 DOI: http://dx.doi.org/10.3886/xxx

Proceedings of EDDI13
5th Annual European DDI User Conference
December 2013, Paris, France

Encoding Provenance of Social Science Data: Integrating PROV with DDI

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

Abstract

Provenance is a key component of evaluating the integrity and reusability of data for scholarship. While recording and providing access provenance has always been important, it is even more critical in the web environment in which data from distributed sources and of varying integrity can be combined and derived. The PROV model, developed under the auspices of the W3C, is a foundation for semantically-rich, interoperable, and web-compatible provenance metadata. We report on the results of our experimentation with integrating the PROV model into the DDI metadata for a complex, but characteristic, example social science data. We also present some preliminary thinking on how to visualize those graphs in the user interface.

Keywords: Metadata, Provenance, DDI, eSocial Science.

1 Introduction

For the past 50 years, quantitative social science has been built on a shared foundation of data sources originating from survey research, aggregate government statistics, and in-depth studies of individual places, people, or events. Underlying these data is a well-established infrastructure composed of an international network of highly-curated and metadata-rich archives of social science data such as ICPSR (Inter-University Consortium for Political and Social Research) and the UK Data Archive. These archives continue to play an important role in quantitative social science research. However, the emergence and maturation of ubiquitous networked computing and the ever-growing data cloud has introduced a spectacular quantity and variety of new data sources into this mix. These include massive social media data sources such as Facebook, Twitter, and other online communities, which when combined with more

[1] School of Information, University Of Michigan, Ann Arbor, Michigan USA.
[2] Cornell Institute for Social and Economic Research, Cornell University, Ithaca, New York USA.
[3] School of Industrial and Labor Relations, Cornell University, Ithaca, New York USA.

traditional data sources, provide the opportunity for studies at scales heretofore unimaginable. This paradigm shift has been described by Gary King, a Harvard political scientist, as the *social science data revolution*, which is characterized by a "changing evidence base of social science research" (King, 2011a, 2011b).

These huge changes in both the quantity and nature of data in quantitative social science have created with King calls an "infrastructural challenge" (King, 2011b). This challenge is not unique to social science; data-centric scholarship is becoming increasingly popular across the disciplinary spectrum, from physical and life sciences to engineering to the humanities (American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences, 2006; Atkins et al., 2003; Daw et al., 2007). .Addressing the specific infrastructural needs of each of these diverse fields while, at the same time building a common infrastructure across the breath of scholarship, has become a major challenge of the 21st century (Paul N. Edwards et al., 2013).

The successful development and adoption of a data infrastructure for the emergent social science paradigm faces two notable challenges. The first of these is need to address confidentiality and cloaking of data elements (Abowd, Vilhuber, & Block, 2012), which we addressed in (Lagoze, Block, Williams, Abowd, & Vilhuber, 2013). A substantial portion of the data commonly used for quantitative social science are confidential because they associate the identities of the subjects of study (e.g., people, corporations, etc.) with private information such as income level, health history, and the like. Confidentiality is important in a number of other data domains such as health informatics, but a particularly interesting twist in social science is the existence of disclosure limitations not only on the data, but also on the metadata. These may include statutory disclosure restrictions on statistical features of the underlying data, such as extreme values, and even prohibitions on the disclosure of variables names themselves. In (Lagoze, Block, et al., 2013), we described a method for encoding appropriate disclosure attributes in DDI metadata.

Another challenge in the development of data infrastructure for social science is the importance of and complexity of data provenance. Even before the emergence of data-rich online social networks, many of the data underlying social science research were embedded in complex provenance chains composed of inter-related private and publicly accessible data and metadata, multithreaded relationships among these data and metadata, and partially-ordered version sequences. The combination of these factors and others often makes it difficult to understand and trace the origins of data that are the basis of a particular study. The results are barriers to the essential scholarly tasks of testing research results for validity and reproducibility, creating a substantial risk of breach of the scientific integrity of the research process itself. It also presents an often insurmountable barrier to data reuse, which is fundamental to the incremental building of research results in a scholarly field (Zimmerman, 2008).

The increasing tendency to mix traditional archival-based data with Web-based, more-informal data calls for an approach to the provenance problem that embraces a generic information architecture perspective. As indicated by the increasing momentum of efforts like linked open data (Heath & Bizer, 2011), architecturally supported silos separating interdisciplinary data are not addressing the demands of 21st-century research. The need for a "web-wise" solution to the provenance issue (Cheney, Chong, Foster, Seltzer, & Vansummeren, 2009) was the inspiration for the W3C (World Wide Web Consortium) initiation of an international effort to develop an extensible, semantically-based, and practical solution for encoding provenance. The PROV documents "define a model, corresponding serializations and other supporting definitions to enable the interoperable interchange of provenance information in heterogeneous environments such as the web" (Paul Groth & Moreau, 2013).

In (Lagoze, Williams, & Vilhuber, 2013), we reported on our initial experimentation with the PROV model for encoding real-world provenance scenarios associated with existing social science data. We also proposed a preliminary method for embedding that provenance information within the metadata specification developed by the Data Documentation Initiative (DDI) (Vardigan, Heus, & Thomas, 2008) the emerging standard for most social science data. We showed that, with some refinements, the PROV model is indeed suitable for the task, and thereby lays the groundwork for implementing user-facing provenance applications that could enrich the quality and integrity of data-centric social science. In this paper, we report on our recent advancements in this work with DDI in the PROV model, which include specifying the nature of the XML expressing provenance that could be incorporated into DDI and experimenting with visualizations of the semantics expressed in those encodings. This completes the planning phase of our work in this area, which will be followed by an implementation stage that we hope to report on in future papers.

This work is one thread of an NSF Census Research Network award (Abowd et al., 2012). A primary goal of this project is to design and implement tools that bridge the existing gap between private and public data and metadata, that are usable to researchers with and without secure access, and that make proper curation and citation of these data possible. One facet of this larger project, which provides a development context for the work reported in this paper, is an evolving prototype and implementation of the Comprehensive Extensible Data Documentation and Access Repository (CED²AR). This is a metadata repository system that allows researchers to search, browse, access, and cite confidential data and metadata, and the provenance thereof, through either a web-based user interface or programmatically through a search API.

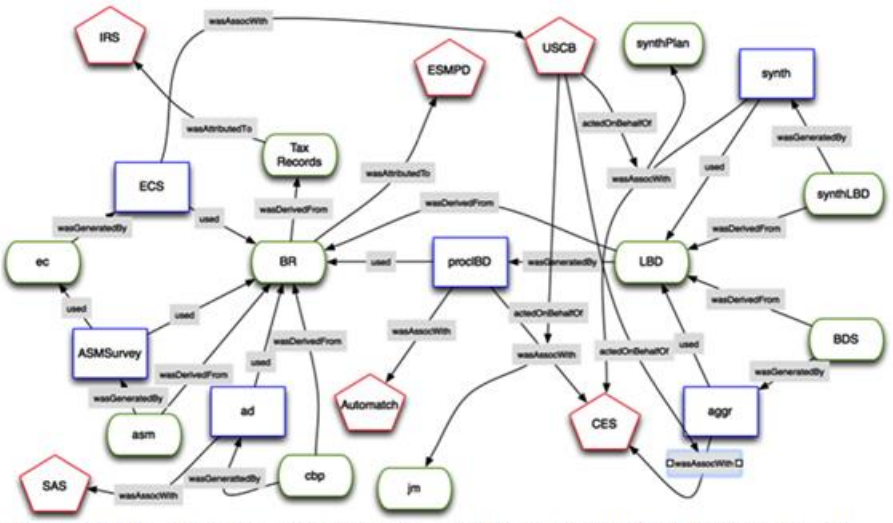
2 Applying the PROV Model to a Social Science Scenario

The W3C PROV model is fully described in a family of documents (Paolo Missier, Khalid Belhajjame, & James Cheney, 2013) that cover the data model, ontology, expressions and various syntaxes, and access and searching. The model is based the notion of *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. In addition to these building blocks, the PROV model describes a set of relationships that can exist between them that express attribution, delegation, derivation, etc. Space limitations prohibit further explanation of the model and this paper assumes that the reader has a working familiarity with PROV.

In (Lagoze, Williams, et al., 2013), we applied the PROV model to the two frequently-used social science data products; Longitudinal Business Data (LBD) and the Longitudinal Employer-Household Dynamics (LEHD) data sets. The remainder of this paper builds on this work and explains it in the context of the LBD example. The example, illustrated in Figure 1, is somewhat simplified for legibility and does not represent the full provenance graph as it would be constructed in a production-quality system. Our diagramming convention is the same as that used in the W3C PROV documentation; oval nodes denote entities, rectangular nodes denote activities, and pentagonal nodes that agents. The provenance graph shown in **Error! Reference source not found.** is paired with a declaration of its component entities, activities, and agents encoded in PROV-N, a functional notation meant for human consumption (Moreau & Missier, 2013). Although our work includes an encoding of relationships among these objects in the same notation, space limitations of this paper prohibit the inclusion of these full descriptions.

As the figure indicates, the Census Bureau's Longitudinal Business Database (LBD) is one component of a complex provenance graph. The LBD is derived entirely from the Business Register (BR), which is itself derived from tax records provided on a flow base to the Census Bureau by the Internal Revenue Service (IRS). The methodology to construct the LBD from snapshots of the BR is described in (Jarmin & Miranda, 2002), and it is being continually maintained (updated yearly) at the Census Bureau. Derivative products of the LBD are the Business Dynamics Statistics (BDS) an aggregation of the LBD (Haltiwanger, Jarmin, & Miranda, 2008) and the Synthetic LBD (Kinney et al., 2011), a confidentiality-protected synthetic microdata version of the LBD. However, the LBD and its derivative products are not the only statistical data products derived from the BR. The BR serves as the enumeration frame for the quinquennial Economic Censuses (EC), and together with the post-censal data collected through those censuses, serves as the sampling frame for the annual surveys, e.g., the Annual Survey of Manufactures (ASM). Aggregations of the ASM and EC are published by the Census Bureau, confidential versions are available within the Census RDC's. Furthermore, the BR serves as direct input to the County Business Patterns (CBP) and related Business Patterns through aggregation and disclosure protection mechanisms




```
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

3 Integrating DDI and PROV

DDI has emerged as the standard for encoding metadata for social science datasets. Currently there are two threads of development in the DDI community. The 2.X branch, commonly known as DDI-Codebook, primarily focuses on bibliographic information about an individual data set and the structure of its variables. The 3.X branch, commonly known as DDI-Lifecycle, is designed to document a study and its resulting data sets over the entire lifecycle from

conception through publication and subsequent reuse. Some of the semantics of DDI-Lifecycle overlap and sometimes conflict with the PROV semantics specified by the W3C.

We argue that a DDI-centric approach to provenance, such as that taken by DDI-Lifecycle, might be inadvisable in the emerging scholarly environment where integration of traditional archival data with Web-based data (King, 2011b) is increasingly becoming the norm. We have decided to take the approach of working within the simpler DDI-Codebook framework and embedding the web architecture-aware PROV metadata within the individual data set-specific DDI records. Such an approach also offers the advantage of being more amenable to exposure of DDI metadata in a web-visible manner such as that specified by the linked open data initiative (Bizer, Cyganiak, & Heath, 2007).

The overall design approach taken is modular as illustrated in Figure 2. Only the metadata related to the specific data set is stored in its respective DDI record, which then links via a URI to the PROV metadata stored in other DDI records. This modular approach is similar to that proposed by the W3C PROV group in the "bundles" recommendation (Moreau & Lebo, 2013); as stated in the specification the bundles model is "useful for provenance descriptions created by one party to bring to provenance descriptors created by another party." Furthermore, "such a mechanism would allow the 'stitching' of provenance descriptions together". This is exactly our goal, to express within the DDI for specific data set only its provenance dependencies and independently allow data sets to then express derivation from that existing data set fire their own provenance bundle. The full provenance graph for a specific application instance can then be reconstructed dynamically by combining these individual subgraphs, i.e., "stitching" them together.

The `<relStdy>` element in DDI 2.5 provides a useful place to encode provenance data specific to the respective data set. As documented in the DDI 2.5 schema[4], this field contains "information on the relationship of the current data collection to others (e.g., predecessors, successors, other waves or rounds or to other editions of the same file). This would include the names of additional data collections generated from the same data collection vehicle plus other collections directed at the same general topic. Can take the form of bibliographic citations."

[4] http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd

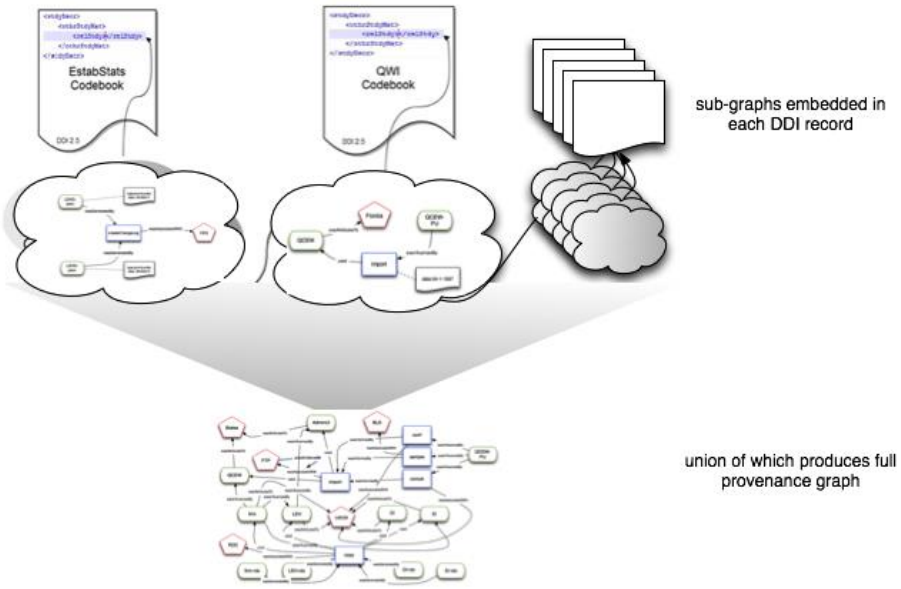


Figure 2. Storing provenance subgraphs related to a given resource within the <relStudy> element in the corresponding DDI metadata. That subgraph links, by resource, to other subgraphs located in other codebooks and ancillary entities (e.g., plans, ages) to allow dynamic generation of the entire provenance graph.

In our previous paper (Lagoze, Williams, et al., 2013) we explored encoding the PROV module in RDF/XML. However, since there is no constraining schema for RDF/XML, this would require wrapping that description within a CDATA tag in order to not interfere with schema compliance testing of the entire DDI description. In this paper, we explore what we consider a much more sensible approach; that is, leveraging the XML encoding of PROV semantics (Moreau, 2013), and then making minor change to the DDI 2.5 schema to instruct validators to evaluate the PROV subtree within the constraints of the PROV XML schema. We note that the decision to use either the XML or RDF/XML encoding may be influenced by current work within the DDI community to develop an RDF encoding for DDI metadata that could then easily accommodate RDF-encoding of provenance metadata (Bosch, Cyganiak, Gregory, & Wackerow, 2013; Kramer, Leahey, Southall, Vampras, & Wackerow, 2012).

The remainder of this section illustrates a number of these PROV/XML encoded bundles that are components of the full LBD provenance graph illustrated in Figure 1. The XML shown in each of the figures does not include a number of details of the full graph due to space limitations.

```
<prov:document xmlns:prov="http://www.w3.org/ns/prov#"
    xmlns:cdr="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#" xmlns:dct="http://purl.org/dc/terms/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"  xmlns:xsd="http://www.w3.org/2001/XMLSchema">

    <!-- Plans -->
    <prov:plan prov:id="cdr:procLBDPlan">
        <prov:location xsi:type="xsd:anyURI">
            http://www.vrdc.cornell.edu/info7470/2007/Readings/jarmin-miranda-2002.pdf</prov:location>
        <prov:type>prov:Plan</prov:type>
        <dct:title>Process LBD Plan</dct:title>
    </prov:plan>
    <prov:plan prov:id="cdr:synthPlan">
        <prov:location xsi:type="xsd:anyURI">
            http://www2.vrdc.cornell.edu/news/wp-content/uploads/2011/02/discussion_paper_101943.pdf</prov:location>
        <prov:type>prov:Plan</prov:type>
        <dct:title>Process LBD Plan</dct:title>
    </prov:plan>

    <!-- Agents and Responsibility -->
    <prov:agent prov:id="cdr:USCB">
        <prov:type>prov:Organization</prov:type>
        <foaf:givenName>United States Census Bureau</foaf:givenName>
    </prov:agent>
    <prov:agent prov:id="cdr:Automatch">
        <prov:type>prov:SoftwareAgent</prov:type>
        <foaf:givenName>Automatch</foaf:givenName>
    </prov:agent>
    <prov:agent prov:id="cdr:CES">
        <prov:type>prov:Organization</prov:type>
        <foaf:givenName>Center for Economic Studies</foaf:givenName>
    </prov:agent>
</prov:document>
```

Figure 3. Common XML fragment containing shared entities

3.1 Encoding cross-module entities

The XML document in Figure 3 defines the set of entities that are shared across the other provenance bundles. As will be illustrated below, these the entities defined in this document are selectively included into those bundles through the use of an XML `<include>` tag with an `xpointer` attribute. The entities defined here are:

- Plans
  - procLBDPlan: the process LBD plan
  - synthPlan: the synthetic LBD plan
- Agents
  - UCSB: United States Census Bureau
  - Automatch: the respective software agent
  - CES: Center for Economic Studies

```
<prov:document xmlns:prov="http://www.w3.org/ns/prov#"  xmlns:xi="http://www.w3.org/2001/XInclude"
    xmlns:cdr="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#" xmlns:dct="http://purl.org/dc/terms/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"  xmlns:xsd="http://www.w3.org/2001/XMLSchema">

    <xi:include href="ProvXIncludes.xml" xpointer="xpointer(//prov:plan)" />
    <xi:include href="ProvXIncludes.xml" xpointer="xpointer(//prov:agent)" />

    <!-- Entities -->
    <prov:entity prov:id="cdr:BR">
        <dct:title>Business Register</dct:title>
    </prov:entity>

    <!-- Activities -->
    <prov:activity prov:id="cdr:maintainElectronicVersion"/>

    <!-- Association and Attribution -->
    <prov:wasAssociatedWith>
        <prov:activity prov:ref="cdr:maintainElectronicVersion"/>
        <prov:agent prov:ref="cdr:ESMPD"/>
    </prov:wasAssociatedWith>
    <prov:wasAttributedTo>
        <prov:entity prov:ref="cdr:BR"/>
        <prov:agent prov:ref="cdr:ESMPD"/>
    </prov:wasAttributedTo>
    <prov:actedOnBehalfOf>
        <prov:delegate prov:ref="cdr:USCB"/>
        <prov:responsible prov:ref="cdr:ESMPD"/>
        <prov:activity prov:ref="cdr:maintainElectronicVersion"/>
    </prov:actedOnBehalfOf>

    <!-- Usage and Generation -->
    <prov:used>
        <prov:activity prov:ref="cdr:maintainElectronicVersion"/>
        <prov:entity prov:ref="cdr:BR"/>
    </prov:used>

</prov:document>
```

Figure 4. Business Register (BR) provenance subgraph in PROV-XML.

3.2 BR provenance

Figure 4 shows the XML document defining the provenance particular to the Business Register (BR) entity. As is indicated, the BR is created by a process where the Economic and Statistical Methods Programming Division (ESMPD maintains the electronic version on behalf of the US Census Bureau (USCB).

3.3 LBD provenance

Figure 5 shows the provenance dependencies of the Longitudinal Business Database (LBD). As indicated, the LBD is derived from the Business Register (BR); the URI of which joins it to the provenance graph for the Business Register defined in Figure 4. This derivation involves a number of other agents both organizational (CES acting on behalf of the Census Euro) and software (AutoMatch), and the enactment of an established plan (proLBDPlan).

```
<prov:document xmlns:prov="http://www.w3.org/ns/prov#"  xmlns:xi="http://www.w3.org/2001/XInclude"
    xmlns:cdr="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#" xmlns:dct="http://purl.org/dc/terms/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"  xmlns:xsd="http://www.w3.org/2001/XMLSchema">

    <xi:include href="ProvXIncludes.xml" xpointer="xpointer(//prov:plan)" />
    <xi:include href="ProvXIncludes.xml" xpointer="xpointer(//prov:agent)" />

    <!-- Entities -->
    <prov:entity prov:id="cdr:BR">
        <dct:title>Business Register</dct:title>
    </prov:entity>
    <prov:entity prov:id="cdr:LBD">
        <dct:title>Longitudinal Business Database</dct:title>
    </prov:entity>

    <!-- Activities -->
    <prov:activity prov:id="cdr:procLBD"/>

    <!-- Revision and Derivation -->
    <prov:wasDerivedFrom>
        <prov:generatedEntity prov:ref="cdr:LBD"/>
        <prov:usedEntity prov:ref="cdr:BR"/>
    </prov:wasDerivedFrom>

    <!-- Association and Attribution -->
    <prov:wasAssociatedWith>
        <prov:activity prov:ref="cdr:procLBD"/>
        <prov:agent prov:ref="cdr:USCB"/>
        <prov:plan prov:ref="cdr:procLBDPlan"/>
    </prov:wasAssociatedWith>
    <prov:wasAttributedTo>
        <prov:entity prov:ref="cdr:LBD"/>
        <prov:agent prov:ref="cdr:CES"/>
    </prov:wasAttributedTo>
    <prov:wasAttributedTo>
        <prov:entity prov:ref="cdr:LBD"/>
        <prov:agent prov:ref="cdr:Automatch"/>
    </prov:wasAttributedTo>
    <prov:actedOnBehalfOf>
        <prov:delegate prov:ref="cdr:USCB"/>
        <prov:responsible prov:ref="cdr:CES"/>
        <prov:activity prov:ref="cdr:synthesizeLBD"/>
    </prov:actedOnBehalfOf>

    <!-- Usage and Generation -->
    <prov:used>
        <prov:activity prov:ref="cdr:procLBD"/>
        <prov:entity prov:ref="cdr:BR"/>
    </prov:used>
    <prov:wasGeneratedBy>
        <prov:entity prov:ref="cdr:LBD"/>
        <prov:activity prov:ref="cdr:procLBD"/>
        <prov:time>2012-03-02T10:30:00</prov:time>
    </prov:wasGeneratedBy>

</prov:document>
```

Figure 5. Longitudinal Business Database (LBD) provenance subgraph in PROV-XML

```
<prov:document xmlns:prov="http://www.w3.org/ns/prov#" xmlns:xi="http://www.w3.org/2001/XInclude"
    xmlns:cdr="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#" xmlns:dct="http://purl.org/dc/terms/"
    xmlns:foaf="http://xmlns.com/foaf/0.1/">

    <xi:include href="ProvXIncludes.xml" xpointer="xpointer(//prov:plan)" />
    <xi:include href="ProvXIncludes.xml" xpointer="xpointer(//prov:agent)" />

    <!-- Entities -->
    <prov:entity prov:id="cdr:LBD">
        <dct:title>Longitudinal Business Database</dct:title>
    </prov:entity>
    <prov:entity prov:id="cdr:SYNLBD">
        <dct:title>Synthesized Longitudinal Business Database</dct:title>
    </prov:entity>

    <!-- Activities -->
    <prov:activity prov:id="cdr:synthesizeLBD"/>

    <!-- Revision and Derivation -->
    <prov:wasDerivedFrom>
        <prov:generatedEntity prov:ref="cdr:SYNLBD"/>
        <prov:usedEntity prov:ref="cdr:LBD"/>
    </prov:wasDerivedFrom>

    <!-- Association and Attribution -->
    <prov:wasAssociatedWith>
        <prov:activity prov:ref="cdr:synthesizeLBD"/>
        <prov:agent prov:ref="cdr:USCB"/>
        <prov:plan prov:ref="cdr:synthPlan"/>
    </prov:wasAssociatedWith>
    <prov:wasAttributedTo>
        <prov:entity prov:ref="cdr:SYNLBD"/>
        <prov:agent prov:ref="cdr:USCB"/>
    </prov:wasAttributedTo>

    <!-- Usage and Generation -->
    <prov:used>
        <prov:activity prov:ref="cdr:synthesizeLBD"/>
        <prov:entity prov:ref="cdr:LBD"/>
    </prov:used>
    <prov:wasGeneratedBy>
        <prov:entity prov:ref="cdr:SYNLBD"/>
        <prov:activity prov:ref="cdr:synthesizeLBD"/>
        <prov:time>2012-03-02T10:30:00</prov:time>
    </prov:wasGeneratedBy>

</prov:document>
```

Figure 6. Synthesized Longitudinal Business Database (synLBD) provenance subgraph in PROV-XML

3.4 synLBD provenance

Figure 6 shows the XML defining the provenance graph for the Synthesized Longitudinal business database (synLBD). As indicated, the synLBD is a derivation of the LBD, the URI of which joins it to the provenance graph of that entity defined in Figure 5. This derivation is performed under the auspices of the Census Bureau according to the plan synthPlan.

4 Conclusion and Future Work

In a series of three papers, of which this is the third, we have investigated and proposed solutions for two fundamental issues in the curation of quantitative social science data; confidentiality and provenance. In (Lagoze, Block, et al., 2013), we described a method for embedding field-specific and value-specific cloaking in DDI metadata. In (Lagoze, Williams, et al., 2013), we described the applicability of the W3C-developed PROV model for encoding the complex provenance chains characteristics of social science data. We also explored the embedding of an RDF/XML encoding of that provenance declaration within DDI. This encoding anticipates ongoing work in the DDI community on a full RDF encoding of DDI semantics. In this paper, we investigated an alternative XML encoding of the PROV metadata and the modularization of that description in separate provenance bundles.

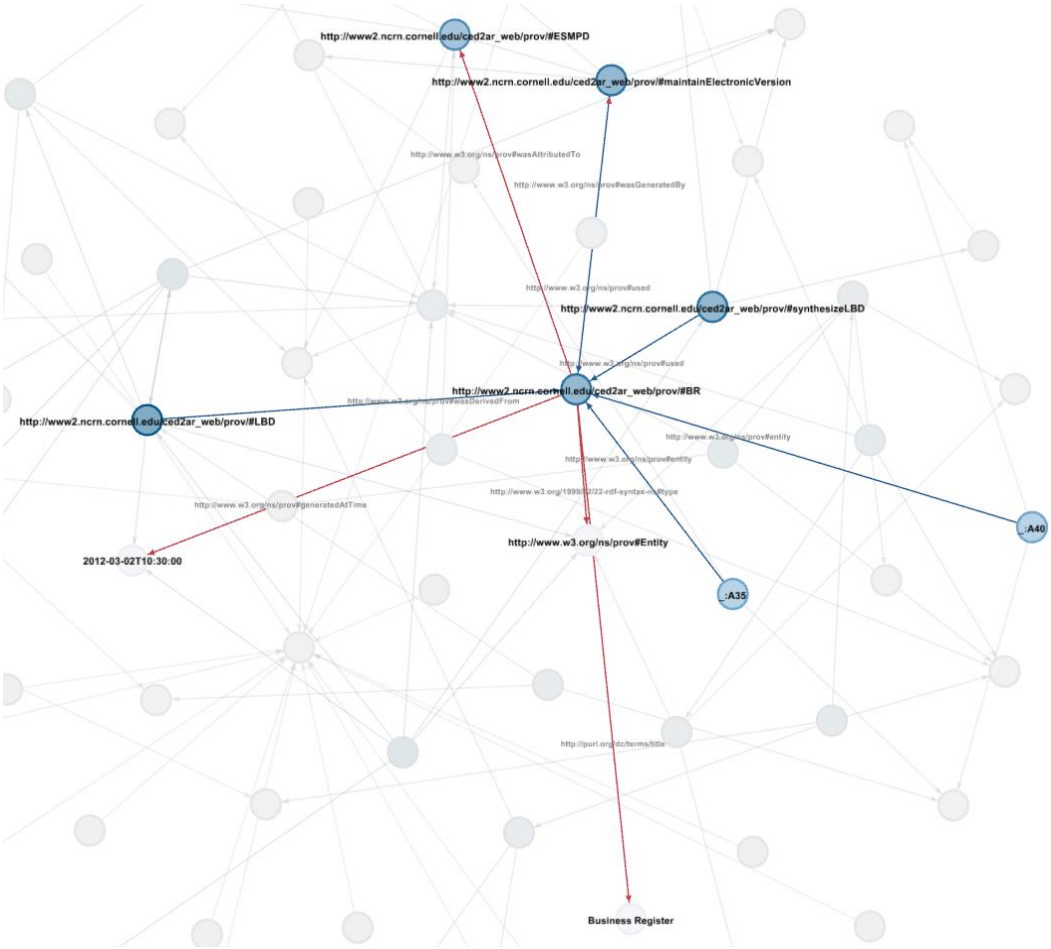


Figure 7. Prototype visualization of a provenance graph fragment embedded in context.

Although we have implemented some preliminary prototypes of this work, our future work focuses on the full production-level implementation within the CED2AR system. One relevant design issue is user visualization and exploration of provenance graphs. Initial thinking on this is illustrated in Figure 7. We anticipate first release of our implementation in 1st quarter 2014 and look forward to interactions with the DDI and related communities to refine this work.

5 Acknowledgements

We acknowledge NSF grants SES 9978093, ITR 0427889, SES 0922005, SES 1042181, and SES 1131348. Thanks to Ben Perry for his help with visualizations.

6 References

Abowd, J., Vilhuber, L., & Block, W. (2012). A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs. In J. Domingo-Ferrer & I. Tinnirello (Eds.), *Privacy in Statistical Databases (LNCS 7756)* (Vol. 7556, pp. 216–225). Springer Berlin / Heidelberg. doi:10.1007/978-3-642-33627-0_17

American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. (2006). *Our Cultural Commonwealth: The final report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences*. ACLS. Retrieved from http://www.acls.org/cyberinfrastructure/cyber.htm

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., … Wright, M. H. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure. National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure. Retrieved from http://www.nsf.gov/od/oci/reports/CH1.pdf

Bizer, C., Cyganiak, R., & Heath, T. (2007). How to Publish Linked Data on the Web. Free University of Berlin. Retrieved from http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

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.

Cheney, J., Chong, S., Foster, N., Seltzer, M., & Vansummeren, S. (2009). Provenance. In *Proceeding of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications - OOPSLA '09* (p. 957). New York, New York, USA: ACM Press. doi:10.1145/1639950.1640064

Daw, M., Procter, R., Lin, Y., Hewitt, T., Ji, W., Voss, A., … Crouchley, R. (2007). Developing an e-Infrastructure for Social Science. In *Proceedings of e-Social Science'07*. Ann Arbor.

Haltiwanger, J., Jarmin, R. S., & Miranda, J. (2008). Jobs Created from Business Startups in the United States. Retrieved from http://www.census.gov/ces/pdf/BDS_StatBrief1_Jobs_Created.pdf

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *1*(1), 1–136. doi:10.2200/S00334ED1V01Y201102WBE001

Jarmin, R., & Miranda, J. (2002). *The Longtitudinal Business Database*. Retrieved from https://www.census.gov/ces/pdf/CES-WP-02-17.pdf

King, G. (2011a). The Social Science Data Revolution. *Horizons in Political Science*. Cambridge, MA: Harvard University. Retrieved from http://gking.harvard.edu/files/gking/files/evbase-horizonsp.pdf

King, G. (2011b). Ensuring the data-rich future of the social sciences. *Science (New York, N.Y.)*, *331*(6018), 719–21. doi:10.1126/science.1197872

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, *79*(3), 362–384. Retrieved from http://econpapers.repec.org/RePEc:bla:istatr:v:79:y:2011:i:3:p:362-384

Kramer, S., Leahey, A., Southall, H., Vampras, J., & Wackerow, J. (2012, September 1). Using RDF to describe and link social science data to related resources on the Web: leveraging the Data Documentation Initiative (DDI) model. Data Documentation Initiative. doi:10.3886/DDISemanticWeb01

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

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

Moreau, L. (2013). *PROV-XML: the PROV-XML Schema*. Retrieved from http://www.w3.org/TR/prov-xml/

Moreau, L., & Lebo, T. (2013). *Linking across Provenance Bundles*. Retrieved from http://www.w3.org/TR/2013/NOTE-prov-links-20130430/

Moreau, L., & Missier, P. (2013). *PROV-N: The Provenance Notation*. Retrieved from http://www.w3.org/TR/2013/REC-prov-n-20130430/

Paolo Missier, Khalid Belhajjame, & James Cheney. (2013). The W3C PROV family of specifications for modelling provenance metadata. In *EDBT/ICDT '13* (pp. 773–776). Genoa: ACM Press. doi:10.1145/2452376.2452478

Paul Groth, & Moreau, L. (2013). *PROV-Overview: An Overview of the PROV Family of Documents*. Retrieved from http://www.w3.org/TR/prov-overview/

Paul N. Edwards, Steven J. Jackson, M. K. Chalmers, Geoffrey C. Bowker, Christine L. Borgman, David Ribes, ... Scott Calvert. (2013). Knowledge Infrastructures: Intellectual Frameworks and Research Challenges.

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a Standard for the Social Sciences. *The International Journal of Digital Curation*, *3*(1).

Zimmerman, A. (2008). New Knowledge from Old Data Sharing and Reuse of Ecological Data. *Science Technology Human Values2*, *33*(5), 631–652.

Appendix A: Full provenance graph expressed in RDF/XML

```
<?xml version="1.0" encoding="utf-8"?>
<!-- $ID $URL -->
<rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema/"
  xmlns:prov="http://www.w3.org/ns/prov#"
  xmlns:cdr="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:nso="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

  <!-- Entities -->
  <prov:Entity rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#BR"
   dcterms:title="Business Register">
   <prov:generatedAtTime
    rdf:datatype="http://www.w3.org/2001/XMLSchema/dateTime"
    >2012-03-02T10:30:00</prov:generatedAtTime>
   <prov:wasAttributedTo
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#ESMPD"/>
   <prov:wasGeneratedBy

rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#maintainElectronicVersion"
    />
  </prov:Entity>

  <prov:Entity rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#LBD"
   dcterms:title="Longitudinal Business Database">
   <prov:generatedAtTime
    rdf:datatype="http://www.w3.org/2001/XMLSchema/dateTime"
    >2012-03-02T10:30:00</prov:generatedAtTime>
   <prov:wasAttributedTo
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#CES"/>
   <prov:wasGeneratedBy
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#procLBD"/>
   <prov:wasDerivedFrom
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#BR"/>
  </prov:Entity>

  <prov:Entity rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#SYNLBD"
```

```
   dcterms:title="Synthesized Longitudinal Business Database">
   <prov:generatedAtTime
    rdf:datatype="http://www.w3.org/2001/XMLSchema/dateTime"
    >2012-03-02T10:30:00</prov:generatedAtTime>
   <prov:wasAttributedTo
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#USCB"/>
   <prov:wasGeneratedBy
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#synthesizeLBD"/>
   <prov:wasDerivedFrom
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#LBD"/>
  </prov:Entity>
  <prov:Entity
   rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#synthPlan">
   <rdf:type rdf:resource="http://www.w3.org/ns/prov#Plan"/>
   <rdfs:comment xml:lang="en">See
    http://www2.vrdc.cornell.edu/news/wp-
content/uploads/2011/02/discussion_paper_101943.pdf
    for more detail.</rdfs:comment>
  </prov:Entity>
  <prov:Entity
   rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#procLBDPlan">
   <rdf:type rdf:resource="http://www.w3.org/ns/prov#Plan"/>
   <rdfs:comment xml:lang="en">See
    http://www.vrdc.cornell.edu/info7470/2007/Readings/jarmin-miranda-2002.pdf
    for more detail.</rdfs:comment>
  </prov:Entity>

  <!-- Agents -->
  <prov:Agent rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#USCB"
   foaf:name="United States Census Bureau">
   <rdf:type rdf:resource="http://www.w3.org/ns/prov#Organization"/>
  </prov:Agent>
  <prov:Agent
   rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#Automatch"
   foaf:name="Automatch">
   <rdf:type rdf:resource="http://www.w3.org/ns/prov#SoftwareAgent"/>
  </prov:Agent>
  <prov:Agent rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#CES"
   foaf:name="Center for Economic Studies">
   <rdf:type rdf:resource="http://www.w3.org/ns/prov#Organization"/>
```

```
  <prov:actedOnBehalfOf
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#USCB"/>
</prov:Agent>
<prov:Agent rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#ESMPD"
  foaf:name="Economic Statistical Methods and Programming Division">
  <rdf:type rdf:resource="http://www.w3.org/ns/prov#Organization"/>
  <prov:actedOnBehalfOf
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#USCB"/>
</prov:Agent>

<!-- Activities -->
<prov:Activity
  rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#synthesizeLBD">
  <prov:qualifiedAssociation>
    <prov:Association>
      <prov:agent
        rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#USCB"/>
      <prov:hadPlan
        rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#synthPlan"
      />
    </prov:Association>
  </prov:qualifiedAssociation>
  <prov:qualifiedUsage>
    <prov:Usage>
      <prov:entity
        rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#BR"/>
    </prov:Usage>
  </prov:qualifiedUsage>
  <prov:used rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#BR"/>
  <prov:wasAssociatedWith
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#USCB"/>
</prov:Activity>

<prov:Activity
  rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#procLBD">
  <prov:qualifiedAssociation>
    <prov:Association>
      <prov:agent
        rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#USCB"/>
      <prov:hadPlan
```

```
      rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#procLBDPlan"
     />
    </prov:Association>
   </prov:qualifiedAssociation>
   <prov:qualifiedAssociation>
    <prov:Association>
     <prov:agent
      rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#Automatch"/>
     <prov:hadPlan
      rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#procLBDPlan"
     />
    </prov:Association>
   </prov:qualifiedAssociation>
   <prov:qualifiedUsage>
    <prov:Usage>
     <prov:entity
      rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#LBD"/>
    </prov:Usage>
   </prov:qualifiedUsage>
   <prov:used rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#LBD"/>
   <prov:wasAssociatedWith
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#USCB"/>
  </prov:Activity>

  <prov:Activity
   rdf:about="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#maintainElectronicVersion">
   <prov:qualifiedAssociation>
    <prov:Association>
     <prov:agent
      rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#ESMPD"/>
    </prov:Association>
   </prov:qualifiedAssociation>
   <prov:qualifiedUsage>
    <prov:Usage>
     <prov:entity
      rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#BR"/>
    </prov:Usage>
   </prov:qualifiedUsage>
   <prov:used rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#BR"/>
   <prov:wasAssociatedWith
```

```
    rdf:resource="http://www2.ncrn.cornell.edu/ced2ar_web/prov/#ESMPD"/>
  </prov:Activity>

</rdf:RDF>
```