REPORT NO. DOT-TSC-OST-74-17

# THE DOT NATIONAL COUNTY COMPONENT CONVERTER FILE: PROSPECTS, PROBLEMS FEASIBILITY

Pamela Werner

AUGUST 1974

FINAL REPORT

Prepared for
U.S. DEPARTMENT OF TRANSPORTATION
OFFICE OF THE SECRETARY
Office of the Assistant Secretary for
Policy, Plans and International Affiars
Washington DC 20590

## NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

**Technical Report Documentation Page**

| | | |
|---|---|---|
| 1. Report No.<br>DOT-TSC-OST-74-17 | 2. Government Accession No. | 3. Recipient's Catalog No. |

4. Title and Subtitle

THE DOT NATIONAL COUNTY COMPONENT CONVERTER FILE: PROSPECTS, PROBLEMS, FEASIBILITY

5. Report Date

August 1974

6. Performing Organization Code

7. Author(s)

Pamela Werner

8. Performing Organization Report No.

DOT-TSC-OST-74-17

9. Performing Organization Name and Address

Urban Systems Laboratory  
Massachusetts Institute of Technology  
Cambridge MA 02139*

10. Work Unit No. (TRAIS)

OP517/R5809

11. Contract or Grant No.

DOT-TSC-692

12. Sponsoring Agency Name and Address

U.S. Department of Transportation  
Office of the Secretary  
Office of the Assistant Secretary for Policy, Plans and International Affairs  
Washington DC 20590

13. Type of Report and Period Covered

Final Report  
December 1973 to April 1974

14. Sponsoring Agency Code

15. Supplementary Notes

*Under contract to: U.S. Department of Transportation  
Transportation Systems Center  
Kendall Square  
Cambridge MA 02142

16. Abstract

Systematic review of factors affecting the feasibility of developing a county component geocoding converter file is made. Discussion of staged evaluation of such a file is presented.

17. Key Words

Geocoding,  
County Components

18. Distribution Statement

DOCUMENT IS AVAILABLE TO THE PUBLIC THROUGH THE NATIONAL TECHNICAL INFORMATION SERVICE, SPRINGFIELD, VIRGINIA 22151.

| | | | |
|---|---|---|---|
| 19. Security Classif. (of this report)<br>Unclassified | 20. Security Classif. (of this page)<br>Unclassified | 21. No. of Pages<br>58 | 22. Price |

# PREFACE

Work toward the development of a national geocoding capability under U. S. Department of Transportation auspices is now in its fourth year. The initial focus of this work, performed principally at the Urban Systems Laboratory, MIT, and the Charles Stark Draper Laboratory, Inc., was the design and operationalization of a county level converter. That task is now essentially completed. Files 1 and 2, the county converter and associated mapping capability are operational.

This paper, by Pamela Werner, and some ongoing exploratory studies at the Draper Laboratory mark the beginning of the second major stage in the development of a national geocoding converter. Work in this second stage will center on the subcounty level, a level requiring work of far greater complexity than that at the county level.

Because this paper is, in effect, the opening document of a new stage, it is most important that its purpose and function be clearly understood. The function of the paper is to state, as systematically as is possible at this time, the issues that will have to be resolved prior to each phase of subcounty converter development. It is, in short, not a paper designed to provide answers or resolve issues previously raised, but one that raises questions and articulates issues. As such, the paper will hopefully circulate among those interested in or concerned with File 3 development, generating responses and statements of views that will assist in a more precise formulation of future work on this file.

H. W. Bruck
Principal Investigator
Transportation Information Systems Project
Urban Systems Laboratory, MIT

# CONTENTS

## LIST OF ILLUSTRATIONS

# 1. INTRODUCTION: THE DOT GEOCODING PROGRAM

When the U. S. Department of Transportation became concerned with national geocoding systems, the success of the urban geocoding program conducted by the Bureau of the Census and supported to a large extent by DOT and other Federal agencies had already been established. The 1962 statutory requirement for a continuing metropolitan transportation planning process, establishment of the Census Advisory Committee on Small Area Data in the summer of 1965, and the beginning of the Census Use Study a year later, in 1966, were the prime stimuli for the forward thrust in urban geocoding. By 1971 a fairly standardized universe of urban geographic base files utilizing the Dual Independent Map Encoding (DIME) system had been implemented in almost all 267 SMSAs. A battery of DIME related computer programs, including UNIMATCH, GRIDS, CRAM and DACS had been developed and were available for use (See Figure 1).

In the United States comparable incentives for the development of national geographic base files have not existed. Indeed, national geographic coding is characterized by that fundamental chaos that results when a multitude of special purpose systems is brought into being by users needing to deal with the pressing urgencies of the moment. Unlike micro-level geocoding, which has exhibited a tendency towards standardization, uniformity and convergence of various approaches towards one major urban geocoding concept, macro-level geocoding tends towards divergence, specialization, and multiplicity of structures.

| PROGRAM | FUNCTION |
|---|---|
| ADMATCH<br>(Address Matcher) | An address matching system that provides the capability of geographically coding computer readable records containing street addresses. The system compares the addresses on input data records (after standardization with a preprocessor) with the address ranges in a reference file. A "match" occurs when the street names are judged identical or equivalent and when the address falls within the defined range. Any or all geographic codes from the reference file may be attached to the matched data records. |
| UNIMATCH<br>(Universal Matcher) | An improved matching system that has many capabilities not available in ADMATCH such as the ability to handle building names, street intersections, and non-address matching. It is a generalized record linkage system which will compile, assemble, and execute a file matching system tailored to the specific needs of the user. |
| Grid Related Information Display System (GRIDS) | A computer mapping system developed for producing character printed maps from detailed data. GRIDS has a flexible user oriented language and has several mapping options available. |
| DIME Area Centroid System (DACS) | A flexible computer system for locating centroids and calculating areas (square feet or acres) for blocks, block groups, census tracts, etc., from DIME files. Centroid location is required primarily as input to GRIDS and other map generating computer packages. |
| Network Allocation of Population to Shelter (NAPS) | A system developed by the Census Use Study for the Office of Civil Defense. It assigns the population to their closest fallout shelter up to the shelter capacity limit. It uses the DIME file to determine the shortest path to the shelter and uses the Third Count Census Summary Tape to establish population demand. |
| Computer Resource Allocation Model (CRAM) | A computer tool, based on the DIME file, that is designed for use in facilities planning (schools, service agencies, recreational facilities). It is a generalized facility location system that allocates demand among the set of available facilities according to their capacity to supply services, street accessibility, and access time. CRAM is an outgrowth of the more limited NAPS system. |

Source: Bureau of the Census, Census Use Study: DIME Workshops, May 1973.

Figure 1. DIME Related Software Packages

These fundamental differences in the development of urban and national geocoding dictated the approach taken by DOT in its efforts to achieve improvements in the utilization of national geocoding systems. Though a single, standard, DIME-like national geocoding system would seem desirable, the extent of investments in the many existing national systems and the lack of a geographic common denominator at the macro-level comparable to block face at the micro-level, make the comprehensive code approach infeasible. Therefore, DOT has focused its efforts on the development of a geographic code conversion capability. This code conversion capability is intended to facilitate the association of variously coded macro-level data files and to provide a flexible geographic cross referencing system for the benefit of users.

Drawing upon recommendations made at the National Geocoding Conference in 1971, and a systematic examination of existing systems, the DOT national geocoding program was designed to center on four potential converter files, as follows:

> File 1 is a county file. It contains a record for every county and county equivalent in the United States, identifying for each county all major county codes and the associated county aggregate, state or regional codes.
>
> File 2 is a county coordinate file. This is a DIME-structured file containing "link-node" records for all county segment boundaries in the United States. It is intended to provide mapping and graphic display capabilities at thc county level.
>
> File 3 is a county component file. It contains a record fcr each county and county equivalent in the United States, identifying for that county all major subcounty area, place, point, city or location codes.
>
> File 4 is an international file containing the geographic codes used to designate countries and other national entities such as territories, zones and principalities.

File 1, the county file, and File 2, the county coordiante file, have been constructed and are available from the Transportation Systems Center, Cambridge, Massachusetts. A brief review paper surveying the potential universe of codes for inclusion in the international file (File 4) was prepared while Files 1 and 2 were being constructed. On the basis of recommendations contained in that survey, further work on File 4 has been postponed. Officials of DOT decided to concentrate instead on feasibility studies concerned with the construction of the subcounty or county component file, File 3.

The purpose of this paper is to present an overview of the feasibility of constructing a converter file at the subcounty geographic level. It will deal with the nature of subcounty geocoding systems in general, the conceptual framework for File 3, and potential file structures for each progressive stage in the construction of the subcounty converter file. The latter section includes a discussion of the various problems anticipated at each stage.

The paper is intended as an opening statement for further discussion and consultation concerning File 3. Therefore, the structures presented should be regarded as tentative and suggestive rather than as firm recommendations. Furthermore, this paper is a prologue to development of limited computer test structures that are essential prior to preparation of final cost estimates and ultimate feasibility determination.

## 2. THE NATURE OF SUBCOUNTY GEOGRAPHIC SYSTEMS

At the coarsest levels of national geocoding, the first and second order political subdivisions of the United States serve as a relatively universal framework for ordering data geographically. Although these subdivisions have been debunked as irrationally delineated spatial units with a few geographic or functional bases for existence, and threatened by numerous reorganization proposals, the current pattern of states and counties is well established.[1] It is generally accepted as a viable system of geographic reference. The geocoding difficulties at these levels are therefore not noncompatibility of areal units or limited specialized usage of the units coded, but primarily are differences in the numerous codes now used to identify each state and county in the United States. There are at least twenty-five established sets of state and county codes used by government and private industry in the United States today.

Below the county level, however, there are a number of obstacles to interfacing the various geocoding systems. There is a lack of definition, both semantic and geographic, of subcounty units. There is the problem of variation between urban and rural land use activities, settlement patterns, and population densities which create great disparaties in the size of the spatial units coded both within a single system and among the various systems. The relative instability of subcounty units

[1] In addition to the numerous statewide proposals to abandon traditional counties in favor of more administratively effective spatial units, such as the planning regions of Connecticut, a new system based on 38 states has been proposed recently by G. Etzel Pearcy, a professor of geography at California State University, Los Angeles.

and frequent changes affecting this universe pose other problems primarily in the area of update and maintenance. And, finally, there is the lack of an areal common denominator which could provide for the flexible disaggregation and reaggregation capabilities considered highly desirable attributes in urban geocoding.

Beyond trade names and acronyms, a mundane but typical example of the semantic difficulties and lack of definition which have plagued national geocoding at the subcounty level are the various usages of words such as place and point. For many months, the American National Standards Institute (ANSI) debated the meaning of the word, place. The committee appointed to deal with this issue finally concluded that no formal definition or criterion could be formulated to the satisfaction of all committee members and, consequently, the following, almost meaningless phraseology was adopted: "A place is that which is recognized as a named place by a segment of the public."

The word point is no less promiscuously used, according to perspective of purpose. To a member of the ANSI point location committee, it is an imaginary, nondimensional dot on the surface of the earth which represents a Cartesian, State Plane or Universal Transverse Mercator coordinate fix. This is the classic Euclidian definition. To a member of the American Trucking Associations or the Association of American Railroads, a point represents a siding, a loading dock or waybill station, in short, an entry in the Standard Point Location Code file. To the Census of Transportation, it represents a geographic centroid of a municipality, one of approximately 5,600 such "Key Points" in the United States.

These two terms--place and point--as well as others, such as location, site and area, have developed multiple and ambiguous meanings. They are specifically user dependent and therefore difficult, if not impossible, to interface or compare with any real precision.

| | GEOGRAPHIC IDENTIFICATION | GEOGRAPHIC DEFINITION |
|---|---|---|
| NOMINAL | Location Designation | Boundary Description |
| ORDINAL | Hierarchical/Grid Code | Local Coordinates |
| CARDINAL | Coordinate Related Code | Global Coordinates |

Figure 2: Taxonomic Matrix of Geocoding

There is also a notable lack of geographic definition for the various units coded at the subcounty level. As illustrated in Figure 2, a fundamental distinction in geocoding is the distinction between geographic identification and geographic definition. Coding systems which consist of a set of unique designations for undefined or implicitly defined locations are geoidentifying systems. Coding systems which provide explicit boundary delineations for the units coded are geodefining systems. While none of the geographic identification systems provide for explicit boundary delineation of the units coded, some provide implicit boundary definitions, others provide no definition at all. Federal Information Processing Standards (FIPS) codes for the counties and county equivalents of the United States, for example, comprise an external index. These codes refer to a set of areal units which have explicit boundaries, but the code itself does not define them. Boundaries for the referents of most subcounty units, such as the ANSI place code, on the other hand, have not been established and, therefore, these codes do not provide even an implicit areal definition of the units coded.

The other major distinctions among geocoding systems are the differences between nominal, ordinal and cardinal geographic references. Strictly defined, a nominal geographic code is the name of a location. However, for the purpose of classifying geocoding systems, nominal codes include names, name abbreviations, mnemonic truncations of names and certain numerical codes. In this case, the word nominal derives meaning not from its root word, "name," but from describing a designation, alpha or numeric, which does not indicate or imply a spatial relation between itself and other similar designations.

Nominal geocodes do not indicate any spatial relationships among the units coded. In contrast, ordinal and cardinal geocodes do. Ordinal geocodes indicate the relative positions of coded units within a spatially related system. A hierarchically structured geocoding system such as the basic one employed by the Bureau of the Census, for example, indicates that a certain census tract is located within a certain minor civil division which is located in a certain county and state. Cardinal geocodes indicate the absolute positions of coded units within a spatially related and incrementally scaled system. The network of meridians and parallels, known as longitude and latitude, is the best example of a cardinal geocoding system.

Existing subcounty geocoding systems fall into all three categories, nominal, ordinal and cardinal, the majority being ordinal geographic references in that they are hierarchical in nature. This does not affect the nature of subcounty geocoding as severely or detrimentally as the characteristic lack of definition for subcounty units. But the preponderance of ordinal systems can complicate the procedure of matching lists of subcounty entities because of variation among hierarchies. ANSI place codes, for example, combine two geographic levels and are ordered alphabetically by name of place within states. GSA city codes, on the other hand, combine three levels, place by county by state, and SPLC has an entirely different hierarchical structure which orders points by county or county section, by state or state section, by regions.

Among the minority of national geocoding systems with subcounty entities that have either implicit or explicit geodefinition many of the interface problems occur as a result of the variance between the size of spatial units coded. In view of the areal disparity between states such as Rhode Island and Alaska, or between counties such as San Bernadino, California (20,117 square miles) and Camp, Texas (192 square miles) this problem is not solely related to subcounty geocoding. At the subcounty level, however, the problem is an order of magnitude greater simply because of the number of entities involved. Furthermore, at the subcounty level, the variance in size of the units coded is a more rigorous reflection of urban/rural land use patterns and unlike counties or states, will change as the land use pattern is altered.[1]

The Zone Improvement Program (ZIP) instituted by the U. S. Postal Service and Congressional Districts are two systems of spatial organization which illustrate the size range of the areal units coded in a single geocoding system. The geographic units of both these systems are delineated primarily on the basis of population densities. The size of ZIP Zones is constrained by the logistics of actually delivering the mail. In consequence, the ZIP system has a much finer spatial grain, but notable size disparity between units still exists. In Michigan, for example, there are approximately 1,500 ZIP Zones covering 566,817 square miles while Alaska has only about 200 zones covering 566,432 square miles. Congressional Districts range in size from whole states to a small number of city blocks.

While the size variance of states and counties did not hinder the construction of National Geocoding Converter File 1, the county level converter system, it will pose a problem in the construction of File 3, the subcounty converter. From the county level up, most existing national geocoding systems are neatly hierarchical, the county providing a "least common denominator" spatial unit. Below the county level, there is no unit which, like county, is relatively common to all nationwide information systems. As already noted, one major difficulty at the lower level is a lack of either implicit or explicit geodefinition. Areal

[1] Although counties generally tend to be smaller in metropolitan areas and larger in rural areas there are a great number of exceptions as in the case of the two counties cited above. The 20,117 square mile San Bernadino County is part of a Standard Metropolitan Statistical Area while Camp County, Texas which comprises only 192 square miles is rural in nature.

units such as places, having neither spatial definition nor an acceptable verbal definition, obviously cannot be the base for national geographic base files of the quality and utility needed. Even when these subcounty units having no geodefinition are eliminated, subcounty units having implicit or explicit geodefinition are defiantly nonconformal and/or are pieces of systems such as Congressional Districts with their already noted size range.

It might be possible to resolve the difficulties posed by Congressional Districts by utilizing the third order political subdivisions of the United States, the Minor Civil Division or Census County Division (MCD/CCD) as a subcounty standard. Except in areas of high population densities, Congressional District boundaries conform with MCD boundaries. Thus, Congressional Districts could be entered into a subcounty converter file in a manner analogous to the method used to deal with New England SMSAs in the county converter.[1] Regardless of size, Congressional Districts could then be related uniformly to MCD units, either as sections of MCDs or, in most cases, as MCD aggregates up to the county and even state level. Since SMSAs respect MCD boundaries without exception, such a structure would also completely define SMSAs, even in New England.

Relating less then county size Congressional Districts to MCDs in this fashion would be relatively simple. However, if a third subcounty system such as ZIP Zones is added, the situation becomes exceedingly complex. Figures 3 and 4 serve to illustrate this point. Figure 3 is a map of ZIP Zone and Congressional District boundaries in the City of Detroit (due to the variances in scale it is impossible to display

[1] See documentation for National Geocoding Converter File 1 for a full description of the structure of the county converter and of the method developed for including the New England SMSAs. National Geocoding Converter File 1: Structure and Content, ed. Santo LaTores. DOT-TSC-OST-73-44,I. (Cambridge, Mass.: Transportation Systems Center, U.S. Department of Transportation), 1974.

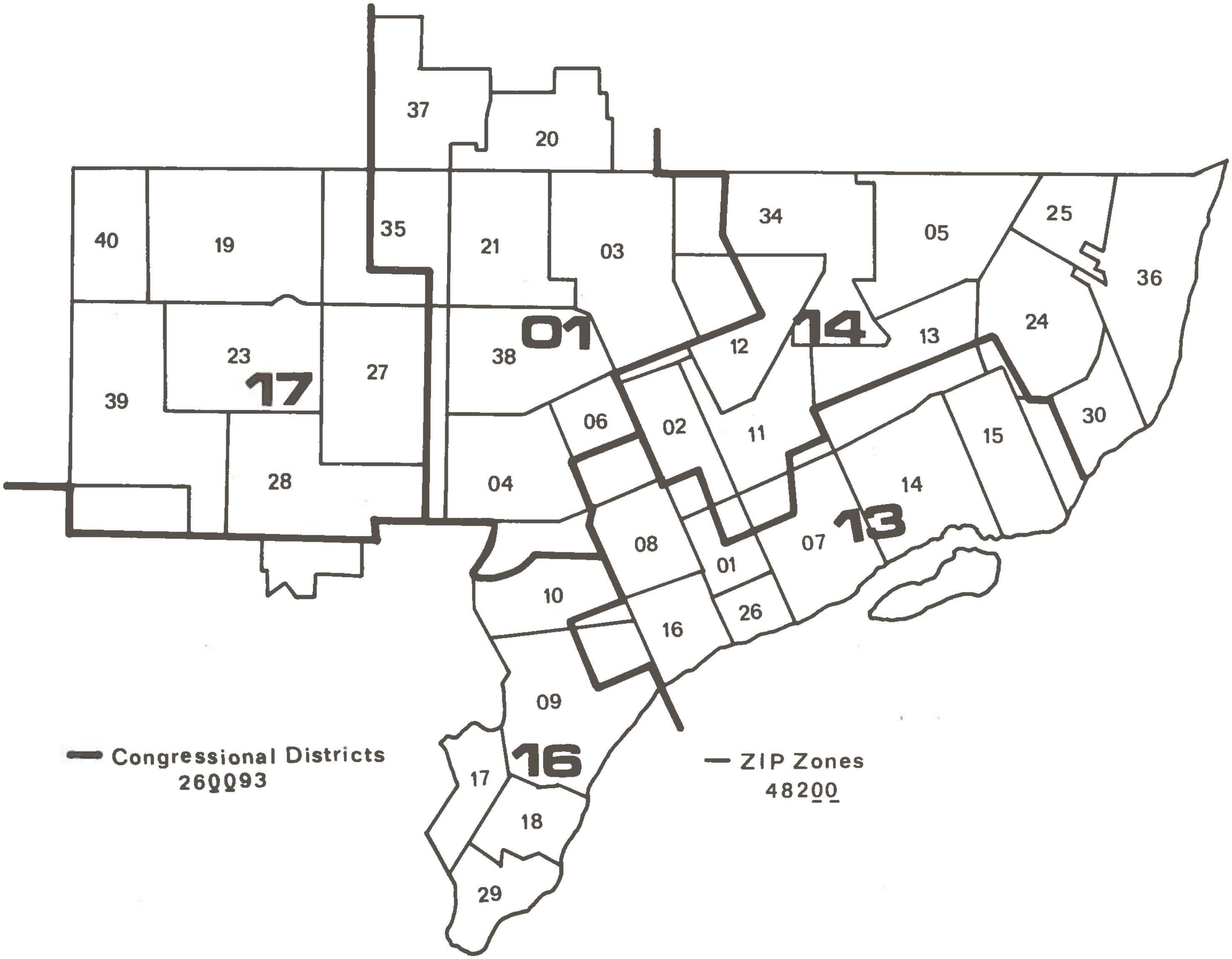

Figure 3. ZIP Zones and Congressional Districts in Detroit

Congressional District, ZIP Zone and MCD boundaries at this level with any clarity). Detroit is subdivided into 36 ZIP Zones and 5 Congressional Districts shown partially on this map, only 3 MCDs are actually divided by two or more districts. The Detroit MCD is divided into five parts, Warren Town is divided into two parts and Southgate into two. In contrast the boundaries of ZIP Zones violate both MCD and Congressional District boundaries in well over 50 instances.

Figure 4 is a map of Franklin, Hampshire, Hampden, Worcester, Norfolk and Middlesex Counties in Massachusetts, displaying county, MCD and Congressional District boundaries with ZIP Zone boundaries at the three digit level. While the Congressional District boundaries do not respect county boundaries, they do conform strictly with MCDs. The ZIP Zones, however, even at the three digit level respect neither.

In concluding this general discussion on the nature of subcounty geocoding systems, there is one other important characteristic to consider. This is the relative instability of subcounty geopolitical entities. According to the 1972 Boundary and Annexation Survey conducted by the Bureau of the Census, 1,471 places, 29 percent of the places responding to the survey, reported boundary changes occurring during the period January 2, 1970 through January 1, 1971; 1,518 places, or 30 percent, reported boundary changes during the period January 2, 1971 through January 1, 1972.[1] Survey results show that, during both time periods, the great majority of places (approximately 71 percent) experiencing boundary

[1] In the Boundary and Annexation Surveys conducted by the Bureau of the Census boundary change information is requested from all incorporated places of 2,500 or more inhabitants.

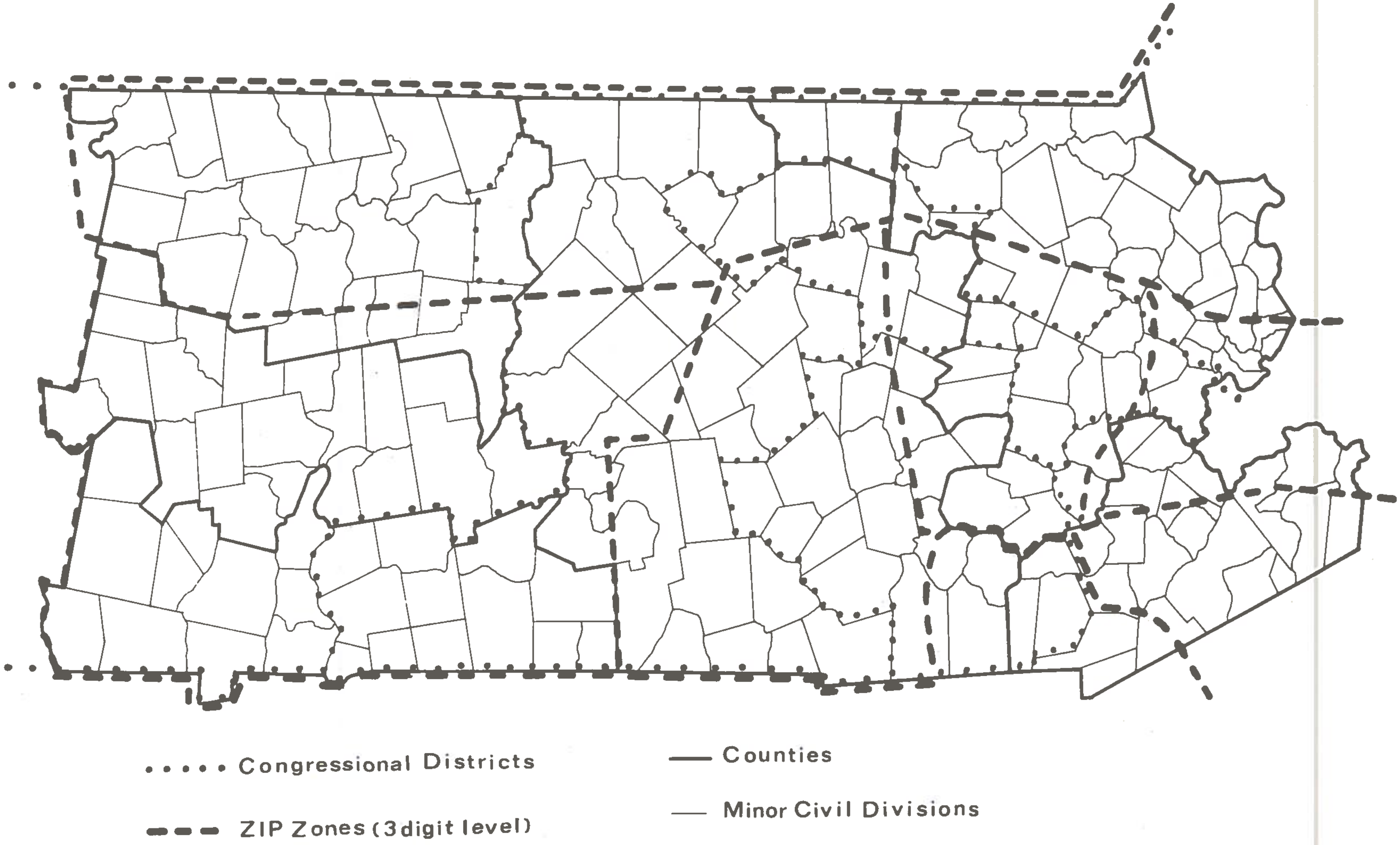

Figure 4. Overlapping Boundaries in Western Massachusetts

changes were those with populations of less than 20,000. For the two year period covered by the surveys, 9,622 annexations, 154 detachments and 27 other types of boundary changes were reported. Overall, places in 43 states reported annexations, with annexations made by places in Illinois, California and Texas accounting for approximately 40 percent of all annexations.

In addition to changes in corporate unit boundaries, numerous changes in corporate status took place during the period covered by the two surveys: 155 new municipalities (in 32 states) received certificates of incorporation, and 6 places which had not been actively functioning as governments were reactivated. In 11 states, 37 places disincorporated, 16 places (in 9 states) no longer exist as separate entities because they merged with other municipalities and three city-county or equivalent consolidations occurred.

Although the results of the 1973 Boundary and Annexation Survey have not yet been released, officials of the Bureau of the Cënsus expect the recent rate of change to continue, if not increase, over the next two years. This estimate is, in part, an anticipation of the impact of revenue sharing on the status of local governments. In order to qualify for an entitlement many previously unorganized areas will be applying for general government status as incorporated places, townships or municipalities.

In addition to all such newly delineated units, local governments which straddle county or state lines will be split for revenue sharing purposes. A large number of consolidations, mergers and annexations are also anticipated because, in many cases, such restructuring qualifies governments for a revenue sharing entitlement increase.

The long term future of the revenue sharing legislation and its impact on subcounty geopolitical units is unclear. It may be that once the initial spurt of incorporation has passed, the newly created local governments will begin to consolidate and, in many areas of the United States where counties are small to medium in sized units, the counties may subsume local governments.

Many of the other subcounty geographic entities, such as Congressional Districts, ZIP Zones and SPLC points, also undergo significant annual changes. Between the 92nd and 93rd Congress, for example, 40 states altered the boundaries of their Congressional Districts in conformity with constitutional and statutory agreements. And, currently, the SPLC Policy Council is considering the purging of many marginally important parts from their files. Thus, a county component converter file would require frequent extensive updating from a number of sources. It is even possible that long term trends in geopolitical configurations at the subcounty levels will eventually necessitate structural changes in the file. Due to the relative instability of subcounty spatical units, the construction of a county component converter file will involve a commitment to substantial and continuous update activity.

## 3. THE FRAMEWORK FOR A COUNTY COMPONENT CONVERTER FILE

The proposed DOT county converter file is divided into three separate, progressive phases:

> File 3A is a listing of the names and codes associated with the various subcounty spatial units (place, point, etc.) contained in a selected universe of national geocoding systems. As illustrated in Figure 5, the names and codes for each set of subcounty entities is listed in alphabetic order by state and county. Each set is completely independent with no implied nominal or geographic relationship between sets other than location of the entities listed, in whole or in part, within the boundaries of the given county.
>
> File 3B is a nominally matched and consolidated listing of the names and codes contained in File 3A, ordered alphabetically by state and county. Figure 6 depicts such a matched listing. The location called Baldwinville, for example, which appears in the File 3A sets for the Census, GSA, Geoloc and SPLC subcounty entities is listed once in File 3B along with the four different associated codes. File 3B contains codes which have been only nominally matched and does not imply any geographic conformality among the subcounty units.
>
> File 3C contains a list of all the various county component entities spatially related at the subcounty level in as exact a geographic frame of reference as the nature of the units permit. The actual configuration of this file is undetermined. Each county component might be related to a single existing set of subcounty units such as the Minor Civil Divisions. Or, several types of computer programs for polygon analyses might be utilized to extrapolate and define areas of spatial conformality.

The development of File 3 progresses from a very simple, nominal inventory of subcounty entities towards the delineation and definition of any existing geographic conformity among the complex of overlapping spatial units. As the file evolves through phases A,B, and C, the cost and effort involved increase dramatically. While the first phase of File 3 could be constructed in a short time with relatively minimal effort, the

| 25027 MASSACHUSETTS MA WORCESTER COUNTY | | | | | | | |
|---|---|---|---|---|---|---|---|
| CENSUS | | GSA | | GEOLOC | | SPLC | |
| Ashburmham Center | 250270140 | Athol | 200270040 | Auburn | 25 AREZ | Albee Corners | 145760 |
| Ashburnham Town | 250270150 | Auburn | 200270050 | Baldwinville | 25 AUVZ | Albeeville | 145653 |
| Athol Center | 250270190 | Baldwinville | 200270070 | Clinton | 25 DYVS | Ashburnham | 145112 |
| Athol Town | 250270200 | Barre | 200270079 | Douglas | 25 FJTW | Ashburnham Twp | 145110 |
| Auburn Town | 250270220 | Barre Plains | 200270080 | Fort Devens | 25 HEHL | Athol | 145320 |
| Baldwinville | 250270260 | Blackstone | 200270110 | Gardner | 25 HSEZ | Auburn | 145480 |
| Barre Center | 250270290 | Brookfield | 200270150 | Leicester | 25 NBHR | Baldwinville | 145181 |
| Barre Town | 250270370 | Charlton | 200270185 | Leominster | 25 NCLD | Ballard Hall | 145231 |
| Berlin Town | 250270400 | Clinton | 200270220 | Milford | 25 QFXQ | Barbar | 145460 |
| Blackstone Town | 250270410 | Dudley | 200270263 | Rutland Heights | 25 UPHB | Barrack Hill | 145381 |
| Bolton Town | 250270430 | East Brookfield | 200270272 | Spencer | 25 UYVU | Barre | 145361 |
| Boylston | 250270480 | East Douglas | 200270282 | Warren | 25 YKNN | Barre Falls | 145351 |
| Brookfield Center | 250270550 | Fisherville | 200270340 | Webster | 25 YQQE | Barre Twp | 145360 |
| Brookfield Town | 250270560 | Fitchburg | 200270350 | Westboro | 25 YSNM | Bartletts Village | 145738 |
| Charlton | 250270660 | Gardner | 200270390 | Worcester | 25 ZHAP | Barton | 145740 |
| Clinton | 250270790 | Gilbertville | 200270420 | | | Bayview | 145740 |
| Cordaville | 250270855 | Harvard | 200270467 | | | Beacon Park | 145740 |
| Douglas Town | 250270970 | Holden | 200270480 | | | Bean Porridge Hill | 145161 |
| Dudley Town | 250271000 | Hopedale | 200270510 | | | Bearfoot | 145760 |

Figure 5. Illustrative Example of File 3A Structure

| 25027 MASSACHUSETTS MA WORCESTER COUNTY | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Name | Census | GSA | GEOLOC | SPLC | MCD | SMSA | UA | Type |
| Albee Corners | ---------- | ---------- | ------- | 145760 | 025 | 9240 | ---- | 0 |
| Albeeville | ---------- | ---------- | ------- | 145653 | 030 | 6480 | ---- | 0 |
| Ashburnham Center | 250270140 | ---------- | ------- | 145112 | 005 | ---- | ---- | 5 |
| Ashburnham Town | 250270150 | ---------- | ------- | 145110 | 005 | ---- | ---- | 4 |
| Athol Center | 250270190 | ---------- | ------- | ------ | 010 | ---- | ---- | 5 |
| Athol Town | 250270200 | 200270040 | ------- | 145320 | 010 | ---- | ---- | 4 |
| Auburn Town | 250270220 | 200270050 | 25 AREZ | 145480 | 015 | 9240 | 3110 | 4 |
| Baldwinville | 250270260 | 200270070 | 25 AUVZ | 145181 | 250 | ---- | ---- | 5 |
| Ballard Hall | ---------- | ---------- | ------- | 145231 | 100 | 9240 | 3110 | 4 |
| Barber | ---------- | ---------- | ------- | 145460 | 015 | 9240 | 3110 | 4 |
| Barrack Hill | ---------- | ---------- | ------- | 145381 | 020 | ---- | ---- | 0 |
| Barre Center | 250270280 | ---------- | ------- | 145361 | 020 | ---- | ---- | 5 |
| Barre Falls | ---------- | ---------- | ------- | 145351 | 020 | ---- | ---- | 0 |
| Barre Plains | ---------- | 200270080 | ------- | 145363 | 020 | ---- | ---- | 0 |
| Barre Town | 250270290 | 200270079 | ------- | 145360 | 020 | ---- | ---- | 4 |
| Bartletts Village | ---------- | ---------- | ------- | 145738 | 025 | 9240 | ---- | 0 |
| Bayview | ---------- | ---------- | ------- | 145740 | 025 | 9240 | ---- | 0 |
| Beacon Park | ---------- | ---------- | ------- | 145740 | 025 | 9240 | ---- | 0 |
| Bean Porridge Hill | ---------- | ---------- | ------- | 145161 | 250 | ---- | ---- | 0 |

**Figure 6. Illustrative Example of File 3B Structure**

third phase of File 3 may well be too costly and too time consuming to complete. The critical factors involved would be the current and expected levels of resources available for development of this file and the degree of geographic explicitness desired on the File 3C level. At this time there is no doubt that achieving the final phase of File 3 will require large scale commitment on the part of DOT officials. It is, however, important to note that even if all three phases cannot be implemented, each phase is separable and independent, with value as a geocoding product in and of itself.

## 3.1 File 3A: An Inventory of County Components

As described above, the first phase of the proposed DOT County Component Converter File (File 3A) is an inventory of all subcounty entities contained in a selected universe of national geocoding systems. The file structure is a simple one consisting of a fixed set of elements which identify a county within a state, a number of subsets which list, in alphabetic order, the names and codes of the subcounty entities located within the given county. Each national geocoding system included in the file comprises one subset of names and codes for every county (See Figure 7).[1]

The first set of decisions to be made in connection with this phase of File 3 concerns the universe of national geocoding systems to be included. Figure 7 illustrates the record layout for 17 possible subsets

[1] Congressional Districts and Standard Location Areas are the only exceptions among the systems suggested for inclusion into File 3. Since Congressional Districts and Standard Location Areas are not named entities, only code designations would appear in the subset for these systems.

Fixed Set:

| FIPS Code State/County | State Name | State Abbr. | County Name |
|---|---|---|---|
| 5 | 20 | 2 | 28 |

Subsets:

| Census Place Name | Census Place Code | | IBM City Name | IBM City Code |
|---|---|---|---|---|
| 28 | 9 | | 28 | 9 |

| DUN City Name | DUN City Code | | GSA City Name | GSA City Code |
|---|---|---|---|---|
| 28 | 9 | | 28 | 9 |

| RS Place Name | RS Place Code | | ANSI Place Name | ANSI Place Code |
|---|---|---|---|---|
| 28 | 8 | | 28 | 7 |

| Geoloc Place Name | Geoloc Place Code | | SPLC Point Name | SPLC Point Code |
|---|---|---|---|---|
| 28 | 6 | | 28 | 6 |

| ZIP Zone Name | ZIP Zone Code | | Congressional District | | SLA Code |
|---|---|---|---|---|---|
| 28 | 5 | | 6 | | 11 |

| PICADAD Place Name | PICADAD Key Point | | Airport Place Name | FAA Code |
|---|---|---|---|---|
| 28 | 4 | | 28 | 3 |

| UA Name | UA Code | | COE Port Name | COE Port Code |
|---|---|---|---|---|
| 28 | 4 | | 28 | 5 |

| FT Port Name | FT Port Code | | OSAI Port Name | OSAI Port Code |
|---|---|---|---|---|
| 28 | 4 | | 28 | 3 |

See Appendix for explanation of abbreviations

Figure 7. Potential Record Layout for File 3A

Seven of these systems are represented in the DOT County Converter File (File 1). The remainder were either redundant or non-applicable at the county level and therefore excluded from File 1. The place codes assigned for the purpose of revenue sharing, for example, are prefixed by the FIPS state and county codes already included in File 1. Other systems, such as the FAA airport codes, are strictly subcounty designations.[1]

Subcounty geographic systems can be divided into two basic groups: those entities which are geographically undefined (this includes "point" and undefined "place" codes); and those entities with geographic definition. Each group can be further subdivided according to structural similarities of system coverage and other characteristics in the following manner.

1. Geographically Undefined Entities

A. Census, IBM, DUN, and GSA and revenue sharing codes are state/county/place hierarchies, each with a two digit state code, a three digit county code and a three or four digit place code that uniquely identifies a named but geographically undefined location on a one to one basis.

B. ANSI and Geoloc codes are two level hierarchies which designate a state in the first two digits and a place within that state in the last four or five digits. These codes uniquely identify a named but geographically undefined location on a one to one basis.

[1] The Federal Aviation Administration maintains several location identification systems all of which provide geographic reference at the place level. The three character identifiers discussed in this paper are assigned to those U. S. airports, heliports, and seaplane bases on which there is established a manned FAA air traffic control facility, a terminal air navigational aid within the airport or which receive DOD airlift service or scheduled route air carrier service. Other code series are assigned to waypoints, navigational fixes, etc.

C. FAA airport codes, Foreign Trade, U. S. Corps of Engineers and OSAI port codes are essentially named but geographically undefined place designations with a limited, selective universe of entities. The Foreign Trade and Corps of Engineers codes are hierarchical in nature (district/port) but do not conform to either state or county boundaries. FAA and OSAI codes are simple, three digit nominally geocoded designations.

D. SPLC and PICADAD are point codes. More accurately they can be described as the point associated codes for named places. In both systems a single point number can be associated with several named places within relatively close geographic proximity. Thus the number of named places included in these files exceeds the number of coded points. The PICADAD point code is a non-hierarchical set of serial numbers limited to a universe of approximately 56,000 "key points." The universe of SPLC point codes is expandable, and is a hierarchy ordered by state <u>or</u> state section and county <u>or</u> county section.

2. Geographically Defined Entities

A. Congressional Districts, ZIP Zones and Urbanized Areas are geographically defined areal units delineated primarily on the basis of population densities. They vary considerably in size, both within and among the three systems. The areas defined do not necessarily conform to county (or Minor Civil Division) boundaries although in some cases they do. This is the most structurally heterogeneous group of codes: Congressional Districts respect all state but not all county or MCD boundaries; ZIP Zones are hierarchically structured but respect neither state, county nor MCD boundaries; Urbanized Areas are not hierarchically structured at all and, unlike Congressional Districts or ZIP Zones, do not include the entire area of the United States.

B. Standard Location Area (SLA) codes are the last four digits of the strictly hierarchical (Region/State/Area/County) National Location Code. These codes were assigned under a complex set of criteria which delimit spatial units on the basis of population densities without crossing unit boundaries established by the Bureau of the Census.[1] Standard

[1] Refer to: U.S. Department of Commerce, Bureau of the Census, <u>National Location Code</u> prepared by the Office of Civil Defense and the Office of Emergency Planning, 1962, for a full description of the criteria for Standard Location Areas. Also, <u>Geographic Codes for Region, State, Area and County</u>, Office of Emergency Preparedness (IS6-111), February, 1971 or <u>A Survey of National Geocoding Systems</u>, U. S. Department of Transportation, 1972, for further explanation of the current status of the NLC coding system.

Metropolitan Statistical Areas (SMSAs) also fall within this category. In the New England States, SMSAs are composed of whole subcounty spatial units (MCDs) rather than counties. Although SMSAs are included in the DOT County Converter File, they might also be considered for inclusion in File 3.[1]

C. Minor Civil Divisions and Census County Divisions are geopolitical subcounty units which can be treated simply as county components or might serve as an hierarchical level between county and place. They are established as first order political subdivisions of counties by the states or by the Bureau of the Census in cooperation with a state.

The purpose of this long excursion into the structure and purpose of the various national geocoding systems which include subcounty entities is to call attention to the variety of systems to be considered for inclusion in File 3 and to highlight the critical importance of the decision to include or exclude. Although there is no absolute necessity to include codes contained in File 3A in File 3C, each system considered for inclusion should be evaluated for pertinence to each stage in the development of the entire file and not simply for its suitability to the first phase. The peculiarities of each geocoding system and the difficulties which these may create in any phase of File 3 development must be taken into consideration fully before a universe of codes can be selected.

After selecting a universe of national geocoding systems for File 3, the necessary code sources and authorization for use of these codes must be obtained. Although this appears as a very straightforward endeavor, a number of problems may be encountered in the process. Of the 17 systems listed in Figure 7, code sources for only the 5 systems included in File 1 have been obtained and the currentness of these files is

[1] In order to define all SMSAs accurately, either File 1 will have to be changed so that MCDs in New England are dealt with as county equivalents, or File 3 will have to be ordered on an MCD basis to accomodate them.

uncertain.

The other system code sources selected for inclusion in File 3 will have to be obtained. Those not obtained thus far include machine readable sources for IBM, DUN, GSA, Geoloc, SLA, Revenue Sharing and ANSI codes. While some of these codes are public information and can be obtained for the cost of copying the file, sources for systems such as the Dun and Bradstreet place code will cost as much as $500.00 for one file and copyright issues have yet to be resolved.

Obtaining all the required source tapes may be further complicated by the fact that certain files are not yet complete. The Federal Aviation Administration is in the process of compiling a machine readable random access file containing all the FAA location identifiers (including the three character airport codes). This file is scheduled for completion within the next few months but may not be immediately available for public distribution. Also, the ANSI place codes are not currently on file in machine readable form. Furthermore, there is some question as to the availability of this system for use in File 3 since Rand McNally (one of the organizations involved in preparing the ANSI place code system) may wish to maintain the codes as proprietary.

Certain codes with a very limited universe of entities should be entered manually as File 3A subsets. These include all three port code systems and possibiy the Urbanized Area units.

Because of the simplicity of its structure, there is only one processing problem anticipated in the construction of File 3A. This will be the sorting of the subcounty entities by county in the several systems which have no county reference. These systems include ANSI, Geoloc, Congressional Districts, the airport and port codes. As indicated previously, the universe of entities coded in all three port code systems is small enough to be sorted by machine, probably on a name matching basis. The matching process will be complicated by the number of duplicate place names existing in each state, or, as in the case of FAA codes, the United States as a whole. Since they are not assigned names, Congressional Districts cannot be matched to counties in such a manner. And, although the majority of the Congressional Districts are listed accurately by county on MEDList, MEDList contains only the lowest numbered district for counties with more than one district. Therefore some level of manual checks,editing and coding cannot be avoided even on machine sorting at this level of File 3, and there will be additional difficulties involved in sorting all of the subsets by county.

When it is completed, File 3A will comprise several subsets of subcounty entities ordered alphabetically within counties. The length of each subset will vary considerably. One set of subcounty codes may include as many as several hundred entities while another set for the same county may consist of only a few or, even, a single entity. There will be a great deal of redundancy, not only in the repetition of place names among the sets, but also in the state/county prefixes assigned to many place codes. There will be approximately 40 fixed sets with only one or, possibly, two entities in each subset because independent cities have equivalent status. And, there will be subsets with no subcounty entities in all but a very few counties.

### 3.2 File 3B: Nominally Matched County Components

In the second phase of the proposed DOT County Component Converter File (File 3B), the subsets generated for File 3A are to be compared and matched by name. As illustrated in Figure 6, a county component name is listed once along with all the codes assigned to that name in any of the subsets. All names, even singular, unmatched names, contained in File 3A are retained in File 3B. The file is to be ordered alphabetically by state and county with the same fixed set of elements contained in File 3A.

The focal point of file processing at this stage is the application of the Universal Matcher System (UNIMATCH) developed by the Bureau of the Census. UNIMATCH is a generalized record linkage system which will compile, assemble, and execute a file match. It has the capability of handling building names, street intersections and other non-address matching such as place names. The generality of UNIMATCH is achieved by permitting user specification of fields to be compared, the nature and content of the comparisons (e.g., character, numeric, or parity comparisons on fixed fields or intervals), the significance to be attached to a success or failure to compare, and finally the action to be taken depending upon the level of success of comparison. Such action might include further comparisons or copying of selected fields from the reference file to the data file. UNIMATCH may be used in conjunction with another Bureau of the Census program called UNISTAND (Universal Standardizer), a generalized record standardization system designed to provide the user with a capability of standardizing fields subject to uncertainty and variability. UNIMATCH and UNISTAND provide an efficient name matching capability ideally suited to the processing requirements of File 3B.

Due to the importance of individual user specifications in the UNIMATCH system, as many place name matching problems as can be anticipated should be accomodated in the specifications for processing File 3B. This will contribute to the accuracy of the matches accomplished and hopefully reduce the number of unmatched entities along with the level of manual verification required. The more obvious place name matching problems to be considered include the following:

-Matching acceptable spelling variations commonly used to designate the same place (e.g., Marlborough or Marlboro)

-Matching incorrectly spelled place names and providing for correction

-Matching place names that have been truncated in order to minimize length of field

-Matching place names with directional or title prefixes, (East Medford or San Miguel)

-Matching place names with municipal suffixes (e.g., center, town or village)

-Matching all possible name or name component abbreviations (e.g., E. Medford and East Medford)

-Matching place names containing non-alpha characters such as dashes and apostrophes

Clearly, it may not be desirable to deal with these and the many other potential matching problems by programming the matching routine to accomodate all of them. In many cases doing so could create more problems than it solves and certain difficulties are simply unavoidable. The fact that there are a great number of places with the same name (names like Springfield, Jackson, Reading, etc. are very common) will inevitably result in incorrect matches, especially when processing geocoding systems that are not hierarchically ordered by state, or state and county.

In addition to the name matching difficulties described briefly above, there are several other potential problem areas to be considered and a number of important decisions to be made in connection with them. Determining the manner in which inter-county places will be entered on File 3B is one of these. The Bureau of the Census deals with such places by listing the place name two or more times in the MEDList file, depending upon the number of counties in which it is partially located, with the word part in paranthesis next to the name each time. The county code prefix changes, but the place code remains the same. Thus Fort Devens, Massachusetts, is listed in Worcester County coded 250271305, and again in Middlesex County coded 250171305.

Multiple listings of this type could easily be incorporated in File 3B from the Census MEDList tape. Other source files are not so easily handled in this manner because not all source files indicate whether a coded entity crosses county boundaries. In the Geoloc file for instance, Fort Devens is a single entry with no indication that this place is located in both Middlesex and Worcester Counties. By comparing the names for all inter-county places listed on the Census tape for a match in any other source and then creating the necessary number of duplications in each set this problem could be partially resolved. There remain, however, many other inter-county areal entities, such as ZIP Zones, which are not listed in the Census file, and therefore cannot be matched and flagged for additional listing by machine.

An alternative strategy for dealing with inter-county entities would be to list them only once. The national geocoding systems containing so called point locations are easily dealt with in this manner because, theoretically, a point has no real dimension and therefore must be clearly assignable to one county or another. Other subcounty entities with either explicit or implicit dimensions could also be assigned to one county or another simply by establishing some standard set of criteria for such allocation. Either the entity is assigned to the county in which the center of population or geographic centroid of the areas is located. This will require the deletion of any existing multiple listings and it will involve working with an extensive set of maps in order to determine the county assignments on a place by place basis. The trade off entailed in the application of this strategy is the generalization or forcing of the geographically defined units. An entity like Fort Devens, for example, which is actually located in two counties would be listed only once and seem to be located entirely within a single county as on the Geoloc File. Therefore, a combination of the two methods outlined, such as single listings for dimensionless entities and multiple listings for areally defined entities might be considered as a viable compromise solution.

Incorporating all Congressional Districts, ZIP Zones, and Standard Location Areas into File 3B structure poses yet another problem. In urban areas, these units are generally much smaller than any other entities listed in File 3B and, therefore, even a nominal match on a one to one basis is impossible. For example, within the City of Boston, which is represented in most national geocoding systems as a single coded unit named Boston, there are 4 unnamed Congressional Districts, approximately thirty ZIP Zones coded under local post office or street names such as Kenmore Square and Hanover Street, and at least thirty unnamed Standard Location Areas.

If any of these three national geocoding systems are selected for inclusion in File 3B, the codes involved will require some form of special matching treatment in densely populated areas. They could be incorporated into File 3B as code ranges. Thus the file would indicate that Congressional Districts 932507 through 932509 and 932511, and ZIP Zones 02108 through 02138 and 02210 are located in whole or in part within the City of Boston. The procedure would be somewhat awkward in that the field lengths will vary and the manual labor involved in generating these ranges will be extensive, but it is not an infeasible solution to the problem.

It would be inappropriate at this stage of the discussion concerning the feasibility of File 3B to speculate in detail about other potential problem areas. Certainly there will be some unanticipated difficulties and others, which could be mentioned in this paper, may never actually materialize. Until the universe of codes involved has been selected and detailed source file specifications are obtained by DOT, further discussion of potential problems at this level would be counterproductive and confusing. Airport codes, for example, may or may not require special treatment depending upon the final structure of the file which the FAA is in process of compiling. If the airport serving the Boston metropolitan area is entered into the file as Logan BOS, this coding may result in a mismatched or unmatched record. Most other files will have no matching record for a place named Logan in Massachusetts or, possibly even a record for some totally different entity named Logan. If the airport is listed as Boston BOS, the SPLC record for Logan Airport will go unmatched, while in all other cases the airport will be matched with the City of Boston. A number of such equivocal situations could be cited and are fruitless to pursue at this time.

There is, however, one final aspect of File 3B which should be mentioned and it concerns the structure of the file. Figure 8 illustrates one potential record layout for File 3B which includes all of the national geocoding systems discussed in this paper. Some of the fields may eventually be eliminated and others may be added, but a majority of those fields will be generated in the place name matching process, with exceptions previously noted. Three other fields will also have to be incorporated into File 3B in an alternative fashion because, MCDs, SMSAs, and Urbanized Areas are named entities, they are not comparable with any other list of place names in the appropriate alphabetic order among them; or listed separately at the end of the list of place names; or, as illustrated in Figure 6 each place name could be assigned the appropriate MCD, SMSA or Urbanized Area code. However,

**Fixed Set:**

| FIPS Code<br>State/County | State Name | State Abbr. | County Name |
|---|---|---|---|
| 5 | 20 | 2 | 28 |

**Subset:**

| Place Name | Census<br>Place Code | IBM<br>City Code | DUN<br>City Code | GSA<br>City Code | RS<br>Place Code | ANSI<br>Place Code |
|---|---|---|---|---|---|---|
| 28 | 9 | 9 | 9 | 9 | 8 | 7 |

| Geoloc<br>Place Code | SPLC<br>Point Code | ZIP<br>Zone Code | Congressional<br>District | PICADAD<br>Key Point | MCD/CCD<br>Code | SMSA<br>Code |
|---|---|---|---|---|---|---|
| 6 | 6 | 5 | 6 | 4 | 3 | 4 |

| UA Code | Place Type | SLA Code | FAA Code<br>Airports | FT<br>Port Code | OSAI<br>Port Code | COE<br>Port Code |
|---|---|---|---|---|---|---|
| 4 | 1 | 11 | 3 | 4 | 3 | 5 |

See Appendix for explanation of abbreviations

Figure 8. Potential Record Layout for File 3B

doing so involves some of the same assignment difficulties discussed in connection with inter-county places carried to an inter-Minor Civil Division level.

Between the first and second phase of development, the problems involved in the construction of the proposed DOT National County Component Converter File increase by an order of magnitude. The overwhelming problem at the second level is the geographic variation both within and between national subcounty geocoding systems. The file can easily accomodate the universe of so called place and point codes, but the other important district and zone codes are difficult to relate to any standard. On one hand, some systems can be treated as code ranges within places and, on the other hand, a number of places are located within larger grain, but still subcounty, systems. DOT officials will have to decide just how much of this variation can be accomodated efficiently in one county component file.

## 3.3 Spatial Relationship Among County Components

In the third phase of the proposed DOT County Component Converter File, the primary objective will be relating all of the various county component entities contained in File 3 spatially as well as nominally in as exact a geographic frame of reference as the nature of units permit. Since the universe of elements to be included in File 3C is, to a large extent, dependent upon the results obtained in Files 3A and 3B, it is impossible at this time to visualize what an actual record layout for File 3C might be. It is possible, however, to discuss several generalized issues involving the structure and potential uses of File 3C.

The most critical set of decisions to be made concerning the file are: what type of spatial relationships among subcounty entities should (or can) the file be designed to reflect and should the geographic relationships established be ordinal or cardinal in nature?

If an ordinal system is decided upon, the universe of entities included in File 3C could be related to an established set of areally defined subcounty units in a hierarchical manner. Relating all subcounty units to the first order political and administrative subdivisions of each county (MCD or CCD) is one alternative. A county component would then be defined as located within a certain MCD, within a certain county, within a certain state, or as an aggregate of MCD units within the same hierarchy. The geographic relationships thus established would be strictly relative relationships. For some coded entities, such as ZIP Zones and Urbanized Areas, the spatial match between the subcounty entities and MCDs will have to be forced areal approximations but for the majority of place and point locations, establishing an MCD frame of reference is a fairly simple process. In essence, this file would be a "universal" MEDList combining and relating a number of national geocoding systems in the manner used to relate the various Census codes. While this alternative does not approach the geographic precision and does not permit the types of computerized analyses which a coordinate related system can provide, an ordinal MCD related file of this type would be comparatively easy to construct. In fact, if all the systems to be incorporated in File 3C have already been included on File 3B, it will involve little more than a resorting of the file to move from phase B to phase C.

A major advantage of a coordinate related geocoding system is that a number of computerized spatial analyses, such as radial searches and cluster analysis, can be performed. This capability increases the flexibility and utility of the file substantially. However, if a cardinal or coordinate related system is decided upon there are problems and options to be considered. The principal obstacle in assigning coordinates to the subcounty units under consideration in this paper is the size variation among the units to be coded. The point and place locations pose relatively few difficulties, because they can be represented by a single set of Cartesian coordinates. This holds true even for place locations which may have implicit spatial dimension, because, at a national scale, the area is generally inconsequential and is adequately represented by a centroid point. A number of national geocoding files containing place listings, e.g., the MEDXYList and PICADAD, have centroid coordinates associated with each place.

Thus assigning centroid coordinates for all point and place entities contained in File 3C could be accomplished through a series of matching processes. In such a process, coordinate related source files would be ranked according to the presumed accuracy of the coordinate readings and the extent of the File 3 universe contained in each file. Then the source files would be matched against File 3B and coordinates assigned to each place and point matching a source file entity until only those entities contained in files with no coordinate reference were left. These remaining points and places would be assigned coordinates manually.

Large geographic units, however, cannot be represented by a single set of coordinates. It would be necessary to digitize the boundaries of these units in order to provide adequate coordinate descriptions of the areas involved. Doing so would involve creating a DIME structured link-node boundary file for at least Minor Civil Divisions and ZIP Zones (perhaps Urbanized Areas and Congressional Districts as well) analogous to the county boundary file, File 2. The subcounty boundary file would be an order of magnitude larger than File 2, containing some 100,000 to 200,000 areal units as compared to the approximately 3,000 counties contained in the county boundary file. The utility value of such a link-node boundary file for county components would be very high and is well beyond the resources now available for File 3.

Therefore, some form of compromise between a strictly ordinal and strictly cardinal system may be the most viable alternative serving to maximize the utility and minimize the cost of File 3C. All point and place codes adequately represented by a single set of coordinates would have cardinal identification and all other coded entities would merely be geoidentified within the county/state hierarchy. This approach provides for a certain limited capability for computerized spatial analyses among the point associated entities and relates the remainder to these coordinate defined locations in a relative or ordinal manner.

One final alternative approach to the development of File 3C is conceivable. This approach more nearly resembles an upward or aggregated extension of urban geocoding rather than a downward or disaggregated extension of national geocoding because it requires the use of an extremely small scale areal unit, the enumeration district, as the national "block face" or building block unit. Conceivably, the enumeration district (ED) could ensure all of the flexibility in national geocoding that the block face has provided in urban geocoding.

Enumeration districts are small administrative areas established by the Bureau of the Census and used for the collection and tabulation of population and housing data. On the average each ED contains approximately 250 housing units with a population ranging from about 750 people in non SMSA areas to 1,500 people in SMSA areas.[1] The two most important criteria in the delineation of enumeration districts are: 1) the population contained in a single ED is limited to an adequate enumerator workload; and 2) the ED generally does not cross the boundary of a city, township, minor civil division or other areas (except census blocks) for which census data are to be tabulated.

Because enumeration districts do not generally cross the boundaries of other geopolitical and geostatistical areas for which the Bureau of the Census tabulates data, the ED could provide an effective "building block" unit for the various geodefined subcounty entities contained in File 3C. Congressional Districts, SMSAs, Standard Location Areas, minor civil divisions and census county divisions, incorporated places, urbanized areas and entire counties can be defined in terms of ED aggregates and thus, spatially related in the proposed DOT County Component Converter File with a relatively high degree of accuracy and great deal of geographic flexibility. The fact that EDs have already been assigned centroid coordinates of latitude and longitude further enhances the utility of the enumeration district unit.

[1] In 1970 enumeration districts in areas enumerated by the mail-out/mail-back canvass were replaced by slightly modified enumeration districts called block groups. A block group is a combination of contiguous city blocks having an average population of about 1,000 people. Block groups are for all statistical purposes the equivalent of conventional enumeration districts within the census-by-mail areas.

Enumeration districts are, however, very small scale units. There are over 250,000 EDs (as compared with 36,000 MCDs and CCDs, and 3,100 counties) in the United States and the DOT County Component Converter File would have to be expanded considerably in order to accomodate this large universe of entities. Furthermore, the future stability of the ED unit is uncertain despite the fact that the Bureau of the Census is attempting to hold the 1970 EDs fairly constant. Finally, it is unclear whether such a fine grained geographic base level is justifiable in a national geographic reference file. Excess geographic detail can only make the system awkward and inefficient to use.

Whatever structure is decided upon, the completion of File 3C will be dependent upon access to certain resources. One of the most important resources would be an extensive set of maps. A select set of detailed subcounty maps may be necessary reference material during the initial stages of File 3 development, but File 3C will require a far more exhaustive catalogue of reference maps. It may even be necessary to have ready access to a number of map series, including the Rand McNally Commercial Atlas, the United States Geologic Survey Topographic Maps, the FHWA State Highway Maps, and the Metropolitan Map Series of the Bureau of the Census. Among other uses, these maps would be used to verify the location of subcounty entities, to resolve any source file conflicts, to identify those marginal SPLC points which have no mapped reference and to assign missing coordinate readings to point and place locations. Cartographic reference material is indispensable to the geocoding program, especially at the File 3C level.

Although each phase of the proposed DOT County Component Converter is viewed as a separable, independent file, they do constitute a progression of which File 3C is the culmination. Each of the two preceeding files contribute and are preparatory to the ultimate configuration of File 3C. The successes and failures of Files 3A and 3B will indicate the difficulties involved in the final preparation of File 3C. It is impossible at this time to speculate on how extensive these might be. Therefore, in closing this brief, very general discussion on the feasibility of File 3C, there is only one important issue to reiterate. In order to perform the types of computerized spatial analyses which seem desirable and for which there is a definite user demand, as many county components as possible should be coordinate related and defined. Where it is currently infeasible to provide such geodefinition, consideration should be given to achieving this goal.

## 4. CONCLUSION: PREPARATION FOR FILE 3

This paper has presented an overview of the proposed DOT County Component Converter File and enumerated some of the issues and problems involved in the implementation of the three phases planned for the file. It is a first cut exposition on the subject, intended as a basis for discussion and more detailed investigations of the feasibility of File 3. Further development will need to focus on the preparation and compilation of limited test structures for each phase of the file. On the basis of the test results, issues raised in this paper can be resolved and cost estimates prepared for the full implementation of the file.

Figure 9 is a schematic illustration suggesting a development process for File 3. The sequence is initiated with a general discussion and the selection of codes for possible inclusion in the file; test structures are designed and compiled; cost estimates prepared for each segment of File 3; and, the process ends with a final review to determine the ultimate feasibility of constructing a nationwide county component converter file at each of the three levels.

In view of the varied nature of subcounty geographic systems and the potential cost of obtaining the numerous source files for these systems, one of the most important phases in the development of File 3 will be a thorough investigation of all the geocodes considered for inclusion in the converter prior to actually acquiring the source files. This would involve establishing contact with the agencies and organizations

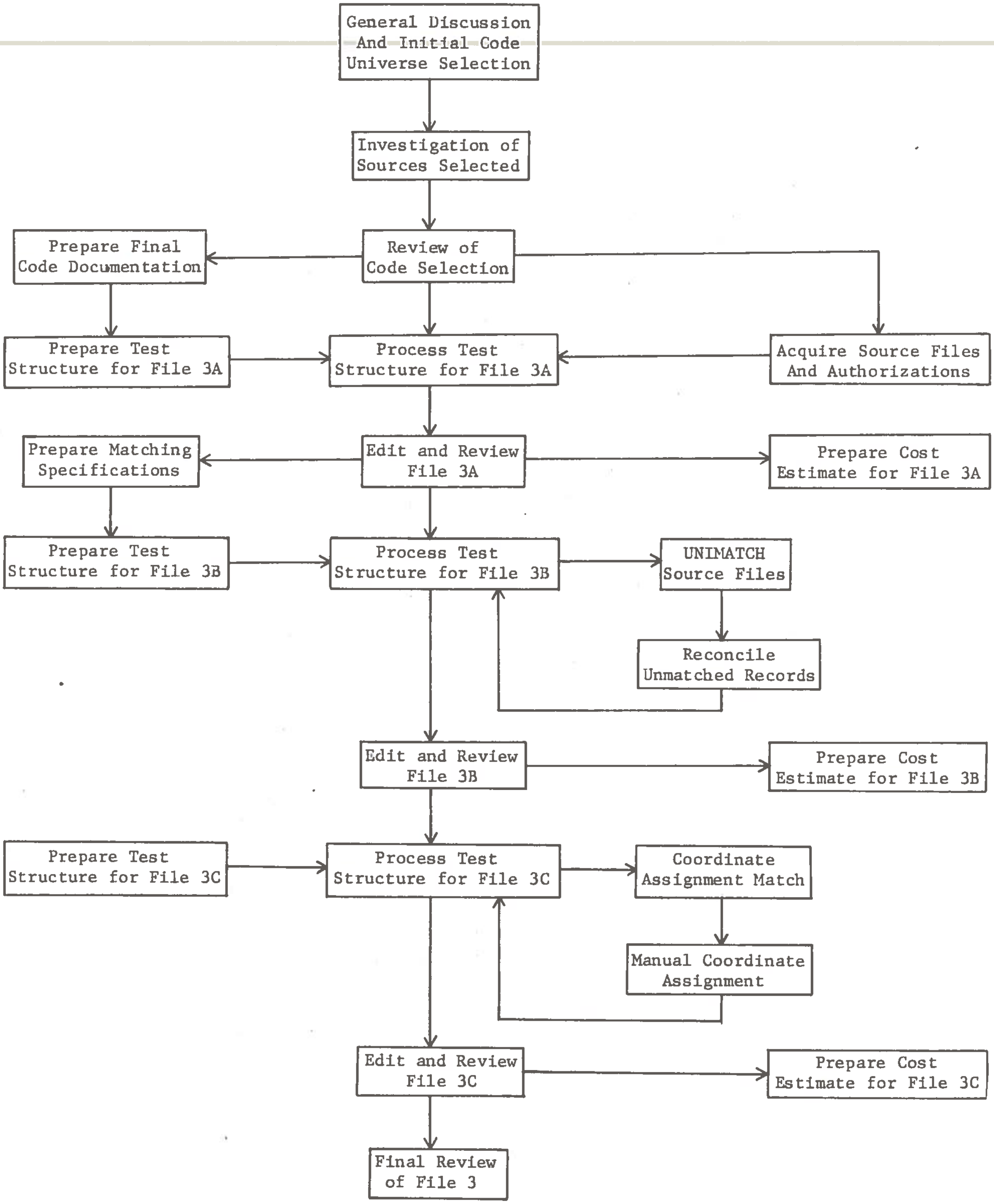
General Discussion And Initial Code Universe Selection
Investigation of Sources Selected
Prepare Final Code Documentation
Review of Code Selection
Acquire Source Files And Authorizations
Prepare Test Structure for File 3A
Process Test Structure for File 3A
Prepare Matching Specifications
Edit and Review File 3A
Prepare Cost Estimate for File 3A
Prepare Test Structure for File 3B
Process Test Structure for File 3B
UNIMATCH Source Files
Reconcile Unmatched Records
Edit and Review File 3B
Prepare Cost Estimate for File 3B
Prepare Test Structure for File 3C
Process Test Structure for File 3C
Coordinate Assignment Match
Manual Coordinate Assignment
Edit and Review File 3C
Prepare Cost Estimate for File 3C
Final Review of File 3

Figure 9. Schematic Diagram of File 3 Development

responsible for each of the systems under consideration and soliciting cooperation in the compilation of detailed information on the source files available. Such information would include record layouts and machine specifications for each file, an estimation as to how "clean" and up-to-date the files are and any anticipated problems in using the source in the construction of File 3. Also, these investigations should compile some information about the types of data associated with each set of codes, current utilization, and frequency of update and any future geocoding programs or improvements planned by the agency.[1]

As Figure 9 suggests, the results of these investigations should be reviewed and the initial code selection reevaluated on the basis of the information presented concerning each of the source files. At this point in the development of File 3, it may be necessary to eliminate a number of the national geocoding systems considered for inclusion in the file. A certain basic set of point and place codes which, as previously noted, can be incorporated into File 3 with ease will probably constitute the core of the file. Each additional system considered for inclusion will have to be evaluated carefully on the basis of its value to the file, the structural alterations necessary to incorporate and relate it to other codes, the degree of clarity or precision which could

[1] There are indications of several important changes or improvement programs taking place in house at agencies sponsoring large national geocoding systems e.g., the development of an FAA geocode master file, the possibility of an SPLC file purge, an extensive activity at the U. S. Postal Service to provide ZIP Code user packages along the lines of the Census Use Study.

be achieved in relating it to other codes, and the cost of accomodating the system in the converter. After these trade offs have been weighed in relation to each of the three phases of File 3 and some determination has been made as to which sets of codes are feasible for inclusion at which phases, then a final code documentation can be prepared for use in designing the test structures and the appropriate source files and authorizations obtained. In many ways this is the most crucial phase in the development of File 3.

The test structures for each phase of File 3 should be geographically limited, but include all of the geocoding systems selected for inclusion at each level. This will provide a sufficiently limited test structure and enable the DOT to evaluate all of the source files involved.
The criteria for selection of geographic areas to be tested should be based on the need to include both rural and urban settlement patterns encompassing the range of geographic reference systems which vary according to population densities. This could be accomplished by selecting a single, relatively large and diverse state or by selecting a number of counties throughout the United States.

In conclusion, there is no doubt that the proposed DOT County Component Converter File is an ambitious project with many potential difficulties the extent of which cannot be determined until the actual test structures are completed. The file progresses from a very simple nominal inventory of subcounty entities towards delineation and definition of any existing geographic conformity among a complex of overlapping spatial units. As the file evolves through phase A, B, and C the cost and effort involved increases dramatically. Therefore, it is imperative that the cooperation of all the agencies sponsoring codes to be included in File 3 be solicited and their advice encouraged. The U. S. Postal Service and the Bureau of the Census in particular would make desirable allies.

## APPENDIX: A Glossary of Codes for File 3

| ITEM | FILE | LENGTH | SOURCE | DEFINITION |
|---|---|---|---|---|
| Census Place Name | AB | 28 | Medlist, Bureau of the Census | The full alphabetic name assigned to places by the Bureau of the Census |
| Census Place Code | AB | 9 | Medlist, Bureau of the Census | The combination two digit numeric state, three digit numeric county (FIPS) and four digit numeric place code authorized by the Bureau of the Census |
| IBM City Name | A | 28 | IBM New City Guide, IBM | The full alphabetic name assigned to places (cities) by IBM |
| IBM City Code | AB | 9 | IBM New City Guide, IBM | The combination two digit numeric state, three digit numeric county, and four digit numeric place (city) codes authorized by IBM |
| DUN City Name | A | 28 | Dun and Bradstreet | The full alphabetic name assigned to places (cities) by Dun and Bradstreet |
| DUN City Code | AB | 9 | Dun and Bradstreet | The combination two digit numeric state, three digit numeric county and four digit numeric place (city) codes authorized by Dun and Bradstreet |
| GSA City Name | A | 28 | General Services Administration | The full alphabetic name assigned to places (cities) by the General Services Administration |
| GSA City Code | AB | 9 | General Services Administration | The combination two digit numeric state, three digit numeric county (now superceded by FIPS) and four digit numeric place (city) codes authorized by the General Services Administration |
| GEOLOC Name | A | 28 | Department of Defense | The full alphabetic name assigned to places (locations) by the Department of Defense |

Appendix (cont'd)

| ITEM | FILE | LENGTH | SOURCE | DEFINITION |
|---|---|---|---|---|
| State/County Code | AB | 5 | National Geocoding Converter File 1, Department of Transportation | The two digit numeric code for states (including D.C.) and three digit numeric code for counties and county equivalents in the United States as authorized by the (FIPS) Federal Information Processing Standard Register |
| State Name | AB | 20 | National Geocoding Converter File 1, Department of Transportation | The full alphabetic spelling of state names as authorized by the FIPS Register |
| State Abbreviation | AB | 2 | National Geocoding Converter File 1, Department of Transportation | The two digit alphabetic abbreviation designating each of the 51 first order political subdivisions of the United States including the 50 states and the District of Columbia |
| County Name | AB | 28 | National Geocoding Converter File 1, Department of Transportation | The full alphabetic county or county equivalent name as authorized by the FIPS Register |
| MCD/CCD Name* | B | 28 | Medlist, Bureau of the Census | The full alphabetic name of Minor Civil Division and Census County Divisions as authorized by the Bureau of the Census |
| MCD/CCD Code | B | 3 | Medlist, Bureau of the Census | The three digit numeric code for each Minor Civil and Census County Divisions as authorized by the Bureau of the Census |
| SMSA Code | B | 4 | Medlist | The four digit numeric code designating Standard Metropolitan Statistical Areas as defined and authorized by the Office of Management and Budget |

*Minor Civil Divisions are the primary political and administrative subdivisions of a county. In the 21 states for which MCDs are not suitable, the Census Bureau has established Census Civil Divisions in lieu of MCDs.

Appendix (cont'd)

| ITEM | FILE | LENGTH | SOURCE | DEFINITION |
|---|---|---|---|---|
| GEOLOC Code | AB | 6 | Department of Defense | The combination two digit numeric state (same as FIPS) and four digit alpha place (location) authorized by the Department of Defense |
| ANSI Place Name | A | 28 | American National Standards Institute | The full alphabetic name assigned to places by the American National Standards Institute |
| ANSI Place Code | AB | 7 | American National Standards Institute | The combination two digit numeric state and five digit numeric place code authorized by the American National Standards Institute |
| RS Place Name | A | 28 | Department of the Treasury | The full alphabetic name assigned to places receiving revenue sharing money by the Department of the Treasury |
| RS Place Code | AB | 8 | Department of the Treasury | The combination two digit numeric state, three digit numeric county and three digit numeric place codes authorized by the Department of the Treasury |
| SPLC Point Name | A | 28 | SPLC Master File, American Trucking Associations | The full alphabetic name assigned to places (points) by the American Trucking Associations and Association .of American Railroads |
| SPLC Point Code | AB | 6 | SPLC Master File, American Trucking Associations | The full hierarchical place (point) code authorized by the American Trucking Associations and the Association of American Railroads |
| PICADAD Place Name | A | 28 | PICADAD Place File, Transportation Division, Bureau of the Census | The full alphabetic name assigned to places by the Transportation Division, Bureau of the Census |
| PICADAD Key Point | AB | 4 | PICADAD Place File, Transportation Division, Bureau of the Census | The four digit numeric key point code associated with each place authorized by the Transportation Division, Bureau of the Census |

Appendix (cont'd)

| ITEM | FILE | LENGTH | SOURCE | DEFINITION |
|---|---|---|---|---|
| ZIP Code Place Name | A | 28 | U. S. Postal Service | The full alphabetic name assigned to ZIP code places by the U. S. Postal Service |
| ZIP Code Number | AB | 5 | U. S. Postal Service | The full five digit ZIP code number authorized by the U.S. Postal Service |
| Congressional District | AB | 6 | Medlist, Bureau of the Census | The combination two digit numeric state code, two digit congressional district number and two (should be expanded to three for future) digit congress number |
| Airport Place Name | A | 28 | Federal Aviation Administration | The full alphabetic name of place in which airport is located |
| FAA Airport Code | AB | 3 | Federal Aviation Administration | The three digit mnemonic alpha code assigned to airports with scheduled freight or passenger service assigned the Federal Aviation Administration |
| FT Port Name | A | 28 | Foreign Trade Division, Bureau of the Census | The full alphabetic name assigned to ports by the Foreign Trade Division, Bureau of the Census |
| FT Port Code | AB | 4 | Foreign Trade Division, Bureau of the Census | The combination two digit customs district number and the two digit numeric port code authorized by the Foreign Trade Division, Bureau of the Census |
| COE Port Name | A | 28 | U. S. Corps of Engineers | The full alphabetic name assigned to ports by the U. S. Corps of Engineers |
| COE Port Code | AB | 5 | U. S. Corps of Engineers | The combination one digit region, one digit disdistrict and three digit port code authorized by the U. S. Corps of Engineers |

Appendix (cont'd)

| ITEM | FILE | LENGTH | SOURCE | DEFINITION |
|---|---|---|---|---|
| OSAI Port Name | A | 28 | Department of Transportation | The full alphabetic name assigned to 20 coastal ports by the Office of Systems Analysis and Information, Department of Transportation |
| OSAI Port Code | AB | 3 | Department of Transportation | The three digit numeric code for ports authorized by the Office of Systems Analysis and Information, Department of Transportation |
| Urbanized Area Name | A | 28 | Medlist, Bureau of the Census | The full alphabetic name of Urbanized Areas as authorized by the Bureau of the Census |
| Urbanized Area Code | AB | 4 | Medlist, Bureau of the Census | The four digit numeric code for Urbanized Areas assigned by the Bureau of the Census |
| Standard Location Area Code | AB | 11 | Office of Emergency Preparedness | The combination 7 digit region, state, area, county code and 4 digit Standard Location Area code as authorized by the Office of Emergency Preparedness |
| Place Designation | B | 1 |  | An arbitrarily assigned (or ANSI standard) designation to indicate type of place |

# APPENDIX

A digilent review of the work performed under this contract has revealed no new innovation, discovery, improvement or invention.