The History of Geographic Information Systems: Perspectives from the Pioneers

Timothy W. Foresman, editor
University of Maryland, Baltimore County

To join a Prentice Hall PTR Internet mailing list, point to
http://www.prenhall.com/mail_lists/

Prentice Hall PTR
Upper Saddle River, NJ 07458

CHAPTER 4

Topology and TIGER: The Census Bureau's Contribution

Donald F. Cooke

Introduction

The Census Bureau is conspicuously absent from the roster of chartered mapping agencies of the U.S. government. Nevertheless, it was the bureau and not the U.S. Geological Survey or Defense Mapping Agency which led in implementing topological data structures, developing street address geocoding, and building the first truly useful nationwide general-purpose spatial data set.

The Census Bureau never set out explicitly to accomplish this. How it did is a story of an agency performing its mission with exceptional concern for its constituency, in an ever-changing technological environment and affected at crucial junctures by individuals with extraordinary talents and temperaments. The story that follows describes how luck, timing, and temperament determined the evolution of key GIS technology much more than the orderly, step-by-step progress historians would like to relate. As Mark Twain said, "Of course, truth is stranger than fiction. Fiction, after all, has to make sense."

Background

Moore's law says that computer processing power doubles every 18 months: a 100-fold increase each decade. The U.S. Census Bureau, required by Article I of the Constitution to enumerate the

Donald Cooke is founder and president of Geographic Data Technology, Inc., a supplier of digital maps to commercial markets. Cooke was a member of the Census Bureau team that developed the Dual Independent Map Encoding (DIME) system in 1967. Besides his family, his interests include windsurfing, ice hockey, vintage sports cars, astronomy, and the learning process. *Author's Address:* Donald Cooke, Geographic Data Technology, Inc., 11 Lafayette St., Lebanon, NH 03766. E-mail: don_cooke@mail.gdt1.com.

country every ten years, is periodically torn between the need to keep census statistics comparable from decade to decade and the need to adapt to the alien data processing environment engendered each decade by two orders of magnitude growth of computer technology.

The bureau has always played an important role in the evolution of data processing. The most often-cited instance is the invention of a punch card tabulation system by the Census Bureau, which enabled the bureau to complete processing the 1890 Census before it had finished with the last non-automated (1880) one.

Though the bureau had bought the first non-military digital computer in 1950, the 1960 Census was still—to the outside world—primarily a paper operation. Census made extensive use of computers for internal processing in 1960, including designing and building FOSDIC (Film Optical Sensing Device for Input to Computers) to scan the 1960 questionnaires. But the bureau published 1960 statistics solely in printed tables. Only a handful of intrepid analysts requested 1960 Census results in machine-readable form; these requests were filled on punch-cards.

Preparations for the 1970 Census

For the past three decades, the Census Bureau has shown exemplary sensitivity and care for the needs of users of its data. Much of the intelligence about data user requirements was supplied by a panel of outsiders called the Census Small Area Data Advisory Committee.

This committee recommended formation of two groups in 1966: the Data Access and Use Laboratories (DAULabs), led by Jack Beresford at the bureau's Suitland, MD, headquarters, and the New Haven, CT, Census Use Study, located at the site of the April 1, 1967, dress-rehearsal test of new census enumeration procedures.

The first time that the country was enumerated primarily by mail was in 1970. The bureau planned to buy commercial mailing lists on computer tape, print mailing labels, and mail questionnaires to each household. People were to fill out the forms "in the comfort and privacy of their homes," then drop them in the mail to a regional processing center. Census had mailed out forms in 1960, but enumerators visited each home to pick them up, recording the location—Census Tract and Block—of each household on a map. The 1970 mail-out/mail-back plan promised to be more efficient, but its success hinged on being able to "geocode" each questionnaire to the appropriate census block *without* sending people into the field with maps.

Increasing computer power, geographical requirements of a new enumeration technique, and a sincere desire on the part of the bureau to serve data users better resulted in a permanent change in accessibility of census data on one hand and a vital contribution to GIS technology on the other.

Beresford's DAULabs group facilitated a smooth transition from the paper era of the 1960 Census to the machine-readable era of 1970 and beyond, by setting standards for tape files and documentation, establishing satellite Summary Tape Processing Centers to assist data users, and conducting hundreds of workshops and training seminars. DAULabs' efforts created today's demographic analysis industry. Companies like Claritas, National Planning Data, and Urban Decision Systems were among the first of Beresford's Summary Tape Processing Centers.

In contrast to Beresford's measured success, New Haven was to become the first battleground in a technical and management revolution that kept the census geographic operations in turmoil into the 1980s but ultimately led to creating a nationwide spatial data resource which anchors the nation's spatial data infrastructure and has spawned Business Geographics, currently the fastest-growing segment of GIS.

The New Haven Census Use Study in 1967

The Small Area Data Advisory Committee posed five challenges to the Census Use Study:

1. Could a useful transportation survey be run using a sample of census households, keyed so that the transportation data could be augmented by individual data from decennial census questionnaires?
2. Could the same be done with a health questionnaire?
3. What would it take to be able to generate *any* cross tabulation of census data, taking into account the need to preserve privacy of respondents? How useful would it be to local data users to have this capability?
4. What about matching two data files where the common element in each is a street address? Can this be done efficiently even with varying addressing conventions? Would the ability to geocode data sets by address matching be a useful function?
5. What about computer mapping? Census and local data are inherently spatial; could new computer graphics technologies be used to map the data?

The Census Bureau appointed Caby Smith as study director. Smith was a Mississippian who had joined the bureau as an entry-level typist. He had risen in the ranks and developed a reputation as a "can-do" manager, if a bit of a maverick. Perhaps this assignment would fit his temperament and keep him out of the way of the more sober statisticians and administrators who were busy gearing up for the 1970 enumeration.

Smith's deputy on the New Haven site was George Leyland, a recent Harvard graduate whose wife Mary was a key manager in an IBM-funded urban information system project at the City of New Haven. Smith and Leyland set about to staff the study, Smith bringing on Joyce Annecillo, a shrewd and experienced administrative assistant with skills to meet bureaucratic complexities to come. Leyland hired Bill Maxfield and the author, both of whom were finishing studies at Yale, to work on mapping and special tabulations and engaged Jack Sweeney of Cambridge Computer Associates to tackle the address matching initiative. All told, the New Haven staff stabilized at about a dozen people, with several more commuting occasionally from census headquarters in Suitland.

Using computers at Yale and the City of New Haven, the staff started on the five tasks outlined by the Small Area Data Committee. Although Sweeney had considerable programming experience, neither the author nor Maxfield had formal training in geography or cartography. They both had nominal classroom exposure to FORTRAN but no background in computer

graphics, data analysis, or demographics. In addition, both preferred to plunge in and start programming, eschewing a literature search which might have uncovered, for example, Robert Dial's 1964 Ph.D. thesis on Street Address Conversion System (SACS). Dial's contribution would come a dozen years later when, as an Urban Mass Transit Administration (UMTA) administrator, he provided funding for a critical prototype GIS at the Census Bureau.

The New Haven staff started working with the New Haven Address Coding Guide (ACG), produced by the Census Geography Division with local assistance, as a prototype of the geographic base file needed to geocode mailing addresses in 144 metropolitan areas to be enumerated by mail. The ACG contained block-face records, each of which supplied:

Street Name

ZIP Code

A low-to-high address range

Census tract and block number corresponding to the range of addresses

Conceptually a block face was one side of a city block, a usable definition in regular downtown street patterns but one that fell apart in curvilinear suburban developments. William Fay, chief of the Census Geography Division, had in 1965 described the ACG as an "ideal foundation for a computer mapping file." The Census Bureau's in-house engineering department had built a digitizer (from scratch; digitizers were not yet a marketplace product), and operators had digitized a coordinate measurement at the middle of each block face.

Though Sweeney's fledgling address matcher could use the digitized ACG to assign census codes and coordinates to addresses, Cooke and Maxfield could do little with the ACG to map census data that were summarized to tract or block. The use study requested that the geography division redigitize the New Haven ACG, this time taking two coordinate measurements, one at each end of the block face. This would allow the budding mapping programmers to draw lines from one end of the block face to the other, recreating the street network and displaying the census blocks.

The result was disappointing for two reasons. First, the block face digitizing technique meant that the coordinates of each downtown street intersection were measured eight separate times—and, because of operator variability and drift in the newly-built digitizers, usually there were eight different coordinate readings. Urban maps, while recognizable, looked unacceptably crude. The second problem was more serious. Operators were instructed to digitize both ends of a block face, which became an impossible task in the curvilinear suburbs. One computer plot of a suburban New Haven tract was christened the "ruptured eagle" by George Farnsworth, a use study's Washington staffer—scant reward for the geography division workers who had struggled with the use study's digitizing requirements.

Maxfield, ever optimistic, set about trying to program around the ACG's flaws, searching for and averaging nearby coordinates. The results were better, but everyone shuddered at the

prospect of digitizing each of the roughly four million intersections in ACG areas eight times each, using prototype digitizing boards, then averaging the coordinates. There had to be a better way.

Digitizing efficiency became a focal point. How could one digitize each point just once? Perhaps one could analyze the ACG and come up with lists of intersections to present to the operator. But what about a street that crossed another one, then circled back and intersected it again? What about streets that turned and curved without intersecting?

In early June of 1967, James Corbett of the bureau's Statistical Research Division (SRD) presented the use study staff and Technical Steering Group with a terse and opaque overview of the topology of maps, describing how zero-, one-, and two-cells could be related through incidence matrices. The New Haven staffers did not understand this, but Corbett insisted that it was important. Finally one of the staffers admitted his bewilderment to Corbett and said, "I just want to know if we have to number the nodes." Corbett replied in the affirmative and the logjam was broken.

The procedure is relatively simple. Take a census map which has streets labeled and tracts and blocks numbered. Start anywhere and assign unique numbers to street intersections (nodes) in any order. Number the nodes at the end of dead-end streets. Put nodes anywhere that a street makes an appreciable bend. You do this because these are the points that you will eventually digitize.

Now notice that the nodes define "objects," to use today's terminology, that are straight lines between nodes. Each line segment can have only two nodes and can be between only two census blocks. Even better, all of the information about each line fits on one punch card:

Street Name
From Node
To Node
Left Tract/block
Right Tract/block
Left Address Range
Right Address Range
ZIP Left
ZIP Right

Conventions quickly evolve: If the line segment is part of a street and has addresses, then the "From" node is the one at the low address end of the segment. Otherwise, it doesn't matter. "Left" and "Right" orientation is determined by standing on the "From" node and looking toward the "To" node.

Maxfield and the author numbered the nodes on a map of Tract 1 in New Haven, manually encoded each line segment, keypunched the segments, digitized node coordinates from graph

paper, merged in node coordinates for the proper node numbers, and plotted the resulting file. A couple of zingers appeared due to miscoded node numbers, but a corrected file plotted perfectly, demonstrating the usefulness of check-plotting to detect errors.

Then Corbett's presentation started to make sense: The "zero-cells" were nodes; the "one-cells" were the line segments; the "two-cells" were the blocks. Recording the "From" and "To" nodes was really building the zero/one-cell incidence matrix. Recording the "Left-Right" block numbers built the one/two-cell incidence matrix. Didn't Corbett say you could check the fidelity of the coding by multiplying the incidence matrices?

The researchers couldn't figure out how to multiply incidence matrices. Instead, Cooke wrote a 30-line Michigan Algorithm Decoder (MAD) program which read the Tract 1 database and attempted to chain the line segments together around each census block by linking the nodes together. The program insisted that there were errors—that it could not chain each block, even though the plot check had led to correcting all node errors. But there were still errors in block numbers, which did not appear on the plot. For example, block 101 might be keyed 110 in the "Left" block field on one of the line segments. What the MAD program would see is a missing segment for block 101 and a superfluous segment for block 110. One correction fixed both problems.

Topological editing was born. The existence of a program that could systematically detect and flag clerical coding errors ignited a small but bright hope that it would be possible to create huge mapping databases at the block level in large cities in such a way that you could assure that *all* boundaries of *all* polygons would close without error, a requirement for automated mapping.

In short order, the New Haven staff made mapping files for the entire cities of West Haven and New Haven. Many agencies supported this effort: Don Luria of the IBM/New Haven project supplied clerical staff and Bob Barraclough at Tri-State Transportation Commission (New York) had his staff digitize West Haven (on digitizers that Tri-State had paid ITEK corporation $150,000 to design and build). The State of Connecticut Highway Department ran check-plots. The "ruptured eagle" development in western New Haven now looked like a map, not a joke, giving the Census Use Study success in cartography that one would have expected from the geography division. This contributed to a growing schism between the use study and Census Geography Division.

In August 1967, Farnsworth christened the new process DIME (Dual Incidence Matrix Encoding, later Dual Independent Map Encoding). "Dual" reflected the two incidence matrices; "Independent" was taken by the New Haven Staff to mean without the help of the geography division. On short notice, the author and Maxfield wrote up the DIME process (Cooke and Maxfield 1967), and Barraclough squeezed their DIME presentation into his computer graphics session at the September 1967 Urban and Regional Information System Association (URISA) conference.

By fall of 1967, Sweeney's Admatch program was running well; the health initiative was supplying numerous databases for geocoding and mapping tests; the Census Bureau had processed the April dress-rehearsal census and delivered prototype summary tapes for the use study to map. A Yale administrator declared the Census Bureau a threat to individual privacy and

cut off the use study's account at the computer center. The result was that the mapping effort turned north to Harvard.

The Harvard connection coincided perfectly with the SYMAP boom at the Harvard Lab for Computer Graphics (Chapter by Chrisman). During the fall of 1967, the New Haven staff produced reams of maps with SYMAP and any plotting equipment that was available at Harvard and MIT. Caby Smith sensed an opportunity to capitalize on DIME and took the New Haven staff on a tour of federal agencies, promoting DIME and computer mapping and promising to perform groundbreaking research projects that could be funded through interagency transfer of funds unspent at yearend. This blitz (six New Haven-to-Washington round trips in November 1967 alone for one staffer) assured the finances—and independence—of the Census Use Study.

DIME in the 1970s

The Census Use Study's influence was immediate: Samuel Arms, author of the exquisite and little-known "Map Models" system, who was in the 1967 URISA audience, promised to implement DIME concepts upon his return to Oregon to eliminate sliver problem that plagued his polygon system. Jack Dangermond, founder of ESRI, went so far as to say that the New Haven people "invented topology," the sort of hyperbole one might expect from a generous Californian. Ken Deuker and Ed Horwood attended the URISA DIME presentation, so the innovation was immediately disseminated to the University of Washington as well as the Harvard Lab, which at the time were the two major academic GIS centers.

The Census Use Study had been charged to get maps out of computers; it discovered that the real problem was how to get the maps into the computers in the first place. The primitive data processing environment of the time channeled innovation to useful ends. The 80-column punch-card and focus on geocoding demanded that the line segment be the fundamental object of DIME, in contrast with the polygon focus of virtually every other GIS project (CGIS, the Harvard Lab, PIOS, Map-Models, etc.). The batch processing environment necessitated DIME's topological consistency edits to trap clerical errors, and the topological purity attainable through DIME's edits allowed algorithmic generation of error-free polygon files from DIME files.

The study's chauvinistic competition with the geography division led to recording left and right address ranges in the New Haven DIME file, breaking reliance on the error-prone ACG. The undisciplined "not-invented-here" temperament of census researchers saved them from a possible technical sidetrack which Dial's SACS system might have afforded. Jack Sweeney's mentoring of the New Haven apprentices grounded them in good data processing practice, far more useful to the development of DIME than formal training in geography or analytic geometry. DIME, after all, turned out to be an exercise in data management and processing, not a computer graphics or cartographic problem.

The Census Bureau's response to innovation at its out-of-control research outpost was predictable. Managers all the way up to Associate Director Morris Hanson, who had pioneered use

of computers at the bureau, were drawn into a turf battle against the use study. Caby Smith hired Booz-Allen Hamilton to document the New Haven findings, then put a continent between bureau headquarters and his operation by moving his staff—funded by interagency transfers—to Los Angeles, reconstituting the use study as SCRIS (Southern California Regional Information Study). Smith continued to recruit excellent staff, notably Matt Jaro, who honed Sweeney's Admatch work and expanded it into the more general UNIMATCH. (Sweeney had left in 1968 to start Urban Data Processing, Inc., with the author and Maxfield; Jaro now heads Matchware Technologies Inc.) Another early SCRIS hire was physicist Marvin White, who would make a crucial contribution to DIME a decade later.

Back in Washington, the weight of outside funding on the Census Geography Division forced them to "add DIME features" to the existing ACGs and create DIME files in 90 new areas. Bill Fay, who had championed ACG and resisted DIME, was replaced as geography division chief in 1971. The use study initiated a series of week-long DIME training workshops which competed for mindshare with ACG/DIME meetings sponsored by geography division.

Friction between Smith and the census establishment boiled over in 1974 into a demoralizing, scorched-earth bureaucratic battle complete with FBI investigations of all SCRIS and Census Use Study personnel. Smith sidestepped the fray by founding the National Computer Graphics Association (NCGA), which drew 1,800 attendees to its first conference, 8,000 to its second, and the international Segment-Oriented Referencing System Association (SORSA), while serving out his federal career as a chief scientist at the National Parks Service. He still heads the World Computer Graphics Association, an offshoot of NCGA.

SCRIS returned to Suitland and completed its remaining contracts as the "Center for Census Use Studies," headed by Don Luria from the IBM/New Haven study. Luria, who had also run both the Charlotte, NC, and Wichita Falls, TX, USAC urban information system projects, later moved to New Mexico where he founded the largest catering organization in the state.

DIME becomes TIGER

As the 1980 Census approached, the geography division updated the 1970 ACG/DIME files to make the 1980 GBF/DIME (Geographic Base File) files. The Correction, Update, and Extension (CUE) process was a labor-intensive, batch-oriented process involving thousands of workers at hundreds of local agencies. Turnaround times for updates and edits were measured in weeks, and the bureau was faced with redigitizing all the expanded GBF coverage.

Though batch-mode update and editing were barely feasible, off-line digitizing with no graphical feedback to operators was a nightmare. Fred Broome, the geography division manager in charge of digitizing, faced immense obstacles to progress as the Census Systems Division insisted that all computer processing be done on UNIVAC mainframes. Broome broke into interactive minicomputer technology only by acquiring a DEC PDP-11 through Intergraph Corporation, under the guise of purchasing a digitizing system.

Marvin White, now in the Census Statistical Research Division, faced the same obstacle. He had inherited a prototype on-line DIME file editor called ARITHMICON from Corbett (now

retired) and had run a comprehensive test on the economics of managing a citywide DIME file as an on-line database with an interactive graphics interface. But the systems division's mainframe mandate prevailed, and White's commercial time-sharing account was terminated.

White persevered by calling a friend from his California days, Frank Lockfeld, who ran the Center for Urban Analysis in San Jose. Lockfeld was tired of struggling to maintain the Santa Clara County DIME file as a sequential database and had purchased a Z-8000 Onyx supermicrocomputer as an alternative. He needed software; White needed interactive computer time. The continentwide gap was spanned by the Federal Telephone System. White obtained funding from Bob Dial at UMTA for a pilot on-line DIME demonstration project and had his prototype 2D system running before federal accountants caught up with the outrageous surge in long-distance phone usage.

Though it was used for years by Lockfeld and a Baltimore DIME pioneer, Fred Westerfield, White's 2D system really proved its worth as a design document. It demonstrated in detail an on-line, topologically structured paradigm for managing large map databases—salvation from the purgatory of batch processing of huge spatial databases. How could White get this innovation adopted by the bureau?

The geography division was reeling from problems which had forced cancellation of local quality-control procedures. Another division chief was replaced, this time by a monthly rotation of middle managers. Redistricting battles were unearthing geographic discrepancies between the 1980 Census statistics, paper maps, and the GBF/DIME files. Broome's digitizing project fell hopelessly behind schedule.

Joe Knott, a middle manager at geography division, recognized the value of the 2D model and helped White turn 2D over to Broome at the geography bureau (White left the bureau in 1984 for a "temporary" assignment at ETAK, a car navigation firm; he's still there). Broome immediately implemented enough of 2D to convert the digitizing process to the Direct-Dig on-line paradigm and brought his project to completion on schedule.

The success of 2D technology had a profound effect on a new generation of savvy, computer-literate geography division managers. Acting as the "Coffin Twelve" (a reference to their windowless meeting room), they produced in 1982 a technical manifesto committing the geography division to integrate all of the bureau's spatial knowledge into a single, nationwide, on-line database called TIGER (Topologically Integrated Geographic Encoding and Referencing), organized along the lines of White's 2D prototype. Following successful demonstration of a "Tigger" prototype, Bob Marx, a Minnesota geographer, emerged to lead the geography division to the fulfillment of the DIME vision in the 1990 TIGER files.

Epilogue

The story of TIGER is well documented (Marx 1986). Corbett's persistence in stressing the importance of applied mathematics gave administrators at all levels confidence that the new TIGER technology was a sound investment. Knowledgeable, hands-on geography division managers involved the U.S. Geological Survey in providing rural coverage to extend TIGER to

representing the whole country. Even Marvin White and the author were called back in to contribute as their companies (ETAK and Geographic Data Technology) performed TIGER digitizing and editing contracts between 1986 and 1988.

Marx claims he never wrote a contingency plan in case TIGER proved intractable. He insists that the geography division had no choice but to succeed in implementing TIGER; no other course was technically or administratively feasible. He may be right, but that does not diminish the magnitude of the risk that the geography division undertook in the early 1980s. Its success has put the world's most useful general-purpose spatial database into the hands of more users than any other GIS data resource. The current boom in business geographics is only possible because of the groundwork laid by the Census Geography Division in building TIGER.

Summary

Today's nationwide TIGER file is the backbone of the adoption of GIS in business geographics applications. TIGER frees business users from the drudgery of map digitizing and allows them to concentrate on applying GIS technology to business problems.

TIGER is the serendipitous by-product of a modern computerized census process. The U.S. Geological Survey put its energies into computerizing topographic maps designed by John Wesley Powell a century earlier, but the result (Digital Line Graphs) has not had anywhere near the impact that the Census Bureau's accidental by-product has.

TIGER's precursor, DIME, turned the traditional cartographic paradigm on its head. The classical cartographic process started with photogrammetry and careful scribing of linework, upon which annotation is later applied. In contrast, DIME started by recording all of the annotation, cycling through topological edits and corrections, and finally—almost as an afterthought—inserting digitized node coordinates. The coordinates, formerly the framework of the entire cartographic construct, are simply attributes of the nodes and shape points of TIGER's topological structure.

But the story of DIME is more than one of technology and introduction of the mathematics of topology into managing spatial databases. The DIME idea and its topological data structure were inevitable. DIME emerged by accident of history in New Haven in 1967 scarcely influenced by prior developments.

What is extraordinary in this story is not so much the technical innovation but the importance of key personalities in getting the innovation adopted: Caby Smith forcing the Census Bureau to abandon the ACG for DIME, Marvin White's timely and modest delivery of the technological key to making nationwide TIGER feasible, and Marx's leadership and confidence each made a crucial and effective contribution appropriate to the time.

Bibliography

Broome, Marx, Tomasi, et al. 1990. *Cartography and Geographic Information Systems,* 17 (1), entire issue.

Cooke, D. F., and W. H. Maxfield. 1967. "The Development of a Geographic Base File and its Uses for Mapping." *Proceedings of the Urban and Regional Information System Association (URISA).* Washington, D.C.: URISA.

Marx, R. W. 1986. "The TIGER System: Automating the Geographic Structure of the United States Census." *Government Publications Review,* 13: 181–201.

U.S. Bureau of the Census. 1968–69. *Census Use Study Reports 1–12.*

U.S. Bureau of the Census. 1974. *DIME Comix.*