CISER

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

The Complicated Provenance of American Community Survey Data: How Far will PROV and DDI Take Us?

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

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

Presentation at the 2nd Annual North American DDI User Conference (NADDI14)
Vancouver, British Columbia, Canada
2 April 2014

Cornell University

CISER

Outline

Answering the Q: How far will PROV and DDI take us? We don't know; complicated story!

- Background/Previous Work
- Use Case(s) involving ANCESTRY Variable in ACS
- Technical solutions at File (Dataset) and Variable Level
- Future Work

Questions and Discussion

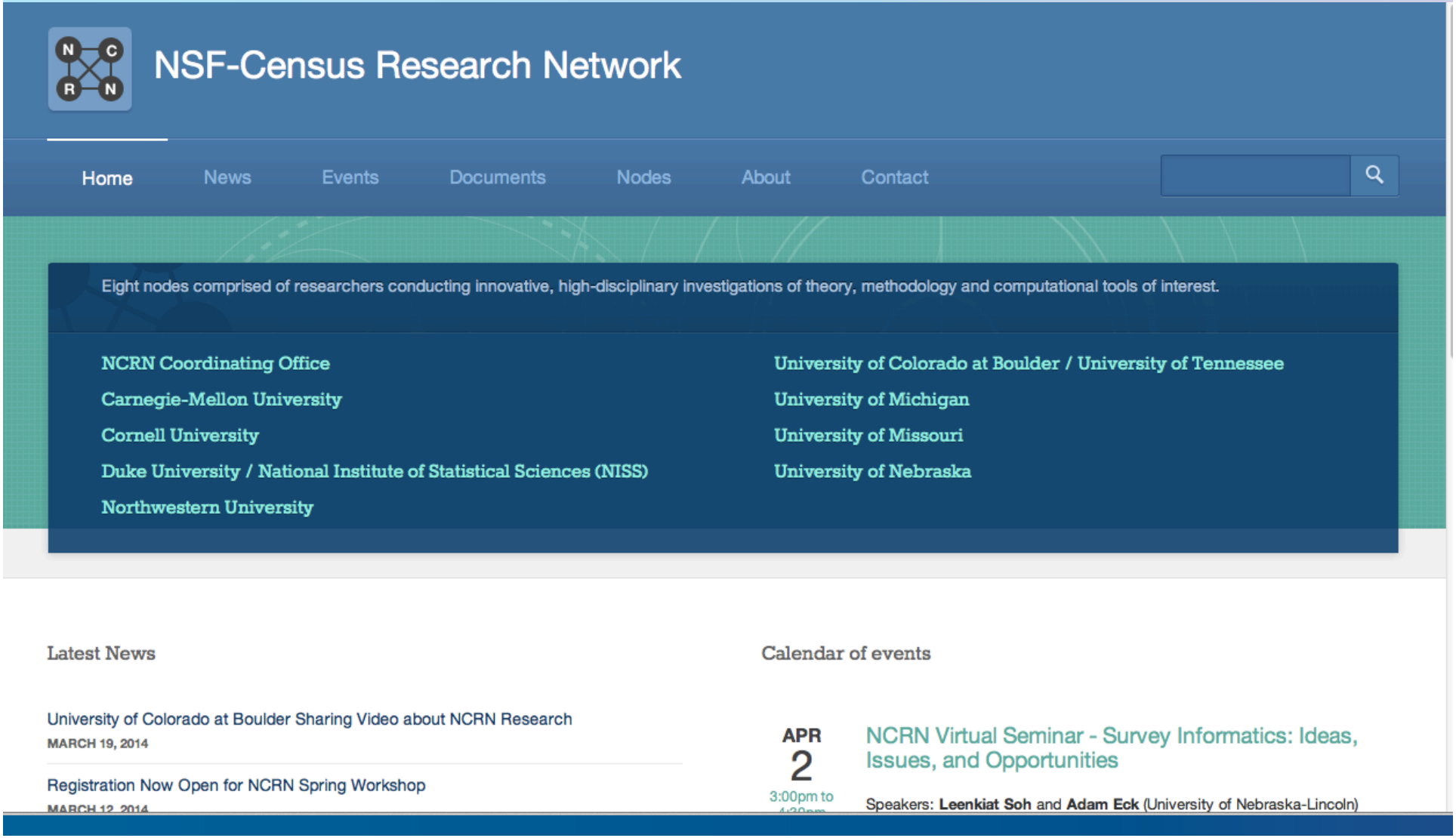
CISER

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

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

CISER

Part of NCRN Research Network



CISER

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

- Method for solving the data curation problem that confronts the custodians of restricted-access research data and the scientific users of such data
- Accommodates physical security and access limitation protocols, and allows for much improved provenance tracking
- Metadata repository system that allows researchers to search, browse, access, and cite confidential data and metadata (via a web-based user interface or programmatically through a search API)

CISER

Proposed a <dataAccs> Solution at EDDI12 in Bergen

NCRN DDI Solution at the Variable Level: <dataAccs>

```
<stdyDscr>
    <citation> [8 lines]
    <dataAccs ID="A1">
        <useStmt>
            <conditions>Public</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A2">
        <useStmt>
            <confDec>To download this dataset, the user must obtain Special Sworn Status from the United States Census Bureau.</confDec>
            <conditions>Confidential</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A3">
        <useStmt>
            <confDec>You're never gonna see this data.</confDec>
            <conditions>Need to know</conditions>
        </useStmt>
    </dataAccs>
</stdyDscr>
```

CISER

Variable Level Solution (continued)

```
<var ID="V1500" dcml="0" files="F3" intrvl="discrete" name="totfam_kids" access="A1">
     <location width="12"/>
     <labl>Total Number of Children in Family</labl>
     <valrng> [2 lines]
     <sumStat type="vald">1000</sumStat>
     <sumStat type="invd">0</sumStat>
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <catgry> [3 lines]
     <varFormat schema="other" type="numeric"/>
 </var>
 <var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A2">
     <location width="12"/>
     <labl>Total Personal Income</labl>
     <valrng> [2 lines]
     <sumStat type="vald">240</sumStat>
     <sumStat type="invd">760</sumStat>
     <sumStat type="min">-278.739</sumStat>
     <sumStat type="max">39515.631</sumStat>
     <sumStat type="mean">1861.779</sumStat>
     <sumStat type="stdev">4015.033</sumStat>
     <varFormat schema="other" type="numeric"/>
 </var>
```

CISER

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

```
<var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A1">
    <location width="12"/>
    <labl>Total Personal Income</labl>
    <catgry>
        <catValu>0</catValu>
        <labl>5-25k</labl>
    </catgry>
    <catgry>
        <catValu>1</catValu>
        <labl>25-75k</labl>
    </catgry>
    <catgry>
        <catValu>2</catValu>
        <labl>75-125k</labl>
    </catgry>
    <catgry>
        <catValu>3</catValu>
        <labl>125-250k</labl>
    </catgry>
    <catgry access="A2">
        <catValu>4</catValu>
        <labl>250k+</labl>
    </catgry>
    <varFormat schema="other" type="numeric"/>
</var>
```

Small tweak to the DDI Codebook Schema would fix this.

CISER

<dataAccs> developments since EDDI12

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

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

CISER

Select Cornell NCRN Publications

Forthcoming. “Lagoze, Carl, Lars Vihuber, Jeremy Williams, Benjamin Perry, and William C. Block, “CED^{2}AR: The Comprehensive Extensible Data Documentation and Access Repository.” In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), London UK, September 2014.

2013 Lagoze, Carl, with William C. Block, Jeremy Williams, John M. Abowd, and Lars Vilhuber. “Data Management of Confidential Data”. In: International Journal of Digital Curation 8.1, pp.265-278. DOI: 10.2218/ijdc.v8il.259

2012 Abowd, John M., Lars Vilhuber, and William C. Block. “A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs.” In: Privacy in Statistical Databases. Ed. By Josep Domingo-Ferrer and Ilenia Tinnirello. Vol. 7556. Lecture Notes in Computer Science. Springer, pp.216-225. DOI: 10.1007/978-3-642-33627-0_17

CISER

Provenance

"data provenance, one kind of metadata, pertains to the derivation history of a data product starting from its original sources" [...] "from it, one can ascertain the quality of the data base and its ancestral data and derivations, track back sources of errors, allow automated reenactment of derivations to update the data, and provide attribution of data sources"*

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

CISER

Provenance and Metadata

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

```
<xs:complexType name="othrStdyMatType">
    <xs:complexContent>
       <xs:extension base="baseElementType">
          <xs:sequence>
             <xs:element ref="relMat" minOccurs="0" maxOccurs="unbounded"/>
             <xs:element ref="relStdy" minOccurs="0" maxOccurs="unbounded"/>
             <xs:element ref="relPubl" minOccurs="0" maxOccurs="unbounded"/>
             <xs:element ref="othRefs" minOccurs="0" maxOccurs="unbounded"/>
          </xs:sequence>
       </xs:extension>
    </xs:complexContent>
</xs:complexType>
```

Downside: No structure. Mostly verbose entries.

CISER

2013 work with PROV

- Explored encoding PROV in RDF/XML* (Required use of CDATA tag to avoid interfering with schema compliance; deemed less promising)
- More recently: exploring W3C PROV Model as basis for encoding provenance metadata in DDI

W3C PROV Model is based upon:

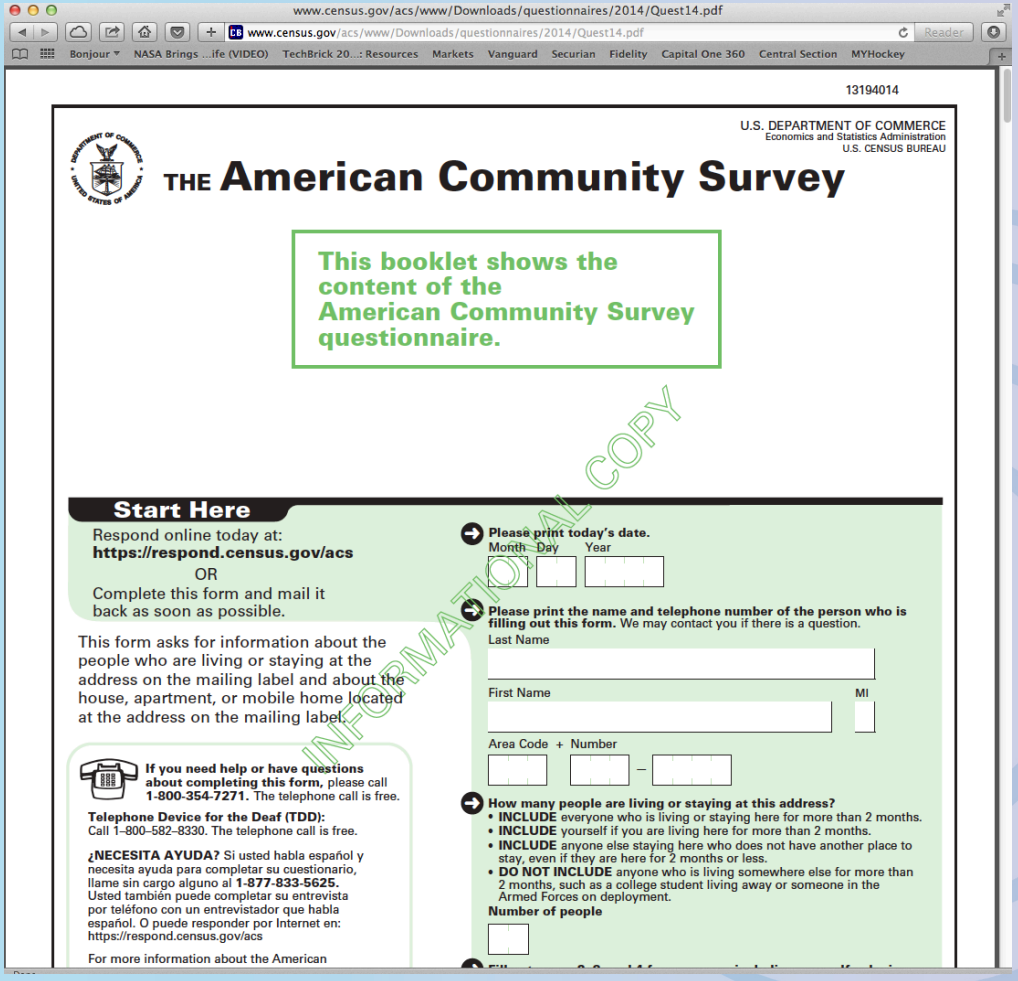
- **entities** that are physical, digital, and conceptual things in the world;
- **activities** that are dynamic aspects of the world that change and create entities; and
- **agents** that are responsible for activities.
- A set of **relationships** that can exist between them that express attribution,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

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

CISER

The American Community Survey (ACS)

- Ongoing statistical survey conducted by the U.S. Census Bureau
- Approximately 250,000 surveys/month (3 million per year)
- Replacement for detailed long-form decennial census


www.census.gov/acs/www/Downloads/questionnaires/2014/Quest14.pdf

13194014

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

THE American Community Survey

This booklet shows the content of the American Community Survey questionnaire.

INFORMATIONAL COPY

Start Here

Respond online today at:
https://respond.census.gov/acs
OR
Complete this form and mail it back as soon as possible.

This form asks for information about the people who are living or staying at the address on the mailing label and about the house, apartment, or mobile home located at the address on the mailing label.

If you need help or have questions about completing this form, please call 1-800-354-7271. The telephone call is free.

Telephone Device for the Deaf (TDD):
Call 1-800-582-8330. The telephone call is free.

¿NECESITA AYUDA? Si usted habla español y necesita ayuda para completar su cuestionario, llame sin cargo alguno al 1-877-833-5625. Usted también puede completar su entrevista por teléfono con un entrevistador que habla español. O puede responder por Internet en: https://respond.census.gov/acs

For more information about the American

Please print today's date.
Month Day Year

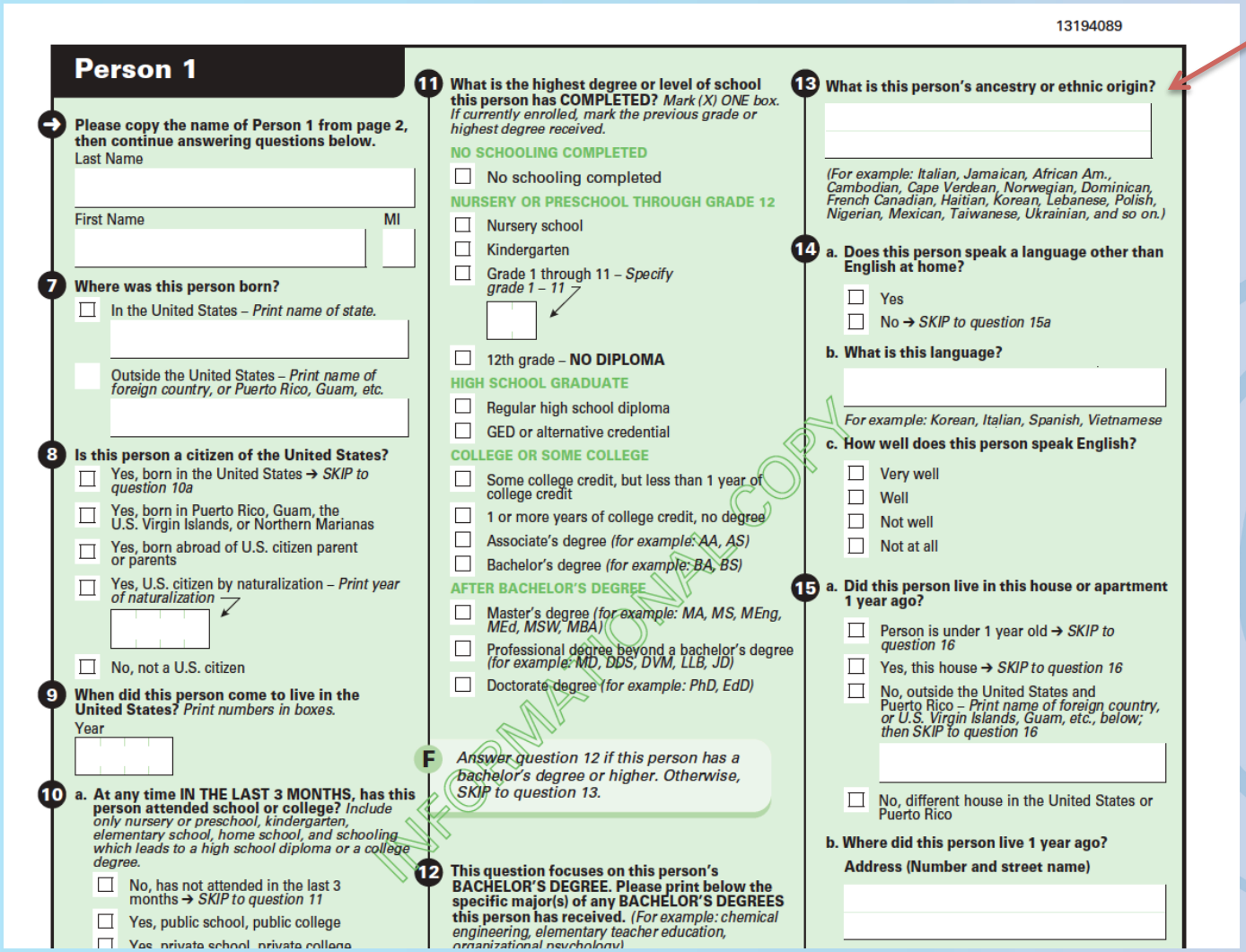
Please print the name and telephone number of the person who is filling out this form. We may contact you if there is a question.
Last Name
First Name MI
Area Code + Number

How many people are living or staying at this address?
- INCLUDE everyone who is living or staying here for more than 2 months.
- INCLUDE yourself if you are living here for more than 2 months.
- INCLUDE anyone else staying here who does not have another place to stay, even if they are here for 2 months or less.
- DO NOT INCLUDE anyone who is living somewhere else for more than 2 months, such as a college student living away or someone in the Armed Forces on deployment.

Number of people

CISER

ACS Question on Ancestry or Ethnic Origin



13194089

Person 1

➔ **Please copy the name of Person 1 from page 2, then continue answering questions below.**

Last Name

First Name MI

7 Where was this person born?

- ☐ In the United States – *Print name of state.*
- ☐ Outside the United States – *Print name of foreign country, or Puerto Rico, Guam, etc.*

8 Is this person a citizen of the United States?

- ☐ Yes, born in the United States ➔ *SKIP to question 10a*
- ☐ Yes, born in Puerto Rico, Guam, the U.S. Virgin Islands, or Northern Marianas
- ☐ Yes, born abroad of U.S. citizen parent or parents
- ☐ Yes, U.S. citizen by naturalization – *Print year of naturalization*
- ☐ No, not a U.S. citizen

9 When did this person come to live in the United States? *Print numbers in boxes.*

Year

10 a. At any time IN THE LAST 3 MONTHS, has this person attended school or college? *Include only nursery or preschool, kindergarten, elementary school, home school, and schooling which leads to a high school diploma or a college degree.*

- ☐ No, has not attended in the last 3 months ➔ *SKIP to question 11*
- ☐ Yes, public school, public college
- ☐ Yes, private school, private college

11 What is the highest degree or level of school this person has COMPLETED? *Mark (X) ONE box. If currently enrolled, mark the previous grade or highest degree received.*

NO SCHOOLING COMPLETED

- ☐ No schooling completed

NURSERY OR PRESCHOOL THROUGH GRADE 12

- ☐ Nursery school
- ☐ Kindergarten
- ☐ Grade 1 through 11 – *Specify grade 1 – 11*
- ☐ 12th grade – **NO DIPLOMA**

HIGH SCHOOL GRADUATE

- ☐ Regular high school diploma
- ☐ GED or alternative credential

COLLEGE OR SOME COLLEGE

- ☐ Some college credit, but less than 1 year of college credit
- ☐ 1 or more years of college credit, no degree
- ☐ Associate's degree *(for example: AA, AS)*
- ☐ Bachelor's degree *(for example: BA, BS)*

AFTER BACHELOR'S DEGREE

- ☐ Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*
- ☐ Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*
- ☐ Doctorate degree *(for example: PhD, EdD)*

F *Answer question 12 if this person has a bachelor's degree or higher. Otherwise, SKIP to question 13.*

12 This question focuses on this person's BACHELOR'S DEGREE. Please print below the specific major(s) of any BACHELOR'S DEGREES this person has received. *(For example: chemical engineering, elementary teacher education, organizational psychology)*

13 What is this person's ancestry or ethnic origin?

(For example: Italian, Jamaican, African Am., Cambodian, Cape Verdean, Norwegian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on.)

14 a. Does this person speak a language other than English at home?

- ☐ Yes
- ☐ No ➔ *SKIP to question 15a*

b. What is this language?

For example: Korean, Italian, Spanish, Vietnamese

c. How well does this person speak English?

- ☐ Very well
- ☐ Well
- ☐ Not well
- ☐ Not at all

15 a. Did this person live in this house or apartment 1 year ago?

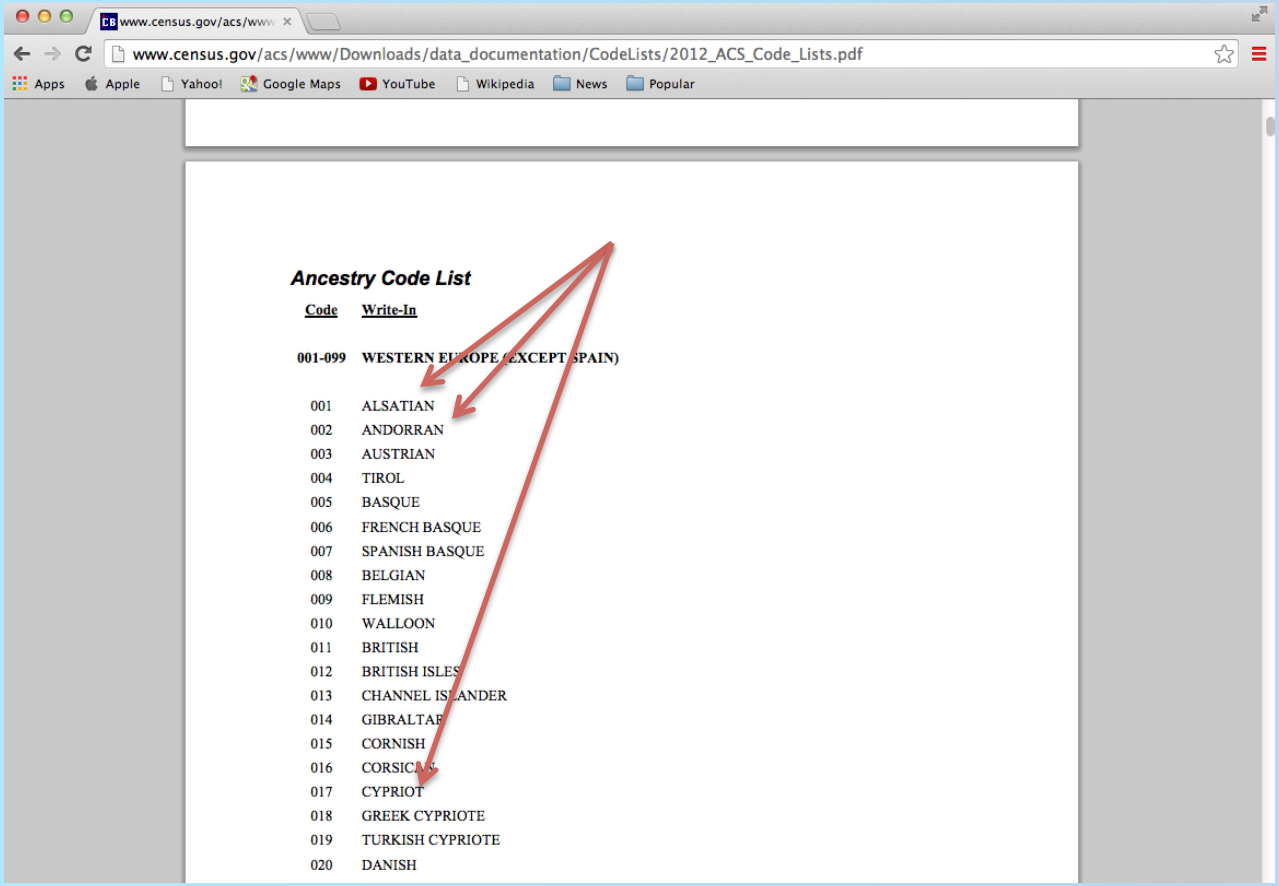
- ☐ Person is under 1 year old ➔ *SKIP to question 16*
- ☐ Yes, this house ➔ *SKIP to question 16*
- ☐ No, outside the United States and Puerto Rico – *Print name of foreign country, or U.S. Virgin Islands, Guam, etc., below; then SKIP to question 16*
- ☐ No, different house in the United States or Puerto Rico

b. Where did this person live 1 year ago?

Address (Number and street name)

CISER

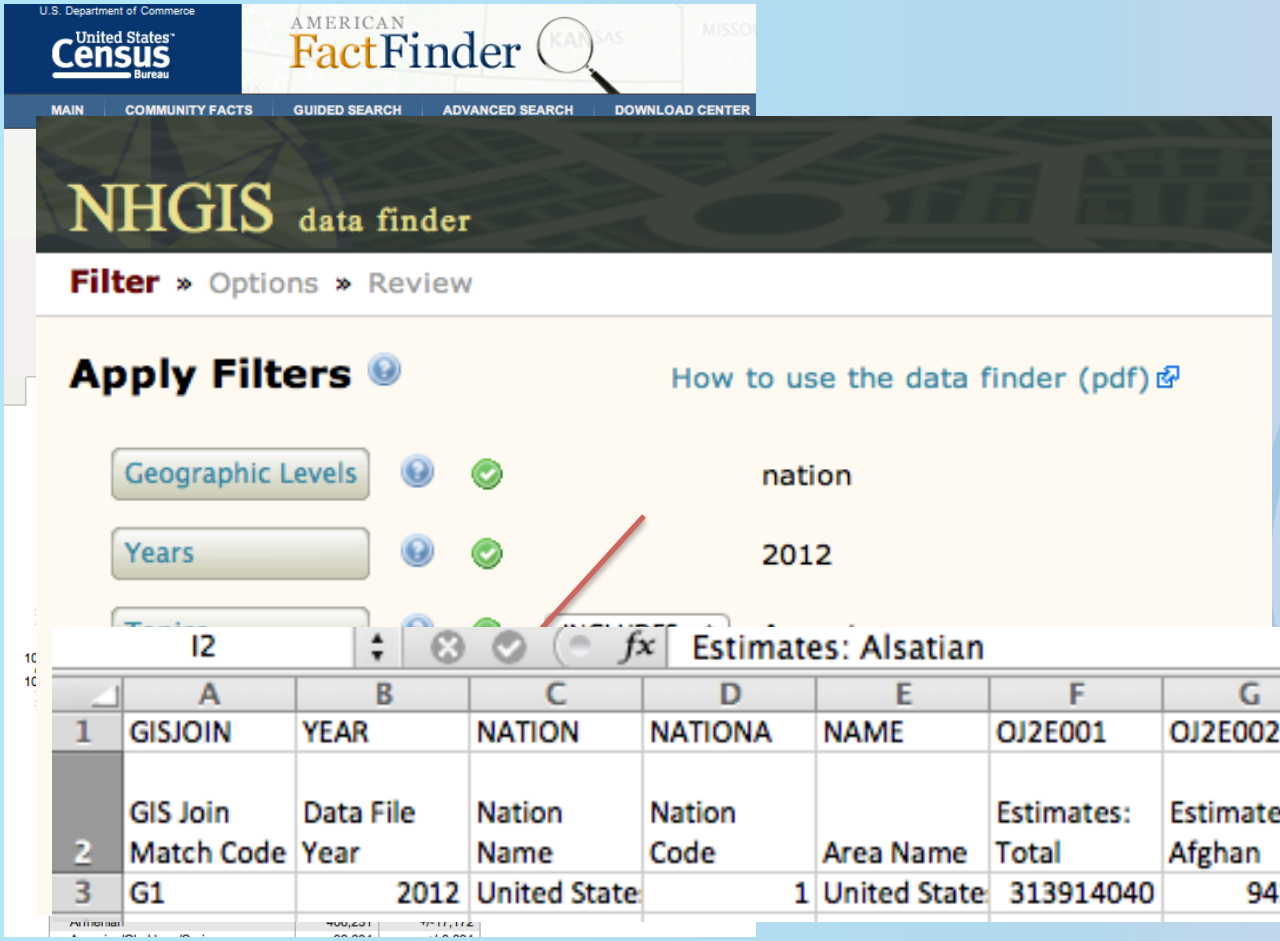
Three Use Cases: Researchers interested in people of Alsatian, Andorran, and Cypriot Ancestry



- U.S. Census Bureau Documentation
- Ancestry Code List
- 2012 ACS

CISER

Multiple Sources of Data originating from the ACS: Examples of Aggregate Data

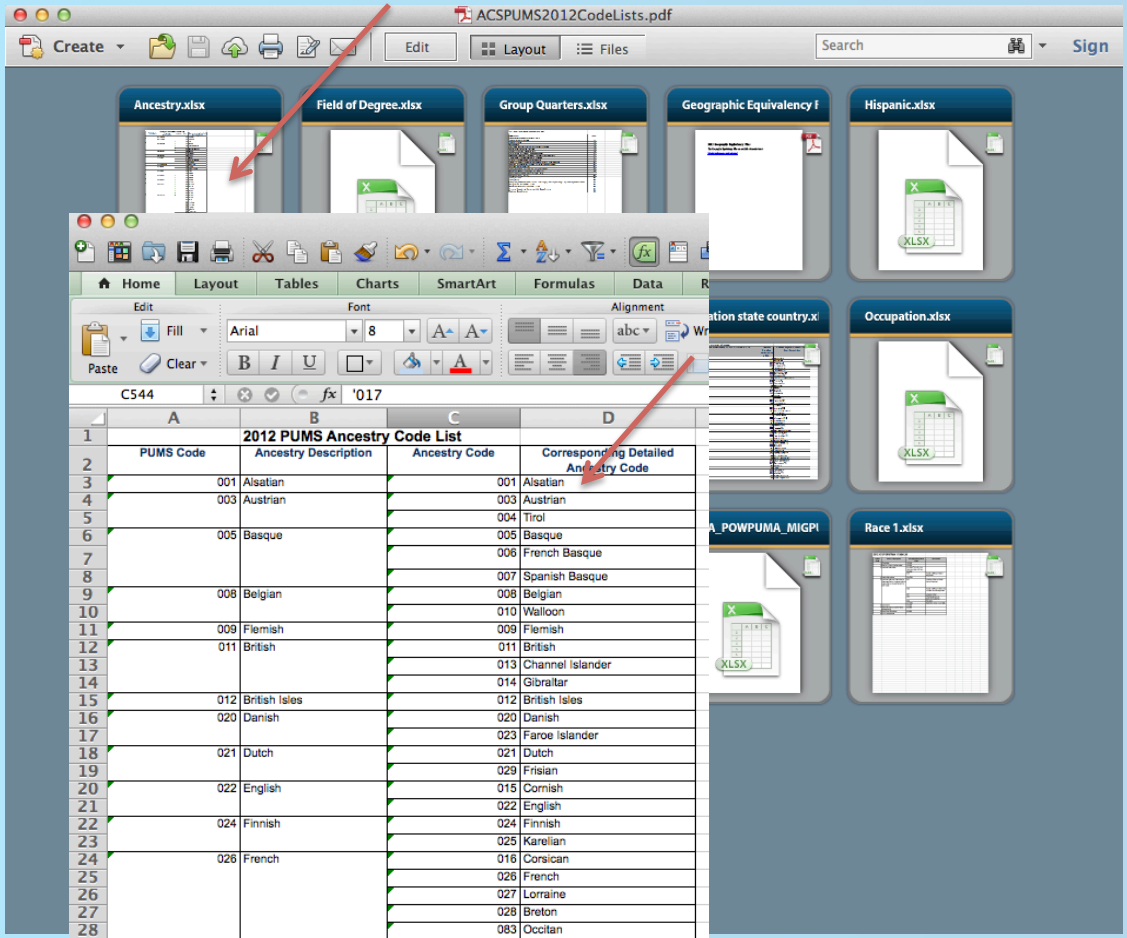


	A	B	C	D	E	F	G	H	I
1	GISJOIN	YEAR	NATION	NATIONA	NAME	OJ2E001	OJ2E002	OJ2E003	OJ2E004
2	GIS Join Match Code	Data File Year	Nation Name	Nation Code	Area Name	Estimates: Total	Estimates: Afghan	Estimates: Albanian	Estimates: Alsatian
3	G1	2012	United State	1	United State	313914040	94244	185696	6626

2012 ACS 1-year Estimate: 6,626 individuals of Alsatian Ancestry living in the United States

CISER

Multiple Sources of Data originating from the ACS: Example of PUMS Microdata



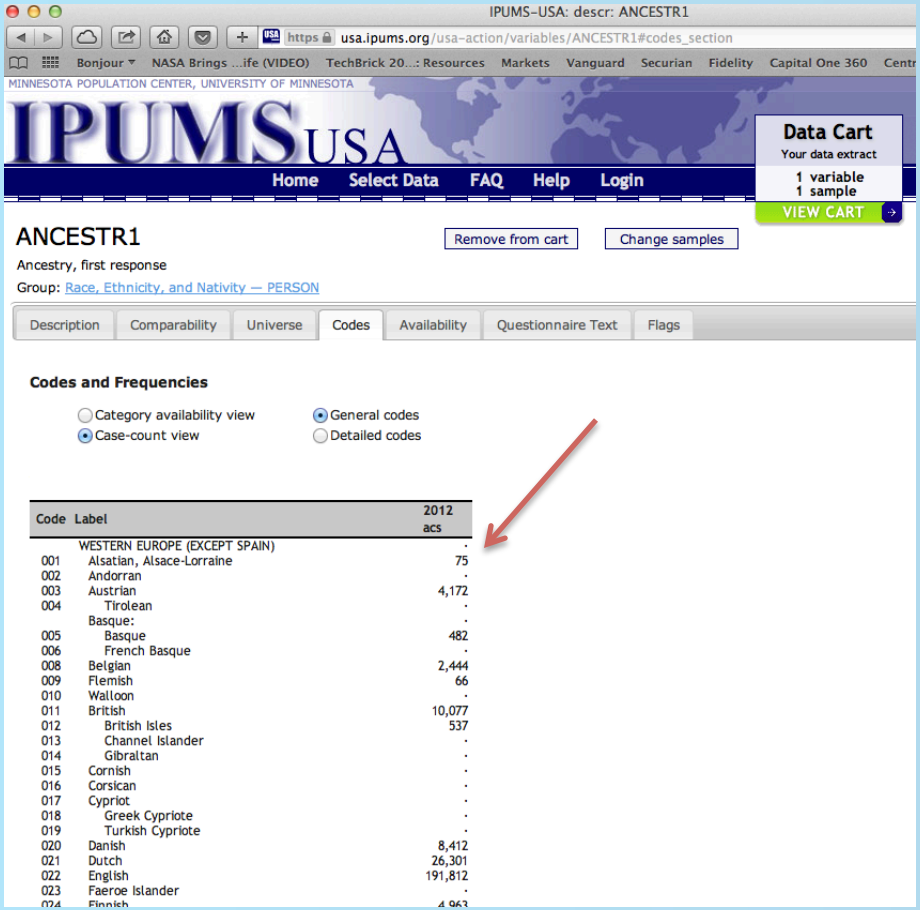
2012 PUMS Ancestry Code List

PUMS Code	Ancestry Description	Ancestry Code	Corresponding Detailed Ancestry Code
001	Alsatian	001	Alsatian
003	Austrian	003	Austrian
		004	Tirol
005	Basque	005	Basque
		006	French Basque
		007	Spanish Basque
008	Belgian	008	Belgian
		010	Walloon
009	Flemish	009	Flemish
011	British	011	British
		013	Channel Islander
		014	Gibraltar
012	British Isles	012	British Isles
020	Danish	020	Danish
		023	Faroe Islander
021	Dutch	021	Dutch
		029	Frisian
022	English	015	Cornish
		022	English
024	Finnish	024	Finnish
		025	Karelian
026	French	016	Corsican
		026	French
		027	Lorraine
		028	Breton
		083	Occitan
032	German	032	German
		033	Bavaria

ACS 2012 PUMS: ANCESTRY Code is 001 for Alsatian

CISER

Multiple Sources of Data originating from the ACS: Example of IPUMS-USA



IPUMS-USA for ACS 2012:

- 001 Alsatian ANCESTRY Code
- 75 cases in the sample

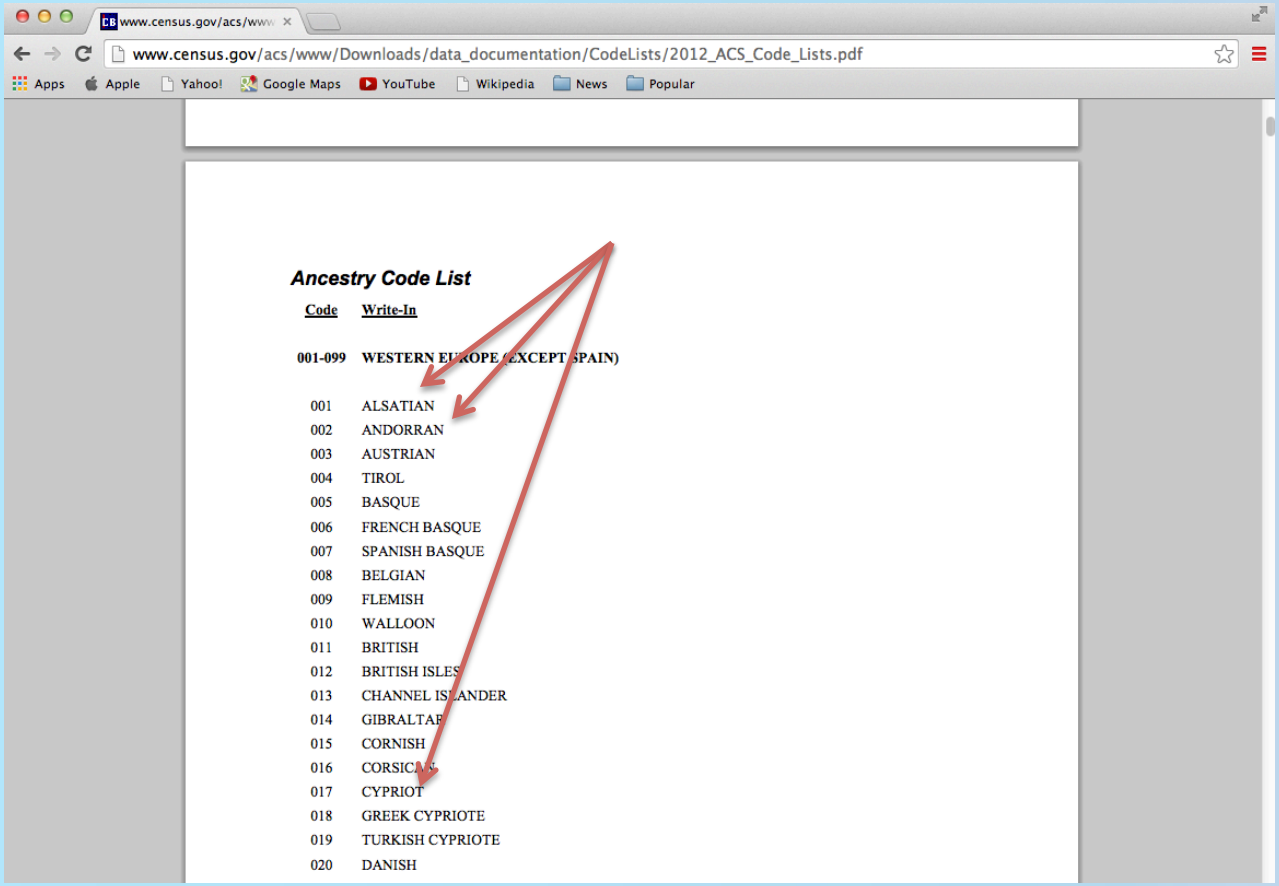
CISER

- Let's review…

	2012 ACS Code List	ACS 2012 PUMS	IPUMS-USA	AFF	NHGIS	
Alsatian	YES (001)	YES (001)	YES (75 cases)	6,626 (est.)	6,626 (est.)	
Andorran						
Cypriots						

CISER

Three Use Cases: Researchers interested in people of Alsatian, Andorran, and Cypriot Ancestry



- U.S. Census Bureau Documentation
- Ancestry Code List
- 2012 ACS

CISER

	2012 ACS Code List	ACS 2012 PUMS	IPUMS-USA	AFF	NHGIS	
Alsatian	YES (001)	YES (001)	YES (75 cases)	6,626 (est.)	6,626 (est.)	
Andorran	YES (002)					
Cypriots	YES (017)					

CISER

U.S. Department of Commerce
United States Census Bureau

AMERICAN FactFinder

Basque	46,874	+/-4,464
Belgian	230,104	+/-8,312
Brazilian	321,544	+/-14,571
British	931,514	+/-17,814

NHGIS data finder

AD	AE	AF	AG	AH	AI
OJ2E025	OJ2E026	OJ2E027	OJ2E028	OJ2E029	OJ2E030
Estimates: Cajun	Estimates: Canadian	Estimates: Carpatho Rusyn	Estimates: Celtic	Estimates: Croatian	Estimates: Cypriot
96349	509952	4944	34844	276808	**6486**

CISER

	2012 ACS Code List	ACS 2012 PUMS	IPUMS-USA	AFF	NHGIS	
Alsatian	YES (001)	YES (001)	YES (75 cases)	6,626 (est.)	6,626 (est.)	
Andorran	YES (002)					
Cypriots	YES (017)			6,486 (est.)	6,486 (est.)	

CISER

MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA

IPUMS USA

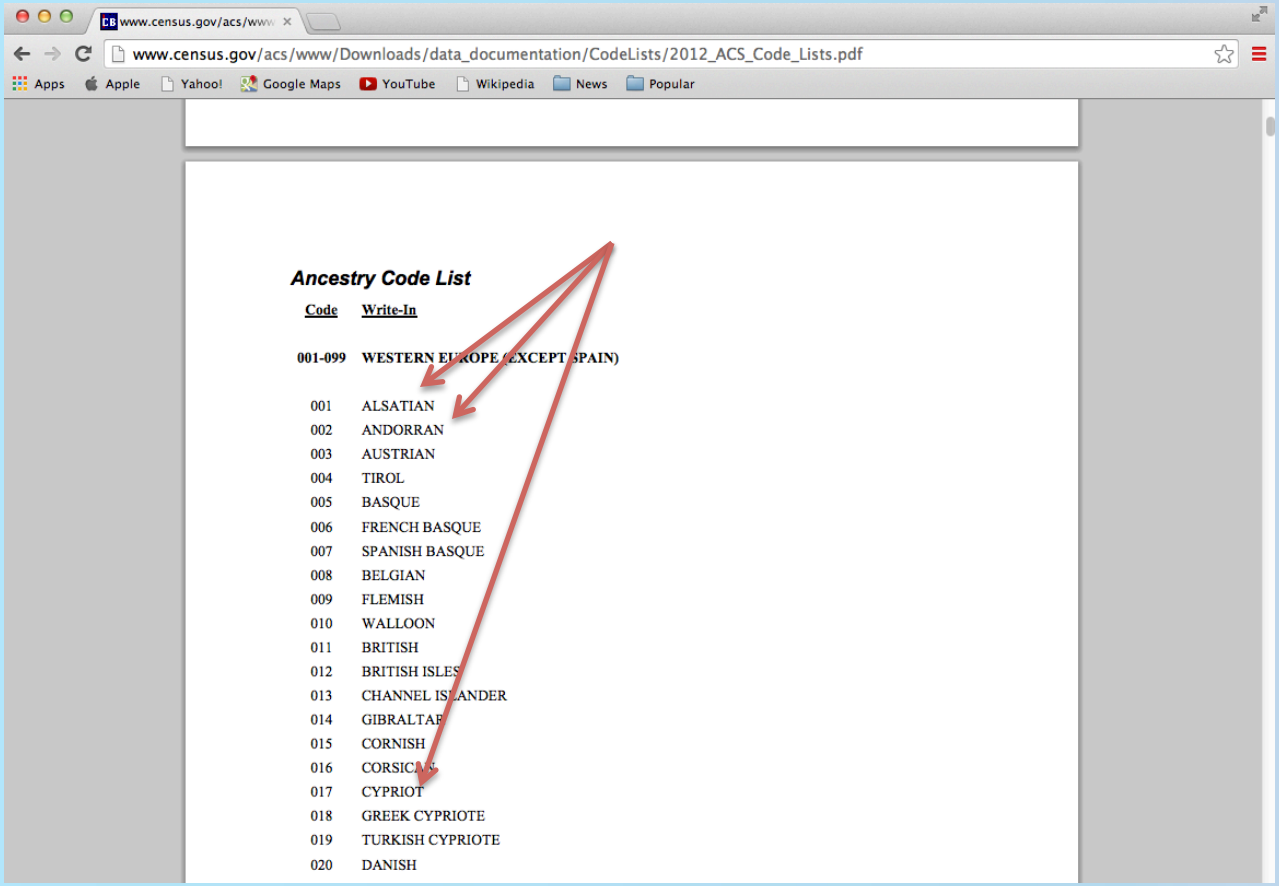
Code	Label	2012 acs
	WESTERN EUROPE (EXCEPT SPAIN)	.
001	Alsatian, Alsace-Lorraine	X
002	Andorran	.
003	Austrian	X
004	Tirolean	.
	Basque:	.
005	Basque	X
006	French Basque	.
008	Belgian	X
009	Flemish	X
010	Walloon	.
011	British	X
012	British Isles	X
013	Channel Islander	.
014	Gibraltan	.
015	Cornish	.
016	Corsican	.
017	Cypriot	.
018	Greek Cypriote	.
019	Turkish Cypriote	.

CISER

	2012 ACS Code List	ACS 2012 PUMS	IPUMS-USA	AFF	NHGIS	
Alsatian	YES (001)	YES (001)	YES (75 cases)	6,626 (est.)	6,626 (est.)	
Andorran	YES (002)					
Cypriots	YES (017)	**NO**	**NO**	6,486 (est.)	6,486 (est.)	

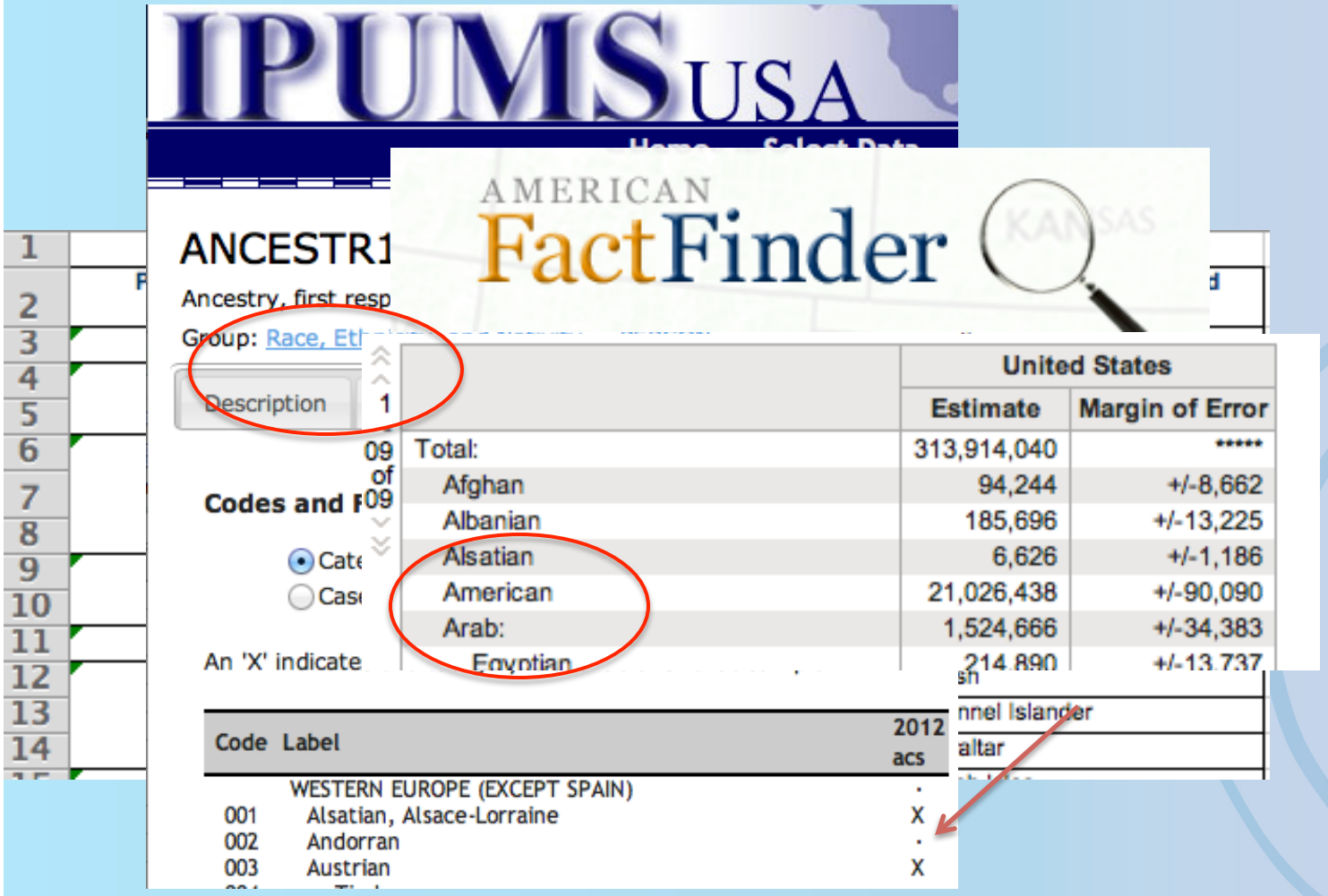
CISER

Three Use Cases: Researchers interested in people of Alsatian, Andorran, and Cypriot Ancestry



- U.S. Census Bureau Documentation
- Ancestry Code List
- 2012 ACS

CISER


IPUMS USA
AMERICAN FactFinder
ANCESTR1
Ancestry, first resp
Group: Race, Eth
Description
Codes and F
Cate
Case
An 'X' indicate
United States
Estimate
Margin of Error
Total: 313,914,040 *****
Afghan 94,244 +/-8,662
Albanian 185,696 +/-13,225
Alsatian 6,626 +/-1,186
American 21,026,438 +/-90,090
Arab: 1,524,666 +/-34,383
Egyptian
Code Label 2012 acs
WESTERN EUROPE (EXCEPT SPAIN) .
001 Alsatian, Alsace-Lorraine X
002 Andorran .
003 Austrian X

CISER

	2012 ACS Code List	ACS 2012 PUMS	IPUMS-USA	AFF	NHGIS	
Alsatian	YES (001)	YES (001)	YES (75 cases)	6,626 (est.)	6,626 (est.)	
Andorran	YES (002)	**NO**	**NO**	**NO**	**NO**	
Cypriots	YES (017)	NO	NO	6,486 (est.)	6,486 (est.)	

CISER

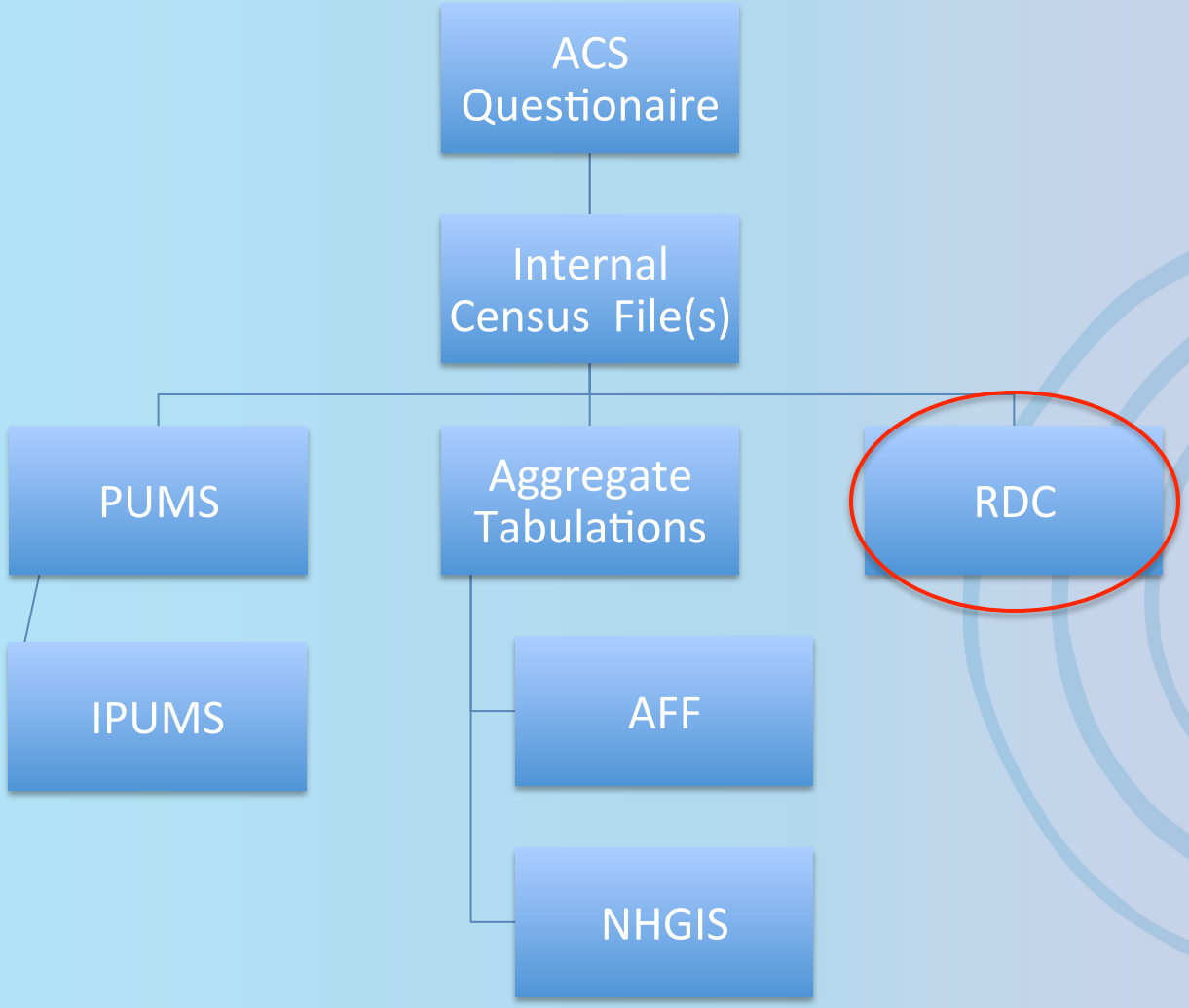
Three Use Cases: Researchers interested in people of Alsatian, Andorran, and Cypriot Ancestry

2012 PUMS Ancestry Code List

...	...	...	...
995	Mixture	995	Mixture
996	Uncodable Entries	996	Uncodable Entries
997	Other Groups	002	Andorran
		017	Cypriot
		018	Greek Cypriote
		019	Turkish Cypriote
		020	Lapp
		021	Liechtensteiner
		022	Manx
		023	Monegasque
		101	Azerbaijani
		107	Ruthenian
		108	Cossack
		117	Finno Ugrian
		118	Mordovian
		119	Voytak
		120	Gruziia
		127	Kalmyk

- U.S. Census Bureau Documentation
- Ancestry Code List
- 2012 ACS

- Simple Provenance of ACS Data Files

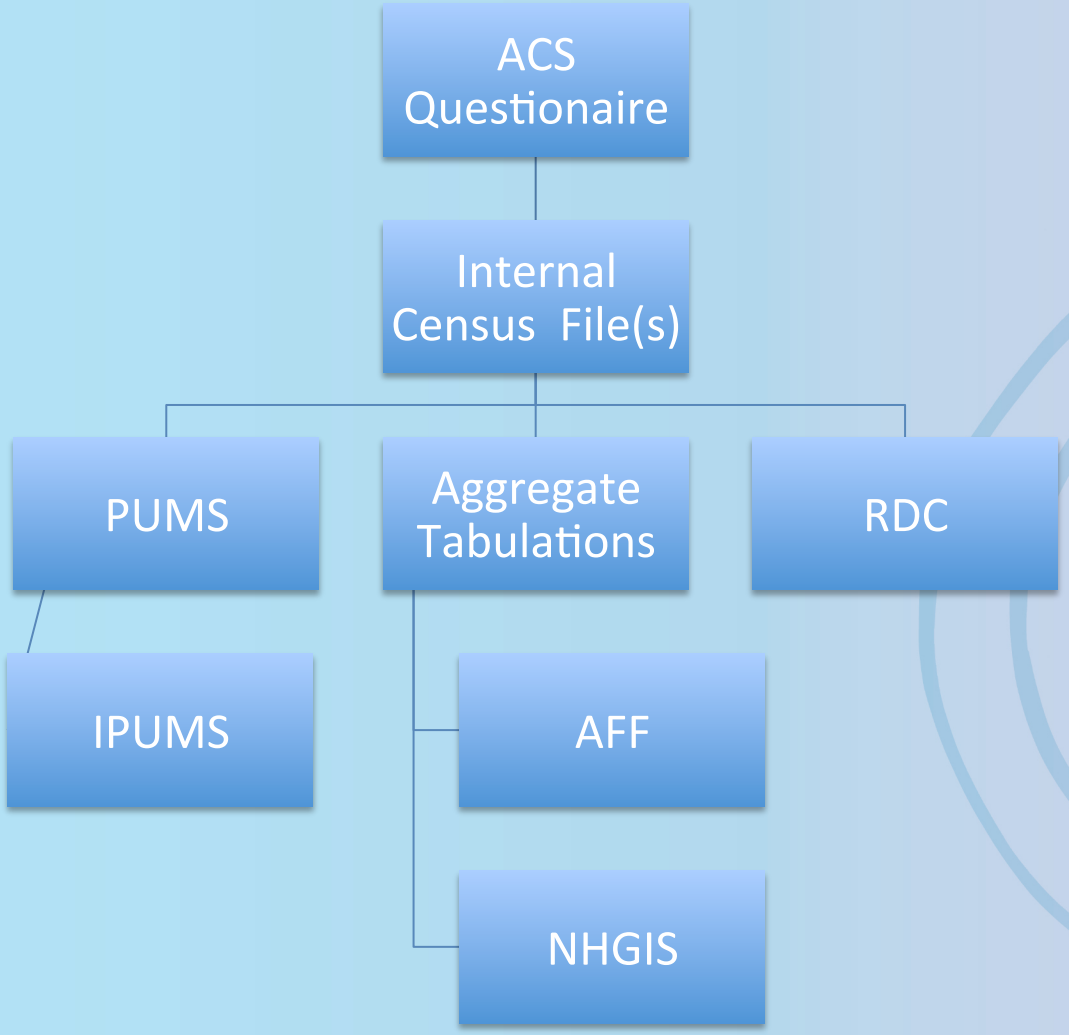


CISER

	2012 ACS Code List	ACS 2012 PUMS	IPUMS-USA	AFF	NHGIS	RDC
Alsatian	YES (001)	YES (001)	YES (75 cases)	6,626 (est.)	6,626 (est.)	Yes
Andorran	YES (002)	NO	NO	NO	NO	?
Cypriots	YES (017)	NO	NO	6,486 (est.)	6,486 (est.)	?

CISER

- Provenance of ACS Data Files



CISER

PROV at the Dataset/File Level

- Mentioned earlier our exploration of encoding PROV in RDF/XML* (CDATA; less promising)
- More recently: exploring W3C PROV Model as basis for encoding provenance metadata in DDI XML using <relStdy>
  - "…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."
- Others working to develop an RDF encoding for DDI metadata that could easily accommodate RDF-encoding of provenance metadata
- Both solutions might be viable; implementation could depend on local preferences

CISER

Embed PROV within <RelStdy>

Step 1 - Material Reference Complex Type with Prov

To include prov:Document within the <relStdy> element, a new complex type called 'materialReferenceWithProvType', which inherits from materialReferenceType can be introduced as follows:

```
<xs:complexType name="materialReferenceWithProvType" mixed="true">
   <xs:complexContent>
      <xs:extension base="materialReferenceType">
         <xs:sequence>
               <xs:element ref="prov:document" minOccurs="0" maxOccurs="1" />
         </xs:sequence>
         <xs:attribute name="provDocURI" type="xs:anyURI" use="optional"/>
      </xs:extension>
   </xs:complexContent>
</xs:complexType>
```

This allows PROV document to be embedded or referenced by URI.

CISER

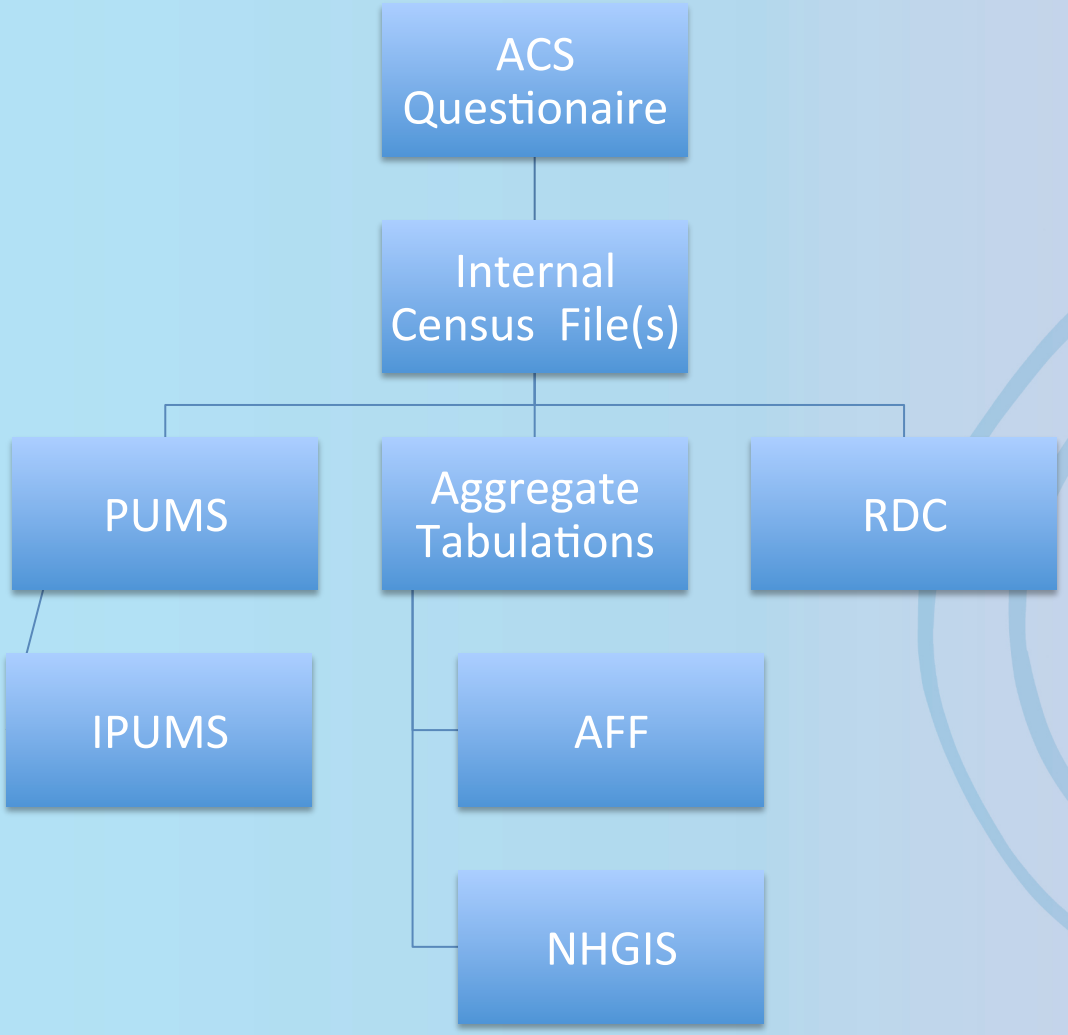
Embed PROV within <RelStdy> (Cont)

Step 2 - Modify the type of relStdy to the new complex type

The relStdy element is changed to inherit from this 'materialReferenceWithProvType', and examples are given

```
<xs:element name="relStdy" type="materialReferenceWithProvType">
   <xs:annotation>
      <xs:documentation>
         <xhtml:div>
            <xhtml:h1 class="element_title">Related Studies</xhtml:h1>
            <xhtml:div>
               <xhtml:h2 class="section_header">Description</xhtml:h2>
               <xhtml:div class="description">Information on the relationship of the current data collection to others  [4 lines]
            </xhtml:div>
            <xhtml:div> [8 lines]
            <xhtml:div> [3 lines]
            <xhtml:div>
               <xhtml:h2 class="section_header">Prov Example - Embedded Provenance Document</xhtml:h2>
               <xhtml:div class="example">
                  <xhtml:samp class="xml_sample"><![CDATA[
                     <relStdy>ICPSR distributes a companion study to this collection titled FEMALE LABOR
                     FORCE PARTICIPATION AND MARITAL INSTABILITY, 1980: [UNITED STATES] (ICPSR 9199).
                     <prov:document>...</prov:document></relStdy>
                  ]]></xhtml:samp>
               </xhtml:div>
            </xhtml:div>
            <xhtml:div>
               <xhtml:h2 class="section_header">Prov Example - Referenced Provenance Document</xhtml:h2>
               <xhtml:div class="example">
                  <xhtml:samp class="xml_sample"><![CDATA[
                     <relStdy provDocURI="http://dx.doi.org/10.1234/exProv123">ICPSR distributes a companion
                     study to this collection titled FEMALE LABOR FORCE PARTICIPATION AND MARITAL INSTABILITY,
                     1980: [UNITED STATES] (ICPSR 9199).</relStdy>
                  ]]></xhtml:samp>
               </xhtml:div>
            </xhtml:div>
         </xhtml:div>
      </xs:documentation>
   </xs:annotation>
</xs:element>
```

- Provenance of ACS Data Files



CISER

Variable Level Provenance

A single attribute is added to the variable type

```
        <xs:attribute name="geoVocab" type="xs:string"/>
        <xs:attribute name="catQnty" type="xs:string"/>
        <xs:attribute name="representationType">
           <xs:simpleType>
              <xs:restriction base="xs:NMTOKEN">
                 <xs:enumeration value="text"/>
                 <xs:enumeration value="numeric"/>
                 <xs:enumeration value="code"/>
                 <xs:enumeration value="datetime"/>
                 <xs:enumeration value="other"/>
              </xs:restriction>
           </xs:simpleType>
        </xs:attribute>
        <xs:attribute ref="prov:ref" use="optional"/>
        <xs:attribute name="otherRepresentationType" type="xs:NMTOKEN" use="optional"/>
     </xs:extension>
  </xs:complexContent>
</xs:complexType>
```

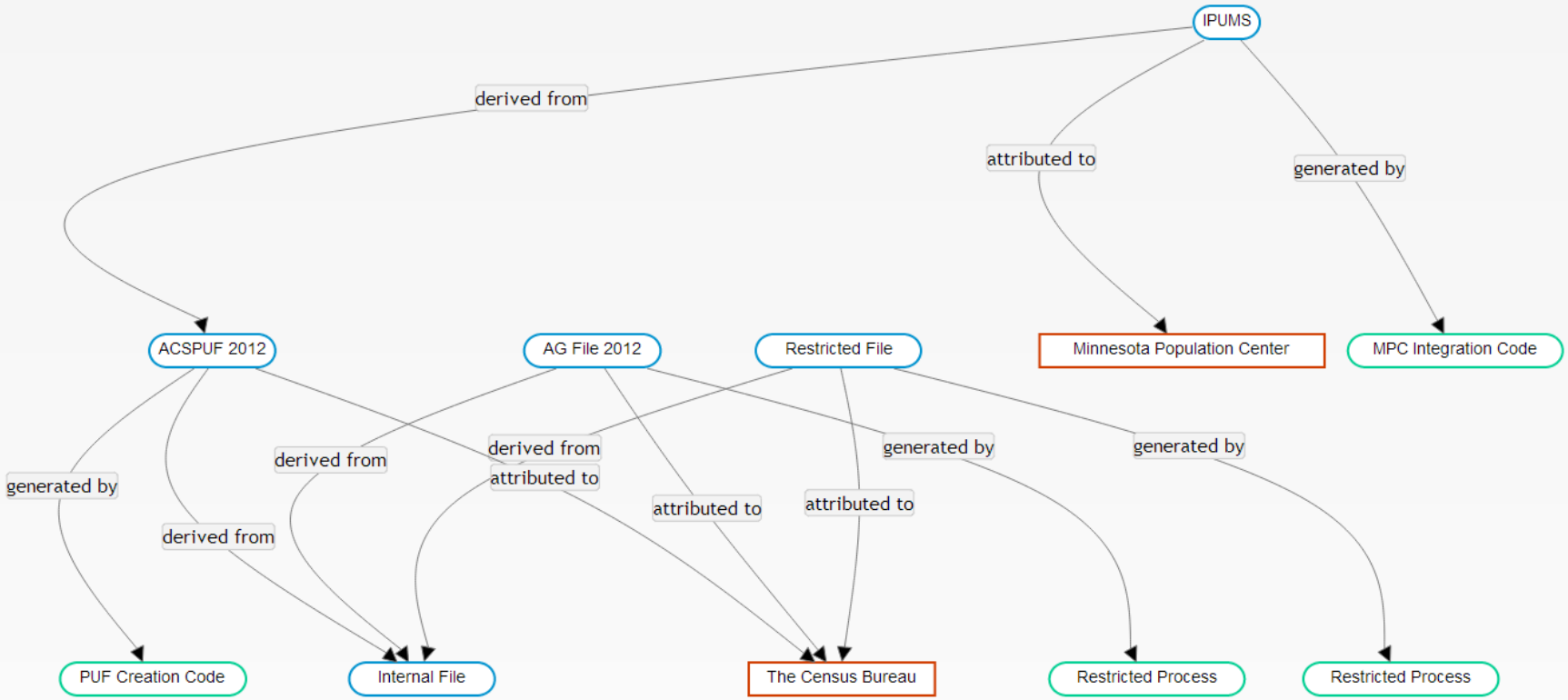
CISER

Variable Level Provenance (cont.)

In our implementation, a variable would reference a prov:Bundle that would be found within embedded prov:Document. Here is an example of a prov:Bundle:

```
<prov:bundle prov:id="ex:bundle1">
   <ex:version>1</ex:version>
</prov:bundle>

<prov:bundleContent prov:id="ex:bundle1">
   <prov:entity prov:id="ex:report1"/>

   <prov:entity prov:id="ex:report2">
      <prov:type xsi:type="xsd:QName">report</prov:type>
      <ex:version>2</ex:version>
   </prov:entity>

   <prov:wasGeneratedBy>
      <prov:entity prov:ref="ex:report2"/>
      <prov:time>2012-05-25T11:00:01</prov:time>
   </prov:wasGeneratedBy>

   <prov:wasDerivedFrom>
      <prov:generatedEntity prov:ref="ex:report2"/>
      <prov:usedEntity prov:ref="ex:report1"/>
   </prov:wasDerivedFrom>
</prov:bundleContent>
```

Prov


IPUMS
derived from
attributed to
generated by
ACSPUF 2012
AG File 2012
Restricted File
Minnesota Population Center
MPC Integration Code
generated by
derived from
derived from
attributed to
generated by
generated by
derived from
attributed to
attributed to
PUF Creation Code
Internal File
The Census Bureau
Restricted Process
Restricted Process

CISER

NCRN Meeting Spring 2014

View | Register

The eight nodes of the NSF-Census Research Network (NCRN) will hold a technical mini-symposium on Thursday and Friday, May 22-23, 2014. The program will comprise two one-half day sessions focused on topics of broad interest to the nodes, the Census Bureau and the federal statistical system. Presentations will highlight research performed by the NCRN nodes, with opportunity for questions and comments from agency researchers.

Thursday, May 22

9:00-9:30 Opening remarks by the Director of the Census Bureau, John Thompson

9:30-12:00 **Session 1: Data Documentation Initiative (DDI) Metadata within the Federal Statistical System: Implementation Challenges, and Provenance Encoding**

CISER

DDI Presentations at NCRN Spring Meeting

Data Documentation Initiative (DDI) Metadata within the Federal Statistical System: Implementation Challenges, Record Linkage, and Provenance Encoding

Jay Greenfield and Sophia Kuan (Booz Allen Hamilton): Describing Adaptive/ Responsive Protocols and Data Processing in DDI

Tim Mulcahy (NORC): DDI and Record Linkage

Jeremy Williams (Cornell University): Encoding Provenance Metadata for Social Science Datasets

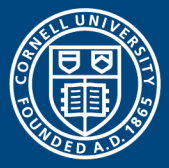
DDI Tools Demonstration

Abdul K. Rahim and Pascal Heus (Metadata Technologies North America)

Thank you!

Questions?

block@cornell.edu
ncrn.cornell.edu



Cornell University