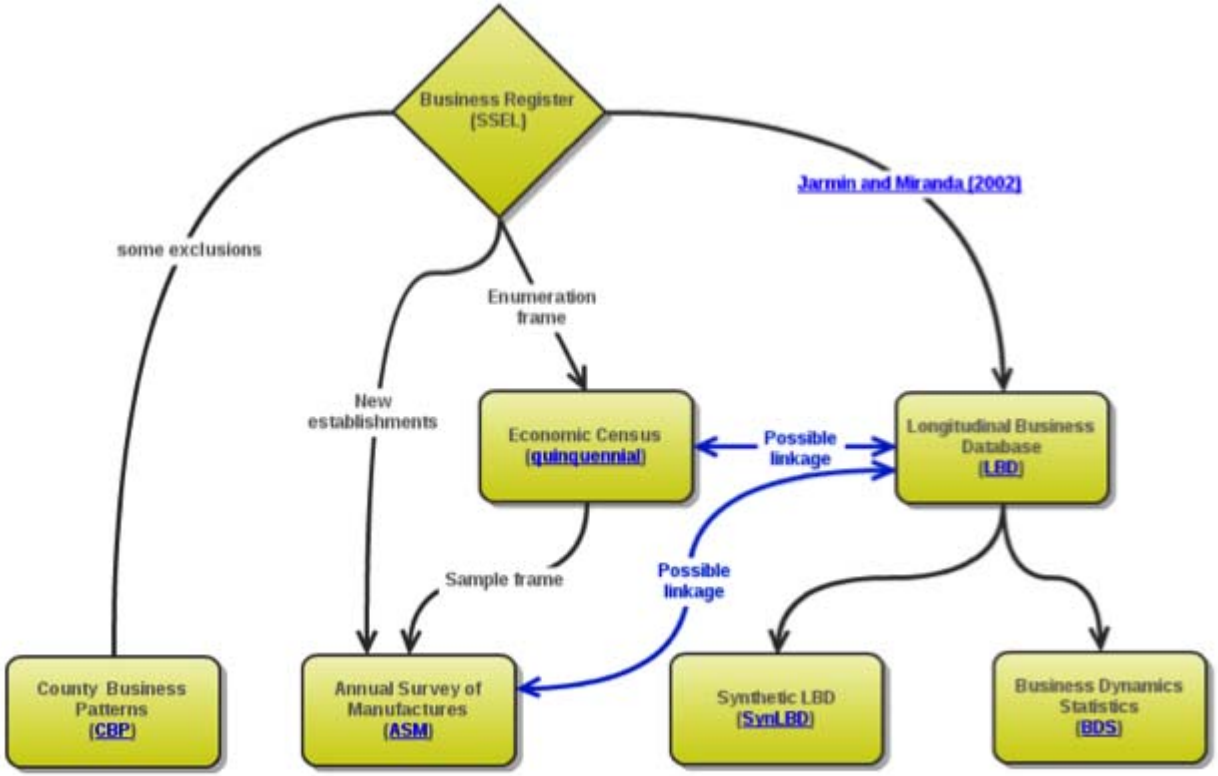
Problem is not limited to economics and social science

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 (16th February 2012)

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

Restricted access data in the provenance chain complicates the curation and knowledge discovery process

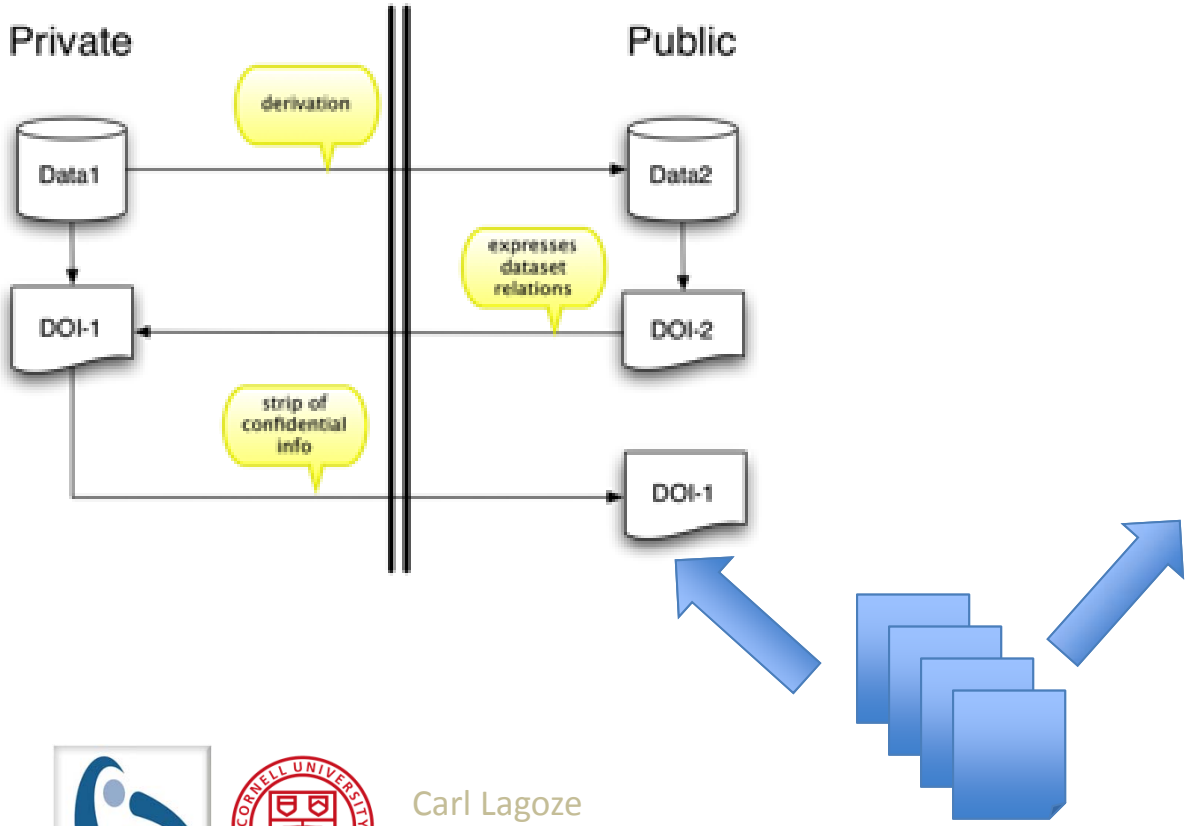


LBD provenance and related data

Carl Lagoze

IDCC – 16 January 2013

Metadata and data, inside and outside the firewall



Carl Lagoze
IDCC – 16 January 2013

Summary of problem

- Inadequate curation of secure data sets
- Inconsistent or nonexistent identification
- Need for selective hiding of data & metadata

Comprehensive Extensible Data Documentation and Access Repository (CED^{2}AR)

- Collect and standardize disparate metadata into a single DDI repository
- Provide a web interface for researchers to access
- Build an API for developers to use
- Use open standards
- Enhanced metadata with cloaking and relationship information

Entity/Identifier Strategy

- DOI's – EZID
- Use-case driven entity definition – metadata level rather than file level
- Opaque DOI - e.g., versions
  - doi:10.5072/FK2M327JW - doi:10.5072/GR2M11PT
  - NOT: doi:10.5072/FK2M327JW/V1 - doi:10.5072/FK2M327JW/V2


Metadata cloaking requirements

- Variable level
- Value level

Expressing cloaking rules

```
<stdyDscr>
    <citation> [8 lines]
    <dataAccs ID="A1">
        <useStmt>
            <conditions>Public</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A2">
        <useStmt>
            <confDec>To download this dataset, the user must obt    </confDec>
            <conditions>Confidential</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A3">
        <useStmt>
            <confDec>You're never gonna see this data.</confDec>
            <conditions>Need to know</conditions>
        </useStmt>
    </dataAccs>
</stdyDscr>
```

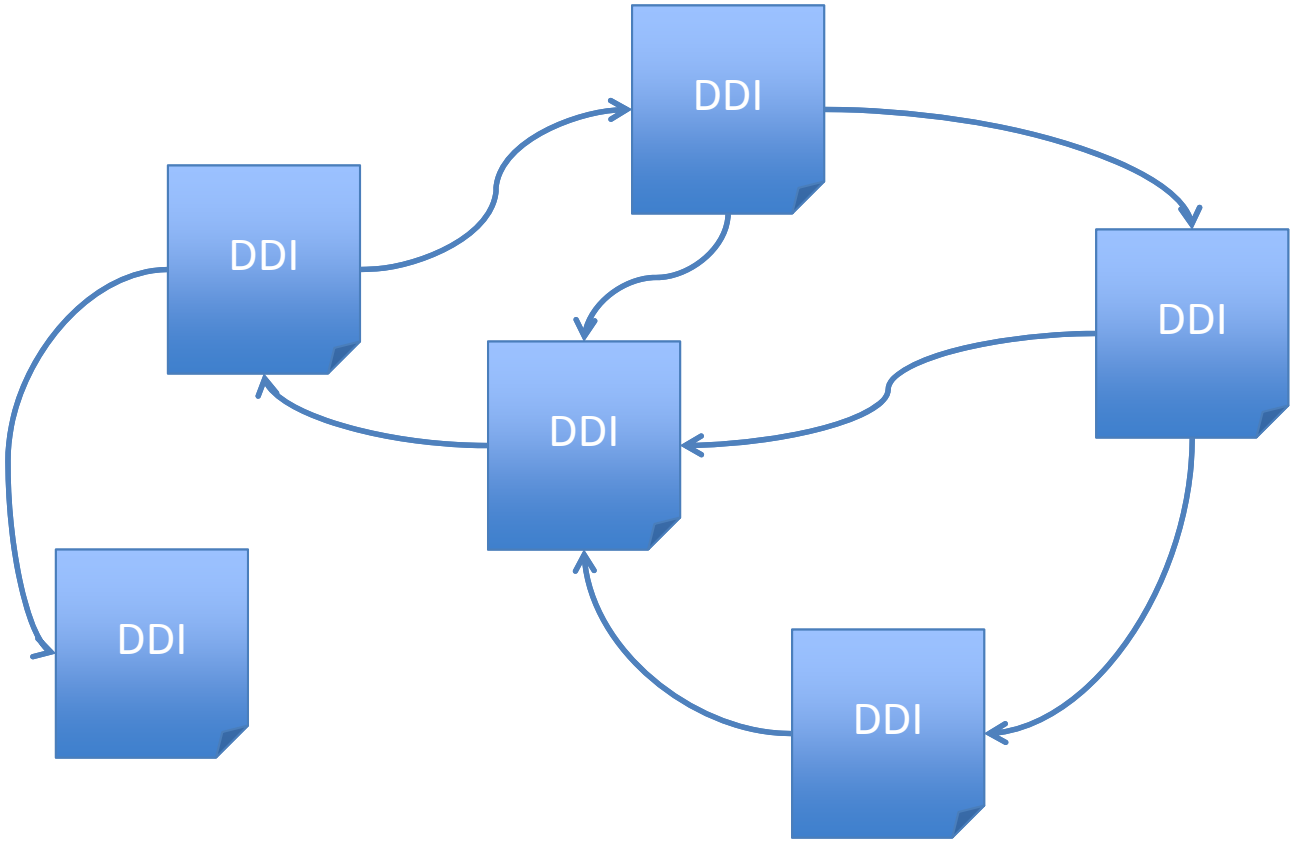
Variable level

```
<var ID="V1500" dcml="0" files="F3" intrvl="discrete" name="totfam_kids" access="A1">
    <location width="12"/>
    <labl>Total Number of Children in Family</labl>
    <valrng> [2 lines]
    <sumStat type="vald">1000</sumStat>
    <sumStat type="invd">0</sumStat>
    <varFormat schema="other" type="numeric"/>
</var>
<var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A2">
    <location width="12"/>
    <labl>Total Personal Income</labl>
    <valrng> [2 lines]
    <sumStat type="vald">240</sumStat>
    <sumStat type="invd">760</sumStat>
    <sumStat type="min">-278.739</sumStat>
    <sumStat type="max">39515.631</sumStat>
    <sumStat type="mean">1861.779</sumStat>
    <sumStat type="stdev">4015.033</sumStat>
    <varFormat schema="other" type="numeric"/>
</var>
```

Value level

```
<var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A1">
    <location width="12"/>
    <labl>Total Personal Income</labl>
    <catgry>
        <catValu>0</catValu>
        <labl>5-25k</labl>
    </catgry>
    <catgry>
        <catValu>1</catValu>
        <labl>25-75k</labl>
    </catgry>
    <catgry>
        <catValu>2</catValu>
        <labl>75-125k</labl>
    </catgry>
    <catgry>
        <catValu>3</catValu>
        <labl>125-250k</labl>
    </catgry>
    <catgry access="A2">
        <catValu>4</catValu>
        <labl>250k+</labl>
    </catgry>
    <varFormat schema="other" type="numeric"/>
</var>
```

Dataset relations
<RelStdy>



Carl Lagoze
IDCC – 16 January 2013