

Appendix A: Draft analysis and migration workflow

Deciding whether to migrate a particular CD

Answers to the following questions can guide decisions about whether migration is necessary or worthwhile:

1. Is the content of the CD already available online?
 - a. If no, consider migrating.
 - b. If yes:
 - i. Is the online version of the content held at a trustworthy archive (ICPSR, NARA, etc.)? If yes, consider not migrating – unless you want to migrate to improve access (e.g. if the content is not freely or easily accessible).
 - ii. Is the online version identical to the CD version (content and/or file formats). If not, consider migrating if it is easy or worth the effort to compare the two versions.
 - iii. Is the online version in non-proprietary formats? If no, consider migrating to preservation formats.
2. Are there any copyright restrictions on any of the CD content? Note that on a single CD, some content may carry copyright restrictions and other content may be public domain – for example, software may have copyright restrictions, but data files may not.
3. Does the CD require use of software that is proprietary, uncommon, or obsolete? If yes, consider migration in order to free the data from software dependency and increase its accessibility. However, if the software enhances access to or presentation of the content, consider emulation as a preservation strategy (instead of or in addition to migration).
4. Is the CD a compilation of content or publications drawn from other sources? Examples are: the National Trade Data Bank (NTDB) and the National Economic Social and Environmental Data Bank (NESE), which include a number of diverse data and text publications on each CD and were packaged with a software interface for accessing the publications; and LandView, which brought EPA and Census data together in a geospatial interface. If the presentation or the compilation is the main point of the CD, and if the content exists elsewhere (either in an online archive or on another CD), it may not be worth migrating these. However, if there is something unique or particularly useful about the presentation, emulation may be a worthwhile preservation strategy.
5. What will it take to migrate the content? Can it be done using the software available on the CD (for example, NCHS SETS software, which allows for bulk export of datasets to non-proprietary formats) or with readily available software (e.g. opening a DBF file in Microsoft Excel and saving as comma-delimited ASCII). Will it require a manual file-by-file process, which is time-consuming and potentially prone to error? Could a script be written to batch process all the files?

Browsing a data CD & converting data to ASCII format

The following procedures apply to CDs in which the primary content is numeric data in some of the most commonly used file formats, which typically arrange the data in tabular format.

- Does the CD automatically start up when you insert the disk?
 - If CD launches a program interface, search the program for instructions on how to access the data.
 - If not, search the CD directory for a readme file, which may provide names and locations of data files, documentation files, or proprietary data-browsing programs.
- If the proprietary software program on the CD is necessary for accessing the data, but fails to run:
 - For DOS-based programs, try using Dosbox software available as a free download at <http://dosbox.sourceforge.net>. This program emulates the DOS environment and allows you to run DOS-based software on newer systems.
 - Look on the publisher's website for updated versions of the program that might run well on newer machines yet still be able to access older data.
 - Look on publisher's website, or search online, for any errors that might have occurred in the process and clues as to how they might be fixed.
 - Try contacting publisher directly about errors encountered with the CD.
 - Try running CD on another computer if it seems to be a physical problem with the disk.
- Browse documentation: Some documentation is accessed through the program interface; otherwise, readme files might direct you towards documentation stored in directories on the CD. Most documentation files are text, word document, or adobe PDF file types. The documentation is often useful for finding export or conversion tools/directions and getting a general idea of what the CD is about.
- Can you convert the files?
 - Some file formats do not require conversion, including files with the following extensions:

File types that usually do not require conversion:	
.asc	ASCII text file
.cmv	ASCII comma-separated value text file
.dat	Data file in ASCII or special format (Usually does not need conversion, but open and observe files in notepad in case it is a special format.)
File types that can be easily converted using "save as" dialog box in Excel:	
.xls	Excel file
.dbf	Database file
Non-ASCII file types:	
	▪ Try opening the data files in notepad: Sometimes data files will have unfamiliar extensions but open up in Notepad and are

	<p>readily recognizable as data tables.</p> <ul style="list-style-type: none"> ▪ Try any conversion tools you may have discovered in documentation. Sometimes proprietary software will provide access to the conversion tool directly from the data-browsing interface; otherwise, the documentation might describe conversion tools located in other directories of the CD. ▪ Check website of publishing company for updated software that might be able to perform conversions, or at least perform conversion in a more efficient manner (faster or in bulk). ▪ Look up file extension online to see if it is really a data file (for example, many files might actually be images). ▪ Some files may be necessary for the presentation and manipulation of the data in the software package on the CD, but not necessary for exported data in a non-proprietary format. For example, Census CDs may include .ndx and .exe files that are for use only with GO software, and need not be included with the migrated data files.
--	--

Metadata and documentation

Migration processes must be documented so that end-users understand 1) that the content they are accessing is not the original version as published on the CD, and 2) what steps were taken to transform the files, in case of discrepancies between the original and migrated versions.

Preservation metadata can describe both migration processes and technical information that will help preserve the files.

Documentation on the original CD may not be appropriate for migrated content – for example, documentation describing how to use the Census GO software will not make sense with Census data files that have been exported to ASCII format. Decisions will need to be made about whether to:

- Migrate and preserve the original documentation to accompany the migrated content files. If yes, how best to indicate that the documentation does not apply to the migrated content?
- Write new documentation to accompany the migrated content.

Needs for software-specific procedures

Some software programs were used on so many depository CDs that specific migration guidelines for those packages would be useful. For example:

GO and Extract (used on Census Bureau CDs)

On some CDs, the data are spread over a number of small files (often in .dbf format) that are meant to be connected by an index (.ndx) file. Can the small files be concatenated into a larger ASCII file, correctly maintaining all the data relationships? Or can the small files be batch extracted and appropriately documented in the absence of the GO software to select data? How will the original documentation need to be revised to correctly describe the migrated output?

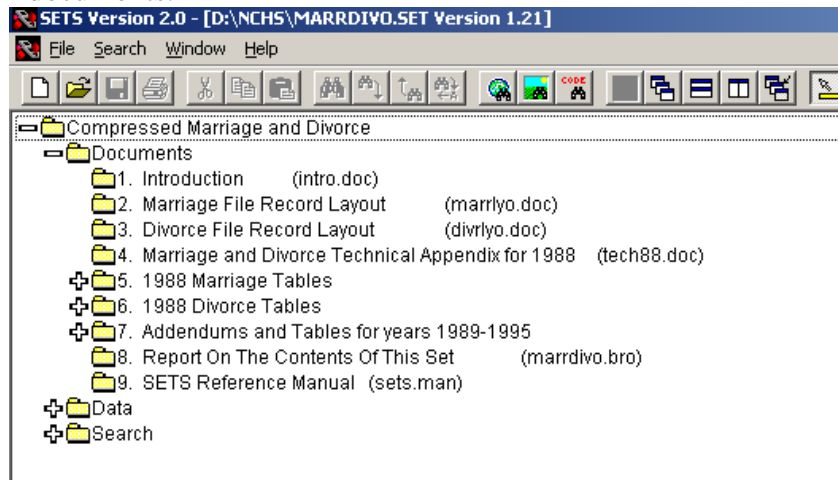
Beyond 20/20 (used on some NCHS and Health Canada CDs)

There does not seem to be a way to do a batch conversion of data from this proprietary format to ASCII.

Appendix B: Detailed analysis of Marriage & Divorce Vital Statistics CD

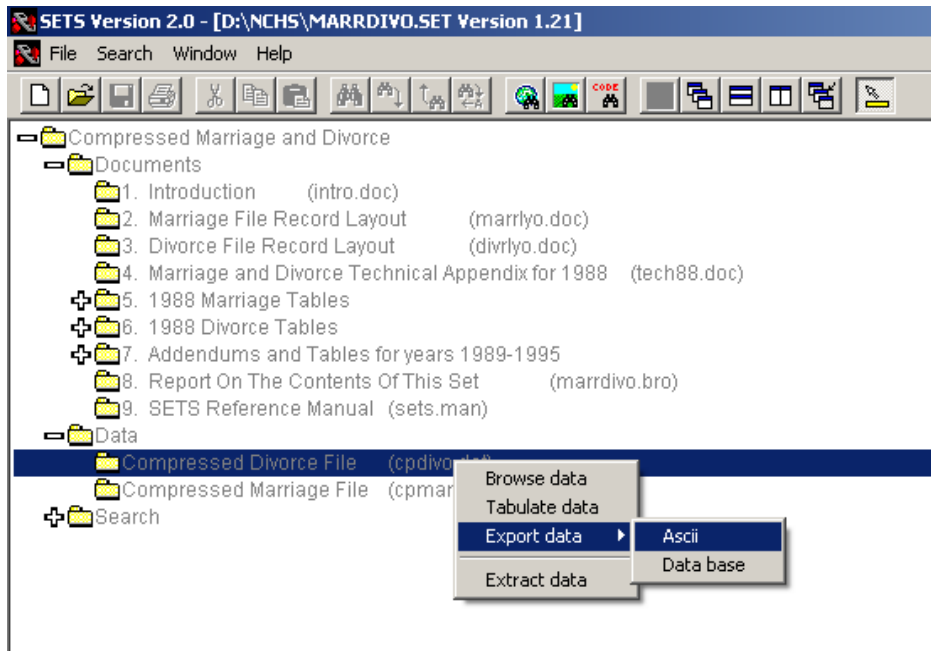
Exporting and Converting Data from
CDC/NCHS CD-ROM Series 21:
Marriage and Divorce Data, 1898-95

- Download and install SETS 2.0 from CDC website
 - <http://www.cdc.gov/nchs/about/otheract/sets/download.htm> for file and installation instructions.
 - *The version of SETS that comes with the CD-ROM, SETS 1.22, had a dated DOS interface and estimated 3 hours to export the divorce data on the CD-ROM. The latest build on the website, SETS 2.0, performed the job in less than 5 minutes for close to 900,000 records. Converting other data sets can be longer, however. One dataset that was estimated to take close to 12 hours on SETS 1.22 was completed in about 40 minutes.*
- Opened the “sets” file on the CD-ROM by going to File > Open, browsing to the CD-ROM drive, opening the folder “nchs”, and clicking on marrdivo.sets.
- Field codes and record details can be accessed via the SETS interface or can be found as Word Documents on the CD-ROM. The SETS 2.0 interface indicates the file name of such documents.

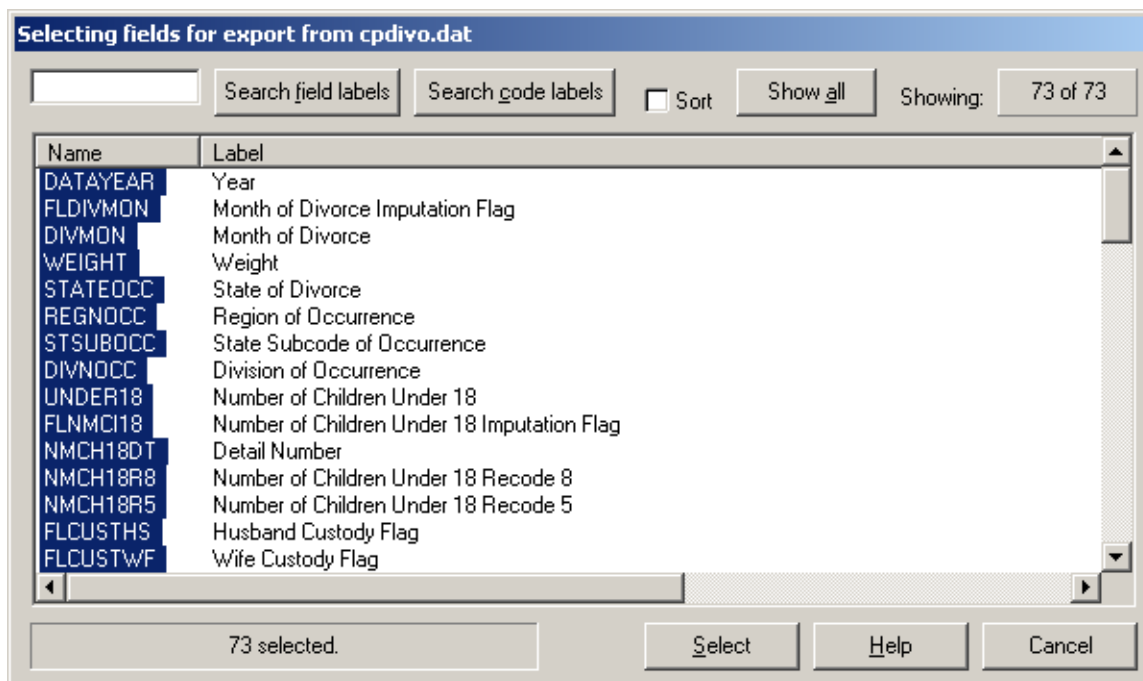


The navigation window above indicates that the marriage file record layout is contained in a file called marrlyo.doc while the divorce file record layout is on divrlyo.doc.

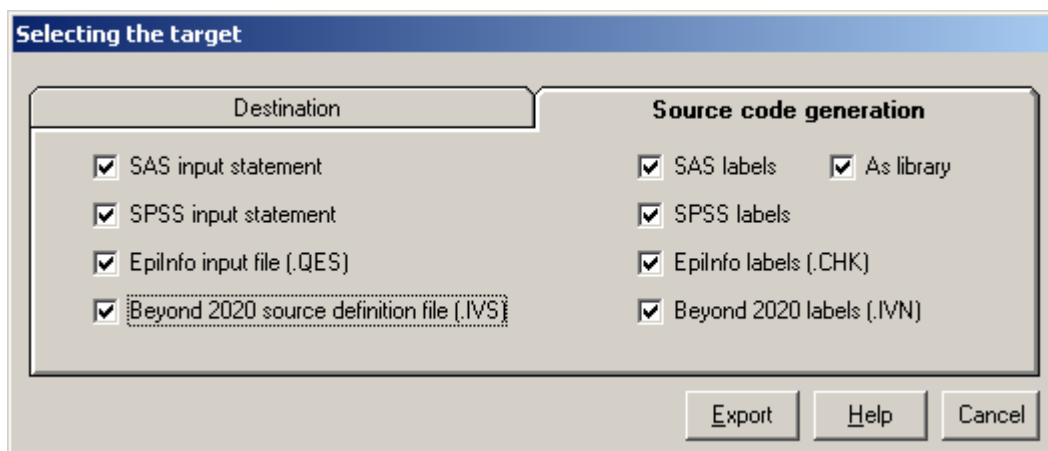
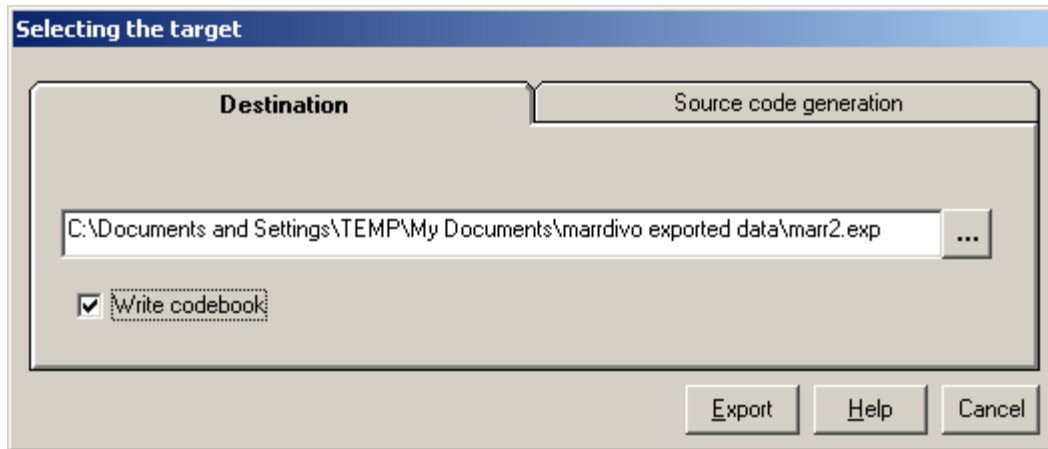
- To export the data, select click on the dataset you want to export and select ascii.



- A new window allows you to select which fields you want to export. To select all fields, click on the label that says “name”.



You have the option of selecting a codebook and whatever other kinds of data you want to export.



After you hit export, the new files are created in the specified destination.

The export function generated several files, including:

- A *.exp* file that contained the raw data.
- The SAS and SPSS outputs, which contained SAS scripts and SPSS input files for reading and labeling the output information.
 - To read the data file using SPSS, enter the full path to the file in the data list file statement, for example: **DATA LIST FILE="C:\My Documents\Data\da2992.txt" /**
 - For help on using SPSS to access data, use the ICPSR website <http://webapp.icpsr.umich.edu/cocoon/ICPSR-FAQ/0062.xml>.
- A *.cbk* file codebook that could be viewed with the notebook program. This codebook was not very useful. While the exported codebook provided field names, it did not provide descriptions of each field or how to interpret the numbers provided.

Appendix C: Metadata samples

MARCXML record:

```
<?xml version="1.0" encoding="UTF-8" ?>
<collection xmlns="http://www.loc.gov/MARC21/slim"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/MARC21/slim
http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd">
<record><leader>02210cas a2200493 a 4500</leader>
<controlfield tag="001">3664004</controlfield>

<controlfield tag="005">20020514115650.0</controlfield>

<controlfield tag="008">950222c19929999dcuar          f0      0eng
d</controlfield>

<controlfield tag="006">m          d f          </controlfield>

<controlfield tag="007">couugu</controlfield>

<datafield tag="010" ind1=" " ind2=" ">
<subfield code="a"> 95645688 </subfield>
<subfield code="z">sn 95027175 </subfield>
</datafield>

<datafield tag="035" ind1=" " ind2=" ">
<subfield code="a">(OCoLC)ocm32033327</subfield>
</datafield>

<datafield tag="035" ind1=" " ind2=" ">
<subfield code="9">FJE8405YL</subfield>
</datafield>

<datafield tag="037" ind1=" " ind2=" ">
<subfield code="b">U.S. Dept. of Commerce, Data User Services Division,
Washington, DC 20233</subfield>
</datafield>

<datafield tag="040" ind1=" " ind2=" ">
<subfield code="a">DGPO</subfield>
<subfield code="c">DGPO</subfield>
<subfield code="d">DLC</subfield>
<subfield code="d">OCoLC</subfield>
<subfield code="d">InU</subfield>
<subfield code="d">DLC</subfield>
</datafield>

<datafield tag="042" ind1=" " ind2=" ">
<subfield code="a">lc</subfield>
</datafield>

<datafield tag="043" ind1=" " ind2=" ">
<subfield code="a">n-us---</subfield>
</datafield>
```

```
<datafield tag="050" ind1="0" ind2="0">
<subfield code="a">HJ275</subfield>
</datafield>

<datafield tag="074" ind1=" " ind2=" ">
<subfield code="a">0154-B-02</subfield>
</datafield>

<datafield tag="086" ind1="0" ind2=" ">
<subfield code="a">C 3.266/3:</subfield>
</datafield>

<datafield tag="245" ind1="0" ind2="0">
<subfield code="a">Consolidated federal funds reports</subfield>
<subfield code="h">[computer file] :</subfield>
<subfield code="b">CFFR.</subfield>
</datafield>

<datafield tag="246" ind1="1" ind2=" ">
<subfield code="i">Title on technical documentation title screen:</subfield>
<subfield code="a">Consolidated federal funds report (CFFR) ... on CD-
ROM</subfield>
<subfield code="f">1984/1993-</subfield>
</datafield>

<datafield tag="246" ind1="3" ind2="0">
<subfield code="a">CFFR</subfield>
</datafield>

<datafield tag="260" ind1=" " ind2=" ">
<subfield code="a">Washington, DC :</subfield>
<subfield code="b">U.S. Dept. of Commerce, Bureau of the Census, Data User
Services Division,</subfield>
<subfield code="c">[1994-</subfield>
</datafield>

<datafield tag="300" ind1=" " ind2=" ">
<subfield code="a">computer laser optical discs ;</subfield>
<subfield code="c">4 3/4 in.</subfield>
</datafield>

<datafield tag="310" ind1=" " ind2=" ">
<subfield code="a">Annual</subfield>
</datafield>

<datafield tag="362" ind1="0" ind2=" ">
<subfield code="a">Fiscal years 1983-1992-</subfield>
</datafield>

<datafield tag="500" ind1=" " ind2=" ">
<subfield code="a">Title from title screen.</subfield>
</datafield>

<datafield tag="515" ind1=" " ind2=" ">
<subfield code="a">Each issue covers 10 fiscal years.</subfield>
</datafield>
```

```
<datafield tag="516" ind1="8" ind2=" ">
<subfield code="a">Written in ISO 9660 format.</subfield>
</datafield>

<datafield tag="516" ind1="8" ind2=" ">
<subfield code="a">Text files written in ASCII format; data files written in
dBASE III+ format.</subfield>
</datafield>

<datafield tag="538" ind1=" " ind2=" ">
<subfield code="a">System requirements: IBM PC or compatible; 640K RAM; DOS
3.3 or higher; MS-DOS Extensions (version 2.0 or higher); CD-ROM
drive.</subfield>
</datafield>

<datafield tag="580" ind1=" " ind2=" ">
<subfield code="a">Contains data from: Consolidated federal funds report.
Volume I, County areas; and: Consolidated federal funds report. Volume II,
Subcounty areas.</subfield>
</datafield>

<datafield tag="650" ind1=" " ind2="0">
<subfield code="a">Grants-in-aid</subfield>
<subfield code="z">United States</subfield>
<subfield code="v">Statistics.</subfield>
</datafield>

<datafield tag="650" ind1=" " ind2="0">
<subfield code="a">Economic assistance, Domestic</subfield>
<subfield code="z">United States</subfield>
<subfield code="v">Statistics.</subfield>
</datafield>

<datafield tag="650" ind1=" " ind2="0">
<subfield code="a">Government lending</subfield>
<subfield code="z">United States</subfield>
<subfield code="v">Statistics.</subfield>
</datafield>

<datafield tag="650" ind1=" " ind2="0">
<subfield code="a">Public contracts</subfield>
<subfield code="z">United States</subfield>
<subfield code="v">Statistics.</subfield>
</datafield>

<datafield tag="655" ind1=" " ind2="7">
<subfield code="a">CD-ROMs</subfield>
<subfield code="2">lcsh</subfield>
</datafield>

<datafield tag="710" ind1="1" ind2=" ">
<subfield code="a">United States.</subfield>
<subfield code="b">Bureau of the Census.</subfield>
<subfield code="b">Data User Services Division.</subfield>
</datafield>

<datafield tag="787" ind1="1" ind2=" ">
```

```

<subfield code="t">Consolidated federal funds report. Volume I, County
areas</subfield>
<subfield code="w">(DLC) 84644630</subfield>
<subfield code="w">(OCOLC)10681892</subfield>
</datafield>

<datafield tag="787" ind1="1" ind2=" ">
<subfield code="t">Consolidated federal funds report. Volume II, Subcounty
areas</subfield>
<subfield code="w">(DLC) 84644150</subfield>
<subfield code="w">(OCOLC)10681827</subfield>
</datafield>

<datafield tag="927" ind1=" " ind2=" ">
<subfield code="a">9512R0</subfield>
</datafield>

<datafield tag="928" ind1=" " ind2=" ">
<subfield code="a">AC11231999</subfield>
</datafield>

<datafield tag="948" ind1=" " ind2=" ">
<subfield code="a">docs,jds</subfield>
</datafield>

</record>
</collection>

```

MODS record:

```

<?xml version="1.0"?>
<modsCollection xsi:schemaLocation="http://www.loc.gov/mods/
http://www.loc.gov/standards/mods/mods.xsd"
xmlns:xlink="http://www.w3.org/TR/xlink" xmlns="http://www.loc.gov/mods/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<mods>
<titleInfo>
<title>Consolidated federal funds reports [computer file] :</title>
<subTitle>CFFR</subTitle>
</titleInfo>
<titleInfo type="alternative" displayLabel="Title on technical documentation
title screen:">
<title>Consolidated federal funds report (CFFR) ... on CD-ROM 1984/1993-
</title>
</titleInfo>
<titleInfo type="alternative">
<title>CFFR</title>
</titleInfo>
<name type="corporate">
<namePart>United States</namePart>
<namePart>Bureau of the Census.</namePart>
<namePart>Data User Services Division.</namePart>
</name>
<typeOfResource>text</typeOfResource>
<genre authority="lcsch">CD-ROMs</genre>

```

<originInfo>
<place>
<code authority="marc">dcu</code>
<text>Washington, DC</text>
</place>
<publisher>U.S. Dept. of Commerce, Bureau of the Census, Data User Services
Division</publisher>
<dateIssued>[1994-</dateIssued>
<dateIssued encoding="marc" point="start">1992</dateIssued>
<dateIssued encoding="marc" point="end">9999</dateIssued>
<issuance>continuing</issuance>
<frequency>Annual</frequency>
</originInfo>
<language authority="iso639-2b">eng</language>
<physicalDescription>
<form authority="marcform">print</form><extent>computer laser optical discs ;
4 3/4 in.</extent>
</physicalDescription>
<note>Title from title screen.</note>
<note>Each issue covers 10 fiscal years.</note>
<note>Written in ISO 9660 format.</note>
<note>Text files written in ASCII format; data files written in dBASE III+
format.</note>
<note>System requirements: IBM PC or compatible; 640K RAM; DOS 3.3 or higher;
MS-DOS Extensions (version 2.0 or higher); CD-ROM drive.</note>
<note>Contains data from: Consolidated federal funds report. Volume I, County
areas; and: Consolidated federal funds report. Volume II, Subcounty
areas.</note>
<subject authority="lcsch">
<topic>Grants-in-aid</topic>
<geographic>United States</geographic>
<topic>Statistics</topic>
</subject>
<subject authority="lcsch">
<topic>Economic assistance, Domestic</topic>
<geographic>United States</geographic>
<topic>Statistics</topic>
</subject>
<subject authority="lcsch">
<topic>Government lending</topic>
<geographic>United States</geographic>
<topic>Statistics</topic>
</subject>
<subject authority="lcsch">
<topic>Public contracts</topic>
<geographic>United States</geographic>
<topic>Statistic</topic>
</subject>
<classification authority="lcc">HJ275</classification>
<classification authority="sudocs">C 3.266/3:</classification>
<classification authority="">C 3.266/3:</classification>
<relatedItem type="related">
<titleInfo>
<title>Consolidated federal funds report. Volume I, County areas</title>
</titleInfo>
<identifier type="local">(DLC) 84644630</identifier>
<identifier type="local">(OCOLC)10681892</identifier>

```

</relatedItem>
<relatedItem type="related">
<titleInfo>
<title>Consolidated federal funds report. Volume II, Subcounty areas</title>
</titleInfo>
<identifier type="local">(DLC) 84644150</identifier>
<identifier type="local">(OCoLC)10681827</identifier>
</relatedItem>
<identifier type="lccn" invalid="yes">95645688</identifier>
<identifier type="stock number">U.S. Dept. of Commerce, Data User Services
Division, Washington, DC 20233</identifier>
<recordInfo>
<recordContentSource>DGPO</recordContentSource>
<recordCreationDate encoding="marc">950222</recordCreationDate>
<recordChangeDate encoding="iso8601">20020514115650.0</recordChangeDate>
<recordIdentifier>3664004</recordIdentifier>
</recordInfo>
</mods>
</modsCollection>

```

Study-level DDI record for a normalized (migrated) dataset:

```

<?xml version="1.0" encoding="UTF-8"?>

<codeBook xmlns="http://www.icpsr.umich.edu/DDI"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:date="http://exslt.org/dates-and-times"
xsi:schemaLocation="http://www.icpsr.umich.edu/DDI
http://www.icpsr.umich.edu/DDI/Version2-1.xsd" version="2.1">
  <docDscr>
    <citation>
      <titlStmt>
        <titl>Birth cohort linked birth/infant death data set</titl>
      </titlStmt>
      <prodStmt>
        <producer>Yale University Social Science Libraries and
Information Services</producer>
        <prodDate date="2006-09-27">2006-09-27</prodDate>
      </prodStmt>
    </citation>
  </docDscr>
  <stdyDscr>
    <citation>
      <titlStmt>
        <titl>Birth cohort linked birth/infant death data set</titl>
      </titlStmt>
      <rspStmt>
        <AuthEnty>National Center for Health Statistics (U.S.)</AuthEnty>
      </rspStmt>
      <prodStmt>
        <producer>U.S. Dept. of Health and Human Services, Public Health
Service, Centers for Disease Control and Prevention, National Center for
Health Statistics</producer>
        <prodDate date="1985">1985</prodDate>
        <prodPlac>Hyattsville, Md</prodPlac>
      </prodStmt>
    </citation>
  </stdyDscr>

```

```

<serStmt>
  <serName>NCHS CD-ROM. Series 20</serName>
</serStmt>
<verStmt>
  <notes>Yale University Social Science Libraries and Information
Services migrated the files in this dataset from their original
formats.</notes>
  </verStmt>
</citation>
<studyInfo>
  <subject>
    <keyword>birth, birth rates, birth records, infant mortality,
live births, medical records, parents, pregnancy, prenatal care, reproductive
history, vital statistics</keyword>
    <topcClas vocab="LOC Subject Headings"
vocabURI="http://www.loc.gov/catdir/cpsolcco/lcco.html">Childbirth -- United
States -- Statistics -- Periodicals</topcClas>
    <topcClas vocab="LOC Subject Headings"
vocabURI="http://www.loc.gov/catdir/cpsolcco/lcco.html">Infants -- United
States -- Mortality -- Statistics -- Periodicals</topcClas>
  </subject>
  <abstract>Provides birth and infant death data from 1985- cohort
arranged in three files: a numerator file (linked records of live births and
infant deaths), a denominator plus file (live birth records offered as
numerator/denominator data sets to facilitate computation of infant mortality
rates), and an unlinked file (infant death records which cannot be linked to
a corresponding birth record).</abstract>
  <sumDscr>
    <timePrd event="single" date="YYYY-MM-DD">temporal coverage
here</timePrd>
    <geogCover>United States</geogCover>
  </sumDscr>
</studyInfo>
<dataAccs>
  <setAvail>
    <origArch>Disseminated by the U.S. Government Printing Office
through the Federal Depository Library Program. CD-ROM held at Yale
University Library, Government Documents & Information Center.</origArch>
  </setAvail>
</dataAccs>
<otherStudyMat>
  <relMat callno="HE 20.6209/4-7:20/">
    <citation>
      <titlStmt>
        <titl>Birth cohort linked birth/infant death data
set</titl>
        <IDNo agency="CtY">3795296</IDNo>
      </titlStmt>
    </citation>
  </relMat>
</otherStudyMat>
</studyDscr>
</codeBook>

```

Portion of file-level DDI record:

```
<fileDscr ID="XLS-TW7CAD.XLS"
URI="http://never.its.yale.edu:8080/fedora/get/ssrs:1284/XLS-TW7CAD.XLS"/>

    <fileDscr ID="XLS-TW8A.XLS"
URI="http://never.its.yale.edu:8080/fedora/get/ssrs:1284/XLS-TW8A.XLS"/>

        <fileDscr ID="XLS-TW8C.XLS"
URI="http://never.its.yale.edu:8080/fedora/get/ssrs:1284/XLS-TW8C.XLS"/>
```

Portion of variable-level DDI record:

```
<dataDscr>
  <var ID="DATAYEAR" name="DATAYEAR">
    <labl>Year of Marriage</labl>
    <sumStat type="vald">1357710</sumStat>
    <sumStat type="invd">0</sumStat>
    <sumStat type="min">1989</sumStat>
    <sumStat type="max">1995</sumStat>
    <sumStat type="mean">1991.9801548195</sumStat>
    <sumStat type="medn">1992</sumStat>
    <sumStat type="mode">1990</sumStat>
    <sumStat type="stdev">2.0091191092648</sumStat>
    <catgry>
      <catValu>1989</catValu>
      <labl>1989</labl>
      <catStat type="freq">198197</catStat>
      <catStat type="percent">14.597889092663</catStat>
    </catgry>
    <catgry>
      <catValu>1990</catValu>
      <labl>1990</labl>
      <catStat type="freq">199493</catStat>
      <catStat type="percent">14.693343939427</catStat>
    </catgry>
    <catgry>
      <catValu>1991</catValu>
      <labl>1991</labl>
      <catStat type="freq">191968</catStat>
      <catStat type="percent">14.139101870061</catStat>
    </catgry>
    <catgry>
      <catValu>1992</catValu>
      <labl>1992</labl>
      <catStat type="freq">190877</catStat>
      <catStat type="percent">14.058745976681</catStat>
    </catgry>
    <catgry>
      <catValu>1993</catValu>
      <labl>1993</labl>
      <catStat type="freq">188670</catStat>
      <catStat type="percent">13.896192854144</catStat>
    </catgry>
    <catgry>
      <catValu>1994</catValu>
```

```

        <labl>1994</labl>
        <catStat type="freq">195584</catStat>
        <catStat type="percent">14.405432677081</catStat>
    </catgry>
</catgry>
    <catValu>1995</catValu>
    <labl>1995</labl>
    <catStat type="freq">192921</catStat>
    <catStat type="percent">14.209293589942</catStat>
</catgry>
</var>
<var ID="REGNOCC" name="REGNOCC">
    <labl>Region of Marriage</labl>
    <sumStat type="vald">1357710</sumStat>
    <sumStat type="invd">0</sumStat>
    <sumStat type="min">0</sumStat>
    <sumStat type="max">4</sumStat>
    <sumStat type="mean">2.4346561489567</sumStat>
    <sumStat type="medn">3</sumStat>
    <sumStat type="mode">3</sumStat>
    <sumStat type="stdev">1.1189638193862</sumStat>
    <catgry>
        <catValu>0</catValu>
        <labl>Possessions</labl>
        <catStat type="freq">43690</catStat>
        <catStat type="percent">3.217918406729</catStat>
    </catgry>
    <catgry>
        <catValu>1</catValu>
        <labl>Northeast</labl>
        <catStat type="freq">282136</catStat>
        <catStat type="percent">20.780284449551</catStat>
    </catgry>
    <catgry>
        <catValu>2</catValu>
        <labl>Midwest</labl>
        <catStat type="freq">342190</catStat>
        <catStat type="percent">25.203467603538</catStat>
    </catgry>
    <catgry>
        <catValu>3</catValu>
        <labl>South</labl>
        <catStat type="freq">419735</catStat>
        <catStat type="percent">30.914922921684</catStat>
    </catgry>
    <catgry>
        <catValu>4</catValu>
        <labl>West</labl>
        <catStat type="freq">269959</catStat>
        <catStat type="percent">19.883406618497</catStat>
    </catgry>
</var>

```

Portion of FOXML object containing Dublin Core:

```
<foxml:datastream CONTROL_GROUP="X" ID="DC" STATE="A">
  <foxml:datastreamVersion ID="DC.0" LABEL="Dublin Code Record"
MIMETYPE="text/xml">
  <foxml:xmlContent>
    <oai_dc:dc
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/">
      <dc:title>Consolidated federal funds reports: CFFR</dc:title>
      <dc:creator>United States Bureau of the Census. Data User
Services Division.</dc:creator>
      <dc:subject>keyword here</dc:subject>
      <dc:subject>Grants-in-aid -- United States --
Statistics</dc:subject>
      <dc:subject>Economic assistance, Domestic -- United States --
Statistics</dc:subject>
      <dc:subject>Government lending -- United States --
Statistics</dc:subject>
      <dc:subject>Public contracts -- United States --
Statistics</dc:subject>
      <dc:publisher>U.S. Dept. of Commerce, Bureau of the Census,
Data User Services Division</dc:publisher>
      <dc:date>1992</dc:date>
      <dc:coverage>temporal coverage here</dc:coverage>
      <dc:rights>restriction here</dc:rights>
    </oai_dc:dc>
  </foxml:xmlContent>
</foxml:datastreamVersion>
</foxml:datastream>
```

Portion of FOXML object containing PREMIS for a single ASCII file in the dataset:

```
<foxml:datastream CONTROL_GROUP="E" ID="DNSTK2.txt" STATE="A">
  <foxml:datastreamVersion ID="DNSTK2.txt.0" LABEL="DNSTK2.txt"
MIMETYPE="text/plain">
  <foxml:contentLocation REF="file:/C:/Documents and
Settings/mikea/Desktop/4409810/normal/ascii_files/DNSTK2.txt" TYPE="URL"/>
  </foxml:datastreamVersion>
</foxml:datastream>
<foxml:datastream CONTROL_GROUP="X" ID="PREMIS-DNSTK2.txt" STATE="A">
  <foxml:datastreamVersion
FORMAT_URI="http://www.loc.gov/standards/premis/v1" ID="PREMIS-DNSTK2.txt.0"
LABEL="PREMIS-DNSTK2.txt" MIMETYPE="text/xml">
  <foxml:xmlContent>
    <premis:object xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:premis="http://www.loc.gov/standards/premis/v1"
xsi:schemaLocation="http://www.loc.gov/standards/premis/v1
http://www.loc.gov/standards/premis/v1/PREMIS-v1-1.xsd">
      <!-- The is at the Object Level category -->

      <premis:objectIdentifier>
```

```

<!-- create namespace convention for FEDORA PID and datastreams -->

<premis:objectIdentifierType>URI</premis:objectIdentifierType>
<premis:objectIdentifierValue>http://never.its.yale.edu:8080/fedora/get/ssrs:
1284/DNSTK2.txt</premis:objectIdentifierValue>
  </premis:objectIdentifier>
  <premis:preservationLevel>1</premis:preservationLevel>
  <premis:objectCategory>File</premis:objectCategory>
  <premis:objectCharacteristics>
    <premis:compositionLevel>0</premis:compositionLevel>
    <premis:fixity>
<premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
<premis:messageDigest>e3928e61e2dd8ffcf5e470aae85ba7b5</premis:messageDigest>
<premis:messageDigestOriginator>SSL/ITS</premis:messageDigestOriginator>
  </premis:fixity>
  <premis:size>791836</premis:size>
  <!-- expressed in bytes -->
  <premis:format>
    <premis:formatDesignation>
      <premis:formatName>text/plain</premis:formatName>
    </premis:formatDesignation>
    <premis:formatRegistry>
      <premis:formatRegistryRole/>
    </premis:formatRegistry>
  </premis:format>
  </premis:objectCharacteristics>
  <premis:originalName>DNSTK2.txt</premis:originalName>
  <premis:storage>
    <premis:contentLocation>
<premis:contentLocationType>URI</premis:contentLocationType>
<premis:contentLocationValue>http://never.its.yale.edu:8080/fedora/get/ssrs:1
284/DNSTK2.txt</premis:contentLocationValue>
  </premis:contentLocation>
  <premis:storageMedium>magnetic
storage</premis:storageMedium>
  </premis:storage>

```

Appendix D: XSLT stylesheet

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
version="1.0"
  xmlns:date="http://exslt.org/dates-and-times"
  xmlns:mods="http://www.loc.gov/mods/v3"
  exclude-result-prefixes="#default mods">

  <xsl:output method="xml" indent="yes" encoding="UTF-8"/>

  <xsl:template match="/">
    <xsl:apply-templates/>
  </xsl:template>

  <xsl:template match="mods:modsCollection">
    <xsl:apply-templates/>
  </xsl:template>

  <xsl:template match="mods:mods">
    <codeBook xmlns="http://www.icpsr.umich.edu/DDI"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.icpsr.umich.edu/DDI
http://www.icpsr.umich.edu/DDI/Version2-1.xsd" version="2.1">
      <docDscr>
        <citation>
          <titlStmt>
            <titl><xsl:call-template name="titl"/></titl>
          </titlStmt>
          <prodStmt>
            <producer>Yale University Social Science Libraries
and Information Services</producer>
            <prodDate>
              <xsl:attribute name="date"><xsl:value-of
select="date:date()"/></xsl:attribute>
              <xsl:value-of select="date:date()"/>
            </prodDate>
          </prodStmt>
        </citation>
      </docDscr>
      <stdyDscr>
        <citation>
          <titlStmt>
            <titl><xsl:call-template name="titl"/></titl>
          </titlStmt>
          <xsl:call-template name="rspStmt"/>
          <prodStmt>
            <xsl:call-template name="producer"/>
            <xsl:call-template name="prodDate"/>
            <xsl:call-template name="prodPlac"/>
          </prodStmt>
          <xsl:call-template name="serStmt"/>
          <xsl:call-template name="verStmt"/>
          <verStmt>
            <notes>
```

<xsl:text>Yale University Social Science Libraries and Information Services migrated the files in this dataset from their original formats. Migration and format details for each file are described below.</xsl:text>

```
</notes>
</verStmt>
<biblCit>citation here</biblCit>
</citation>
<stdyInfo>
  <subject>
    <keyword>keyword here</keyword>
    <xsl:call-template name="topcClas"/>
  </subject>
  <xsl:call-template name="abstract"/>
  <sumDscr>
    <xsl:call-template name="timePrd"/>
    <xsl:call-template name="nation"/>
    <xsl:call-template name="geogCover"/>
  </sumDscr>
</stdyInfo>
<dataAccs>
  <setAvail>
    <origArch>Disseminated by the U.S. Government
    Printing Office through the Federal Depository Library Program. CD-ROM held
    at Yale University Library, Government Documents & Information
    Center.</origArch>
    <fileQnty>file quantity here</fileQnty>
  </setAvail>
  <useStmt>
    <restrctn>restriction here</restrctn>
  </useStmt>
</dataAccs>
<othrStdyMat>
  <relMat>
    <xsl:attribute name="callno">
      <xsl:value-of
select="mods:classification[@authority='sudocs']"/>
    </xsl:attribute>
    <citation>
      <titlStmt>
        <titl><xsl:call-template name="titl"/></titl>
        <IDNo agency="CtY">
          <xsl:value-of
select="mods:recordInfo/mods:recordIdentifier"/>
        </IDNo>
      </titlStmt>
    </citation>
  </relMat>
</othrStdyMat>
</stdyDscr>
</codeBook>
</xsl:template>

<xsl:template name="titl">
  <xsl:value-of select="mods:titleInfo[not(@type)]/mods:title"/>
  <xsl:for-each select="mods:titleInfo[not(@type)]/mods:subTitle">
    <xsl:text>: </xsl:text>
  </xsl:for-each>
</xsl:template>
```



```

                <xsl:text> </xsl:text>
            </xsl:if>
            <xsl:apply-templates/>
        </xsl:for-each>
    </AuthEnty>
</xsl:for-each>
</rspStmt>
</xsl:if>

</xsl:template>

<xsl:template name="producer">
    <xsl:for-each select="mods:originInfo/mods:publisher">
        <producer xmlns="http://www.icpsr.umich.edu/DDI">
            <xsl:apply-templates/>
        </producer>
    </xsl:for-each>
</xsl:template>

<xsl:template name="prodDate">
    <xsl:for-each select="mods:originInfo/mods:*[@encoding='marc' and
@point='start']">
        <prodDate xmlns="http://www.icpsr.umich.edu/DDI">
            <xsl:attribute name="date">
                <xsl:value-of select="node()"/>
            </xsl:attribute>
            <xsl:apply-templates/>
            <xsl:for-each select="mods:originInfo/mods:*[@encoding='marc'
and @point='end']">
                <xsl:if test="node()!='9999' and
not(contains(node(),'u'))">
                    <xsl:text>-</xsl:text>
                    <xsl:apply-templates/>
                </xsl:if>
            </xsl:for-each>
        </prodDate>
    </xsl:for-each>
</xsl:template>

<xsl:template name="prodPlac">
    <xsl:for-each
select="mods:originInfo/mods:place/mods:placeTerm[@type='text']">
        <prodPlac xmlns="http://www.icpsr.umich.edu/DDI">
            <xsl:apply-templates/>
        </prodPlac>
    </xsl:for-each>
</xsl:template>

<xsl:template name="serStmt">
    <xsl:if test="mods:relatedItem[@type='series']">
        <serStmt xmlns="http://www.icpsr.umich.edu/DDI">
            <xsl:for-each select="mods:relatedItem[@type='series']">
                <serName>
                    <xsl:call-template name="titl"/>
                </serName>
            </xsl:for-each>
        </serStmt>
    </xsl:if>
</xsl:template>

```

```

    </xsl:if>
  </xsl:template>

  <xsl:template name="verStmt">
    <xsl:if test="mods:originInfo/mods:edition">
      <verStmt xmlns="http://www.icpsr.umich.edu/DDI">
        <version>
          <xsl:attribute name="type">edition</xsl:attribute>
          <xsl:apply-templates
select="mods:originInfo/mods:edition"/>
        </version>
      </verStmt>
    </xsl:if>
  </xsl:template>

  <xsl:template name="topcClas">
    <xsl:for-each select="mods:subject[@authority='lcsh']">
      <topcClas xmlns="http://www.icpsr.umich.edu/DDI" vocab="LOC
Subject Headings" vocabURI="http://www.loc.gov/catdir/cpsol/lcco/lcco.html">
        <xsl:for-each select="*">
          <xsl:if test="position()=1">
            <xsl:text> -- </xsl:text>
          </xsl:if>
          <xsl:apply-templates/>
        </xsl:for-each>
      </topcClas>
    </xsl:for-each>
  </xsl:template>

  <xsl:template name="abstract">
    <xsl:for-each select="mods:abstract">
      <abstract xmlns="http://www.icpsr.umich.edu/DDI">
        <xsl:apply-templates/>
      </abstract>
    </xsl:for-each>
  </xsl:template>

  <xsl:template name="timePrd">
    <xsl:choose>
      <xsl:when test="not(mods:subject/mods:temporal)">
        <timePrd xmlns="http://www.icpsr.umich.edu/DDI">
          <xsl:attribute name="event">
            <xsl:text>single</xsl:text>
          </xsl:attribute>
          <xsl:attribute name="date">
            <xsl:text>YYYY-MM-DD</xsl:text>
          </xsl:attribute>
          <xsl:text>temporal coverage here</xsl:text>
        </timePrd>
      </xsl:when>
      <xsl:otherwise>
        <xsl:for-each select="mods:subject/mods:temporal">
          <timePrd xmlns="http://www.icpsr.umich.edu/DDI">
            <xsl:attribute name="event">
              <xsl:choose>
                <xsl:when test="@point">
                  <xsl:value-of select="@point"/>

```

```

        </xsl:when>
        <xsl:otherwise>
            <xsl:text>single</xsl:text>
        </xsl:otherwise>
    </xsl:choose>
</xsl:attribute>
<xsl:attribute name="date">
    <xsl:choose>
        <xsl:when test="@encoding='w3cdtf'">
            <xsl:value-of select="node()"/>
        </xsl:when>
        <xsl:when test="@encoding='iso8601'">
            <xsl:if test="string-length(node())=8">
                <xsl:value-of
select="concat(substring(node(),1,4),'-',substring(node(),5,2),'-
',substring(node(),7,2))"/>
            </xsl:if>
            <xsl:if test="string-length(node())=6">
                <xsl:value-of
select="concat(substring(node(),1,4),'-',substring(node(),5,2))"/>
            </xsl:if>
            <xsl:if test="string-length(node())=4">
                <xsl:value-of select="node()"/>
            </xsl:if>
        </xsl:when>
        <xsl:when test="not(@encoding)">
            <xsl:text>YYYY-MM-DD</xsl:text>
        </xsl:when>
    </xsl:choose>
</xsl:attribute>
<xsl:apply-templates/>
</timePrd>
</xsl:for-each>
</xsl:otherwise>
</xsl:choose>
</xsl:template>

<xsl:template name="nation">
    <xsl:for-each
select="mods:subject/mods:geographicCode[@authority='iso3166']">
        <nation xmlns="http://www.icpsr.umich.edu/DDI">
            <xsl:attribute name="abbr">
                <xsl:value-of select="@*" />
            </xsl:attribute>
            <xsl:apply-templates/>
        </nation>
    </xsl:for-each>
    <xsl:for-each
select="mods:subject/mods:hierarchicalGeographic/mods:country">
        <nation xmlns="http://www.icpsr.umich.edu/DDI">
            <xsl:apply-templates/>
        </nation>
    </xsl:for-each>
</xsl:template>

<xsl:template name="geogCover">
    <xsl:if test="mods:subject/mods:geographic">

```

```
        <geogCover xmlns="http://www.icpsr.umich.edu/DDI">
            <xsl:value-of select="mods:subject/mods:geographic"/>
        </geogCover>
    </xsl:if>
</xsl:template>

</xsl:stylesheet>
```