Encoding Provenance Metadata for Social Science Datasets

Carl Lagoze[1], Jeremy Willliams[2], and Lars Vilhuber[3]

[1] School of Information, University of Michigan, Ann Arbor, MI
clagoze@umich.edu
[2]Cornell Institute for Social and Economic Research, Cornell University, Ithaca, NY
jw568@cornell.edu
[3]School of Industrial and Labor Relations, Cornell University, Ithaca, NY
lars.vilhuber@cornell.edu

Abstract. Recording provenance is a key requirement for data-centric scholarship, allowing researchers to evaluate the integrity of source data sets and reproduce, and thereby, validate results. Provenance has become even more critical in the web environment in which data from distributed sources and of varying integrity can be combined and derived. Recent work by the W3C on the PROV model provides the foundation for semantically-rich, interoperable, and web-compatible provenance metadata. We apply that model to complex, but characteristic, provenance examples of social science data, describe scenarios that make scholarly use of those provenance descriptions, and propose a manner for encoding this provenance metadata within the widely-used DDI metadata standard.

Keywords: Metadata, Provenance, DDI, eSocial Science

1 Introduction

Quantitative social science has, for decades, been at the forefront of data-centric research methodologies [1]. An important foundation of this has been 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. This curated data, combined with the ever-increasing volume of social data on the web, offers exciting new research directions for scholars in economics, sociology, demographics, environmental science, health, and other fields.

However, the maturation of cyberinfrastructure for e-social science faces a number of hurdles, some of which are common across all eScholarship efforts, and some of which are exacerbated by or unique to the characteristics of social science data. One of these hurdles, which we described in [2], is the issue of confidentiality; a significant segment of these data 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, etc. Notably, the data are not the only problem, because the metadata may also be subject to disclosure limitation. This may include statutory

Submitted to MTSR 2013.

disclosure restrictions on statistical features of the underlying data, such as extreme values, and even prohibitions on the disclosure of variables names themselves. In [2], we described a method for encoding disclosure attributes in DDI metadata [3].

In this paper, we address the issue of encoding provenance of social science data. A number of characteristics of social science data including the divide between inter-related private and publicly accessible data and metadata, complex multithreaded relationships among these data and metadata, and the existence of partially-ordered version sequences make it difficult to understand and trace the origins data that are the basis of a particular study. This places unacceptable 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.

Recent work undertaken under the auspices of the World Wide Web Consortium (W3C) provides the foundation for a semantically-rich and practical solution for encoding provenance. The PROV documents "define a model, corresponding serializations and other supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments such as the web" [4].

In this paper, we report the results of our experiments with the PROV model for encoding real-world provenance scenarios associated with existing social science data. We also propose a preliminary method for encoding that provenance information within the metadata specification developed by the Data Documentation Initiative (DDI) [3], which is emerging as the de facto standard for most social science data. We show that, with some refinements, the PROV model is indeed suitable for this task, and thereby lays the groundwork for implementing user-facing provenance applications that could enrich the quality and integrity of data-centric social science.

The work reported here is one thread of an NSF Census Research Network award [5]. 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 (CED^{2}AR). 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 The Provenance Problem

As defined by [6], which provides an excellent survey of the data provenance landscape, "data provenance, one kind of metadata, pertains to the derivation history of a data product starting from its original sources ". As they state, this provenance metadata is fundamental to the scientific process because "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".

Prior to the emergence of the Web, data and the means for encoding their provenance were generally siloed in specific applications and domains [7]. The web context has fundamentally changed this partitioned environment by providing accessibility to data from multiple sources, in heterogeneous formats, and of varying integrity and quality [8]. This has increased the importance of semantically-rich and interoperable provenance metadata for understanding the lineage of and integrity of the "mash-ups" that are facilitated by web data [9]. The development of the PROV model, which builds on earlier work on the Open Provenance Model (OPM) [10], leverages developments by the semantic web community, such as RDF and OWL, to provide both the semantics and interoperable encoding that are necessary to express provenance in the web environment.

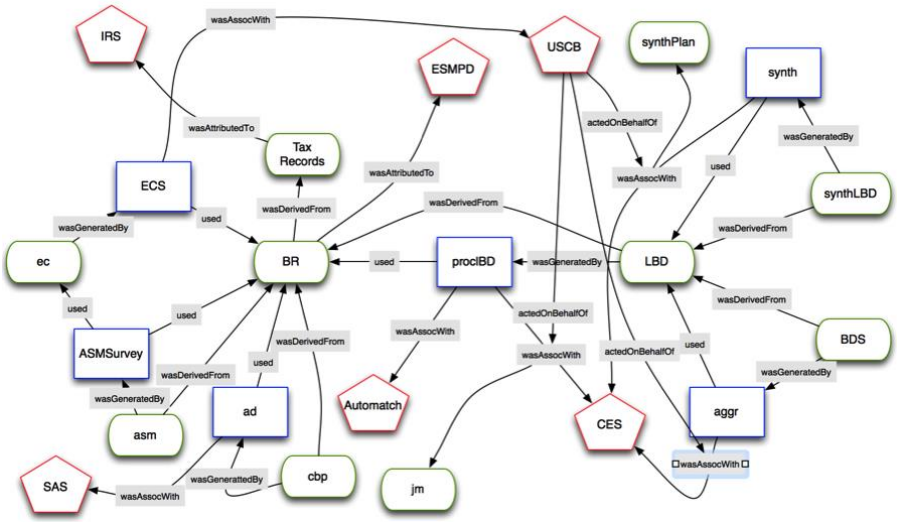
3 Applying the PROV Model to Social Science Scenarios

The W3C PROV Model is fully described in a family of documents [4] that cover the data model, ontology, expressions and various syntaxes, and access and searching. The model is based in the notions 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 three 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 explication of the model and paper assumes that the reader has a working familiarity with PROV.

In our earlier paper [4], we informally described the provenance of the production cycle of two frequently-used social science data products; Longitudinal Business Database (LBD) and the Longitudinal Employer-Household Dynamics (LEHD) database. In this section, we formalize these descriptions using PROV classes and properties. The diagrams that follow are simplified for legibility and do not represent the full graph as it would be constructed in a production-quality system. Our diagramming convention follows that used throughout the W3C PROV documentation; oval nodes denote entities, rectangular nodes denote activities, and pentagonal nodes denote agents. We pair each diagram with a declaration of its component entities, activities, and agents expressed in PROV-N, a functional notation meant for human consumption [11]. 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.

The Census Bureau's Longitudinal Business Database (LBD) is at the center of a complex provenance graph that is illustrated in **Fig. 1**. 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 [12], 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, and the Synthetic LBD, 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 RDCs. Furthermore, the BR serves as direct input to the County Business Patterns (CBP) and related Business Patterns through aggregation and disclosure protection mechanisms.

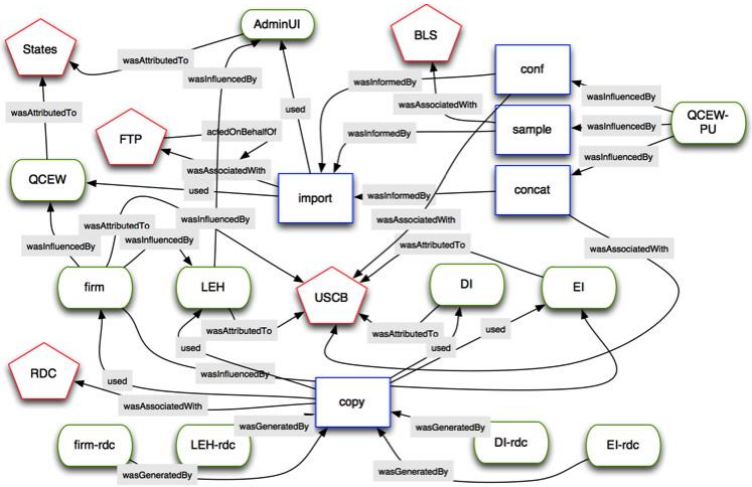



```
entity(cdr:LBD, [prov:type='cdr:dataset', prov:label="Longitudinal Business Data"])
entity(cdr:synthLBD, [prov:type='cdr:dataset', prov:label="Synthetic LBD"])
entity(cdr:BDS, [prov:type='cdr::dataset', prov:label="Business Dynamics Statistics"])
entity(cdr:BR, [prov:type='cdr:dataset', prov:label="Business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

Fig. 1. Longitudinal Business Database (LBD) provenance




```
agent(cdr:States, [prov:type='prov:Organization, prov:label="State Agencies"])
agent(cdr:FTP, [prov:type='prov:SoftwareAgent')
agent(cdr:USCB, [prov:type='prov:Organization, prov:label="US Census Bureau"])
agent(cdr:BLS, [prov:type='prov:Organization, prov:label="Bureau of Labor Statistics"])
agent(cdr:RDC, [prov:type='prov:Organization, prov:label="Research Data Center"])

entity(cdr:AdminUI, [prov:type='cdr:dataset', prov:label="State Admin Unemploy Insur"])
entity(cdr:QCEW, [prov:type='cdr:dataset', prov:label="State Census of Employ & Wages"])
entity(cdr:firm, [prov:type='cdr:dataset', prov:label="Firm/Estab Characteristics"])
entity(cdr:LEH, [prov:type='cdr:dataset', prov:label="Longitudinal Employ History"])
entity(cdr:DI, [prov:type='cdr:dataset', prov:label="Demographic Info"])
entity(cdr:EI, [prov:type='cdr:dataset', prov:label="Ebterprise Info"])
entity(cdr:QCEW-PU, [prov:type='cdr:dataset', prov:label="QCEW Public Use"])
entity(cdr:firm-rdc, [prov:type='cdr:dataset', prov:label="Firm/Estab Characteristics"])
entity(cdr:LEH-rdc, [prov:type='cdr:dataset', prov:label="Longitudinal Employ History"])
entity(cdr:DI-rdc, [prov:type='cdr:dataset', prov:label="Demographic Info"])
entity(cdr:EI-rdc, [prov:type='cdr:dataset', prov:label="Ebterprise Info"])

activity(cdr:import, [prov:label="import state data"])
activity(cdr:conf, [prov:label="confidentiality protection"])
activity(cdr:sample, [prov:label="sample frames"])
activity(cdr:concat, [prov:label="concatenate quarterly QCEW"])
activity(cdr:copy, [prov:label="copy census data to RDC"])
```

Fig. 2. LEHD/QWI Provenance

A similar complex set of relationships exists for the Longitudinal Employer-Household Dynamics (LEHD) Infrastructure files, illustrated in **Fig. 2**. Published since 2003, the Quarterly Workforce Indicators (QWI) are derived from a complex set of combined firm-, job- and person-level files. The key inputs are administrative files from the Unemployment Insurance (UI) system, which are managed by each of the states of the union. The states also maintain an establishment-level set of related files, typically referred to as the Quarterly Census of Employment and Wages (QCEW). A snapshot is sent to the Census Bureau every quarter, where they are combined with historical data from previous quarters, additional demographic information matched from sources at the Census Bureau, and enterprise information from, among other sources, the Business Register. The resulting establishment-level flow statistics are

further aggregated by geographic areas, using disclosure protection methods (noise infusion and suppression). Longitudinal linking and imputation of workplace geography for workers leads to revisions of historical quarters. The entire collection of time series is republished every quarter. Each revision of each file in this system, whether internal or published, has a unique identifier. A snapshot is made of the entire system approximately every four years for use by researchers in the Census Research Data Center, and can be associated with a specific release,

4 Leveraging PROV descriptions for social science research

These formal, machine-readable provenance descriptions serve as the foundation for sophisticated provenance queries by researchers who want to understand the foundations of data they are using. The following scenarios illustrate the utility of queries upon the provenance graph. Due to space limitations, we do not specify the nature of the query mechanism used here (as noted in [13] there are a variety of querying mechanisms for PROV), however in our implementation we are using SPARQL on RDF graphs generated from the PROV descriptions.

4.1 Scenario 1:

An analyst investigating employment trends is making use of the public use version of the Quarterly Census of Employment and Wages (QCEW-PU). She hears from a fellow researcher that some of the data within the QCEW were derived from data collected from the states. Furthermore, she has read about computational problems in Florida during 1995 that affect the integrity of the data collected by that state. She needs to use the June 1999 version of the QCEW-PU and wants to see whether the troublesome Florida data might underlie this synthesized data.

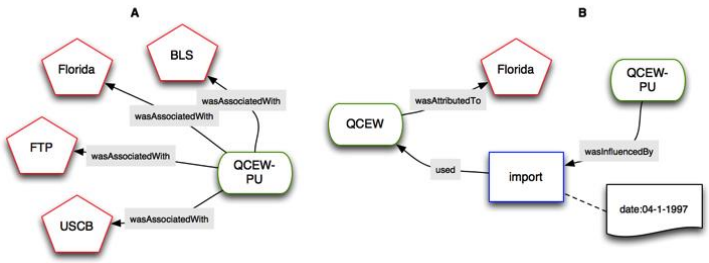


Fig. 3. Discovering the agents directly or indirectly responsible for a data set (left pane A) and the metadata of an activity for which a specific agent is responsible (right pane B)

As illustrated in **Fig. 3** the analyst can first query the provenance system to discover the agents are involved in the provenance chain of the QCEW-PU data set. Upon doing this she sees that the state of Florida is included among a set of agents (left pane A). She zeroes in on the metadata of the activity (import) associated with linkage to

Florida data. Examining that metadata she sees that the import from Florida occurred on April 1, 1997 (right pane B), outside the time period when there were problems with the data from Florida. Relieved, she continues with her research.

4.2 Scenario 2:

A university data archive wants to make data about long-term business trends available to its researchers. The LBD is such a dataset, but is encumbered with confidentiality restrictions that prevent the archive from acquiring the data themselves. The data archivist wants to explore related datasets that are publicly available, and what processes were used to generate them, in order to house them and make them accessible for university use. PROV supports, not only broad queries as demonstrated by the previous scenario, but deep queries into a particular process. In this case the provenance query focuses upon the process by which the Longitudinal Business Database is synthesized into a publicly accessible version.

The query issued by the archivist to the provenance system facilitates the exploration of the data production process by returning the entities that were derived from the LBD, the activities that generated those entities, any events that were connected to those entities, and any plans that were associated. In this case, the plan reveals documentation that will provide further detail to inform the archivist. A diagram of the resulting graph is given in **Fig. 4**.

While some details are needed for the sake of clarity (from this and other graph diagrams in this paper), the diagram in **Fig. 4** shows a bit more of the capabilities of PROV for diving into the details of a particular activity in terms of derivation and generation events, as well as more detailed associations and documentation via plans.

4.3 Scenario 3:

Updates to LEHD files occur quarterly, including updates to previous quarter's inputs. A researcher wants to examine the history of these updates in order to understand the nature of longitudinal data. She wishes to ask the following questions. What process updates existing data with improved data from previous quarters? How do these relate to the annually updated Business Dynamics Statistics?

PROV provides a sub-class of `Entity` called `prov:Bundle` that is in itself an entity and that an entire provenance graph. This can be extremely useful when trying to understand recurring data production processes and their results. If a new PROV graph is produced along with the other metadata that documents the data derivation process, a graph of graphs is possible, enabling the dimension of time to be queried across cycles. This scenario requires such a feature to capture the differences that took place over the quarterly updates to the various files within the provenance graph. The diagram in **Fig. 5** represents two quarterly updates of the LEHD.

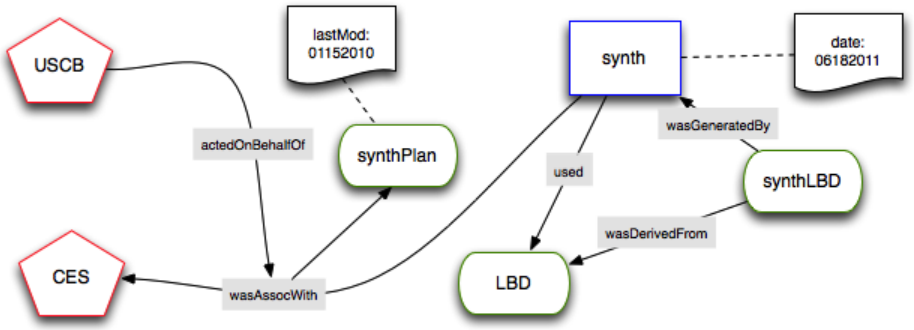


Fig. 4. Discovering the provenance chain of the synthetic LBD data set, including events, plans, and agents and their metadata.

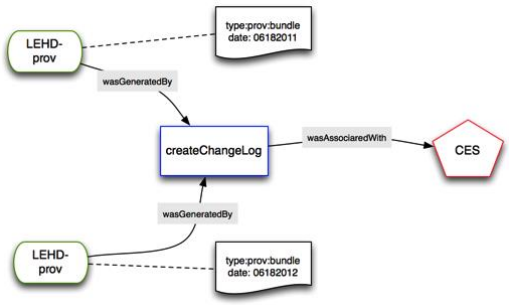


Fig. 5. the prov:Entity subclass, Bundle, is shown. In this case, an activity that creates a change log is discovered for two succeeding quarterly data processing cycles.

The diagram in **Fig. 5** is limited to only two quarters of activity for the sake of illustration. At the scale of decades and centuries of provenance chains, this would be a truly valuable resource to the researcher.

5 Extending PROV Semantics

We found the PROV ontology to be a suitable way of expressing complex provenance chains for the LEHD and LBD. As shown by the scenarios above, the semantics defined by its classes and relations are sufficient for both broad descriptions of database production paths and detailed querying of these paths. We are confident that these two provenance graphs and associated exploration scenarios are adequate exemplars for others in the social science domain, but plan to continue experiments with other data.

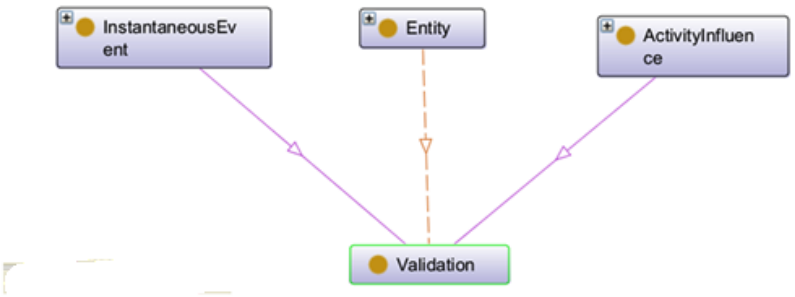


Fig. 6. the inheritance of the proposed Validation class.

We did find in our experiments an area for which the ontology and data model might be specialized for our domain and possibly others. That is, the processes related to *validation*, which plays an important role in the documentation of derived datasets. Querying for actions influenced by validation or events related to validation is arguably a central requirement for social science research, specifically as a means of reporting on measures taken to ensure non-disclosure of confidential data in the production cycle. According to the PROV documentation, the `Plan` class is intended to serve the role of indicating validity throughout the model, and can be used "to validate the execution as represented in the provenance record, to manage expectation failures, or to provide explanations." [14] We found that, for our examples, this way of representing validation is indirect resulting in more complex queries that are slower to perform. We propose an alternative; provide a `Validation` class with related properties within the ontology.

If the ontology were extended to include validation semantics, it could be defined as a specialized `Activity` that provides additional information about how a particular `Entity` was confirmed to be valid. An instance of `Validation` would provide additional descriptions about a `prov:wasValidatedBy` relation linking a `prov:Entity` that was validated to the `prov:Activity` that validated it.

Fig. 6 illustrates the relationship of the proposed `Validation` class to other aspects of the PROV ontology. Our proposal is similar to the manner in which the notion of `PROV:Invalidation` is positioned within the ontology. That is, it inherits both from the `InstantaneousEvent` and `ActivityInfluence` classes. As defined in [14], an `prov:InstantaneousEvent` "happens in the world and marks a change in the world, in its activities and in its entities." A `prov:ActivityInfluence` is the "capacity of an activity to have an effect on the character, development, or behavior of another." While `prov:Invalidation` indicates the negative change of validity status, in our case, `Validation` would indicate a positive change of validity status of a given entity or activity.

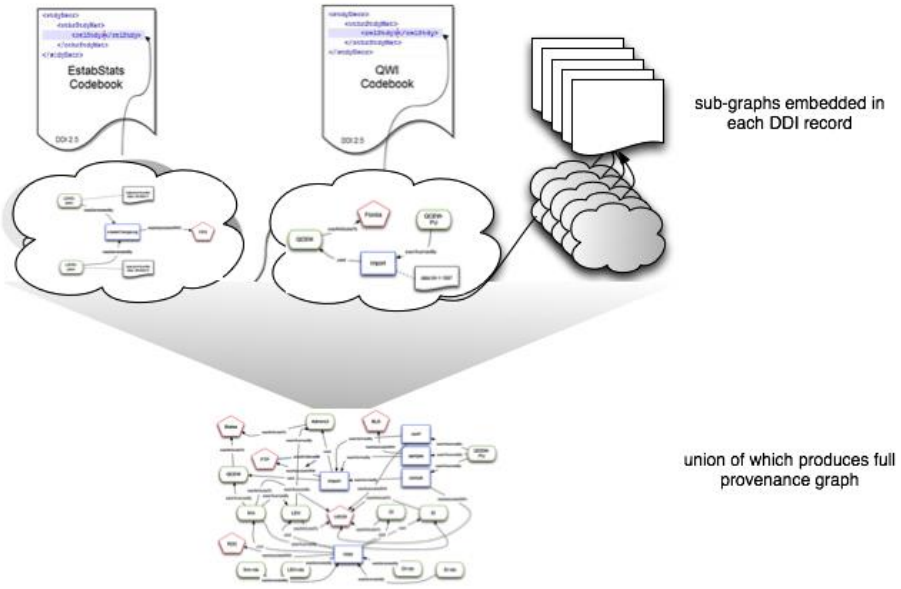


Fig. 7. Storing provenance subgraphs related to a given resource within the <relStudy> element in the corresponding DDI metadata. That subgraph would link, by resource, to other subgraphs located in other codebooks and facilitate dynamic generation of the entire provenance graph.

6 Integrating DDI and PROV

As mentioned earlier, DDI (Data Documentation Initiative) [3] has emerged as the standard for encoding metadata for social science data sets. 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 described in this paper. We argue that rather than encoding provenance in a manner unique to DDI, a better strategy might be to work within the simpler DDI-Codebook context and embed PROV metadata within the individual data set-specific DDI records.

We specify here an easily implemented manner for embedding this information, which we plan to implement. We note that there is an active effort within the DDI community to develop an RDF encoding for DDI metadata that could easily accommodate RDF-encoding provenance metadata [15–17]. As that effort matures, we an-

ticipate that our experiments with provenance and social science data will be a valuable contribution.

For now, the <relStdy> element in DDI 2.5 provides a useful place to encode provenance data specific to the respective data set. The solution is schema-conformant if the PROV metadata is wrapped within a CDATA tag. As illustrated in **Fig. 7** our proposal for doing this is fully modular. Only the metadata related to the specific data set is stored in its respective DDI record, which then links to the PROV metadata stored in other DDI records. The full provenance graph can then be reconstructed dynamically by combining these individual subgraphs.

7 Future Work and Conclusions

Data provenance is an essential aspect of the scholarly process. Researchers need to have access to, explore, and query the lineage of the data that they use for their current research in order to understand and evaluate their integrity, promote reproducibility of the results, and evaluate suitability for future research. This has become especially critical in the web environment where data are sourced not only from established archives, but from many mixed credentialed providers.

In this paper, we have described our initial experiments with the PROV data model, an emerging standard for expressing and encoding provenance, the development of which was sponsored by the W3C. PROV is semantically-expressive and extensible making it a useful platform for encoding provenance in an interoperable and machine-actionable manner. Our experiments described here using PROV to encode some complex, but real, examples from social science research indicate the utility of the model.

Over the next year we plan to build on this work in the following threads. First and foremost, just because it is possible to encode provenance does not mean that is practical. Our modeling efforts described here took a considerable amount of human effort and encoding of tacit knowledge. We plan to work with our partners in the social science data community, especially researchers at ICPSR, who are investigating automatic means of recording provenance within the context of work on DDI-Lifecycle. Second, the utility of these provenance descriptions depends on their comprehensibility by the eventual user. We look forward to contributing our special knowledge of social science researchers and data provenance requirements to general work in the provenance community on visualization of provenance graphs. Finally, as we mentioned earlier in this paper, confidentiality and cloaking of data and metadata are essential within the social science domain. In this vein, we are planning to integrate our confidentiality and provenance work to make it possible for researchers to understand and explore provenance information regardless of their security status via selectively cloaking sensitive aspects of the provenance chain, while exposing all other possible data.

References

[1] M. Daw, R. Procter, Y. Lin, T. Hewitt, W. Ji, A. Voss, K. Baird, A. Turner, M. Birkin, K. Miller, W. Dutton, M. Jirotka, R. Schroeder, G. de la Flor, P. Edwards, R. Allan, X. Yang, and R. Crouchley, "Developing an e-Infrastructure for Social Science.," in *Proceedings of e-Social Science'07*, 2007.

[2] C. Lagoze, W. Block, J. Williams, J. M. Abowd, and L. Vilhuber, "Data Management of Confidential Data," in *International Data Curation Conference*, 2013.

[3] M. Vardigan, P. Heus, and W. Thomas, "Data Documentation Initiative: Toward a Standard for the Social Sciences," *The International Journal of Digital Curation*, vol. 3, no. 1, 2008.

[4] Paul Groth and L. Moreau, "PROV-Overview: An Overview of the PROV Family of Documents," W3C, 2013.

[5] National Science Foundation, "NSF Award Search: Award#1131848 - NCRN-MN: Cornell Census-NSF Research Node: Integrated Research Support, Training and Data Documentation," 2011.

[6] Y. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM Sigmod Record*, 2005

[7] J. Cheney, S. Chong, N. Foster, M. Seltzer, and S. Vansummeren, "Provenance," in *Proceeding of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications - OOPSLA '09*, 2009, p. 957.

[8] Paul Groth, Yolanda Gil, James Cheney, and Simon Miles, "Requirements for Provenance on the Web," *International Journal of Digital Curation*, vol. 7, no. 1. pp. 39–56, 2012.

[9] D. L. McGuinness,P. Fox,P. Pinheiro da Silva,S. Zednik,N. Del Rio,L. Ding,P. West,C. Chang, "Annotating and embedding provenance in science data repositories to enable next generation science applications," *AGU Fall Meeting Abstracts*, vol. 1, 2008.

[10] L. Moreau, J. Freire, J. Futrelle, R. McGrath, J. Myers, and P. Paulson, "The Open Provenance Model," *University of Southampton*, August 2007, pp. 1–30, 2007.

[11] L. Moreau and P. Missier, "PROV-N: The Provenance Notation," W3C, 2013.

[12] R. Jarmin and J. Miranda, "The Longtitudinal Business Database," 2002.

[13] G. Klyne and P. Groth, "Provenance Access and Query," W3C, 2013.

[14] T. Lebo, S. Sahoo, and D. L. McGuinness, "PROV-O: The PROV Ontology," W3C, 2013.

[15] S. Kramer, A. Leahey, H. Southall, J. Vampras, and J. Wackerow, "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, 01-Sep-2012.

[16] T. Bosch, R. Cyganiak, J. Wackerow, and B. Zapilko, "Leveraging the DDI Model for Linked Statistical Data in the Social, Behavioural, and Economic Sciences," *International Conference on Dublin Core and Metadata Applications; DC-2012--The Kuching Proceedings*, Sep. 2012.

[17] T. Bosch, R. Cyganiak, A. Gregory, and J. Wackerow, "DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data," in *Linked Data on the Web Workshop*, 2013.