



Discovery and Recovery of

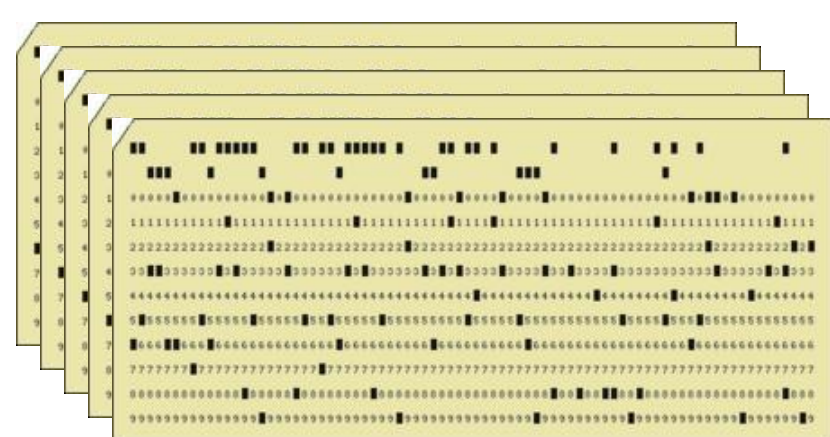
Punched Card Data

Marc Maynard, The Roper Center

Introduction

Data-PASS is a broad-based partnership of six major social science archives in the US, led by the ICPSR. The main goals of the partnership are to identify, locate, acquire and preserve 'at-risk' digital social science data.

While much 'born-digital' data have been acquired through the Data-PASS partnership, from the early stages, it was clear that eventually data stored on punched cards would be located and require recovery efforts. In preparation for this, the ICPSR purchased an IBM card reader. The Roper Center, working with National Opinion Research Center staff, identified nearly 1,200 boxes of punched cards.



Background on Two Case Studies

Cornell Study of Occupational Retirement, 1952-1958

Funded by: Lily Endowment, the National Institute of Mental Health, and the U.S. Public Health Service

Principle Investigator: Gordon Streib

Purpose: Follow the transition from work to retirement covering topics including family, daily activities, work, economic status, pensions, age identity, age stereotypes, retirement plans, health, life satisfaction and adjustment to the retirement transition

Sample: 4,032 workers aged 64 years in 1952 from 259 companies selected across the range of industrial classifications. Four follow-up interviews were conducted in 1954, 1955-56, 1957 and 1958.

Unique Features:

- Gender.** Men and women included.
- Longitudinal Design.** Most prior retirement studies were cross-sectional.
- Health Information.** Medical directors at the companies were interviewed and medical records on the respondents have been retained.
- Food Diaries.** In 1952, but not repeated in the follow-up interviews, 24-hour food diaries were kept by 1,439 men and 201 women in the sample.

Attitudes, Information and Customary Behavior in Health Matters, 1955

(NORC Study 367)

Funded by Health Information Foundation (HIF) and the American Cancer Society.

Principle Investigators: Jacob J. Feldman and Paul B. Sheatsley.

Purpose: To obtain a picture of people's knowledge, attitudes and experiences with respect to health and illness and to a broad range of medical personnel and facilities. Additional questions were asked regarding the Salk Vaccine.

Sample: intensive personal interviews with a national probability sample of 2,379 adults.

Unique Features:

- The HIF/NORC studies were among the first group of surveys to provide data for researchers to examine the federal role in financing health insurance, health care facilities, health care costs, etc.

Discovery

Cards for the Cornell Study were located in the Helen Newberry Library in Chicago, Illinois.

The HIF Study cards were found amidst thousands of boxes in a Chicago warehouse storage area maintained by NORC. Warehouse staff located hundreds of punched card boxes, extracting some 25 traced to this particular 1955 study.



Data Recovery – Methodology

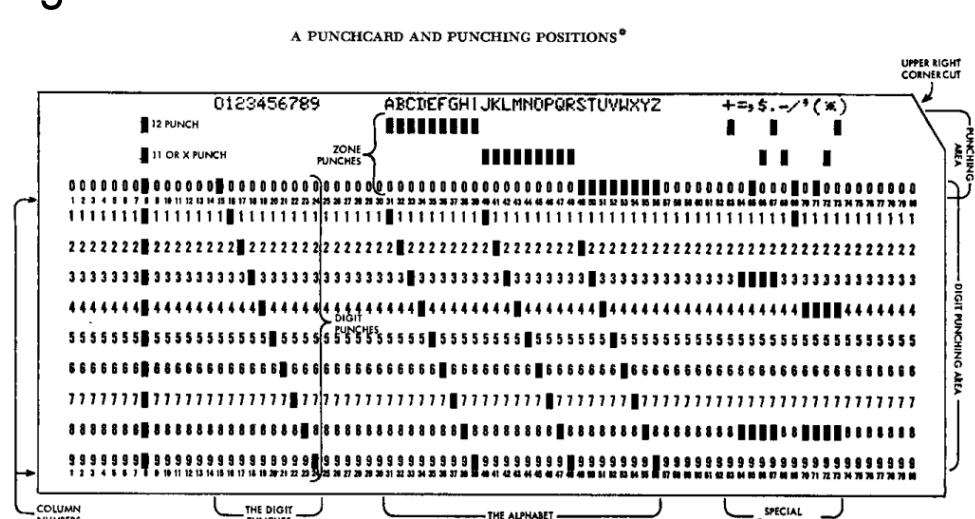
To successfully read the punched cards the Roper Center and ICPSR worked together bringing appropriate equipment and expertise to the project.

- Used a base set of test punched cards to determine output expectations
- Manually "read" the card data to verify the machine read
- Reviewed basic study documentation for key indicators regarding data structure and volume
- Evaluated condition of punched cards box by box (each box holds up to 3,000 cards)
- Attempted to minimize the number of passes (reads) of each card
- "Clean" raw data to enforce the 80-column fixed record length
- Convert raw data to parse multipunched columns and reassign values to new column locations

Punched Cards 101

Typical punch cards have 80 columns, each of the columns contain 12 punch positions (0-9, plus 12 and 11 on the top). Data is entered on the card by punching holes at fixed column and row positions. The punched holes can then be "read" by a card reader and interpreted as numbers, letters, and symbols (see Figure 2).

Figure 1. Illustrated Punched Card



One unique feature of punch cards is their ability to store multiple 'bits' of information in the same column. Some of these combinations form unique values like letters and symbols, but they can also represent multiple, unrelated, discrete values, which is true in our case.

Recovery – Phase 1

Refreshing and Relearning "Old" Skill Sets

1) Re-orient staff with the mechanical nature of the activity and idiosyncrasies of a punch card reader:

- Learning what the lights and buttons meant
- Listening to the vacuum element and rubberized conveyor belt for processing clues
- Monitoring read activity on the monitor
- React to misreads, misfeeds, skipped cards, etc



2) Identify sources of error and define standard methods for addressing them

- **PICK ERROR** -- means that the next card was not 'picked up' properly by the vacuum suction and conveyor belt.
Typical Solution: Identify the last successfully read card, and reset the next few cards, then hit 'RESET' on the card reader. If the same error occurs on restart, there is probably a problem with the physical condition of the card (concaved, bent in such a way that the card cannot be physically grabbed by the machine). Inspect the card, gently reforming it to be picked up by the reader.
- **READ ERROR** -- means that the last card(s) were not read in properly even though they were moved from the input bin to the output bin.
Typical Solution: You have to deal with this in a careful way since this is the main source of skipped/missing cards. Find the last successfully read card and place the rest (in order) back in the input bin. Then hit 'RESET' to read the cards starting from the one after the last successful read.

3) Define the appropriate output format

- Several output formats are support by the WinXread PC interface. In our case the output format had to handle multipunched column data, since a straight column for column ASCII translation would result in missing/deleted data values.

Recovery – Phase 2

Once the card data had been read into a standard ASCII text format (see Figure 2), the multi-punched columns needed to be identified and 'sprayed' out to new column locations. This process required the identification of the column, parsing of the multiple values and reassignment of each value to a new column location (which were serially added to the end of each record). Conversion of values 10, 11 and 12 to 0, X, Y was also carried out.

Figure 2. Raw Data Format from Initial Punch Card Read

```
10/10/10/11/13/ 13/5/10/9/12/ 1 12/ 1 2/2/6/6/2/2/6/5/11/2/4/1/5/1 .....1/4/2/5/12/3/10/ 1 1 1 1 1367/2/
10/10/10/2/1/4/ 13/4/9/9/12/ 1 11/1/7/1/1/5/5/2/1/6/5/11/2/7/1/5/8/ .....1/4/2/5/16/12/3/10/ 1 1 1 1 1367/2/
10/10/10/3/1/13/ 12/4/9/9/12/ 1 11/1/10/1/1/8/6/3/2/5/6/11/2/9/2/4/1 .....1/4/2/5/18/11/5/10/ 1 1 1 1 1367/2/
```

A Perl script was developed to handle these requirements as well as, check for deviations from the expected record length (80) and unexpected characters.

Finally, the punched card data was output in an accessible comma-delimited format that included a first record list of target column numbers to cross-reference with the original questionnaires and documentation (see Figure 3).

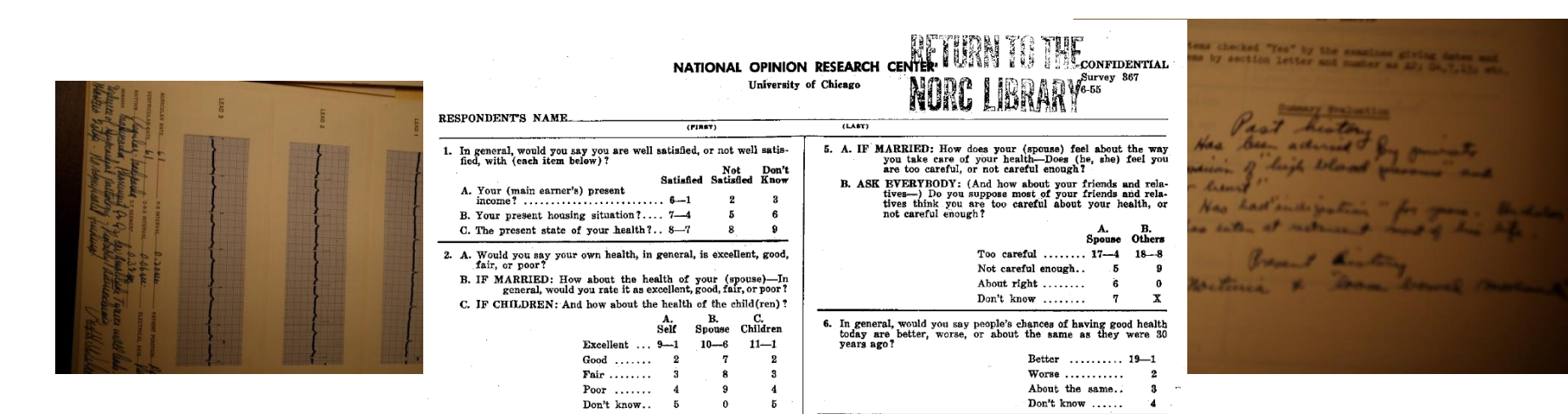
Figure 3. Final Output

```
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31, .....66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,70a,70b,71a,71b,79a,79b,79c
0,0,0,1,1,3,,3,5,0,9,Y,,Y,,2,2,6,6,2,2,6,5,X,2,4,1,5, .....1,4,2,5,,,,,2,5,Y,3,0,3,6,7
0,0,0,2,1,4,,3,4,9,9,Y,,X,,7,1,1,5,5,2,1,6,5,X,2,7,1,5,8, .....1,4,2,5,,,,,2,6,Y,3,0,3,6,7
0,0,0,3,1,3,,2,4,9,9,Y,,X,,0,1,1,8,6,3,2,5,6,X,2,9,2,4, .....1,4,2,5,,,,,2,8,X,5,0,3,6,7
```

On-going Work

The future for both recovery efforts have focused on two main areas:

1. Continue reading and processing all remaining punched cards
2. Inventory, organize and preserve all supporting materials, including documentation, reports, correspondence, etc.



References

Berk, Marc L., Claudia L. Schur, and Jacob Feldman, "Twenty-Five Years Of Health Surveys: Does More Data Mean Better Data?" *Health Affairs*, November/December 2007; 26(6): 1599-1611. retrieved 6/30/2010: <http://content.healthaffairs.org/cgi/content/full/26/6/1599>.

Janda, Kenneth, *Data Processing: Applications to Political Research*. Evanston, Ill.: Northwestern University Press, 1965.

Acknowledgements

Thanks to Bill Hanselman and Jared Lyle of ICPSR and Lois Timms-Ferrara and Cindy Teixeira of the Roper Center for their continued work on this project.

Contact information

Marc Maynard, The Roper Center
University of Connecticut
marc.maynard@uconn.edu