The International Journal of Digital Curation
Volume 8, Issue 1 | 2013

Data Management of Confidential Data

Carl Lagoze,
University of Michigan School of Information

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

John Abowd and Lars Vilhuber,
Labor Dynamics Institute,
Cornell University

Abstract

Social science researchers increasingly make use of data that is confidential because it contains linkages to the identities of people, corporations, etc. The value of this data lies in the ability to join the identifiable entities with external data, such as genome data, geospatial information, and the like. However, the confidentiality of this data is a barrier to its utility and curation, making it difficult to fulfil US federal data management mandates and interfering with basic scholarly practices, such as validation and reuse of existing results. We describe the complexity of the relationships among data that span a public and private divide. We then describe our work on the CED^{2}AR prototype, a first step in providing researchers with a tool that spans this divide and makes it possible for them to search, access and cite such data.

International Journal of Digital Curation (2013), *8*(1), 265–278. http://dx.doi.org/10.2218/ijdc.v8i1.259

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Researchers working at the United States Census Bureau and in Census Research Data Centers (RDCs)[1] have acquired and archived a substantial collection of micro-data on firms, establishments and people. A significant segment of these data are confidential because they contain the identities of the subjects of study (e.g., people, corporations, etc.). However, it this quality that makes these data attractive to young scholars in economics, sociology, demographics, environmental science, health and other fields whose research mandates inherently identifiable data to leverage geospatial relations, genome data, networks of all sorts, linked administrative records, and so on. To carry out this research, these scholars acquire authorized restricted access to the confidential, identifiable data and perform their analyses in secure environments, such as RDCs at the Census Bureau, the Bureau of Labor Statistics, the Internal Revenue Service or the National Center for Health Statistics. A related trend was recently highlighted by research undertaken by Chetty et al. (2012), and is illustrated in Figure 1.

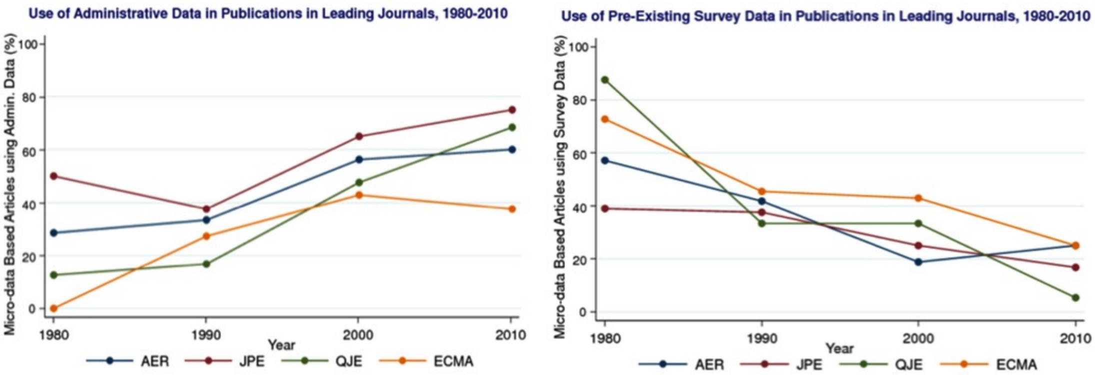


Figure 1. Increasing use of administrative (typically confidential) data in publications (left), contrasted with decreasing use of pre-existing survey data, which is typically public-use (right).

Unfortunately, it is the confidentiality of these data that also has led to what can be called a "curation gap" because the Census Bureau and many other government agencies in the U.S. are prohibited by statute from granting long-term physical custody of their data to archives with well-established data curation practices, such as the Inter-University Consortium for Political and Social Research (ICPSR) or the Integrated Public Use Microdata Series (IPUMS). In other countries, government agencies can sometimes, but not always, leverage similar archives, such as the UK Data Archive and the Australian National Data Service. Confidential data thus differ from public-use data products, which are typically synthesized, aggregated and anonymized derivations of one or more of these confidential data sets, because these repositories can take custody of the public-use data and either ingest, modify or create the metadata that are essential for the curation process.

[1] Research Data Centers (RDCs) are "secure Census Bureau facilities" with 15 locations nationwide that "provide secure access to restricted-use microdata." See: http://www.census.gov/ces/rdcresearch/

The data are not the only problem, because the metadata may also be subject to disclosure limitation. Trivially, when metadata documents statistical features of the underlying data, such as extreme values, they tend to be removed by disclosure limitation protocols. More idiosyncratic are cases where the statutes prohibit publication of critical attributes of the data, such as variable names. IRS rules prohibit this for some variables on certain datasets in use within the U.S. Research Data Centers. Any metadata made available to the public must address these issues, while retaining the broadest information where feasible, i.e., within the secure areas of the statistical agencies, for authorized users.

The divide between private and publicly-accessible data and metadata, and the relationships among them, creates a complex provenance chain, which can make it difficult to understand and trace the origins of particular data. Navigating this provenance chain is made more complicated by inconsistent or non-existent identifier schemes, incomplete metadata, and inconsistent or non-existent mechanisms for expressing relationships among entities.

Poorly documented, unstandardized complexity in the data and metadata, and the underlying curation gap, present a substantial risk of breaching the scientific integrity of the research process itself. The findings that are reported in the peer-reviewed journals are increasingly based on analyses of confidential, restricted-access data, but only public-use data are maintained and curated for open scrutiny. This places often insurmountable barriers to the essential scholarly tasks of testing research results for validity and reproducibility. It also acts as an impediment to the fulfillment of recent directives by U.S. federal funding agencies, such as the National Science Foundation and the National Institutes of Health, that mandate the management, availability and sharing of data produced by funded projects (National Science Foundation, 2011a). To remedy this situation the confidential data themselves must be curated, not just the disclosure-limited, public-use products that this research produces. In addition, the network of linkages among public and private data products and their respective metadata must be visible to both humans and machines. This will give future generations of scientists the ability to scrutinize this work in the same manner that many generations of scientists have been able to with the major public-use data products developed in the last 50 years.

This paper reports on our work to date to address some of these issues within the context of an NSF Census Research Network award titled “Integrated Research Support, Training, and Data Documentation.” (National Science Foundation, 2011b; J. Abowd, Vilhuber & Block, 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 by researchers with and without secure access, and that make proper curation and citation of these data possible. This involves developing a keen understanding of the use-cases that drive our implementation work. The paper therefore begins with a description of two existing provenance and dataset release scenarios. We then describe our design and implementation of an early version of the Comprehensive Extensible Data Documentation and Access Repository (CED^{2}AR), which 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-using and linking to existing archive and provider-generated metadata. While this system does not address our full suite of requirements, it addresses some important issues concerning identifiers, data and metadata confidentiality, and cross data relationships. It does this by leveraging a number of existing technologies and open standards, such as OAI-PMH, DOIs, Dublin Core, and the Data Documentation Initiative (DDI).

Provenance and Confidentiality Scenarios

In this section we describe two example scenarios that illustrate the variety of provenance relationships among existing private and public data products.

LBD Provenance

The Census Bureau's Longitudinal Business Database (LBD)[2] is at the core of many economics papers.[3] It is also at the center of a complex provenance graph that is illustrated in Figure 2. 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)[4], an aggregation of the LBD[5], and the Synthetic LBD, a confidentiality-protected synthetic microdata version of the LBD[6] (Kinney et al., 2011). 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 annual surveys, such as the Annual Survey of Manufactures (ASM)[7]. Aggregations of the ASM and EC are published by the US Census Bureau Center for Economic Research[8] and confidential versions are available within the Census RDCs. Furthermore, the BR serves as direct input to the County Business Patterns (CBP)[9] and related Business Patterns, again, through aggregation and related disclosure protection mechanisms (such as noise infusion (Evans, Zayatz & Slanta, 1998), coarsening, and suppression).

Of the component and derivative datasets shown in Figure 2, none currently have a unique and stable identifier. Aggregations of ASM and EC, as well as CBP data, have identifiers within American FactFinder (AFF)[10], and once published, the data tables are rarely changed. In particular, the data for a particular year are normally not revised once published. On the other hand, derivative products of the LBD are time-series (longitudinal) data products. The fundamental longitudinal link established in the

[2] Longitudinal Business Database: https://www.census.gov/ces/dataproducts/datasets/lbd.html

[3] For a list of examples, see: http://goo.gl/KS6ts

[4] Business Dynamics Statistics: https://www.census.gov/ces/dataproducts/bds/

[5] As found at the Business Dynamics Statistics publications page: https://www.census.gov/ces/dataproducts/bds/publications.html

[6] Synthetic LBD: https://www.census.gov/ces/dataproducts/synlbd/index.html

[7] See: http://www.census.gov/econ/progoverview.html

[8] See: http://www.census.gov/econ/census/

[9] County Business Patterns (CBP): http://www.census.gov/econ/cbp/

[10] American FactFinder: http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml

LBD can be revised every time a new year of data is added, since additional information becomes available. Derivative products thus can also change over the entire time series. None of the releases of the BDS, nor the single release so far of the SynLBD, have unique identifiers. Microdata versions of the EC, ASM, BR, and LBD are available in the RDCs. While, in general, users can assume that the archival version of EC and ASM available in the Census Bureau's RDC was used to create the public-use tabulations, without a definitive identifier to prove that link, the relationship between LBD microdata and its derivative products is more difficult to ascertain. In particular, research versions of the EC, ASM and other economic surveys are regularly updated to provide a link to the LBD via a release-specific identifier.

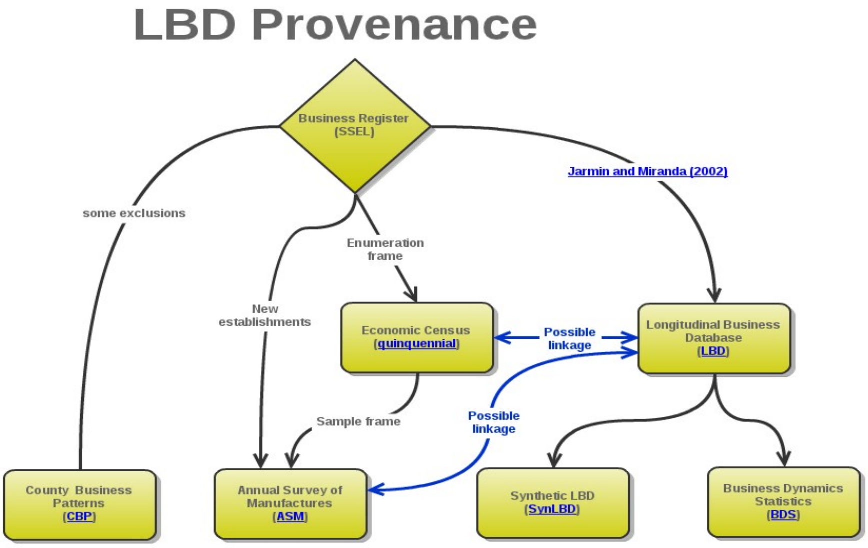

LBD Provenance
Business Register (SSEL)
some exclusions
Jarmin and Miranda (2002)
Enumeration frame
New establishments
Economic Census (quinquennial)
Possible linkage
Longitudinal Business Database (LBD)
Sample frame
Possible linkage
County Business Patterns (CBP)
Annual Survey of Manufactures (ASM)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

Figure 2. LBD provenance.

LEHD-QWI Provenance

A similar complex set of relationships exists for the Longitudinal Employer-Household Dynamics (LEHD) Infrastructure files, illustrated summarily in Figure 3. 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). In theory, the wage numbers in the QCEW can be reproduced by aggregating the UI wage records in the same system. These are administrative, live databases. 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, such as noise infusion and suppression (Abowd et al., 2012). 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. The current release, separated into files of manageable size by tabulation categories, comprises about 3,500 files with an aggregate (compressed) size of about 350GB. Each revision of each file in this system, whether internal or published, has a unique identifier, although users of the published data do not have (easy) access to the data at present. 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, although by the time researchers start using the data, that release has been superseded by subsequent releases multiple times.

To further render the provenance graph complex, the QCEW files (but not the UI wage records) are also provided to the Bureau of Labor Statistics (the original and ongoing sponsor of the QCEW system), and serve as both sample frames and inputs to many tabulations there. Public-use QCEW tabulations exist at the county by industry level, and guide, but do not constrain, the equivalent QWI tabulations. Discrepancies between the two are due to different statistical processing and are complex.

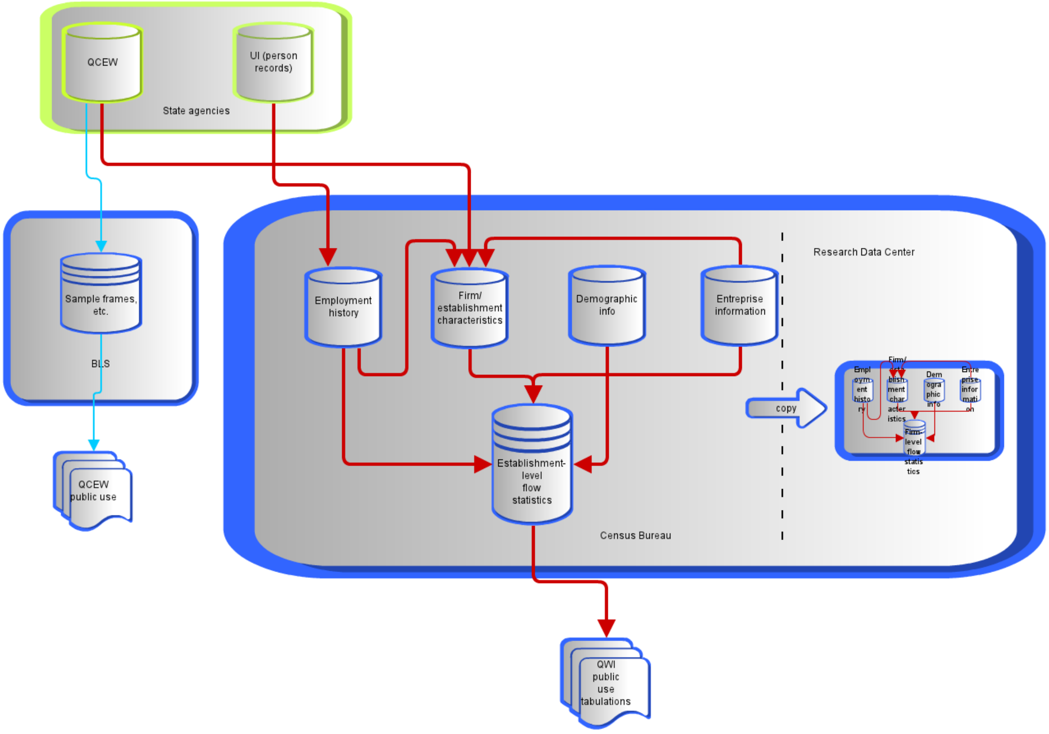


Figure 3. LEHD-QWI provenance.

CED^{2}AR Prototype

The examples above illustrate the complexity of provenance relationships among datasets; a complexity that is further increased by the existence of confidentiality restrictions on parts of the provenance chain. Researchers have a clear need for an easy-to-use tool with which they can navigate this complexity. It should allow them to search and browse metadata federated from several locations spanning the private/public divide. Furthermore, it should adapt to their presence in either restricted or unrestricted locations.

The CED²AR prototype is a first step in the creation of this tool. CED²AR is a metadata discovery tool that ultimately links back to the original data's custodian, whether that custodian has put the data in the public domain (e.g., The American Community Survey, ACS) or has restricted access to the data so it is only accessible within secure areas (e.g., confidential Decennial Census of Population and Housing data). In addition, it merges both private and public use metadata from multiple sources.

The CED²AR system accommodates at least two important scenarios of confidential data and/or metadata, which are illustrated in Figure 4. 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 project, derived from the original the Decennial Census public-use data files that may still be obtained in their unmodified form from the Census Bureau.

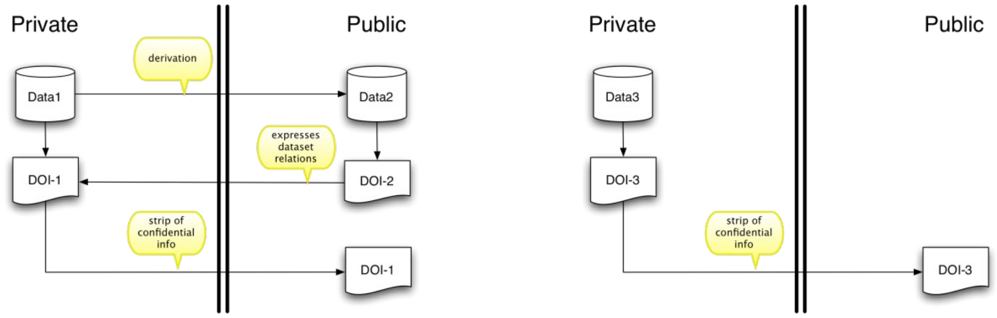

Private
Public
derivation
Data1
Data2
expresses dataset relations
DOI-1
DOI-2
strip of confidential info
DOI-1
Private
Public
Data3
DOI-3
strip of confidential info
DOI-3

Figure 4. 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.

Second, via a well-defined metadata strategy, CED²AR addresses the confidentiality of some components of the metadata. In practice, as applied to the U.S. Census Bureau, a full and complete metadata set will live within the protected and secure area of the statistical agency, and a synchronization protocol will prune the data of its confidential elements, making available a (verifiably) releasable 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 will allow the internal, confidential metadata to also benefit from such enhancements.

CED²AR leverages the DDI metadata standard specified by the International Data Documentation Initiative alliance (DDI)[11]. DDI originated in 1995 and is the most advanced and widely used metadata standard for social science data. It is used by

[11] Data Documentation Initiative: http://www.ddialliance.org/

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), the World Bank, and others.

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 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 datasets over the entire lifecycle from conception through publication and subsequent reuse. The current latest version of this branch is 3.1. 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. We have decided to implement CED^{2}AR using DDI 2.5 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.

The remainder of this section describes the key components of the CED^{2}AR prototype.

Identifiers for Data

Unique, permanent identifiers for data are essential for the purposes of citation, reuse, access and curation. Thus, our first task in the CED^{2}AR design and implementation was the formulation of a consistent unique identifier strategy. This entails a number of conceptual, technical, policy and syntactic decisions that are outlined in the subsections below.

Entity definition

Throughout this paper we have repeatedly used the common term “dataset” informally and without attention to its precise meaning. As Renear, Sacchi and Wickett point out: “the notion of ‘dataset’ found in the literature cannot itself be provided with a precise formal definition” (Renear, Sacchi & Wicket, 2010). Yet, any identifier strategy must include decisions about the nature of the entities to which identifiers are being attributed. The decision is heuristic, user-driven (i.e., what do the users of the system conceptually consider to be a dataset, motivated in part by that which they wish to cite), and application-specific, rather than technical and algorithmic.

In our particular case, because our system is metadata-driven, we are following the rule that an externally/globally-identified dataset 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 with 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.

We anticipate that at some point in the future we may want to implement the notion of an aggregation of a set of related (i.e., by version, or other relationship type) datasets within a single named entity; a container of datasets with the unique global identifier. Later in the project, we may implement this notion using a container technology, such as OAI-ORE (Lagoze et al., 2008). As described later, we do encode relationships among datasets in DDI metadata.

Identifier technology

The Digital Object Identifier (DOI)[12] is a well-established identifier infrastructure, especially in the academic publishing sector. Essentially, DOIs are a managed identifier space built on top of the Handle System (Sun, Lannom & Boesch, 2003), a technology for distributed, persistent and unique naming of digital objects. 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 (Pollard & Wilkinson, 2010) 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, conforming to the entity definition strategy in CED^{2}AR described previously. By leveraging DataCite, we join a growing community of data providers and can interoperate at the identifier level with those other data providers.

Minting new identifiers

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[13] 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.

Identifier syntax

The generic syntax for a DOI is `doi:<nameAuthority>/<itemID>`, where the `<nameAuthority>` is fixed and assigned by the DOI granting authority – in our case EZID. It is then up to us to assign the `itemID` for each of our data sets. We have decided to not imprint any semantics in the `itemID` (i.e., make them opaque). The alternative would have been to make this string a human-readable name that hints at the contents of the data, or even encoding other metadata into the `itemID`, such as date of creation or version number (discussed below). Our perspective is that such an approach may lead to future problems when dataset names change meaning or are used for other purposes, yet are still attached to legacy names or dates. We decided on the simple method of hashing a date timestamp representing the moment of creation of

[12] International DOI Foundation: http://www.doi.org/

[13] EZID: http://n2t.net/ezid/

the DOI, that will result in an opaque string that will lead to a DOI such as: `doi:10.5072/FK2M327JW`.

Identifiers and dataset versions

In our implementation context, and in many others, datasets typically go through a number of versions, either driven by corrections to existing datasets or periodic resampling, as in the case with decennial census data in the U.S. Although we decided, as described above, to make our DOIs opaque, we reflected on the wisdom of imprinting a version number syntactically in the DOI. One common practice is to separate the version number from the main DOI via the familiar `/` character. In this way, we could indicate versions of the same dataset in our DOI in the manner of `doi:10.5072/FK2M327JW/v1`, `doi:10.5072/FK2M327JW/v2`, etc. Note that the exact syntax of the version number encoding practice, or whether to employ one at all, is up to the organization that mints the DOI, since the entire `itemID` has no functional meaning within the DOI/Handle System mechanism.

We decided against this syntactic practice for a number of reasons. First, and corresponding with our reason to mint opaque DOIs in the first place, it is our belief that metadata should not be encoded in the identifier, but should be clearly stated within the actual curated metadata. Thus, our version relationships are encoded in the DDI as described below. Second, there is substantial ambiguity about the distinction between a version of an existing object and an entirely new object (i.e., when do you assign a new DOI or append a version number to an existing DOI?). The library community is familiar with this problem and has developed heuristics usable in the bibliographic context (Carlyle, 2006), which do not transfer well to our domain. Since the assignment of a DOI to an object is essentially a permanent act, we believe that the best approach is a conservative one that avoids imprinting semantics in the syntactic structure, since the semantics may change over time. For example, two entities that have a sequential version relationship may have an intermediary version inserted at a later time. In the end, the immutability of a persistent naming scheme conflicts with the mutability of object relationships.

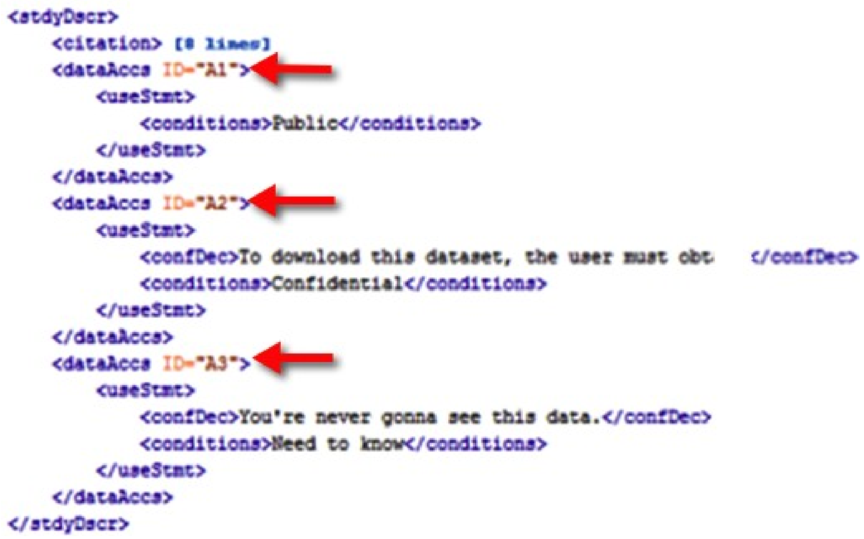
Minting a DOI for other providers' datasets

Most of the data that are included in the CED^{2}AR system come from other providers who have not (yet) assigned DOIs to those data. The process of assigning our own DOI to those data, under our own naming authority, does not preclude that provider from later assigning its own DOI to those data. That is, a single entity may have more than one DOI. The notion of equivalence between two DOIs is out of scope of the DOI system, but could be expressed in the metadata associated with those two DOIs.

Expressing Data-Hiding in DDI

Our initial CED^{2}AR 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 is the hiding of extreme values of variables, which is required by many statistical organizations to protect the anonymity of data. The second is the hiding of the entire existence of variables themselves, a requirement of the IRS mentioned earlier.

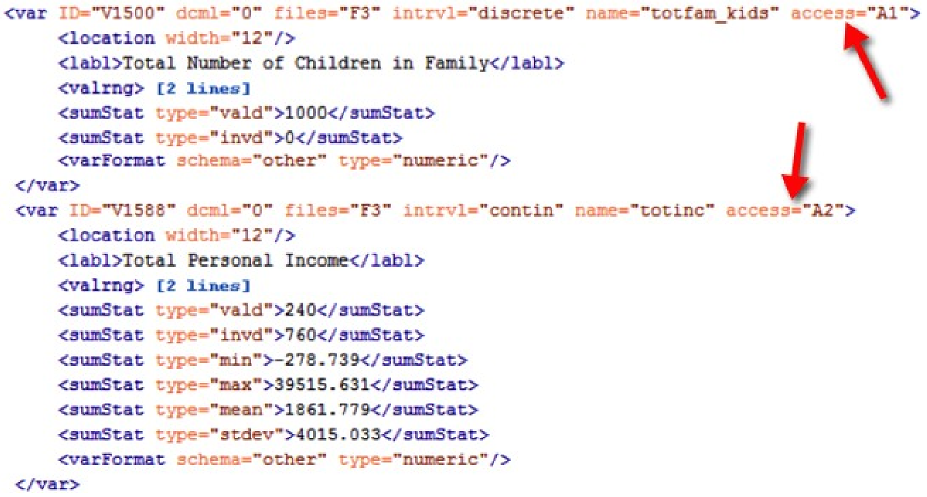
DDI already includes two structural components that accommodate the second form of data hiding. The first is the `<dataAccs>` element, which is nested within the `<stdyDscr>` element – one of the eight main structural branches nested within the root `<codeBook>` element of DDI 2.5. It is possible to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

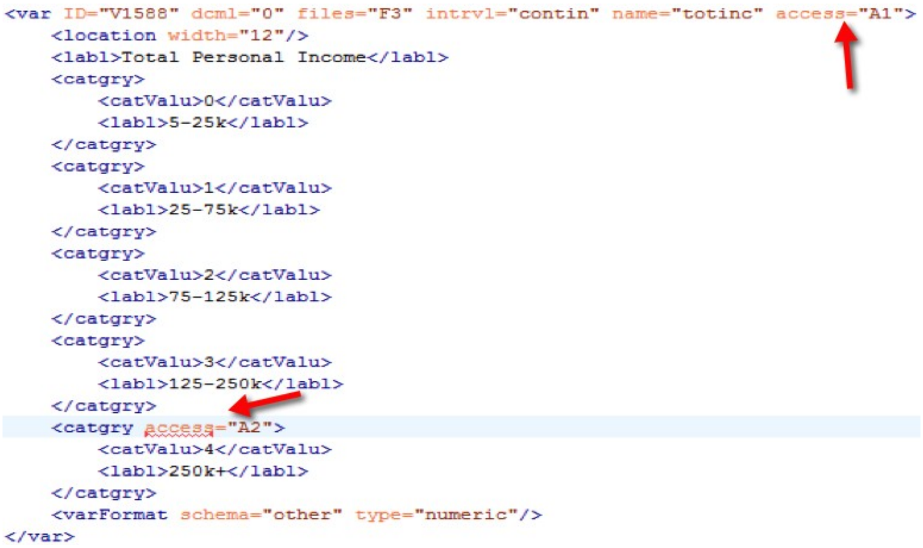
Figure 6 shows the application of the hiding rules defined in Figure 5 to specific variables through the use of the `access` attribute. As shown, the variable `totfam_kids` is public, as defined by rule A1, and the variable `totinc` is private, as defined by rule A2, and therefore should be stripped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

The other form of hiding, namely for extreme values of variables, is not accommodated by the current specification of DDI 2.5. We therefore propose a minor change to the DDI codebook schema that would permit the attachment of the `access` attribute at the `<catgry>` element, in addition to its current allowance in the `<var>` element (as well as at a number of other places outside the scope of this paper). 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`. We are proposing this small tweak to the schema to the International DDI Alliance, since we think that would be of general benefit across the DDI application space.



```
<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.

Expressing Dataset Relations in DDI

As described earlier in this paper, provenance and the relationships between datasets are an important aspect of our problem domain. The examples we described earlier contain numerous relationship types, such as versioning, derivation, part of, etc. Our goal is to encode these relationships in a machine-readable manner in the metadata so that a user searching or browsing data in the user interface can be made aware of the network of relationships that a respective dataset has to other datasets. In the future, we are planning to investigate the notion of an inheritance model, whereby metadata expressed in a "parent" dataset can be passed down to "child" datasets, such as in a version chain where the title of the dataset may not change over several versions.

The DDI 2.5 specification already has a facility for expressing dataset relations through the `<citation>` element that can be included at several locations in the DDI tree, but most usefully for our purposes within the `<relStdy>` element, which describes the relationship between a respective dataset and one or more other datasets. As noted in the DDI 2.5 specification, the value of the `<citation>` element can be any term from the Dublin Core vocabulary,[14] which includes a number of relationship types such as part/whole, version, provenance, etc. While these are not sufficiently

[14] Dublin Core Vocabulary: http://dublincore.org/documents/dcmi-terms/

expressive over the long run, they will be useful for our initial implementation work while we were refer to Dublin Core community and DDI community for a more expressive way of expressing the network of relationships that datasets can have with others.

Conclusions and Future Work

The prototype described in this paper is just a first step towards addressing the complex scenarios outlined at the beginning of the paper. However, it provides a solid foundation for our future work. Issues to be addressed in this future work include a refinement of the granularity of our metadata hiding techniques, implementation and fine-tuning of dataset relations and provenance expression, and the inclusion of more data and corresponding metadata into the system. In this way, we hope to address a long-standing need of social science researchers for a generally available tool for searching, accessing, traversing and citing confidential data.

References

Abowd, J.M., Gittings, K., McKinney, K.L., Stephens, B.E., Vilhuber, L. & Woodcock, S. (2012). *Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time series.* Retrieved from http://ideas.repec.org/p/cen/wpaper/12-13.html

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

Carlyle, A. (2006). Understanding FRBR as a conceptual model: FRBR and the bibliographic universe. *Library Resources Technical Services*, *50*(4), 264–273. doi:10.5860/lrts.50n4.264

Chetty, R. (2012). *The transformative potential of administrative data for microeconometric research*. Retrieved from http://conference.nber.org/confer/2012/SI2012/LS/ChettySlides.pdf

Evans, T., Zayatz, L. & Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics*, *14*(4), 537–551.

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

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. doi:10.1111/j.1751-5823.2011.00153.x

Lagoze, C., Van De Sompel, H., Nelson, M.L., Warner, S., Sanderson, R., Johnston, P. et al. (2008). *Object re-use and exchange: A resource-centric approach.* Indianapolis. Retrieved from http://arxiv.org/abs/0804.2273

National Science Foundation. (2011a). GPG chapter II. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

National Science Foundation. (2011b). NSF award search: Award#1131848 - NCRN-MN: Cornell Census-NSF Research Node: Integrated research support, training and data documentation. Retrieved from http://www.nsf.gov/awardsearch/showAward?AWD_ID=1131848

Pollard, T.J. & Wilkinson, J.M. (2010). Making datasets visible and accessible: DataCite's first summer meeting. *Ariadne*, 64. Retrieved from http://www.ariadne.ac.uk/issue64/datacite-2010-rpt

Renear, A.H., Sacchi, S. & Wicket, K.M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the 73rd ASIS&T Annual Meeting*. Pittsburgh, PA. Retrieved from http://www.asis.org/asist2010/proceedings/proceedings/ASIST_AM10/openpage.html

Sun, S.X., Lannom, L. & Boesch, B. (2003). Handle System overview. IETF RFC 3650 Retrieved from http://www.ietf.org/rfc/rfc3650.txt