TAPE STORY TAPESTRY: HISTORICAL RESEARCH WITH INACCESSIBLE DIGITAL INFORMATION TECHNOLOGIES

SHANE GREENSTEIN

ABSTRACT: Shane Greenstein's essay describes the author's search for computer tapes inventorying the federal government's computer equipment. The difficulties of the hunt illustrate an ironic consequence of technical change: digital information technologies make it incredibly easy to destroy what could not have been gathered without its invention in the first place. The essay observes that many of the control and access mechanisms taken for granted with well-established storage media do not exist for machine-readable data.

Bruce Bruemmer's commentary reflects upon the meaning of Greenstein's experience for archivists. Bruemmer explores how Greenstein views the "keepers" of information and the implication of that view for archivists.

Digital technology has lowered the cost of storing data and increased the amount of research data available, but digital technologies alone are not sufficient to make this data useful. In the first place, stored data has to be inherently "useful" for answering a worthwhile research question. Second, just because information is less costly to store does not necessarily make it more accessible, especially for future research efforts.

This essay has two simple observations. First, if machine-readable data is to be made accessible to future generations, then it must be systematically preserved and organized for use. We take these measures for granted when dealing with other storage media, but seldom have they been applied to machine-readable data. Second, digital technology makes it incredibly easy to destroy what could not have been gathered without its invention in the first place.

The result is that much useful information stored in machine-readable form is being lost to future generations of researchers. Often when data's original purpose is fulfilled, the user has little incentive to protect the data for historical purposes. If and when data is placed in the public domain it is not organized for the benefit of historical researchers. Despite all the *formal mechanisms* designed to help researchers retrieve the more traditional original archival sources, researchers find that unsystematic *informal communication* is essential for retrieval of digital information that is not kept with historical interests in mind.

I personally experienced the consequences of these problems in the spring of 1987 and was fortunate enough to succeed despite them. In August of 1987 I received thirteen computer tapes in the mail. These were the last remaining copies of data recorded between 1971 and 1983.

The hunt for these tapes is quite a story, in which a number of improbable events lead to success—personal contacts, chance encounters, near misses, 11th-hour reprieves, and a maze of federal government agencies. Like any decent mystery, this story also has its share of heroines and heros, blind alleys and red herrings, and an ending full of lessons for the wise.

The lessons make the story worth telling. I found the tapes, but not by using a special index, not with the aid of some of the best reference librarians on the Stanford University campus, and not through a formal records management system. All those mechanisms failed because the data never was properly stored for future generations because its historical value was never recognized. I found the tapes because I was fortunate enough to contact a number of the people who had been associated with the creation and use of the data and who were kind enough to help with my search.

The Unwinding of the Tape Story

Since the late 1950s, when it was a relatively easy task, until the present, when it became nearly impossible, some federal government agency has done an annual inventory of the entire government's computer equipment. The rationale varied, and so did the format of the survey, but the basic task did not. Each year every federal agency and department completed a survey in which it itemized and categorized its automatic data processing (ADP) equipment. The oversight agency tabulated the response and published a summary. Each year's inventory is complete, carefully assembled, and informative.

These inventories contain a historical record of the purchasing and management patterns of the largest buyer of computer equipment in the United States. It is valuable evidence of the technological history of computers and the computer industry as a whole.

For example, they document the growth of total federal government holdings from 531 "processors" in 1960 to 18,474 in 1982. The inventories show the significance of the government's early involvement in purchasing computer equipment when the industry was young and technologically immature. The record of system configurations tracks the changing technological norms of computer equipment use in the country. Later inventories confirm the commonly held perception that the government's systems have aged, heavily burdened by older investments that were not replaced as rapidly as in private firms.

The inventories were recorded in two forms. Each year the complete inventory was recorded, updated, and disseminated on machine-readable magnetic tape. In addition, some of the information was published and distributed in book form to a limited number of buyers throughout the country, and to government depository libraries.

None of these details was known to me when this project began, but they are essential for understanding my efforts to retrieve this data. As is explained below, the books were relatively "easy" to find if you knew where to look. Much of my effort focused on finding copies of the data tapes, whose very existence was in doubt throughout my search.

My involvement with the government's use of computer equipment was serendipitous. I was a graduate student in economics at Stanford University, and had chosen to become a specialist on economic issues in the computer industry. I often found myself doing research in the Johnson Library of Government Documents in Stanford's main library. I had learned through frequent visits that it was home to some of the best reference librarians and reference materials on campus. Perhaps not by coincidence, it was also where most of the most important statistical tables were stored.

While wandering through the stacks one afternoon, I found some infrequently used softcover books, each about the size of a telephone directory. The binding cracked when each was opened, and it was still possible to smell the scent of ink on the pages.

The books were the published versions of the inventories of federal government computer equipment described above. They included not only aggregate summaries of government holdings, but detailed descriptions of each piece of equipment: its type, make, and manufacturer, and the office and geographic location at which each model was held. In two more days of systematic searching, I found that the library held not just one book, but an almost complete set of inventories from 1960 to 1983. (The series ended in 1983 because legislation changed the nature of data collection and halted the publication of the series). This data could support more than just one research paper, it could support an empirical research program for years.[1]

As the historical record began to grow, I wondered how it all got to Stanford's shelves. Stanford's role as a federal depository library partially explains it. All Government Printing Office publications are supposed to be deposited there.

The second part of the answer is what might be called "the necessity for assiduous archivists." There is no substitute for a librarian who diligently collects and saves every conceivable document, even to the point of filling the shelves with material of no present value. It must be done indiscriminately because it is not obvious today what the historian of tomorrow will find valuable and the collection must be organized so that individual documents can be found.

Everyone takes for granted that assiduous librarians serve a useful purpose when information is stored in books. They not only collect, but also organize and categorize those books. As this story will show, the age of machine-readable information has not made a place in data processing center libraries for people like Joan Loftus, government documents librarian for the Johnson Library. If this generation wants to preserve and organize computerized information for future generations, it will be foolish to not have her there in some capacity.

Close inspection of the printed data, especially the summary tables, led to my first disappointment. The published inventories were incomplete, containing only a fraction of the information originally collected in any given year—perhaps a third. The summaries made clear that the remainder of the inventory information had originally resided on a computer database at the government office that collected and assembled the inventories That office was the General Services Administration (GSA). Naturally, I called GSA to see if the original historical data could be made available.

I first spoke with GSA employees who had worked with the *current* (1986) version of this inventory data. They were generally helpful and friendly. They were accustomed to talking to computer firms who wanted information about current machine-readable files. Vendors use this information for marketing and sales efforts. GSA employees knew how to get the names and numbers of agency offices and the complete list of an office's *present* processor and peripheral holdings. Indeed, if I was ever willing to *buy this information,* no matter what I wanted to do with it, they would be delighted to provide it.

After our initial interaction, my questions quickly left the familiar territory and entered a zone of talk sans communication. I could imagine the eyes of the GSA liaison glaze over whenever I used the word *history.* "Historical data?" she said, "Why, the old inventories were thrown out when we moved offices." "Old data tapes?" she said, "Why, we just *update* our database every year. We don't keep track of any of the changes. We don't keep backups from past years. The government only needs the present year's data." And "What happens to old versions? No one knows. Why, most people have only worked in this office a few years."

Because government employees had no direct use for historical information, the current year's tape contained little data of pure historical relevance. The present data supplemented little of what was published in the older inventories. Data of historical interest was preserved only by accident.

Automation has exacerbated the problem. Computerized data tables are so easy to update, change, or delete. *Unless the table is designed to record explicit changes in the database* (an audit trail), researchers cannot usually infer from the current data anything about its past. This database of computer inventories was not designed to track changes. Virtually all the useful historical information had been eliminated because current usage did not require it and because it was so easy just to update and be done with it.

How is this problem usually solved? In the past, information was stored in books, and several years of books would allow researchers to isolate important changes across time. In this case, however, the books that had been kept were not quite up to the task. A complete print-out of the data tapes from previous years could have supplied historical information, but the printouts were incomplete, and thus inconclusive. No one had thought to send each year's tape to any federal depository (or its equivalent for machine-readable data).

Several other characteristics of the published books motivated me to search for the data tapes. In addition to the fact that what was published was less than 40% of all the information that was originally collected, the published form was expensive to analyze. How was the relevant data going to be entered into a computer for statistical analysis when the source documents were as thick as telephone books? The cost of putting the data in machine-readable form for statistical analysis was more than I could afford.

I pressed on in the search for copies of the data tapes and a sequence of seemingly improbable events began.

My investigation branched into many paths. One led to the International Data Corporation (IDC), a well-respected private firm that regularly surveys the computer industry. Maybe they purchased and kept old copies of the tapes? Since the data was noncurrent, I naively thought, why wouldn't they mind lending a copy to a purely academic project?

In pursuing this avenue, I wasted the time of a helpful, though anonymous, Washington IDC representative. In truth, the conversation wasted her time, but not mine. She said that if IDC kept such tapes, they would be at the head office in Framingham, Massachusetts. She gave me the number of the person who would know the answer. Oh, and incidentally, she suggested that I might get a good overview of some issues in the area from a recent publication of the (generally respected) Office of Technology Assessment (OTA), a congressional research office. They had just published a book with the odd title *Federal Government Information Technology: Management, Security, and Congressional Oversight.*[2]

The suggestion to contact the IDC office in Framingham turned out to be one of many false leads. "IDC no longer collects or sells such information," said the man in Massachusetts who had no time for me, the non paying supplicant. That answer should not have surprised me. People in the business of selling information about an industry have no reason to care about historical data unless there is a market for it. If there is, the information costs money.

In an elementary economics course, IDC would not be vilified for its practices. Instead, the preceding anecdote would be used to illustrate the conceptual difference between the social and private incentives firms face when providing a public good—historical data in this case. In the classroom, economists would say that the private incentives for collecting historical data were less compelling than the social incentives. In a classroom this observation would lead to a simple policy prescription: since no private firm is motivated to provide historical data that society values, then a benevolent government-sponsored agency should do so.

Unfortunately, I already knew that the simple prescription for a benevolent agency had not been followed. Because the inventory has value to vendors who sell to the government, federal employees directly responsible for it had adopted an attitude similar to IDC's. The government did not keep historical data because they could not regularly make money by selling it. Has the whole government lost its mind? I thought. What an attitude for the government employees to take! Fortunately, it would turn out that some government employees did have a bit more foresight.

The OTA publication with the odd title opened up another avenue to pursue. The booklet—it will come as no surprise that this was on the Stanford government documents library shelves—contained an extensive and current bibliography, with references to much of the important recent work, government and academic, written on federal automatic data processing (ADP). One citation caught my attention. A five-year old special report from the National Bureau of Standards included in its title the phrase *Federal ADP: A Compilation of Statistics.*[3]

The title held promise. I found the publication—on the government document library's shelves, of course. Sure enough, the author had used the GSA database extensively in compiling her statistics and comparing them with statistics on private firms' holdings.

At the time I had hoped that the writer, Martha Mulford Gray, could be helpful locating other copies of this data. In secret I hoped that she might just have the data, even though I suspected that such an outcome was unlikely. It would have been too easy, a departure from the pattern so far. The goal was to locate

Martha Gray, but how does one locate a federal government employee who worked at a particular division in the National Bureau of Standards (NBS) five years ago? I did the simple thing and called her old office.

Ms. Gray's old office was identified in the front of the publication. The federal telephone directories indicated the office still existed. I prepared to call all the numbers in that office until I reached someone who knew her. So I chose an arbitrary phone number and asked for her. "Oh, why you have the wrong number," the first voice in Gaithersburg, Maryland replied," her number is ———."

From the first time we talked to the last, Ms. Gray was the type of government contact every researcher dreams of finding. She knew almost everything about this database, and where to find the answer if she didn't have the answer. As it turned out, from 1977 to 1982 she had written four analyses of the compiled statistics, only the last of which had been cited in the OTA publication, and she had a good sense of the data's accuracy and limits.[4]

This would be the first of many times that personal recollection of the data would aid my understanding more than did printed matter, but future generations will not have the luxury of being able to talk to Martha Gray about this generation's computer tapes. This luxury demonstrates what might be called *the necessity of informal contacts.* There is no substitute for talking to someone who has lived through the event you are studying or who worked directly with the data you need. These people often understand the anomalies and idiosyncracies in their data. There is no substitute for someone who really knows what is going on.

The first time I inquired about the existence of the tapes, the answer from Maryland went something like this, "Hmm, we used to have some copies, but were going to get rid of them. Maybe we can just give them to you." But as I noted earlier, that would have been too easy. Sure enough, when we discussed it again the next week she said "I have some bad news. Those tapes were thrown out *last month.* They were running out of storage space in the basement, so they asked to get rid of them." Ms. Gray then realized that the tapes in the basement probably had been the last existing copy of the database in government custody.

What happened at NBS suggests that two factors affect the preservation of such data files. First, a person with a sense of value—sentimental, economic, or other—will preserve it. However, as soon as physical control of a dataset passes from the original user to someone else, the probability of destruction increases. Second, material is put in archives by people who think they are making history, not necessarily by people who quietly go about their business, changing the world in which we live. Those who value history can never start collecting data too soon, or making an effort to preserve material that will be of obvious use to future researchers.

Martha Gray took the episode as a lesson. She vowed to start sending materials to the National Archives or Charles Babbage Institute, an institution specializing in computer history. Perhaps that is something of a victory. One convert was made.

Ms. Gray generously sent me *her own printouts* of parts of the database for 1971 to 1979. Her printouts and the published sources together constituted about 50% of the original database for the 1970s. She also suggested contacting several private firms to see if they still had copies of tapes they had bought in the past.

Her idea looked like a long shot. Why should such a firm keep the old information after buying the next year's inventory? I followed through on the suggestion, despite my pessimism, because it was my last chance. I inquired of archivists and librarians at companies that had large interests in the inventory: Digital Equipment Corporation, the Department of Defense, IDC (again) —because one of my dissertation advisors had received information suggesting that they might have the data—and IBM, whose archives is located in Valhalla, New York. All unproductive.

The inquiry at IBM met with particular skepticism. How does one ask for this information over the phone from IBM, a corporation that learned from frequent litigation to play its cards close to its chest? One friend sarcastically suggested the following question "Excuse me, would you happen to have any information on a set of government tapes that might help me demonstrate that IBM had market power?"

I decided to pursue each path to the end, propelled, in retrospect, by what might best be described as irrational obsessiveness. After following several red herrings, I focused on one last possibility—that IBM really did have the tapes. Following this avenue to its logical end led to the most incredible in an already improbable sequence of events.

The archivist in Valhalla gave me the number of the IBM library in Washington, an office run by Ms. Nan Farley. In the first of many conversations, Ms. Farley did for me all she ever did for me, though it was more than enough. In response to *THE TAPE* question, she replied that IBM had bought tapes of the government's inventory in the 1970s and had entered them as evidence in the antitrust trial (the justice department's suit against IBM began in 1969 and lasted until 1981, when the government withdrew its complaints). She knew this to be true, she continued, because she had been the representative for IBM who travelled annually to GSA to pick up the tapes. IBM had bought the tapes, she assured me, but it was anyone's guess what had happened to them after they were entered as trial exhibits.

The trial record involved literally hundreds of thousands of records, and, as a consequence, it is quite impossible to locate anything without exhibit numbers. Ms. Farley said she would try to find me IBM's exhibit number for the tapes. It was still a long shot, but why not try it, I thought. So I waited, and phoned her periodically to remind her not to forget me.

My patience ran out after a month of reminder phone calls. Perhaps there was another way to get that exhibit number, such as directly from accounts of the trial or from the 2nd District Court in New York. Perhaps the tapes were odd enough that someone familiar with the trial record would recognize them right away. At this stage of the search it was worth a shot, even though it was a long shot.

I first consulted the two computer industry histories that based much of their analysis on information in the U.S. v. IBM trial exhibits. Neither mentioned the Federal inventory tapes.[5] Then I called the federal court in New York.

After a few awkward calls and a few unbelievable days of uninterrupted busy signals, I finally found a clerk of the court who was familiar with the IBM record. She had one of those fascinating Brooklyn accents, with vowels somewhere between the throat and nasal passages, and I had a hard time concentrating on her words. She was definitely someone who had lived in New

York all her life and the confidence of her answer seemed to finish the search. The court had kept three rooms full of documents, she asserted, but when the trial was dismissed all nonpaper evidence was sent back to the lawyers. The court most certainly did not now have any tapes.

The search then shifted to the defendant's lawyers. The law firm of Cravath, Swain, and Moore had quite a reputation. This firm had successfully defended IBM against the government antitrust law suit.

As a general rule, firms that have been through multimillion dollar antitrust suits are usually reluctant to surrender their proprietary information. Finding the tapes would require literally getting inside the law offices and the corporation; however, there was no conceivable way a stranger like me would be able to call IBM's law firm and hope to get the kind of cooperative response that was needed to resolve the issue. At the end of the trail and burning to know whether these tapes still existed or had been destroyed, I took the actions of a desperate man. I contacted the *only man* in the nation who could possibly find those tapes—Dr. Frank Fisher.

Fisher is a well-known and widely respected professional economist and MIT professor who was IBM's chief economic witness in the antitrust trial. He prepared many of IBM's economic arguments and published the two books mentioned earlier, along with a series of articles inspired by the issues in the case. If there was anyone in the profession who could open a door at the law firm or at IBM, it had to be Frank Fisher.

How, though, does one go about asking a well-respected total stranger like Frank Fisher for a favor? It just so happened that Professor Fisher had attended Harvard as an undergraduate at the same time as Paul David, a Stanford economics professor who sat on my dissertation committee.

Professor David had frequently stated that he could ask his friend for help if IBM's cooperation was needed for something I was doing. When I finally reached the end of my options, it seemed appropriate to ask Paul David to ask Frank Fisher for a *BIG* favor.

Professor Fishers story is also interesting, involving several telephone calls through the law office and across divisions of the corporation to the litigation chief, Eugene Takahashi. The bottom line is this: Professor Fisher made a direct hit. IBM not only held copies of tapes for most of the years of interest (1967-1979 and 1983), but IBM was willing to donate copies of them to Stanford so I could do my research.

For all intents and purposes, the end was a happy one. The hunt was successful. I wrote my thesis using the data. In the world of research, this is equivalent to living happily ever after.[6]

A retrospective view of the situation inevitably leaves me dazed. The last existing evidence of the history of changes in the government's complete inventory of computer equipment did not reside with any government agency, with any government archive, or with any firm that specialized in selling information about that industry. Instead, it resided in the hands of a private firm that bought the tapes for an antitrust trial and its own internal use. That firm kept the tapes for no apparent reason, and was gracious enough to donate them for research after being prodded by an economist they hired for their own antitrust defense. He, in turn, just happened to be good friends with one of my thesis advisors.

I located and obtained those tapes not by using a special index, not by consulting some of the best reference librarians on Stanford campus, and not by

browsing in a formal library collection. Those conventional mechanisms failed because the information never was properly stored for future generations. I obtained the required documents because I contacted a number of the people associated with the development and use of that information and they were kind enough to help me.

All this leads to my two maxims concerning research with computerized information. The first I call *the necessity of complementary investments.* Just because data is less costly to store does not necessarily make it more accessible. If computerized data is to be accessible for historical research, digital storage itself is not enough. The digitally stored data must be systematically preserved and organized for use, We take these measures for granted when dealing with other storage media, but seldom have they been applied to machine-readable data.

The difficulties I encountered retrieving this data also lead to the second maxim, which I call *the necessity of informal contacts.* In a world dominated by electronic storage instead of books, informal and personal interaction still provide the background necessary to locate, assemble, and understand machine-readable information. Future generations may not have it so easy, because all the potential heroes of their hunts will be gone.

ABOUT THE AUTHOR: Shane Greenstein is presently assistant professor of economics at the University of Illinois at Urbana-Champaign. This article was written while he was a graduate student at Stanford University. He has written on procurement of mainframe computer systems by federal agencies and on the economic consequences from improvements in computer technologies.

NOTES

1. General Services Administration, *Automatic Data Processing Activities Summary in the United States Government, 1972-1982* (Washington, D.C.: GSA, 1986); and General Services Administration, *Automatic Data Processing Equipment Inventory, 1960-1983* (Washington, D.C.: GSA, 1986).
2. Office of Technology Assessment, *Federal Government Information Technology: Management, Security and Congressional Oversight* (Washington, D.C.: U.S. Congress, Office of Technology Assessment, 1986). I would like to thank Jody Greenstein, Sheryl Horowitz, Paul David, Bruce Bruemmer, and Frank Boles for useful comments on earlier drafts. All the participants in the events in this story naturally receive my deepest gratitude. The Center for Economic Policy Research, Stanford University, provided financial support for research connected with locating this data. The Charles Babbage Institute and the National Science Foundation provided support for related dissertation research.
3. Martha Mulford Gray, *Federal ADP Equipment: A Compilation of Statistics—1981* (Washington, D.C.: National Bureau of Standards, 1982).
4. In addition to *Federal ADP Equipment,* see also Martha M. Gray, *An Assessment and Forecast of ADP in the Federal Government* (Washington, D.C.: National Bureau of Standards, 1981); *Computers in the Federal Government: A Compilation of Statistics—1978* (Washington, D.C.: National Bureau of Standards, 1979); *Computers in the Federal Government: A Compilation of Statistics* (Washington, D.C.: National Bureau of Standards, 1977).
5. Franklin M. Fisher, John J. McGowan, and Joen E. Greenwood, *Folded, Spindled, and Mutilated: Economic Analysis and US vs. IBM* (Cambridge: MIT Press, 1983); and Franklin M. Fisher, James W. McKie, and Richard B. Mancke, *IBM and the US Data Processing Industry: An Economic History* (New York: Praeger Publishers, 1983).
6. Shane M. Greenstein, "Computers, Compatibility, and Economic Choice" (Ph.D dissertation, Department of Economics, Stanford University, 1989).