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

Encoding Provenance of Social Science Data: Integrating PROV with DDI

Carl Lagoze,[1] Jeremy Williams, [2] Lars Vilhuber, [3] and William C. Block,[2]

[1] School of Information, University of Michigan
[2] Cornell Institute Social and Economic Research, Cornell University
[3] Labor Dynamics Institute, Cornell University

Presentation at the 5th Annual European DDI User Conference (EDDI13)
Paris, France
4 December 2013

Cornell University

Outline

The Problem: replication vs. security

Background/Previous Work

Integrating PROV with DDI

Questions and Discussion

Our motivation: replication of research results critical to scientific advancement

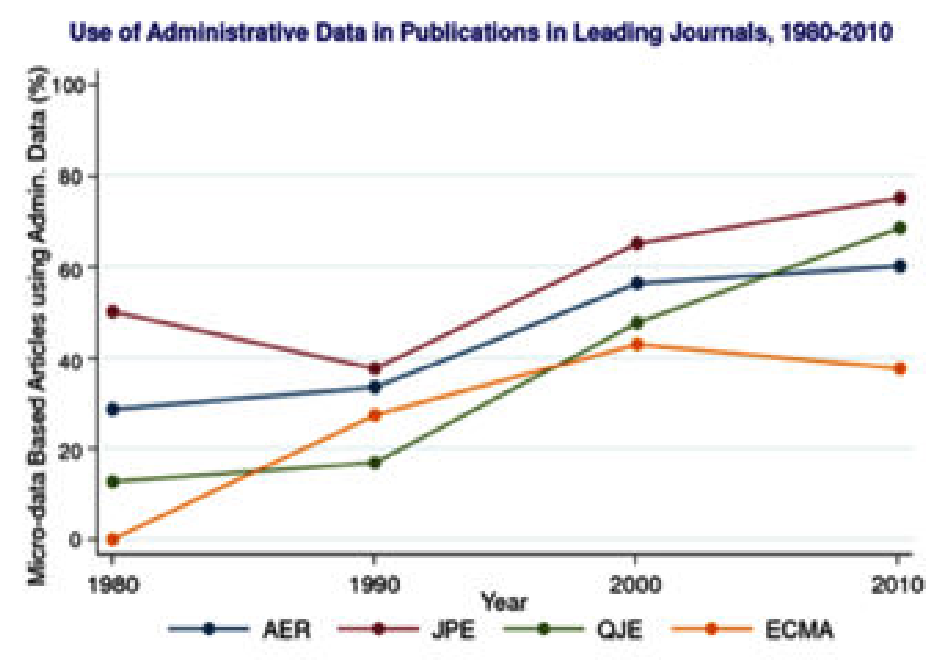
- Scientific Method 101: theory, hypothesis testing, replication
- Data inputs, methods, and computation: *all* essential elements of the scientific approach
- Encouraging: Increased interest in archiving scientific data inputs (e.g., academic journals, U.S. funding agencies, Data Management Plans, etc.)

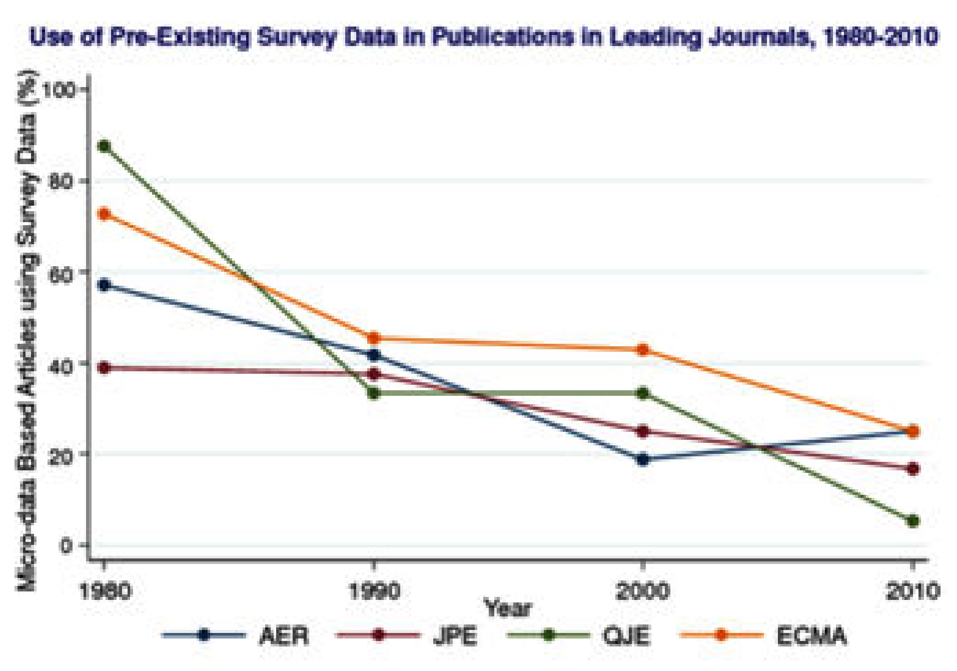
CISER

Problem: Replication relies on good data curation (difficult in a secure environment)

- By definition: Access is *Restricted*
- Lack of data curation blocks future discovery and access
- Replication of results becomes increasing difficult
- Critically important! The Scientific Method depends on the ability to replicate the results of research

Increasing number of scholars pursuing research programs that mandate inherently identifiable data





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)

Familiar problem to EDDI: examples from North America and Europe

- North America:
  - U.S.: Census Bureau (including RDC), Internal Revenue Service, Bureau of Labor Statistics (confidential, public-use data)
  - Canada: Centre for Data Development and Economic Research (CDER: RDC-like for business data); Statistics Canada, Canadian RDC network
- Europe:
  - France: R´eseau Quetelet, Centre d’acc´es s´ecuris ´e
- distant aux donn´ees (CASD)
  - Germany: IAB

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

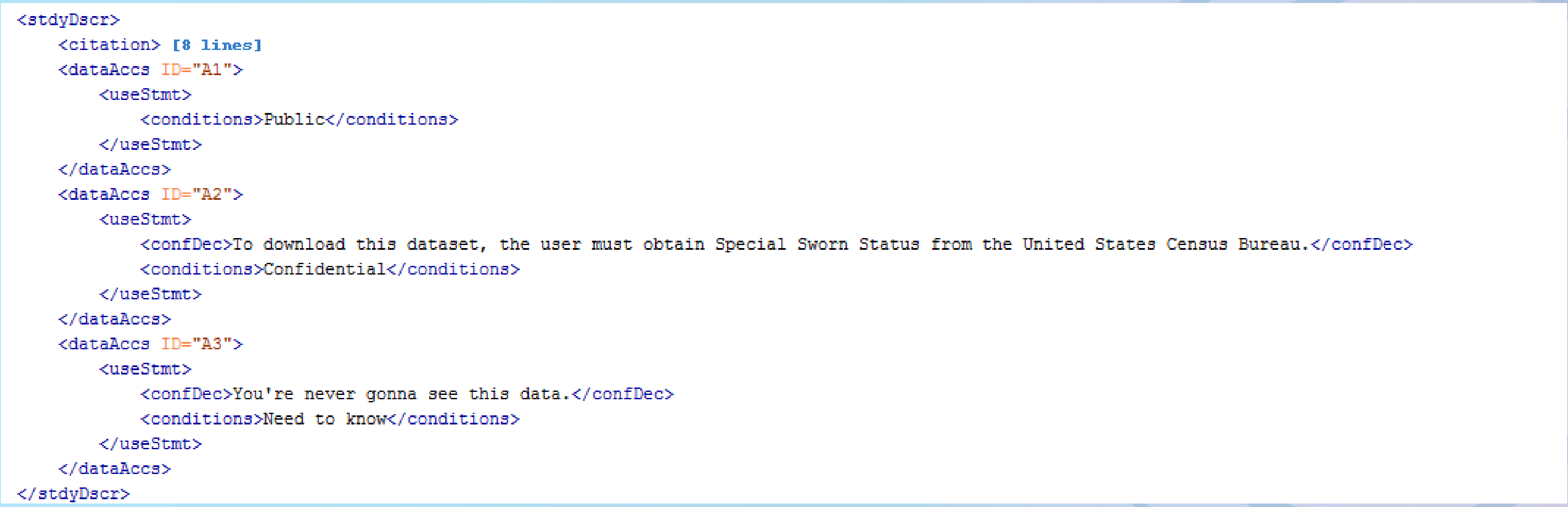
We develop the core of a method for solving the data archive and curation problem that confronts the custodians of restricted-access research data and the scientific users of such data. Our solution recognizes the dual protections afforded by **physical security and access limitation protocols,** and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

NSF-Census Research Network (NCRN) – Cornell Node (“Integrated Research Support, Training and Documentation”)

- CED^{2}AR is one part of this project
- Funded by NSF Grant #1131848.
- For more information, see www.ncrn.cornell.edu.

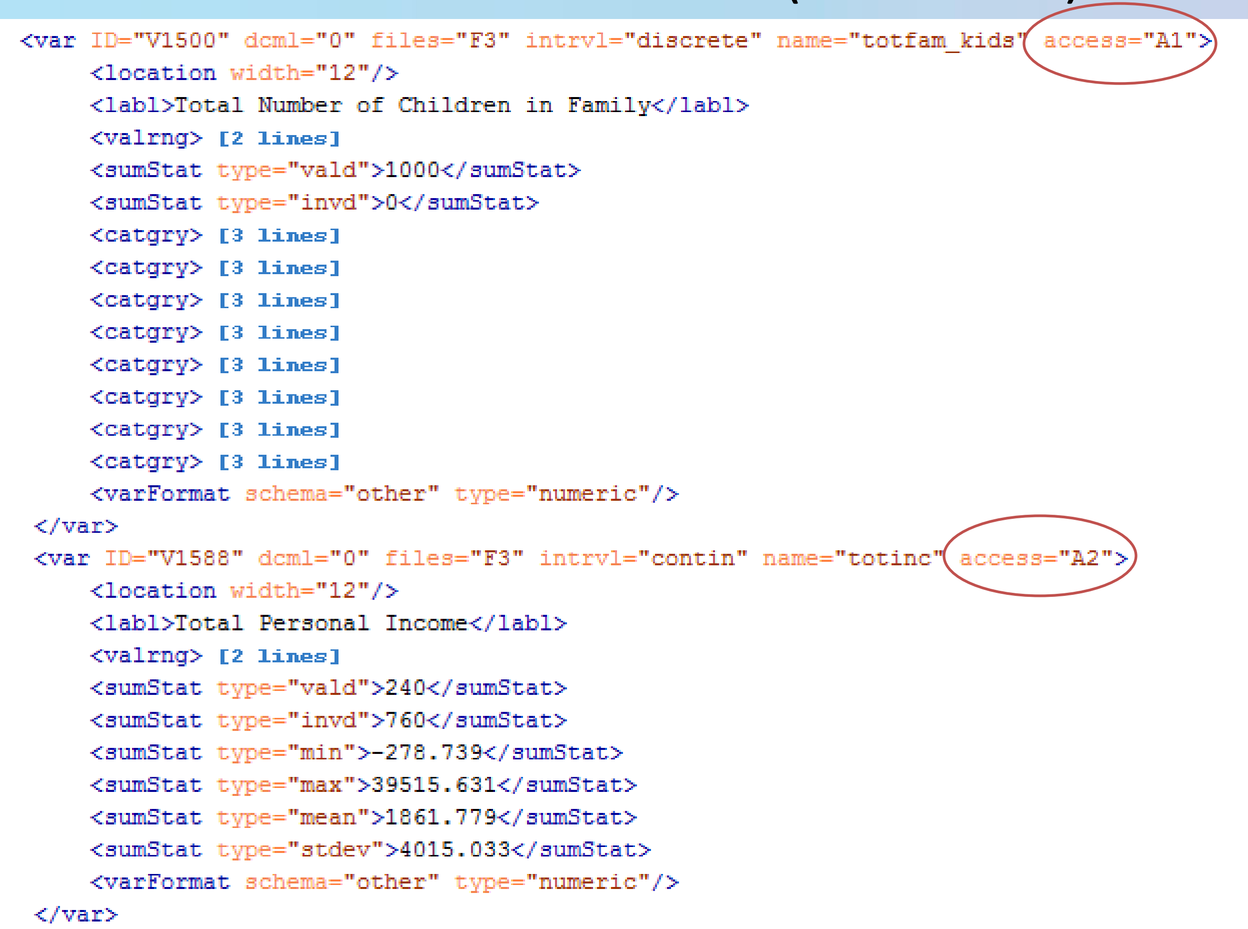
Proposed a <dataAccs> Solution at EDDI12 in Bergen

NCRN DDI Solution at the Variable Level: <dataAccs>



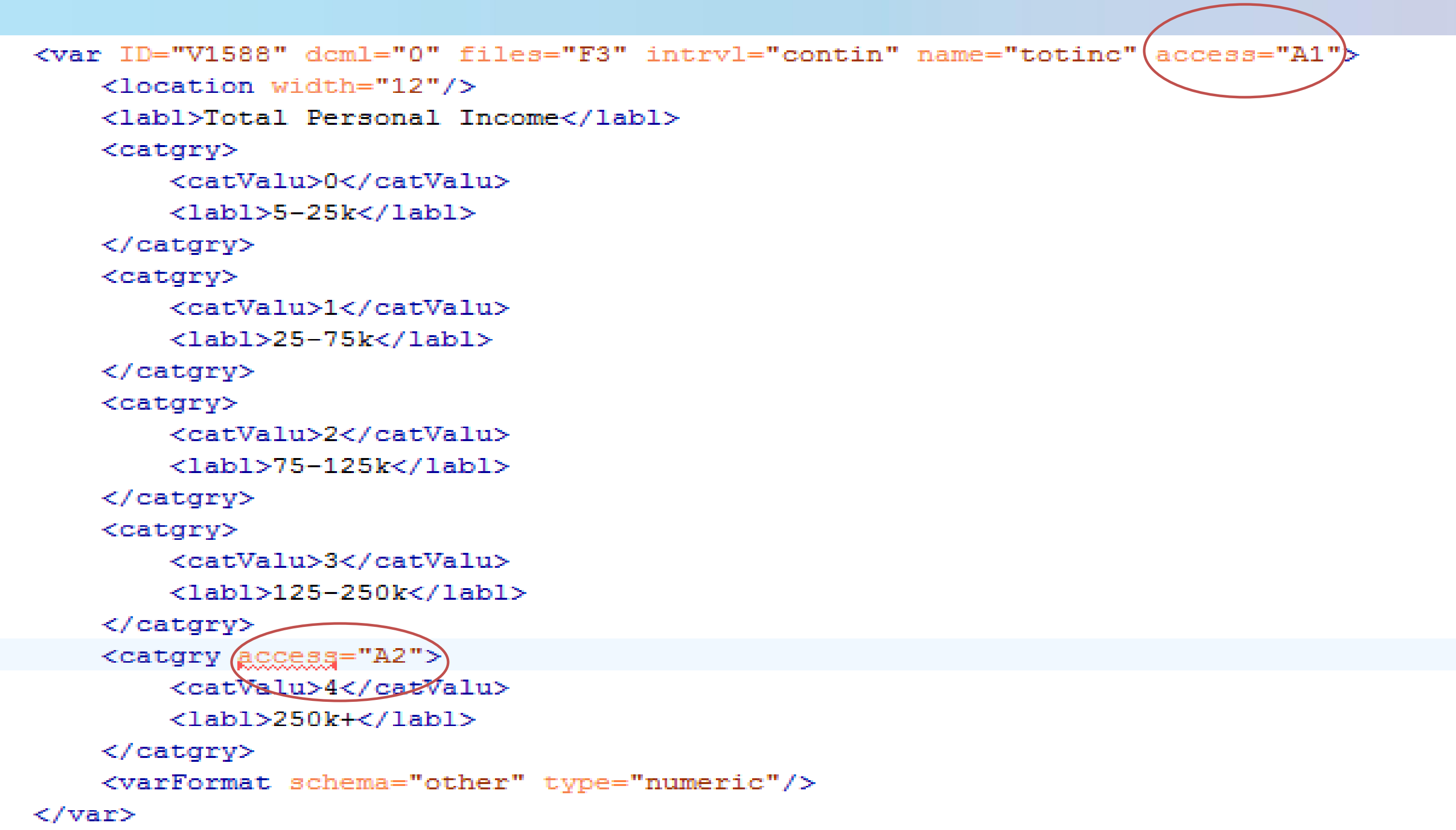
```
<stdyDscr>
    <citation> [8 lines]
    <dataAccs ID="A1">
        <useStmt>
            <conditions>Public</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A2">
        <useStmt>
            <confDec>To download this dataset, the user must obtain Special Sworn Status from the United States Census Bureau.</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 Solution (continued)



```
<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>
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <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>
```

No DDI Solution at the level of a *Value Label*



```
<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>
```

Small tweak to the DDI Codebook Schema would fix this.

<dataAccs> developments since EDDI12

- In Lagoze, Block et.al. (2013) we more completely described the solution for embedding field-specific and value-specific cloaking in DDI Metadata*
- Proposed formal change to DDI 2.5 (April 2013)
- Brought modified “DDI 2.5.NCRN” schema online for testing (Fall 2013)
- Look forward to DDI Technical Implementation Committee taking up our proposal; we have learned a lot since EDDI12

*Lagoze, C., Block, W., Williams, J., Abowd, J. M., & Vilhuber, L. (2013). Data Management of Confidential Data. In *International Data Curation Conference*. Amsterdam.

Provenance

“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”*

*Simmhan, Plale, and Gannon, “A survey of data provenance in e-science,” ACM Sigmod Record, 2005

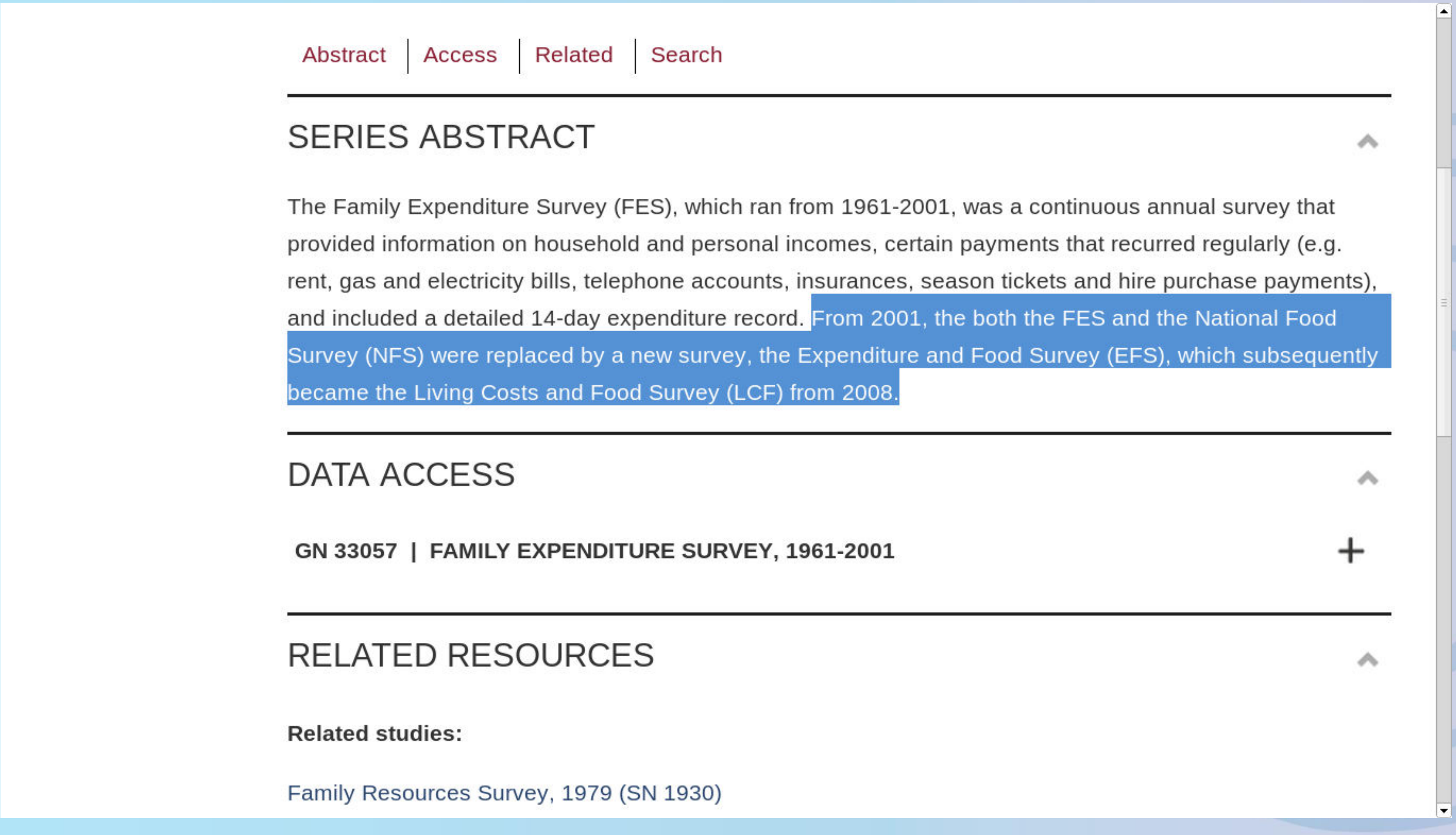
Provenance and Metadata

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

```
<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.

Provenance example from the UK Data Archive



Abstract | Access | Related | Search

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)

Earlier 2013 work (Lagoze, Williams, et. al) explored encoding PROV in RDF/XML*

- Required use of CDATA tag to avoid interfering with schema compliance; deemed less promising
- Interested in the RDF encoding for DDI effort (e.g., Bosch, Cyganiak, Gregory, and Wackerow 2013)**

*Lagoze, C., Williams, J., & Vilhuber, L. (2013). Encoding Provenance Metadata for Social Science Datasets. In *7th Metadata and Semantics Research Conference*. Thessaloniki.

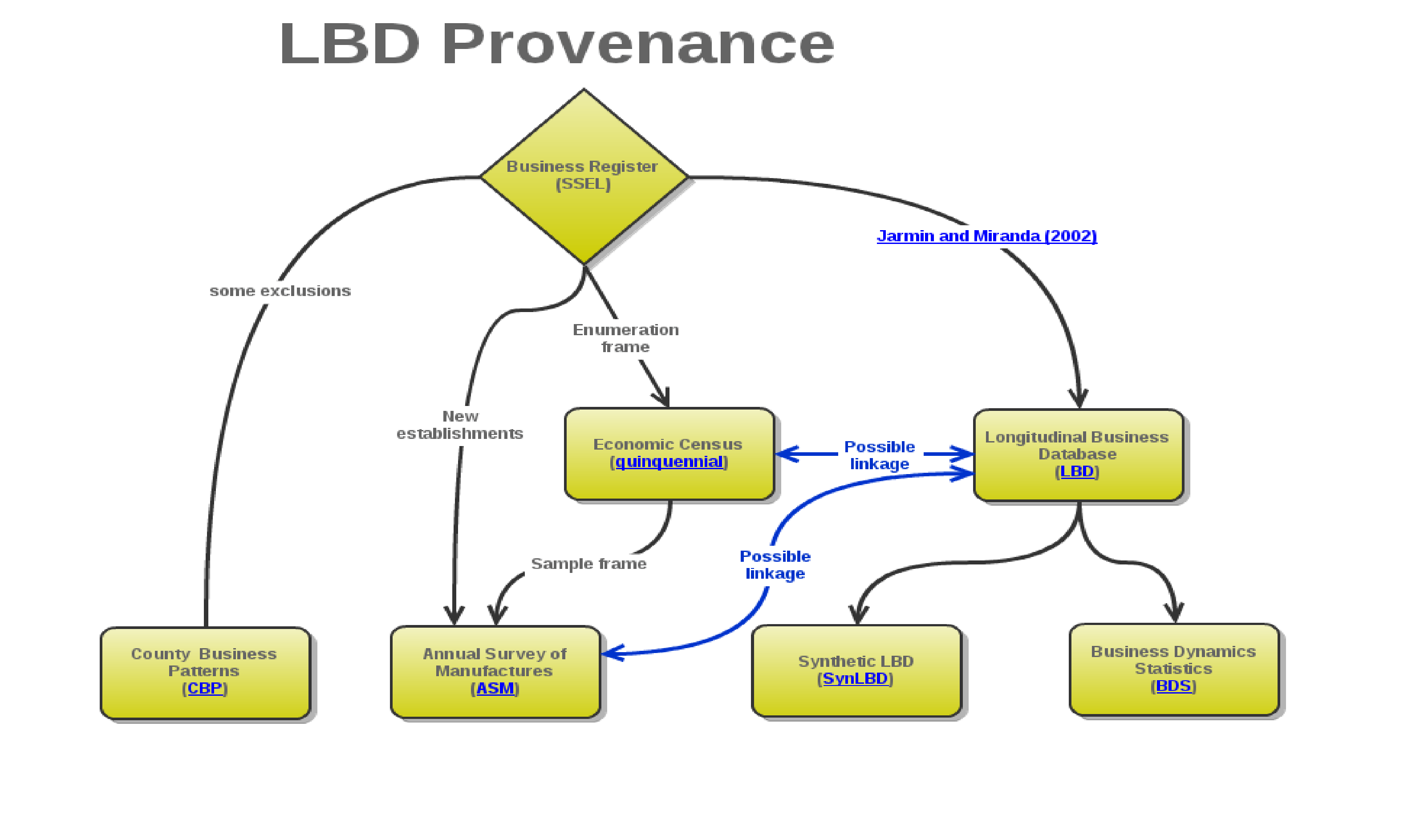
**Bosch, T., Cyganiak, R., Gregory, A., & Wackerow, J. (2013). DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data. In *Linked Data on the Web Workshop*. Rio de Janeiro

More recently: exploring W3C PROV Model as basis for encoding provenance metadata in DDI

W3C PROV Model is based upon:

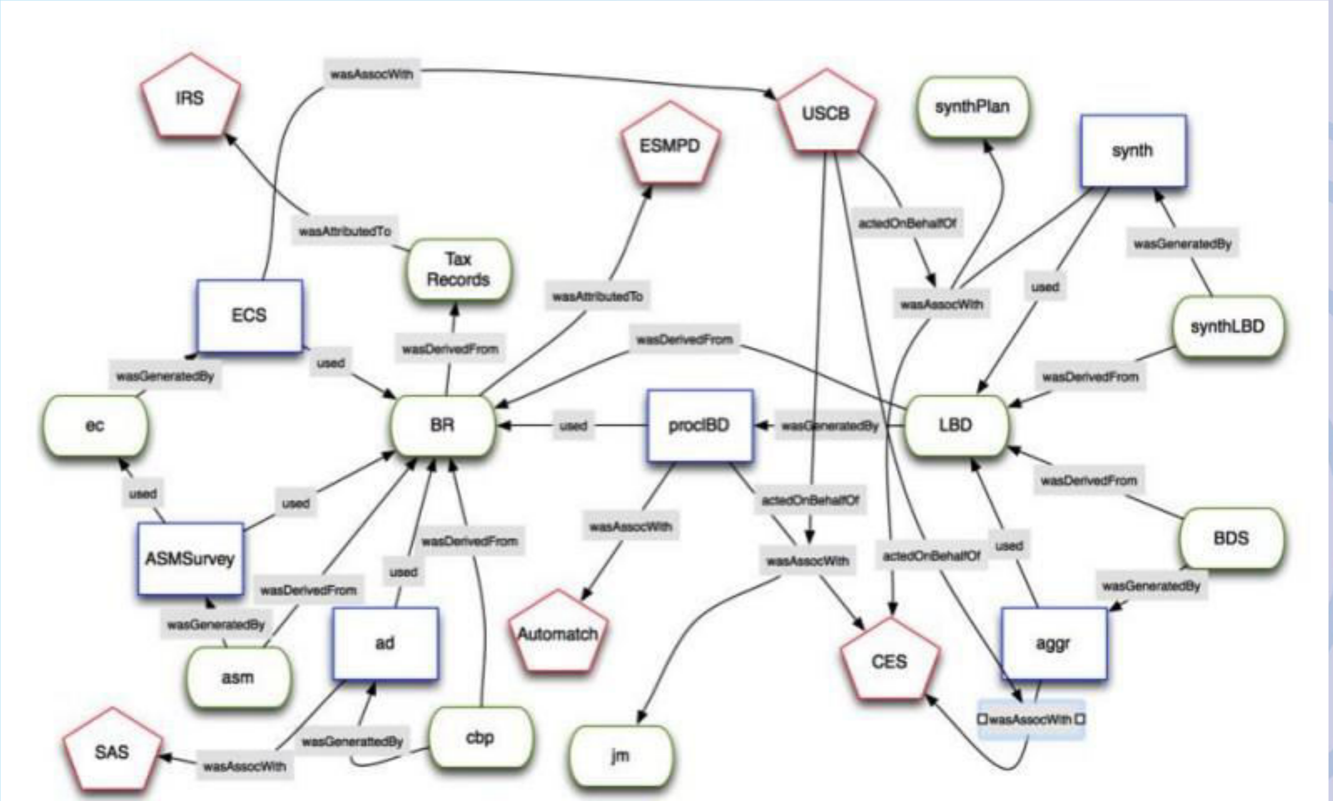
- **entities** that are physical, digital, and conceptual things in the world;
- **activities** that are dynamic aspects of the world that change and create entities; and
- **agents** that are responsible for activities.
- A set of **relationships** that can exist between them that express attribution, delegation, derivation, etc.

Provenance of Longitudinal Business Database (LBD)


LBD Provenance
Business Register (SSEL)
some exclusions
Jarmin and Miranda (2002)
Enumeration frame
New establishments
Economic Census (quinquennial)
Possible linkage
Longitudinal Business Database (LBD)
Sample frame
Possible linkage
County Business Patterns (CBP)
Annual Survey of Manufactures (ASM)
Synthetic LBD (SynLBD)
Business Dynamics Statistics (BDS)

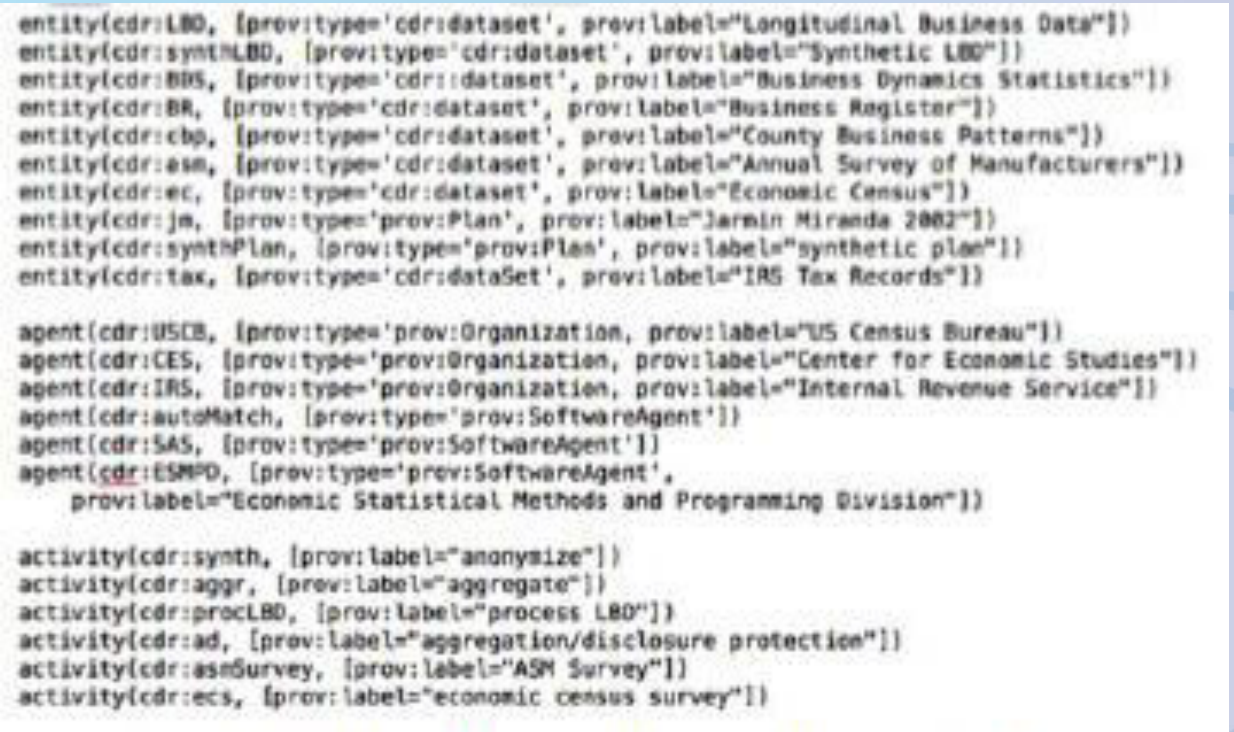
LBD Provenance expressed in PROV

- Oval Nodes: Entities
- Rectangular Nodes: Activities
- Pentangular Nodes: Agents



LBD Provenance expressed in PROV

- Oval Nodes: Entities
- Rectangular Nodes: Activities
- Pentangular Nodes: Agents

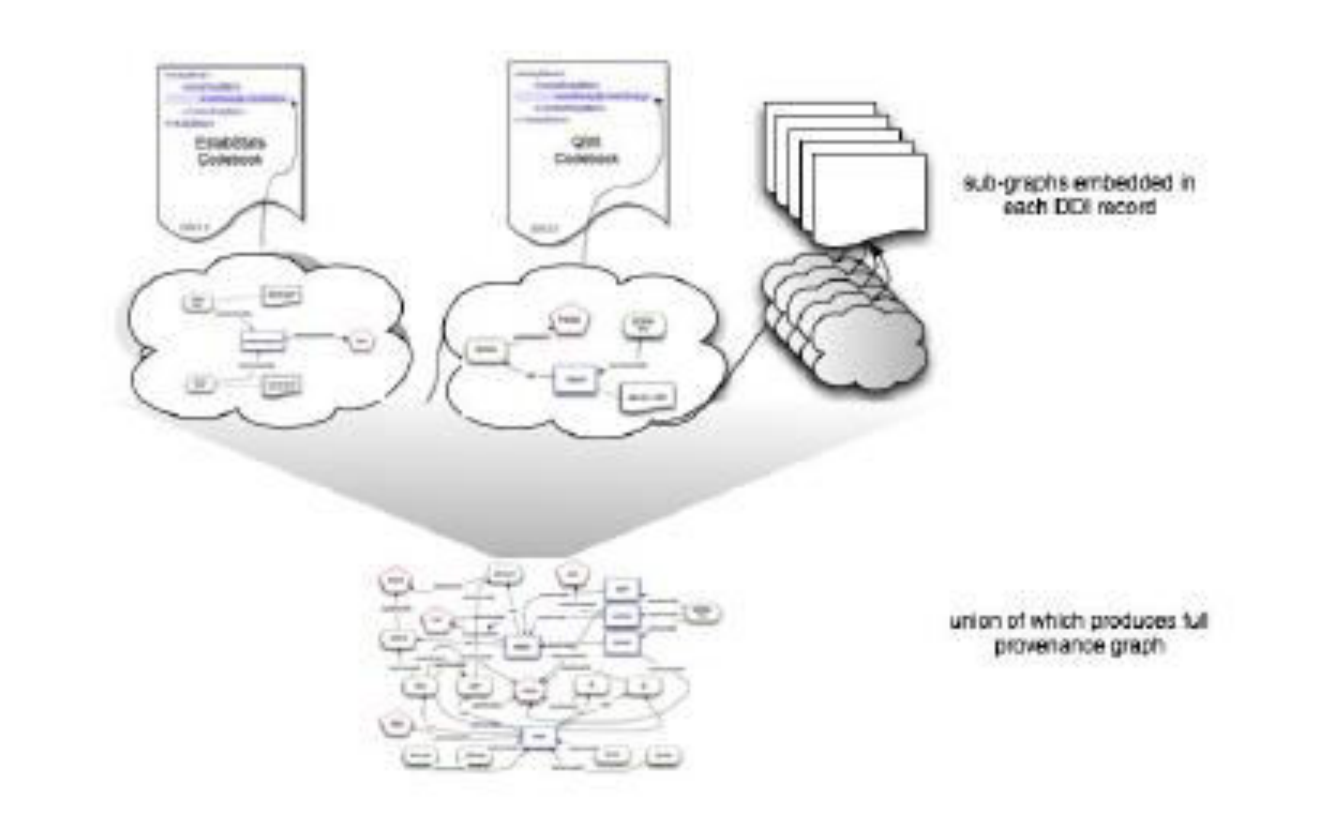


```
entity(cdr:LBD, [prov:type='cdr:dataset', prov:label="Longitudinal Business Data"])
entity(cdr:synthLBD, [prov:type='cdr:dataset', prov:label="Synthetic LBD"])
entity(cdr:BDS, [prov:type='cdr::dataset', prov:label="Business Dynamics Statistics"])
entity(cdr:BR, [prov:type='cdr:dataset', prov:label="Business 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"])
```

Figure 1. Longitudinal Business Database (LBD) provenance graph

Modular Approach; Stitching Provenance Bundles together

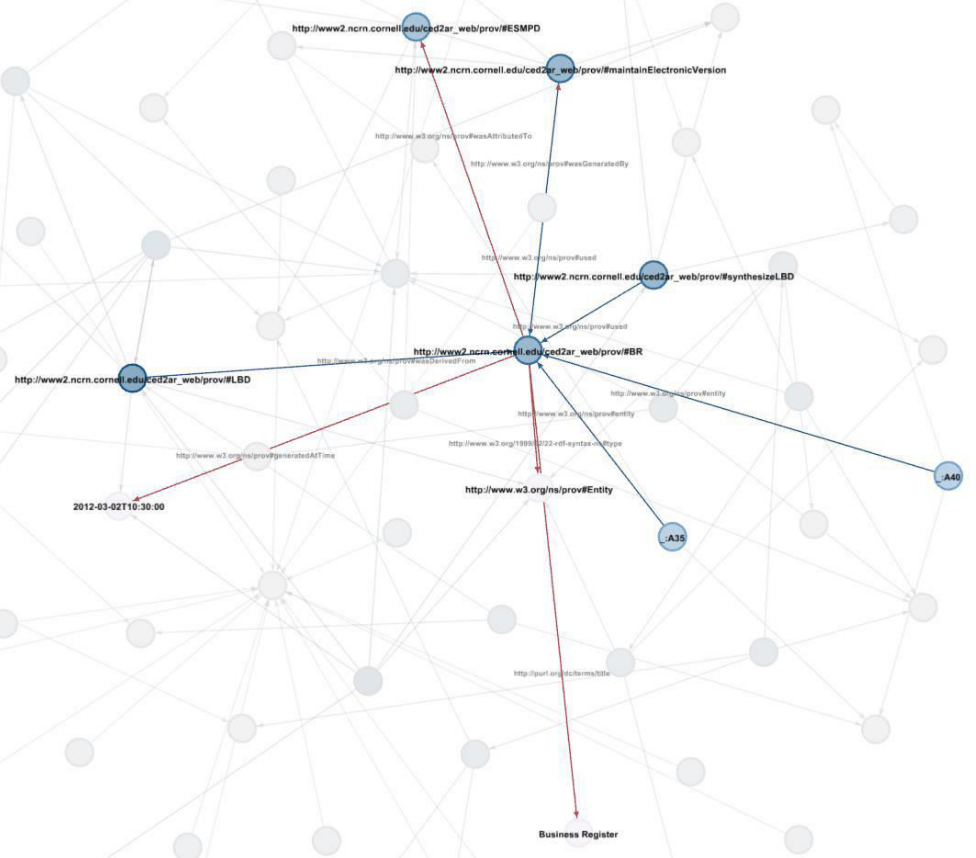
- Provenance subgraph “bundles” link, by resource, to other subgraphs
- Allows dynamic generation of entire provenance graph


Codebook
QWI Codebook
sub-graphs embedded in each DDI record
union of which produces full provenance graph

Future Work: User Visualization and Exploration of Provenance Graphs

- Draft release early 2014

- Look forward to continued interaction with DDI community as we continue to refine CED^{2}AR


http://www2.ncrn.cornell.edu/ced2ar_web/prov/#ESMPD
http://www2.ncrn.cornell.edu/ced2ar_web/prov/#maintainElectronicVersion
http://www.w3.org/ns/prov#wasAttributedTo
http://www.w3.org/ns/prov#wasGeneratedBy
http://www.w3.org/ns/prov#used
http://www2.ncrn.cornell.edu/ced2ar_web/prov/#synthesizeLBD
http://www.w3.org/ns/prov#used
http://www2.ncrn.cornell.edu/ced2ar_web/prov/#BR
http://www.w3.org/ns/prov#wasDerivedFrom
http://www2.ncrn.cornell.edu/ced2ar_web/prov/#LBD
http://www.w3.org/ns/prov#entity
http://www.w3.org/ns/prov#entity
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/ns/prov#generatedAtTime
_:A40
http://www.w3.org/ns/prov#Entity
2012-03-02T10:30:00
_:A35
http://purl.org/dc/terms/title
Business Register

Merci!

Questions?

block@cornell.edu
ncrn.cornell.edu