CED²AR: The Comprehensive Extensible Data Documentation and Access Repository

Carl Lagoze
School of Information
University of Michigan
Ann Arbor,, MI
+1-734-763-1569
clagoze@umich.edu

Lars Vilhuber
School of Industrial & Labor Relations
Cornell University
Ithaca, NY
+1-607-330-5743
lars.vilhuber@cornell.edu

Jeremy Williams
CISER
Cornell University
Ithaca, NY
+1-607-255-4801
jw568@cornell.edu

Benjamin Perry
CISER
Cornell University
Ithaca, NY
+1-607-255-4801
bap63@cornell.edu

William C. Block
CISER
Cornell University
Ithaca, NY
+1-607-255-4801
block@cornell.edu

ABSTRACT
We describe the design, implementation, and deployment 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 through either a web-based user interface or programmatically through a search API, all the while re-reusing and linking to existing archive and provider generated metadata. CED²AR is distinguished from other metadata repository-based applications due to requirements that derive from its social science context. These include the need to cloak confidential data and metadata and manage complex provenance chains.

Categories and Subject Descriptors
H.3.7 [**INFORMATION STORAGE AND RETRIEVAL**]: Digital Libraries – *collection, dissemination, standards.*

General Terms
Design, Standardization

Keywords
Metadata, standards

1. INTRODUCTION
Facilities for the sharing, access to, dissemination of, and curation of data have become an increasingly essential component of the scholarly process. The surge in importance of *open data* for scholarship is reflected in the number of highly-funded initiatives across a broad spectrum of domains. These include the DataONE [29], SEAD [37], and Data Conservancy [27] projects funded through the NSF DataNet program; the emergence of a number of national-level initiatives such as the Australian National Data Service [41]; the proliferation of general-purpose data repositories such as Dryad [11] and figshare [14]; and a host of domain-specific data repositories [7]. In addition, government funding mandates such as NSF and NIH requirements for *data management plans* and policy-level calls for open sharing of data from federally-funded projects [35] add momentum to efforts to build facilities to make those data available.

For over 50 years, quantitative social science has been at the forefront of this so-called *data-centric science*. This research 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. The foundation for this research 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. A significant segment of these source micro-data are confidential because they contain the identities of the subjects of study; e.g., people, corporations, etc. Access to these data is restricted and requires authorized access (via a process similar to obtaining a security clearance) to secure environments known as Research Data Centers (RDC's). In addition, researchers have had unrestricted access to public-use data products, which are synthesized, aggregated, and anonymized derivations of one or more of these confidential data sets.

This traditional research paradigm is being challenged because of the rapidly changing context in which this research takes place. The emergence and maturation of ubiquitous networked computing and the ever-growing data cloud has introduced a spectacular quality and variety of new data sources into this mix. These include massive social media data such as Facebook, Twitter, and other online communities, and an increasing number of open-access social science data repositories[1] [7], which when combined with more traditional data sources, provide the opportunity for studies at scales and complexities heretofore unimaginable. The specificity of the source micro-data (i.e., the references to identities of the subjects of study) is precisely what makes it possible to combine it with these nontraditional data sources. The result has been increasing use over the past two decades of source micro-data (typically confidential) data in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Conference'10, Month 1–2, 2010, City, State, Country.
978-1-4799-5569-5/14/$31.00 ©2014 IEEE.

[1] http://dev.openicpsr.org

publications contrasting with decreasing use of pre-existing survey data (typically public-use) [6]. This new research model has been described by Gary King, a Harvard political scientist, as the *social science data revolution* [18,19].

The confidentiality of these micro-data has led to what can be called a *curation gap*. This arises from the fact that the Census Bureau and many other government agencies in the US are prohibited by statute from granting long-term physical custody of these confidential data to archives with well-established data curation practices such as IPCSR. This is in contrast to the public-use data, which these repositories can take custody of and either ingest, modify, or create the metadata that is essential for the curation process. However, as noted, the findings that are reported in peer-reviewed journals are increasingly based on analyses of the confidential restricted-accessed data. These barriers to access and absence of curation of essential aspects of the provenance chain of research present insurmountable barriers to the essential scholarly tasks of testing research results for validity and reproducibility. This curation gap presents a substantial risk of breach of scientific integrity of the research process itself.

This paper reports on our work to date to address these issues through the design, implementation, and deployment 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 through either a web-based user interface or programmatically through a search API, all the while re-reusing and linking to existing archive and provider generated metadata. CED^{2}AR leverages a number of existing digital library technologies and open standards such as OAI-PMH, DOI's, Dublin Core, and Data Documentation Initiative (DDI). CED^{2}AR is one project within the context of an NSF Census Research Network award titled "Integrated Research Support, Training, and Data Documentation." [34]

The remainder of this paper is structured as follows. The next Section 2 describes features of the social science data context that distinguish it from other digital library contexts and that are relevant to the design of CED^{2}AR. Section 3 describes the overall design of CED^{2}AR, including a description of the target user audience, data sources, system design considerations, and user interface considerations. Section 4 describes our approach to the confidentiality and provenance issues raised in section 2. Section 5 reflects on what we've learned from our deployment experience. We close, in section 6, with a description of future work.

2. QUANTITATIVE SOCIAL SCIENCE

At first approximation, CED^{2}AR is one among many metadata repository systems that are addressable both from a human-facing web UI and machine-oriented API. Other examples of this well-know digital library paradigm include DataONE [29], NSDL [22], OAIster [13], and Public Library of America [36]. In this section we describe three aspects of the social science data context that distinguish CED^{2}AR from other similar applications.

2.1 DDI Metadata

Compared to a number of other digital library domains, quantitative social science is at a relatively advanced stage of metadata development. The DDI metadata format [42], specified by the DDI Alliance[2], originated in 1995 and is one of the most advanced and widely used metadata standard for social science data. It has emerged as a de facto standard and is used by many social science data organizations and projects around the world, including: the Australian Social Science Data Archives, the European Social Survey, the General Social Survey, ICPSR, The Institute for the Study of Labor (Germany), and the World Bank.

There are currently two existing version branches of DDI. The 2.x branch, commonly known as DDI-Codebook, is the more lightweight of the two branches, primarily focusing on bibliographic information about a dataset and the structure of its variables. The current latest version of this branch is 2.5. 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. The current latest version of this branch is 3.2. Version 2.5 was designed for relatively easy upgrade to the version 3 branch. Both versions are expressible in XML and are defined via an XML schema. There is ongoing work on an RDF expression of DDI-Lifecycle, and subsequent publishing of DDI metadata as Linked Open Data [3,4,21]. We decided to implement CED^{2}AR using DDI 2.5 expressed in XML for a number of reasons including existing tools support, lower complexity, adequate functionality, and the promise of easy upgrade to version 3 if that were deemed necessary in the future.

Working within the context of well-established metadata standards has both advantages and problems. It certainly helps avoid "reinventing the wheel". But, as has been documented elsewhere [26,28], any metadata specification strikes a balance between flexibility, extensibility, specificity, and ease of entry. Adding new functionality to a community-consensus metadata standard can be difficult and can, at least temporarily, require going "out of band" (aka violating the standard). Finally, no matter how well-designed a metadata format, it is subject to the imperfections of the human metadata creators. As we will describe later**,** our own experience in CED^{2}AR crosswalking existing metadata records to our own DDI 2.5 demonstrates the complexities of metadata interoperability.

2.2 Confidentiality and Cloaking

Confidentiality and cloaking of selective information in a context-aware manner is a key requirement of any repository of quantitative social science data [1]. 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 variable names themselves.

As a result, our design of the CED^{2}AR system must accommodate two important scenarios of confidential data and/or metadata, which are illustrated in Figure 1. First, more than one version of a single dataset may coexist in both the public and private spheres, with different sets of metadata. A value-added provider may have enhanced the data, or manipulated it in some fashion. A good example is the homogenized datasets provided by the IPUMS (Integrated Public Use Microdata Series) project (http://www.ipums.org0 They are derived from the original Decennial Census data files, which are maintained in their unmodified form by the Census Bureau. Second, a data set may exist only within the protected and secure area of the statistical

[2] http://www.ddialliance.org/

agency, along with a full and complete metadata description. A synchronization protocol may prune that metadata of its confidential elements, making it available as a verifiably released public version of the metadata. If the public version of the metadata is enhanced, for instance by users or IPUMS, synchronization back across the firewall of the secure area should allow the internal, confidential metadata to also benefit from such enhancements.

In [23], we described a method to accommodate this requirement by encoding appropriate disclosure attributes in DDI metadata. We summarize these results in Section 4.1.

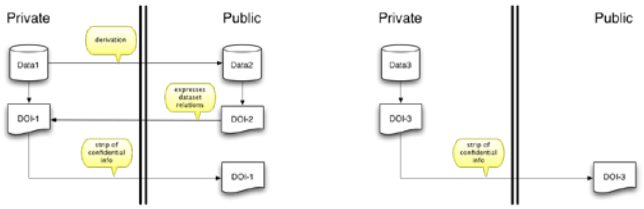

Figure 1. Two scenarios of confidential data and/or metadata. On the left, a data set exists in both a public and private (filtered and possibly enhanced) version, each with its own metadata, public and private, respectively; in addition, a filtered version of the private metadata is exposed publicly. On the right, only a single private data set exists with its own private metadata that is then filtered to the outside for the public use.

2.3 Provenance

Our work on CED²AR has also had to take into account the complex provenance of the data that we wish to make available to our research users, and the importance of that provenance for the integrity of the resulting research. 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 [43].

The complexity of provenance only increases when these traditional archival-based data are mixed with web-based more-informal data. Furthermore, as indicated by the increasing momentum of efforts like linked open data [16], architecturally-supported silos separating discipline-specific data, each addressing essential requirements, like provenance, are not addressing the demands of 21st-century research, which is increasingly interdisciplinary. The need for a "web-wise" solution to the provenance issue [5] 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" [12].

In [24,25] we reported on our work with the PROV model for encoding real-world provenance scenarios associated with existing social science data and for embedding that provenance information within XML-encoded DDI metadata. We summarize those results in Section 4.2.

3. DESIGN AND IMPLEMENTATION

3.1 Target User Audience

The requirements of CED²AR have been driven mainly by two primary user archetypes, which are described in this section. The first includes researchers and service providers (data librarians, archivists, etc.) searching for data germane to a given investigation, and the second involves data producers seeking to disseminate their data for future research.

CED²AR was designed, in large part, for users who are interested in a given topic and want to find data related to it within the United States Census Bureau. The infrastructure needed to meet this requirement is not trivial, as metadata describing social science data can be sparse and unstructured. This condition is exacerbated when many of the most relevant datasets contain confidential information and are therefore only available through restricted access to protect the identities of the subjects therein. A means to standardize metadata about both public-use and restricted-access datasets is necessary to facilitate cross-dataset search functionality without disclosing any information that would compromise privacy. Researchers who find a given dataset require a means to explore its composition to determine whether it is congruent to the goals of their investigation, as well as to discover the details of other datasets that are related to it by topic and/or provenance. Notably, the features that accommodate these scenarios also enable researchers to review the integrity of data products to assure the quality of scientific findings.

While standardizing existing public and restricted-access metadata into a searchable repository meets a prevailing need in the research community, CED²AR also provides mechanisms and workflows by which data producers can deposit and describe their data in a manner that will make it discoverable, accessible and

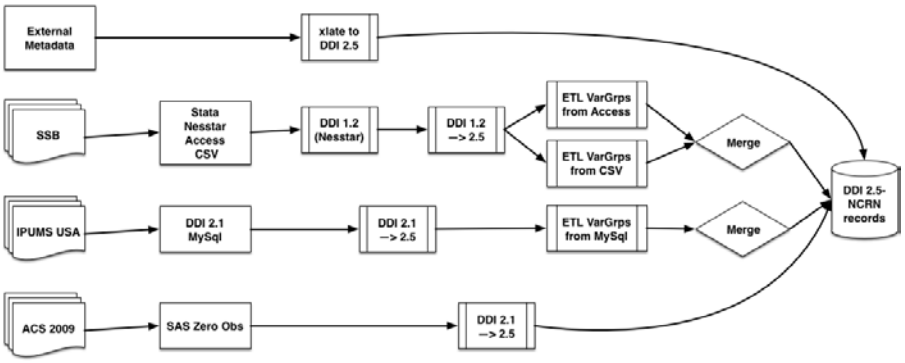

Figure 2. CED²AR metadata workflow overview

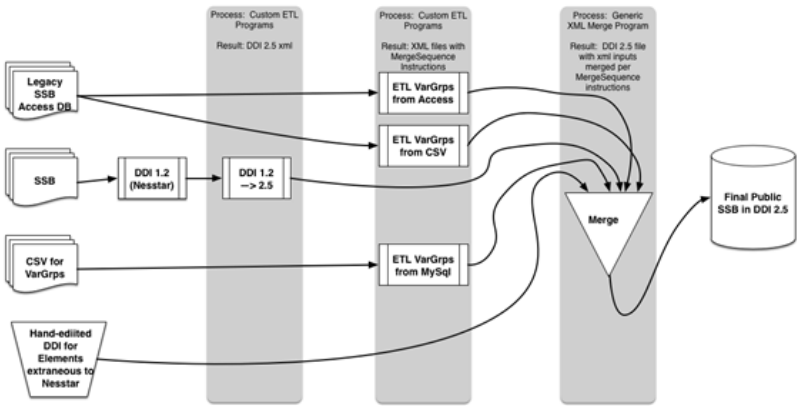

Figure 3. SSB (SIPP Synthetic Beta) Workflow.

comprehensible to future researchers. Sections 3.2-3.4 take a deeper dive into these features, providing some concrete examples to flesh out the approach adopted by the CED²AR team.

3.2 Data Sources

CED²AR is a metadata repository that ingests data from a growing number of disparate sources. The metadata from these sources can be challenging, as they often are sparse, structurally inconsistent, and/or are structured by an amalgam of schemas and formats. The extraction, transformation, and loading (ETL) of these datasets is accomplished by leveraging existing tools and technology whenever possible, but has required custom tool development as well. These tools have been constructed in a modular manner and can be combined to constitute metadata connectors that (at least partially) automate the process for a given data product. Figure 2 provides a sampling of metadata sources that have been standardized and ingested into the CED²AR repository. Two examples may serve to illustrate the challenges one faces when performing ETL.

The SSB (SIPP Synthetic Beta) is a "product that integrates person-level micro-data from a household survey with administrative tax and benefit data"[3]. The project had not previously released DDI metadata, and used an internal custom metadata store to generate a PDF as the sole documentation. The metadata for this project was distributed among files in MS Access, csv, and Stata (Data Analysis and Statistical Software)[4] formats. The workflow of the connector for this project is illustrated in Figure 3. In order to generate a first DDI codebook for the SSB, we used Nesstar Publisher[5] to extract some of the metadata from the native Stata files. Nesstar Publisher currently exports DDI 1.2.2, which we then mapped to CED²AR repository standard DDI 2.5 using custom tools, forming the master metadata file for SSB. Custom code was written to produce blocks of DDI-Codebook XML representing variable categories, and variable descriptions that were previously maintained in an Access database as well as a CSV file. These blocks were then merged into the master XML file with the CED²AR XML-Merge program, ultimately creating the base metadata file for the repository.

The ingest of metadata provided by the Integrated Public Use Microdata Series (IPUMS) USA [40] is another good example. IPUMS maintains and integrates an extensive collection of US Census data (Decennial Census and American Community Survey). IPUMS provided our project with metadata in DDI-Codebook format. However, as of 2012, they also use a relational database, which contains more information (on concepts and variable groups) than was exported as DDI. Having obtained an extract of their SQL data, we proceeded to create a custom connector that integrated the information into our copy of the metadata in order to capture the additional information.

The two examples above describe data sources from the public domain. In the future, CED²AR will implement workflows that will automate the secure ingest of metadata about restricted access data into the publicly available repository. This will enable users to get search and discover restricted access data while ensuring non-disclosure of confidential information. Further, plans exist to implement a metadata editing toolset that can help refine datasets, once they meet minimum repository standards.

3.3 System Design Considerations

The core functionality in CED²AR is implemented through three Java web applications. These components are illustrated in Figure 4. The first is a user interface built with Spring MVC using a semi-RESTful web architecture. The second application is an API constructed with the Restlet framework. The third application is a BaseX XML database, containing the DDI codebooks. Everything is run on Apache Tomcat 7. Each application is compiled with Maven to maintain strict dependency control. All data is retrieved through the API, the web frontend has no direct connection to BaseX. In BaseX, codebooks are uniquely identified by a file handle. This is a unique alphanumeric key, which indexes a specific DDI codebook. Variables are referenceable by their name attribute. For data integrity, variable names must be unique within a codebook.

The API is actually split into two web applications. The first is a read-only API, intended for the web frontend to use. The second part is the editing API, which handles uploading, editing and indexing codebooks. The editing API also handles XML schema validation, and sanitizes the XML from poorly-formed or malicious markup.

The web frontend permits searching and browsing through codebooks. Common search interface features are present, such as advanced searching, comparison views, and filtering. Variable grouping and concept categorization is also used when included by the publisher.

[3] http://www.census.gov/programs-surveys/sipp/data.html

[4] http://www.stata.com

[5] http://www.nesstar.com

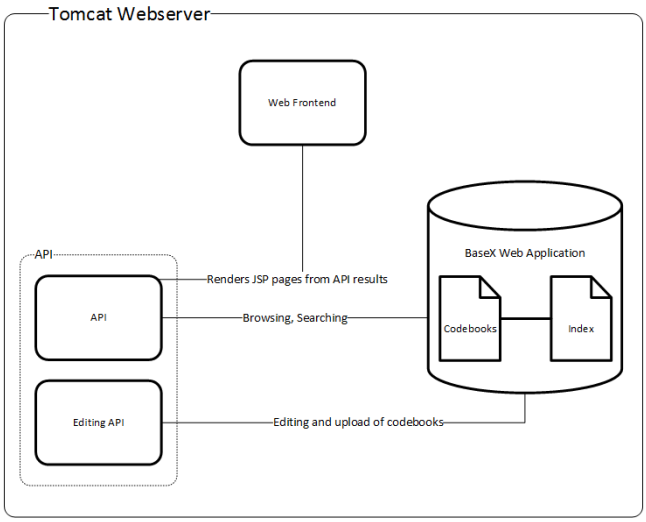

Figure 4. CED²AR implementation architecture

3.4 User Interface Considerations

CED²AR's frontend is built using a combination of Twitter Bootstrap and JQuery. The website scales for any platform, including desktops, laptops, tablets, and smart-phones. All styling works across any modern browser. Functionality increases with the learning curve of our users. For example, like most digital library search engines, power users can employ advanced search options through shortcuts, while new users can use an interactive form. CED²AR includes inline help as well as formal documentation. Navigation is faceted, as users can find the same information through searching, browsing, navigating breadcrumbs and manipulating URLs. In addition, the majority of URLs are concise and human readable. Controls are clustered and lists are brief to reduce memory load. Visual feedback is provided through the use of icons and messages. Aesthetically, the frontend takes a minimalistic approach, by using as little visual stimuli as necessary to inform the user. In addition, by limiting unnecessary features, CED²AR maintains high performance, even with large searches, to keep users engaged. The most current stable version of CED²AR can be found at `http://www2.ncrn.cornell.edu/ced2ar_web/`.

3.5 Identifier strategy

Defining an identity strategy is an essential first step in the design of any content-based system. As described in [8], there are variety of requirements for identifier schemes including persistence, atomicity, uniqueness, etc. However, a necessary precursor step of the choice of an identifier scheme is the definition of the entity to which identifiers are associated. This is particularly difficult and problematic for data because of the imprecise and ambiguous notion of what is a "data set". As [39] point out "the notion of 'data set' found in the literature cannot itself be provided with a precise formal definition". Consequently, the decision about an entity/identifier association is necessarily heuristic, user-driven (i.e., what do the users of the system conceptually consider to be a data set, motivated in part by that which they wish to cite), and application-specific rather than technical and algorithmic.

Because CED²AR is metadata-driven, we are following the rule that an externally, globally-identify data set is one for which we have created a DDI metadata record. This is independent of whether the data exist physically across several files, a case that is well-accommodated by DDI, instances of which can refer to one or more internally-identified data files. The alternative of matching a unique global identifier one-to-one to a data file would not make sense in our situation because these data files do not have a logical correspondence to entities that users care about.

Our initial decision is that CED²AR should use the well-known Digital Object Identifier (DOI) for assigning persistent identifiers to data.. Virtually all academic publishers assign DOIs at the article level in all of their publications. In addition, DOIs are increasingly used to identify data. In this vein, DataCite [38] has emerged as an international consortium that manages DOIs for datasets and that provides or is developing core infrastructure for dataset citation, discovery, and access. Apropos of the last functionality, access, DataCite DOIs resolve to a public landing page for the dataset that contains metadata-derived information about the associated dataset and a direct link to the dataset access method itself. Technically, therefore, the DOI identifies the metadata, which may then provide one or more access points to data files described by the metadata, which conforms to the entity definition strategy in CED²AR. By leveraging DataCite, we join a growing community of data providers and can interoperate at the identifier level with those other data providers.

The DataCite consortium provides two mechanisms for minting new DOIs and registering them with the Handle System. One can either apply to become a full member of the consortium, and then run a Handle System node, or contract with an existing member/service that will then mint DOIs and register them upon request. We determined that the full member route was too complex for our needs. As an alternative, we have decided to use the EZID service[6] provided by the California Digital Library, which is an easy and cost-effective way to maintain and manage DataCite DOIs through a user interface and an API.

We note a serious limitation of DOI's that has been observed by others and which arises from our requirements for the continued development of CED²AR. For the purposes of establishing variable-level provenance, we would like to uniquely identify a variable within a DDI codebook, which we do identify with a DOI in CED²AR. The notion of coining a unique DOI for each variable in each data set is intractable, given the quantity of variables. Instead, we would like to be able to suffix the variable name to the data set DOI and have that suffix "pass-through" to the URL to which the DOI resolves. For example, if `doi:123.56/ds1` resolved to `http://ced2at.org/data/ds1`, we would like `doi:123.56/ds1//varx` to resolve to `http://ced2at.org/data/ds1/varx` without having to register the second DOI. This notion of "suffix pass-through" is implemented within the ARK[7], which we have been exploring for future development.

4. CONFIDENTIALITY & PROVENANCE

In this section we describe our work targeted at these two critical requirements of quantitative social science data.

4.1 Confidentiality Constraints in DDI

Our initial CED²AR implementation supports data-hiding at two levels, which matches the requirements stated earlier and covers most of the needs of our existing data. The first, which is required by many statistical organizations to protect the anonymity of data, is the hiding of statistical or other attributes (extreme values, precise distributions, variable names, etc.). The second is hiding of variables themselves. Hiding can be implemented by visually

[6] http://ezid.cdlib.org

[7] https://wiki.ucop.edu/display/Curation/ARK+Suffix+Passthrough

suppressing the information when displaying the data, or by pruning the DDI XML itself as part of the ETL or metadata synchronization process.

DDI already includes two structural components that accommodate the second form of data hiding. The first is the `<dataAccs>` element, which is nested within the `<stdyDscr>` element, one of the eight main structural branches nested within the root `<codeBook>` element of DDI 2.5. It is possible to list multiple `<dataAccs>` elements, each with unique IDs, and then via the contained `<conditions>` element define a set of hiding conditions. Through the use of a controlled vocabulary for the value of the `<conditions>` element this setting can be machine-readable and hiding therefore can be programmatically controlled. Figure 5 illustrates this showing three hiding rules labeled A1, A2, and A3

```
<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>
```

Figure 5. Using the `<dataAccs>` element to express hiding rules

Figure 6 shows the application of the hiding rules defined in Figure 5 to specific variables through the use of the `access` attribute. As shown, the variable `totfam_kids` is public as defined by rule A1, and the variable `totinc` is private as defined by rule A2, and therefore should be stripped from any metadata record that is exposed outside of the confidential area.

```
<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>
```

Figure 6. Application of hiding rules to specific variables

Hiding of statistical attributes is not accommodated by the current specification of DDI 2.5. We have proposed a minor enhancement to the DDI codebook schema that would permit the attachment of the access attribute to various existing XML elements, such as the `<catgry>` element, in addition to its current allowance in the `<var>` element. This is illustrated in Figure 7 where it is specified that the variable `totinc` is public, according to access rule A1, but the category 4 value (indicating income of $250,000 and above) is confidential according to access rule A2. The full XSD for DDI 2.5-NCRN, against which the code fragment Figure 7 would validate, can be found at `http://www.ncrn.cornell.edu/index.php/projects/ced-ar`, and has been proposed to the DDI Alliance for incorporation into the official DDI-Codebook specification.

4.2 Encoding Provenance in DDI Metadata

As mentioned earlier, provenance is an essential aspect of data and research integrity, and, therefore, an important part of our work on CED^{2}AR. This work is still in its early stages, and this section summarizes the results thus far and plans as we move ahead. The aspects of this work are described in the sub-sections below.

```
<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>
```

Figure 7. Application of hiding at the value level.

4.2.1 Defining the entities in the provenance chain

The current focus of our work is dataset provenance, as opposed to variable-level provenance. We agree that source provenance at the cell (variable) level is a potential issue, but it is a much more complicated issue, not least because the information about variable-level provenance is typically not available to third parties in the desire detail (an ongoing issue of replicability). At this point, we tackle the (in real life) more tractable problem of provenance of datasets first.

4.2.2 Modeling well-known provenance instances

The Census Bureau's Longitudinal Business Database (LBD) is at the core of many economics papers[8]. It is also at the center of a provenance graph that is illustrated in Figure 8, which is useful for understanding provenance in this domain. 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 [17], and it is being continually maintained (updated yearly) at the Census Bureau. Derivative products of the LBD are the Business Dynamics Statistics (BDS) [15], an aggregation of the LBD and the Synthetic LBD, a confidentiality-protected synthetic microdata version of the LBD [20]. 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 and confidential versions are available within the Census RDCs. Furthermore, the BR serves as

[8] http://goo.gl/KS6ts

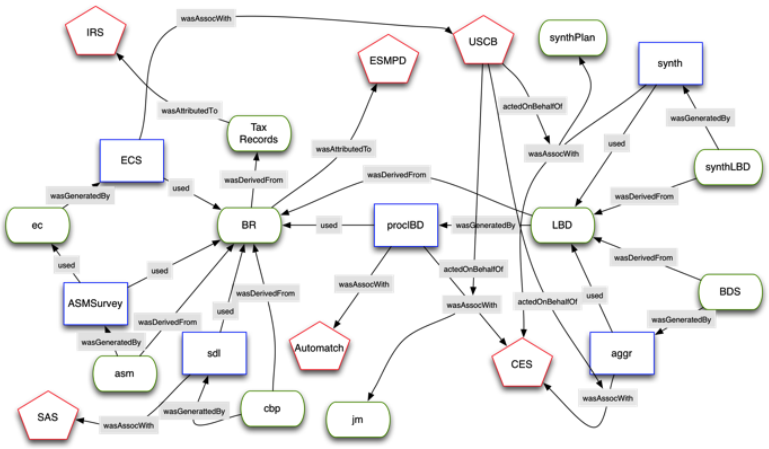

Figure 8. Longitudinal Business Database (LBD) provenance graph

direct input to the County Business Patterns (CBP) and related Business Patterns, again, through aggregation and related disclosure protection mechanisms (noise infusion [9], coarsening, and suppression).

As described earlier, to represent this provenance we leverage the W3C PROV model that is fully described in a family of documents [30] 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. In [25] we show the full encoding of the declarations of the component entities, activities, and agents of the LBD provenance graph in PROV-N, a functional notation meant for human consumption [32]. Three examples of these declarations are:

```
entity(cdr:LBD, [prov:type='cdr:dataset',
       prov:label="Longitudinal Business Data"])
agent(cdr:USCB, [prov:type='prov:Organization,
       prov:label="US Census Bureau"])
activity(cdr:synth, [prov:label="synthesize"])
```

4.2.3 *Modular approach for provenance in DDI*

Our overall design approach taken for encoding this formal representation of provenance in DDI is modular as illustrated in Figure 10. Only the provenance metadata related to the specific dataset 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 [31]; 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 strategy, to express within the DDI for a specific dataset only its provenance dependencies and independently allow datasets to then express derivation from that existing dataset from 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.

4.2.4 *Encoding provenance instances in DDI XML*

The `<relStdy>` element in DDI 2.5 provides a useful place to encode provenance information specific to the respective dataset. As documented in the DDI 2.5 schema[9], 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, which can take the form of bibliographic citations." We have explored this one possibility of integration - wrapping a PROV bundle into the `<relstdy>` element - recognizing that there are other methods. The key issue is to integrate PROV into DDI, Codebook and Lifecycle, in such a way that allows for both data-creator-related processes or workflows, as well as researcher-related (or archive-maintainer-related) workflows and provenance connections.

Apart from deciding where to put the encoding in the DDI, there's the question of how to actually encode the PROV information itself. In [25] 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, an approach we are not fully satisfied with

As an alternative, we are investigating another approach; leveraging the XML encoding of PROV semantics [33] that would require only some focused changes 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 [3,21]. In the end, we believe that there should be two viable and cross-translatable alternatives - a pure RDF approach and the pure XML approach that we propose - that implementers can choose between based on their comfort with each technology.

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

An example of our modular XML encoding is shown in Figure 9, which corresponds to the illustration of Longitudinal Business Database (LBD) provenance illustrated in Figure 8. As indicated, the LBD is derived from the Business Register (BR), the URI of which joins it to the provenance graph for the BR found in another XML PROV module. This derivation involves a number of other agents both organizational (CES acting on behalf of the Census Bureau) and software (AutoMatch) and the enactment of an established plan (procLBD). Note that a number of details are removed from the XML due to space limitations.

4.2.5 Visualizing and recording provenance

Although we have implemented some preliminary prototypes of this work, our future work focuses on the full production-level implementation within the CED^{2}AR system. One relevant design issue is user visualization and exploration of provenance graphs; work that we are currently undertaking. We are exploring a visualization model where the user can traverse the provenance graph in incremental steps, and avoid being overwhelmed by too much information. We anticipate first release of our implementation in 2nd quarter 2014 (in time for conference presentation if this paper were accepted)..

A major barrier to this work that we recognize is the expense, in both time and human effort, of manually collecting and encoding provenance information. An interesting thread of work in the eScience community focuses on automatic collection of provenance information as part of scholarly work flow [2,10]. Specific to DDI, a number of the emerging tools automatically record survey events within a DDI-Lifecycle description. However, the vast stock of legacy data requires alternative approaches, and for many of them, there are few alternatives than human-guided encoding by knowledgeable researchers and data librarians.

5. DEPLOYMENT EXPERIENCE

As noted earlier, our goal with CED^{2}AR is three-fold: to serve as a testing ground for new metadata technologies, to provide a lightweight, easily-deployed metadata viewer, and to display metadata that previously was not publicly available, or wasn't available at all. We are moving ahead on all three dimensions. We have successfully demonstrated the usefulness of two simple enhancements to DDI-Codebook, showcasing it in the application, and applying the enhancements to real metadata. The current state of the web application relies on very few components – a functional Tomcat server and two webapps – and we are working with the Census Bureau to deploy an instance of CED^{2}AR within the restricted-access compute environment in their research data centers. Finally, our own stable instance of CED^{2}AR is the prime location for previously non-existent metadata for the SIPP Synthetic Beta file and the Synthetic LBD. The SIPP Synthetic Beta documentation is being used to teach graduate students about the data. We are working on additional "under-documented" data, growing the available metadata on CED^{2}AR by the month. We have learned quite a lot about the challenges faced by researchers, who typically are not trained as data archivists or information scientists, but who wish to provide metadata on their research outcomes.

```
<otherStdyMat>
   <relMat>https://census.gov?ces/datasets/lbd.html</relMat>
        <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>
        <prov:plan prov:id="cdr:procLBDPlan">
       <prov:location
xsi:type="xsd:anyURI">http://repec.org/paper/0217.html</prov:l
ocation>
          <prov:type>prov:Plan</prov:type>
          <dct:title>The Longitudinal Business Database (Jarmin
& Miranda 2002)</dct:title>
        </prov:plan>
        <prov:activity prov:id="cdr:procLBD"/>
        <prov:wasDerivedFrom>
          <prov:generatedEntity prov:ref="cdr:LBD"/>
          <prov:usedEntity prov:ref="cdr:BR"/>
        </prov:wasDerivedFrom>
        <prov:wasAssociatedWith>
          <prov:activity prov:ref="cdr:procLBD"/>
          <prov:agent prov:ref="cdr:CES"/>
          <prov:plan prov:ref="cdr:procLBDPlan"/>
        </prov:wasAssociatedWith>
        <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:CES"/>
          <prov:responsible prov:ref="cdr:USCB"/>
          <prov:activity prov:ref="cdr:procLBD"/>
        </prov:actedOnBehalfOf>
        <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>
   </relStdy>
</otherStdyMat>
```

Figure 9. Longitudinal Business Database (LBD) provenance subgraph in PROV-XML (Namespace declarations and other details are elided).

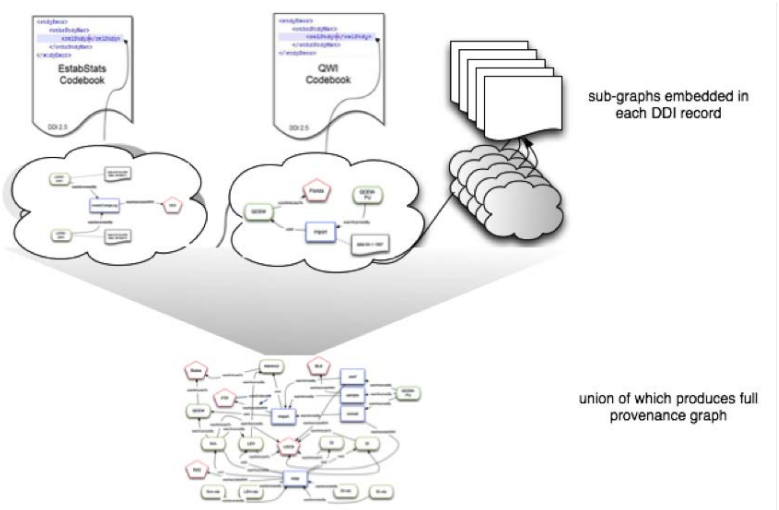

Figure 10. Storing provenance subgraphs related to a given resource within the <relStudy> element in the corresponding DDI metadata. That subgraph links, by resource, to other subgraphs located in other codebooks and ancillary entities (e.g., plans, ages) to allow dynamic generation of the entire provenance

6. FUTURE WORK

In future work, we will continue to explore the possibilities of dynamic and visual exploration of lightly-encoded provenance information, while faced with the challenges of a dispersed and generally incomplete provenance graph. We will also apply the integration of metadata and lightweight provenance encoding to an issue faced by researchers working within restricted-access compute centers: the need to accurately document content and provenance of data and research results that they wish to remove from the restricted-access environment, by showing that such data and research results no longer pose a disclosure risk after they access restrictions have been removed.

ACKNOWLEDGMENTS

We acknowledge NSF grants SES 9978093, ITR 0427889, SES 0922005, SES 1042181, and SES 1131348.

REFERENCES

1. Abowd, J., Vilhuber, L., and Block, W. A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs. In J. Domingo-Ferrer and I. Tinnirello, eds., *Privacy in Statistical Databases (LNCS 7756)*. Springer Berlin / Heidelberg, 2012, 216–225.
2. Barga, R.S. and Digiampietri, L.A. Automatic capture and efficient storage of e-Science experiment provenance. *Concurrency and Computation: Practice and Experience 20*, (2008), 419–429.
3. Bosch, T., Cyganiak, R., Gregory, A., and Wackerow, J. DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data. *Linked Data on the Web Workshop*, (2013).
4. Bosch, T., Cyganiak, R., Wackerow, J., and Zapilko, B. 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*, (2012).
5. Cheney, J., Chong, S., Foster, N., Seltzer, M., and Vansummeren, S. Provenance. *Proceeding of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications - OOPSLA '09*, ACM Press (2009), 957.
6. Chetty, R. The Transformative Potential of Administrative Data for Microeconometric Research. 2012. http://conference.nber.org/confer/2012/SI2012/LS/ChettySlides.pdf.
7. Crosas, M. The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data. *D-Lib Magazine 17*, 1/2 (2011).
8. Duerr, R.E., Downs, R.R., Tilmes, C., et al. On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics 4*, 3 (2011), 139–160.
9. Evans, T., Zayatz, L., and Slanta, J. Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics 14*, 4 (1998), 537–551.
10. Frew, J., Janee, G., and Slaughter, P. Automatic Provenance Collection and Publishing in a Science Data Production Environment - Early Results. *Provenance and Annotation of Data and Processes - Third International Provenance and Annotation Workshop, {IPAW} 2010, Troy, {NY}, {USA}, June 15-16, 2010. Revised Selected Papers*, (2010), 27–33.
11. Greenberg, J., White, H.C., Carrier, S., and Scherle, R. A Metadata Best Practice for a Scientific Data Repository. *Journal of Library Metadata 9*, 3-4 (2009), 194–212.
12. Groth, P. and Moreau, L. *PROV-Overview: An Overview of the PROV Family of Documents*. 2013.
13. Hagedorn, K. OAIster: a "no dead ends" OAI service provider. *21*, 2 (2003), 170–181.
14. Hahnel, M. Exclusive: figshare a new open data project that wants to change the future of scholarly publishing. *Impact of Social Sciences blog*, 2012. http://eprints.lse.ac.uk/51893/1/blogs.lse.ac.uk-

Exclusive_figshare_a_new_open_data_project_that_wants_to_change_the_future_of_scholarly_publishing.pdf.

15. Haltiwanger, J.C., Jarmin, R.S., and Miranda, J. Business Dynamics Statistics: An Overview. *SSRN Electronic Journal*, (2009).

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

17. Jarmin, R. and Miranda, J. *The Longtitudinal Business Database*. 2002.

18. King, G. The Social Science Data Revolution. *Horizons in Political Science*, 2011. http://gking.harvard.edu/files/gking/files/evbase-horizonsp.pdf.

19. King, G. Ensuring the data-rich future of the social sciences. *Science (New York, N.Y.) 331*, 6018 (2011), 719–21.

20. Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., and Abowd, J.M. Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review 79*, 3 (2011), 362–384.

21. Kramer, S., Leahey, A., Southall, H., Vampras, J., and Wackerow, J. Using RDF to describe and link social science data to related resources on the Web: leveraging the Data Documentation Initiative (DDI) model. 2012. http://eprints.port.ac.uk/9029/1/UsingRDFToDescribeAndLinkSocialScienceDataToRelatedResourcesOnTheWeb.pdf.

22. Lagoze, C., Arms, W.Y., Gan, S., et al. Core Services in the Architecture of the National Digital Library for Science Education (NSDL). .in JCDL 2002

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

24. Lagoze, C., Vilhuber, L., Williams, J., and Block, W. Encoding Provenance of Social Science Data: Integrating PROV with DDI. *Proceedings of EDDI13 5th Annual European DDI User Conference*, (2013).

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

26. Lagoze, C. Keeping Dublin Core Simple: Cross Domain Discovery or Resource Description? *D-Lib Magazine 7*, 1 (2001).

27. Mayernik, M.S., Choudhury, G.S., DiLauro, T., et al. The Data Conservancy Instance: Infrastructure and Organizational Services for Research Data Curation. *D-Lib Magazine 18*, 2012.

28. McDonough, J. Structural Metadata and the Social Limitation of Interoperability: A Sociotechnical View of XML and Digital Library Standards Development. *Basilage The Markup Conference 2008*, (2008).

29. Michener, W., Vieglais, D., Vision, T., Kunze, J., Cruse, P., and Janée, G. DataONE: Data Observation Network for Earth — Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. *D-Lib Magazine 17*, 1/2 (2011).

30. Missier, P., Belhajjame, K., and Cheney, J. The W3C PROV family of specifications for modelling provenance metadata. *EDBT/ICDT '13*, ACM Press (2013), 773–776.

31. Moreau, L. and Lebo, T. *Linking across Provenance Bundles*. 2013.

32. Moreau, L. and Missier, P. *PROV-N: The Provenance Notation*. 2013.

33. Moreau, L. *PROV-XML: the PROV-XML Schema*. 2013.

34. National Science Foundation. NSF Award Search: Award#1131848 - NCRN-MN: Cornell Census-NSF Research Node: Integrated Research Support, Training and Data Documentation. 2011. http://www.nsf.gov/awardsearch/showAward?AWD_ID=1131848.

35. Office of Science and Technology Policy. *Increasing Access to the Results of Federally Funded Scientific Research*. Washington D.C., 2013.

36. Peek, R. Digital Public Library of America. *Information Today 29*, (2012), 24.

37. Plale, B., McDonald, R.H., Chandrasekar, K., et al. The SEAD datanet prototype: Data preservation services for sustainability science. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, (2013), 439–440.

38. Pollard, T.J. and Wilkinson, J.M. Making Datasets Visible and Accessible: DataCite's First Summer Meeting. *Ariadne*, 64 (2010).

39. Renear, A.H., Sacchi, S., and Wicket, K.M. Definitions of Dataset in the Scientific and Technical Literature. *Proceedings of the 73rd ASIS&T Annual Meeting*, (2010).

40. Ruggles, S., Alexander, J.T., Genadek, K., Goeken, R., Schroeder, M.B., and Sobek, M. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. 2010.

41. Treloar, A. Design and Implementation of the Australian National Data Service. *International Journal of Digital Curation 4*, 2009.

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

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