
Chapter 4.

RECOVERY OF HISTORICAL U.S. CENSUS BUREAU MICRODATA: SUCCESS TO DATE

B.K. Atrostic, Randy Becker, Todd Gardner, Cheryl Grim, and Mark Mildorf, Center for Economic Studies

Additional years of microdata from the Annual Survey of Manufactures (ASM), Survey of Industrial Research and Development (SIRD), and the Current Population Survey are just a few examples of the valuable data recently recovered by the Center for Economic Studies

(CES). See Text Box 4-1 for highlights of recovered data.

Historic data from over 2,500 tapes were recovered before the Census Bureau's Unisys mainframe computer was decommissioned in 2010. In the 2008 Research Report, CES noted the

many household and business files to be recovered and sought the help of the research community (Becker and Grim, 2009 and Gardner, 2009). Without the intensive recovery effort led by CES during the last half of 2009, important information would have been lost forever.

Text Box 4-1.

RECOVERED DATA: HIGHLIGHTS

The economic and demographic data recovered from the Unisys will be valuable additions to the data already available at the Center for Economic Studies (CES) and the Research Data Centers (RDCs).

Most files will require additional work before they can be used for research purposes, and some may require approval by sponsoring agencies. Examples of data recovered from the Unisys include:

- Earlier years of series already available at CES
 - Censuses of Mining, Retail, Wholesale, and Services
 - Annual Survey of Manufactures (ASM)
 - Survey of Industrial Research and Development (SIRD)
 - Survey of Minority-Owned Business Enterprises
 - Commodity Transport Survey (now called the Commodity Flow Survey)
 - Decennial Census data
 - ~ Puerto Rico sample and complete count files
 - ~ U.S. Possessions sample and complete count files
 - Selected Current Population Survey (CPS) March Supplements
- Series not currently available at CES
 - Agriculture Surveys and Censuses
 - Annual Survey of Oil and Gas
 - Heating Fuel Survey
 - Water Use Survey
 - Survey of Construction
 - Income Surveys Development Program
 - CPS Supplements for months other than March
- New variables for series already available at CES
 - Census of Manufactures (CM) special inquiries data
 - ASM Central Administrative Office data
 - ASM and CM data flags
- Historical analysis files (see Text Box 4-2)
 - Industrial Time Series data
 - Linked CM/SIRD data created for Zvi Griliches



Photo by Lauren Brenner

Occupying 336 square feet, the Unisys IX Clearpath 4400 and 6 tape drives were purchased by the Census Bureau in 1995.

The recovered files have the potential to bring new data series to CES, extend existing data series at CES by a decade or more, and fill gaps in existing data. This potential has already

been proven for the ASM and SIRD. See Text Box 4-2.

The 2009 recovery effort built on work that had been going on at a much lower intensity for

nearly two decades. See Text Box 4-3.

Innovative on-the-fly solutions to problems that hampered earlier data recovery efforts let CES surmount a series of technical, operational, and administrative challenges that each threatened to halt the recovery in mid-process.

While CES led the recovery effort, success on this scale required help from many partners. The project got off the ground because of the strong support of C. Harvey Monk, Jr., then Associate Director for Economic Programs at the Census Bureau.

We also appreciate the support of the Census Bureau's Computer Services Division. They granted us additional time to access the Unisys machine and graciously allowed us to use their office facilities.

Text Box 4-2.

NEWLY AVAILABLE HISTORIC DATA: THE MANUFACTURING SECTOR

The data recovery project has already borne fruit. Annual Survey of Manufactures (ASM) data dating as far back as 1954 have been recovered, and as many as 15 years of additional historical data from the Survey of Industrial Research and Development (SIRD) have been recovered for large firms.¹

The oldest microdata available to researchers at the Center for Economic Studies (CES) and in the Research Data Centers (RDCs) as of December 2009 were manufacturing data from the 1963 Census of Manufactures (CM). Extending the

manufacturing microdata series provides opportunities to study the evolution and behavior of plants and industries over longer time periods and more business cycles.

Industrial Time Series Data

In the 1960s, Census Bureau staff created a longitudinal microdata file using the ASM. Internally, the file was referred to as the Industrial Time Series (ITS). They began with a pilot study using data from 1954–1961. They selected 25 industries containing about 2,500 establishments (Conklin, 1964).²

¹ The recovered data have not been fully converted into research-ready format nor have the data gone through exhaustive checks for completeness.

² Most of the selected industries were capital-intensive and many were primary metals industries.

Continued on page 21

Text Box 4-2.—Con.

Research papers using the data from the pilot study show early evidence of the value of longitudinally linked microdata. Jordan (1965) studies the measurement of capacity utilization, and Schaffer (1968) looks at changes in the structure of manufacturing employment.

CES recovered not only the original pilot study data, but also what we believe to be the 1954–1964 linked ASM data covering all manufacturing industries discussed in Kallek (1982).

The ITS data were abandoned when a new, and ultimately very successful, project to link manufacturing data began in the early 1980s (Kallek, 1982). This project was led by Nancy Ruggles and Richard Ruggles and developed the Longitudinal Establishment Database, a predecessor to the manufacturing microdata available at CES today (Atrostic, 2009).

The ITS data were not included in the 1980s project because of anticipated difficulties in linking to the more recent 1972 data. Given the computing capabilities and other data sources available now, there is a good probability these data can be linked to the currently available manufacturing data at CES.³

Griliches R&D Data

In the mid-1960s, Zvi Griliches—famed Harvard University economist, then at the University of Chicago—was approached by the Census Bureau and the National Science Foundation to work on a project analyzing historical data on industrial R&D.

Griliches led efforts to develop a longitudinally-linked dataset containing 1957–1965 company-level data from the Survey of Industrial Research and Development (SIRD) matched to data from the 1958 and 1963 CM and Enterprise Statistics (ES). Griliches (1980) details this effort and discusses limitations of the data. Using these

linked data, he found a positive relationship between investments in R&D and company productivity.

Griliches returned to the Census Bureau to extend his work in the early 1980s. Unfortunately, the data from his original project had been lost (Griliches, 1986).

With Census Bureau staff doing the hands-on data work, Griliches designed a new dataset. The new dataset contained 1957–1977 firm-level data from the SIRD matched to data from the 1962, 1967, and 1977 ES and data from the 1967 and 1972 CM. The sample was limited to certainty companies from the 1972 SIRD and contains approximately 1,100 companies.

Griliches highlights three findings in his 1986 paper: (1) investments in R&D make a positive contribution to productivity growth; (2) basic research is a more important contributor to productivity than other types of R&D; and (3) privately-financed R&D expenditures appear to be more effective at the company-level than federally-financed R&D expenditures.

The data underlying the findings in Griliches (1986) were recently recovered from the Unisys machine as part of the historical data recovery project.⁴ CES currently has data from the SIRD going back to 1972. The addition of the Griliches sample extends our historical information for many large R&D companies back to 1957.

While the Griliches sample is limited to large companies and key SIRD variables (e.g., total R&D expenditures, number of scientists and engineers), up to 15 additional years of data will now be available for many large R&D companies. The recovered data have been converted into SAS datasets. Foster and Grim (2010) use these newly recovered data to extend the period covered in their analysis of research and development back to the 1950s.

³ See Becker and Grim (2009) for more information on other sources of available historical manufacturing microdata.

⁴ The recovered data do not include the 1977 ES data.

Text Box 4-3.

RECOVERING HISTORICAL MICRODATA: A LONG HISTORY

Before 1998: Microdata recovered from Unisys computers were the source of major CES products, such as the original Longitudinal Research Database (LRD) (Atrostic, 2009 and McGuckin and Pascoe, 1988).

1998–2007: Recovering files became harder. Staff with the necessary skills and institutional memory retired. The Unisys machine became increasingly fragile, prone to crashes, and difficult to bring back on line.

Important data from several censuses and surveys were nevertheless retrieved and some made available for research. Among the data recovered were microdata for the 1960–1990 decennial censuses (Gardner, 2009); the 1955, 1956, and 1965–1971 Annual Survey of Manufactures (Becker and Grim, 2009); and the 1973–1978 Pollution Abatement Costs and Expenditures survey (Becker, 2007).

But moving the thousands of files remaining on the Unisys remained a low priority. CES would need to identify which of those files existed nowhere else, and only one person at CES had extensive experience on 1100/2200 Unisys mainframes.

2008–2009: When the Census Bureau announced the Unisys machine would be

decommissioned in late 2008, CES was able to delay the decommissioning.

2009: CES reached out to the research community. The research potential of microdata remaining on the Unisys were described in a presenta-

Over a decade ago, then-CES researcher, Al Nucci (1998) wrote of the significant technical and institutional challenges in moving historical microdata to modern computing systems:

“In principle, it ought to be a relatively straightforward exercise to obtain these data (i.e., a copy from archival tapes). This is not the case. The Census Bureau’s legacy mainframe is a nonstandard UNISYS (a descendant of the earlier UNIVAC computers). Further, the Census Bureau enhanced the capabilities of these computers with a proprietary file system. Hence the migration of data from earlier years not only has the more traditional problem of missing or incomplete file documentation and deteriorating tape files but requires the use of specialized software, often specially written for each file.

Further, the skills required for these tasks are no longer common at the [Census] Bureau and becoming increasingly rare as time progresses, in addition to the continuing diminution of Census Bureau institutional memory on the contents of earlier files and at their creation.”

tion by B.K. Atrostic at the National Bureau of Economic Research Productivity Program meeting in March 2009, and articles by Becker and Grim, and Gardner in the 2008 CES Research Report. CES actively solicited the views of our Research Data Center partners and, through them, the broader research community. Strong

positive feedback and concrete support gave CES efforts new life.



Photo by Lauren Brenner

One of many drawers full of paper data storage register files, which contain vital information for the recovery of data files stored on the Unisys.

CES assembled team members from a number of professions and divisions across the Census Bureau. Partners from the Research Data Center (RDC) system and the broader research community also joined the data recovery team. See Text Box 4-4.

Recovery is only the first step in creating usable research microdata files. CES programmers are working to develop processes that should greatly streamline the last steps of converting the data from legacy formats and character sets to more readily accessible formats, such as ASCII or SAS datasets. After the conversion is completed, substantial work will remain to assess the

data and make them consistent with data already at CES.

Decommissioning the Unisys closed a long and significant chapter in Census Bureau computing history. Some of the recovered files were likely initially processed on the Census Bureau's UNIVAC I computer, first used by the Census Bureau in 1951. See Text Box 4-5.

DATA RECOVERY PROCESS

The data recovery team overcame significant challenges to recover data from the Unisys:

- Identifying data to be recovered
- Moving data from the Unisys
- Keeping the Unisys running

Identifying Data to be Recovered

The first challenge was to create a list of "rescue worthy" data files on the Unisys.

Information about data stored on the Unisys, such as data storage forms, record layouts, and lists of computer tapes for each "data storage submission" are stored in a paper file called a "register." A data storage register may represent an entire economic census with scores of tapes holding several different files and millions of records, or part of a survey with a single tape holding a single file with only a few hundred records.

Text Box 4-4.

THE DATA RECOVERY TEAM

Center for Economic Studies

B.K. Atrostic
Randy Becker
Jason Chancellor
Joshua Coates (intern)
Henry Cross (intern)
Todd Gardner
Cheryl Grim
Mark Mildorf
Rose Taylor
Ya Jiun Tsai

Demographic Surveys Division

Zelda McBride
Sue Peters

Economic Statistical Methods and Programming Division

Connie Christensen
Stephen Jarvis
Keith Paterno
Robert Penrod
Daniel Vacca

Population Division

Marie Pees

Research Data Center Partner Support

Baruch College, CUNY System
University of California, Berkeley
University of Michigan Interuniversity
Consortium for Political and Social Research
University of Minnesota Population Center



Photo by Lauren Brenner

Center for Economic Studies (CES) Assistant Division Chief for Research Support, Mark Mildorf, with the Unisys in January 2010. As the only CES member of the data recovery team with prior experience on a Unisys, Mark led the charge to download data.

Text Box 4-5.

THE UNISYS: END OF COMPUTING ERA THAT BEGAN WITH UNIVAC I

Decommissioning the Census Bureau's Unisys Clearpath 4400 Mainframe will end a thread of computer usage reaching back to the 1950s. The Census Bureau purchased UNIVAC I, serial number 001, in 1951. It was the first commercial electronic general-purpose data processing computer.

However, the lineage of the Unisys Clearpath is better traced back to the UNIVAC 1100 series of mainframe, specifically the UNIVAC 1105 and 1107 models.

In 1958, the Census Bureau acquired the first two UNIVAC 1105 computers ever sold. In 1963, the Census Bureau purchased two UNIVAC 1107 computers.¹

Since then, the Census Bureau has had one or more 1100 or 2200 series mainframes in near continual operation.

The Unisys Clearpath 4400 is in the same family line as the UNIVAC 1107, sharing hardware

characteristics (such as the 36 bit word) and using the same family of operating systems.

The Census Bureau's UNIVAC I was retired in 1964. The UNIVAC 1105s were retired in 1967, and the UNIVAC 1107s were replaced in 1971. The Unisys Clearpath 4400 is the last Unisys mainframe in operation at the Census Bureau. The Unisys Corporation discontinued support of the Clearpath as of December 31, 2009, and it was decommissioned in 2010.

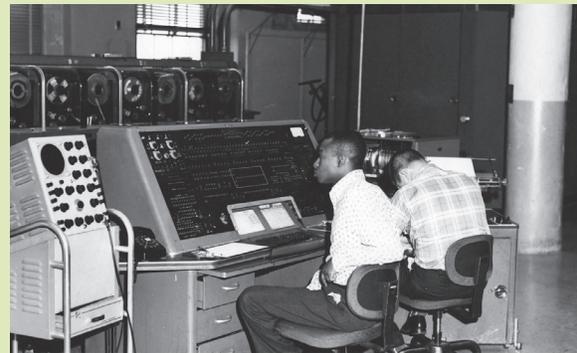


Photo by U.S. Census Bureau, Public Information Office

Built by Remington-Rand, using more than 5,600 computer vacuum tubes; 18,000 crystal diodes; and 300 relays, UNIVAC was the world's first commercial computer.

¹ For more information, see the CNN news article "50th Anniversary of the UNIVAC I," available at <archives.cnn.com/2001/TECH/industry/06/14/computing.anniversary/>.

CES researchers examined over 2,000 registers. Consulting with other subject-matter experts as needed, they set priorities for recovery and created electronic scans of hundreds of important registers.

Moving Data From the Unisys

After the data to be recovered were identified, much work remained. The next step was to move data from the Unisys to a modern computer system.

Moving data from the Unisys was a complicated process with many technical challenges. See Text Box 4-6.

The data had to be converted from a variety of formats to a format that could be copied from the Unisys and read on a modern computer. Initially, the unstructured data were read in variable-by-variable and loaded into SAS datasets. This procedure proved to be painstakingly slow.

The data recovery team made a major change to this procedure

in late October 2009 to emphasize transferring data from the Unisys as quickly as possible. The creation of research-ready SAS datasets was deferred.

This change, together with process improvements and increased staffing, led to a dramatic improvement in the flow of data from the Unisys. In the 6 months prior to the change in processing, the data recovery team moved economic data from approximately 20 registers from the Unisys. In the 4 months

following the change, the data recovery team moved data from over 490 economic registers.

Moving data files from the Unisys was a major production effort. Over 2,500 tapes were processed. The resulting 7,000-plus data files were then copied from the Unisys to Linux computers.

Keeping the Unisys Running

No account of the data recovery effort would be complete without noting the vital role played by Rose Taylor, a retired Census Bureau employee with over 25 years of experience with the Unisys.

Brought back as a contractor, Taylor was essential in keep-



Photo by Lauren Brenner

Rose Taylor, a retired Census employee and key member of the data recovery team, with the Unisys in January 2010.

Text Box 4-6.

TECHNICAL CHALLENGES TO THE DATA RECOVERY EFFORT

Listed below is a small subset of the technical issues that made the data recovery effort such a challenging exercise.

1. *The Unisys operating environment.* Expertise with the Unisys 1100/2200 series is now rare at the Census Bureau. Most of the data recovery work was done by people who had never previously logged on to a Unisys. Of the Center for Economic Studies staff involved with the recovery effort, only one had prior experience on a Unisys, and that experience ended in 1982.
2. *Legacy character sets.* Although some of the data recovered were in the ASCII character set, the vast majority of the data were in the less familiar FIELDATA character set. Developed in the 1950s by the Department of Defense, the FIELDATA character set was only used widely in commercial computing on the Unisys 1100/2200 series. EBCDIC data and data in the Excess-3 character set, a character set not widely used since the 1960s, were also recovered.
3. *Completely unstructured data.* All of the data recovered were stored in files that did not enforce any data structure. To enable users to properly interpret the data, hardcopy record layouts—road maps of the data—were prepared when these files were initially created. In many cases, the record layouts were lost, requiring the data recovery team to develop and implement new methods to move the data off of the Unisys.
4. *Unique record and file formats.* Much of the data were stored in CENSus Input Output (CENIO) files. CENIO is a unique Census Bureau-developed file structure. Data were also retrieved from “Algol Direct Access files” and a “save file” from a System 2000 database.

ing the Unisys running and recovering data.

With Taylor, CES regained expertise in recovering and restarting the machine following frequent system crashes and in managing the flow of computing jobs to avoid conditions leading to crashes. She also understood how programmers had stored files on the Unisys. She solved many puzzles that would have slowed or completely stymied the recovery.

Without Taylor's expertise, it is likely that CES would only have been able to retrieve a handful of files, rather than thousands.

Role of RDC Partners

Help from our RDC partners was also very important to the success of the data recovery project. Our partners at the Berkeley, Michigan, Minnesota, and New York (Baruch) RDCs completed important groundwork to speed the process of converting data pulled from the Unisys into SAS datasets. Our Michigan and Minnesota RDC partners were also instrumental in acquiring the services of Rose Taylor.

MOVING FORWARD

CES needs partners to convert the recovered data files into formats useful for research. Data from hundreds of registers have been successfully copied

from the Unisys. However, as of March 2010, relatively little of the data from the recovered registers has been converted into SAS datasets. Further processing is required to convert the remainder of the recovered data to modern formats.

Record layouts are not available for some data files. These cases will require additional analysis to reconstruct the record layout. File conversion is potentially a painstaking process, but the end result should be to make detailed data available to analyze important economic and social issues.

Researchers interested in partnering with CES to develop these data should contact the authors or send an e-mail to <CES.Data.Recovery.List@census.gov>.

REFERENCES

Atrostic, B.K. 2009. "The Center for Economic Studies 1982–2007: A Brief History." Center for Economic Studies Discussion Paper CES-WP-09-35.

Becker, Randy. 2007. "New Developments With the Pollution Abatement Costs and Expenditures (PACE) Survey." In *Research at the Center for Economic Studies and the Research Data Centers: 2006*. <www.ces.census.gov>

Becker, Randy and Cheryl Grim. 2009. "Recovering Historical Manufacturing Microdata." In *2008 Research Report: Center for Economic Studies and Research Data Centers*. <www.ces.census.gov>

Conklin, Maxwell R. 1964. "Time Series for Individual Plants from the Annual Survey of Manufactures and Related Data." Presented at a Joint Session of the American Economic Association and the American Statistical Association on December 30, 1964.

Foster, Lucia, and Cheryl Grim. 2010. "Characteristics of the Top R&D Performing Firms in the U.S.: Evidence from the Survey of Industrial R&D." Center for Economic Studies Discussion Paper CES-WP-10-33.

Gardner, Todd. 2009. "Expanded and Enhanced Decennial Census Data for Research." In *2008 Research Report: Center for Economic Studies and Research Data Centers*. <www.ces.census.gov>

Griliches, Zvi. 1980. "Returns to Research and Development Expenditures in the Private Sector." In *New Developments in Productivity Measurement and Analysis*, ed. John W. Kendrick and Beatrice N. Vaccara, 419–462. Chicago and London: NBER/University of Chicago Press.

-
- Griliches, Zvi. 1986. "Productivity, R&D, and Basic Research at the Firm Level in the 1970s." *American Economic Review*, 76(1): 141–154.
- Jordan, Willis K. 1965. "The Measurement of Performance Potential in Manufacturing Establishments." Bureau of the Census Working Paper No. 18.
- Kallek, Shirley. 1982. "Objectives and Framework." In *Workshop on the Development and Use of Longitudinal Establishment Data*. U.S. Government Printing Office, Washington, DC.
- McGuckin, Robert H. and George A. Pascoe Jr. 1988. "The Longitudinal Research Database (LRD): Status and Possibilities." *Survey of Current Business*, 68(11): 30–37.
- Nucci, Alfred R. 1998. "The Center for Economic Studies Program to Assemble Economic Census Establishment Information." *Business and Economic History*, 27(1): 249–256.
- Schaffer, William A. 1968. "Changes in the Structure of Manufacturing Employment." Bureau of the Census Working Paper No. 26.

This document is an extract from "2009 Research Report: Center for Economic Studies and Research Data Centers."



The full report is available at <www.census.gov/ces>.

This report has not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau or other organizations. All results have been reviewed to ensure that no confidential information is disclosed.