A NEW APPROACH TO AUTOMATED NAME PLACEMENT

Ümit Başoğlu
CACI, Inc.-Federal
11495 Sunset Hills Road
Reston, Virginia 22090

ABSTRACT

The Automated Name Placement research project was funded by the National Mapping Division of the USGS in September 1981. The purpose of this research project is to determine the feasibility of using computer technology to generate names overlays for map production. There are three components of a comprehensive name placement system: 1) A names data base, 2) type size and style selection procedures, and 3) the placement logic. The original proposal submitted to the National Mapping Division was divided into three phases: 1) creation of a test data base for point, linear and areal names, 2) development of techniques for automated placement of point names, and 3) development of techniques for automated placement of linear and areal names. Work is nearing completion on the first phase which included the creation of a data base of point names along with software to process user requests, select features and produce verification plots. Linear and areal features have also been entered in a second data base following the development of matching algorithms to correlate the names from the Geographic Names Information System (GNIS) with their digital representations from the National Atlas files. This paper discusses the procedures followed during the first phase of this research, shows examples of maps produced by the software and evaluates the future of this area of automated cartography.

INTRODUCTION

Time and space do not permit us to go into an extensive history of the lettering and name placement process. However, name placement has been quite troublesome to cartographers. Advances in technology have helped the map maker in every aspect of the mapping process except for name placement. With the introduction of computers, many cartographic procedures have been automated during the last decade. Name placement is the only process that has not had the advances seen in other areas of cartography. This is a significant problem since over 50% of map preparation time is spent on the name plate, in spite of the fact that various methodologies and procedures have been developed for the placement of names.

Why are we having such difficulty with utilizing computer technology in the name placement process? The answer lies in the complexity of the process and the restrictions of the computers. Although the speed of processing has increased tremendously during the last ten years, the "intelligence" of these machines is still the same. They are as intelligent as we can make them. If we examine the field of automated cartography, we can see that most of the procedures in existence have been developed by non-cartographers. These methods and procedures are straightforward and do not require extensive cartographic training. The name placement process, however, is more complex and does not yield to simple solutions. This is not to say that attempts have not been made. The literature is quite limited but Yoeli,[1] Wilkie[2] and Hirsch[3] are a few of the small group of people who have attempted to automate the name placement process. During the last few years more attention has been given to automated name placement and as we can see in this conference, the importance of the topic is being realized more and more.

We can classify automated name placement into three distinct categories: 1) fully-automated systems, 2) semi-automated systems, and 3) interactive systems. In the fully-automated systems all placement is done with computer processing. The developed software must check for overplotting, place names along linear and within areal features, decide which names to place and where. Semi-automated systems enable some manual interaction such as correcting problems resulting from cases improperly handled by the fully-automated systems. Interactive systems, on the other hand, give the cartographer full control of the placement process. This process is <u>not</u> an automated name placement process but a mechanized manual operation. Our goal is the fully-automated system, but reality for the time being is the semi-automated system.

Some research has been attempted at the fully-automated name placement process. The author's Ph.D. research has shown successful results and the present activity sponsored by the National Mapping Division has yielded good results, although still in its early stages. The present paper will describe the procedures involved in creating comprehensive names data bases, report on the status of the project, and discuss problems encountered and the future of the research.

RESOURCES FOR CREATING THE TEST DATA BASE

Currently, two major efforts are underway to collect comprehensive data for geographic and cartographic purposes. One of these which was described in the earlier paper by Roger Payne[4] is the GNIS effort and the other is the capture of digital cartographic data by the Eastern Mapping Center of National Mapping Division. This effort involves the digitization of the up to eight overlays of data from the 1:2,000,000 scale National Atlas sheets. There are several papers in this conference addressing this data.

These two data bases were the most comprehensive ones available and no user testing was done on either data base similar to the effort that will be described here.

The states of New Mexico and Arizona were chosen as the test area. At the start of the project, we received a file in the Digital Line Graph (DLG) format for this area with eight categories of data: political boundaries, administrative boundaries, roads and trails, railroads, streams, water bodies, cultural features and hypsography. Of these eight categories, the political boundaries overlay was used for outline and reference plotting, the administrative boundaries and water bodies overlays were used during area determination and the streams overlay was used to test the algorithms for linear features.

Three separate files were received from the GNIS data base for the test area: 1) a point names file, 2) linear names file, and 3) areal names file.

The point names file consisted of all populated places within the New Mexico-Arizona area which had been coded with the feature class definition of "ppl" in the GNIS. There were close to 700 names in this file and besides the latitude and longitude of the point, the file contained state and county FIPS codes and 1980 population.

The linear names file contained all the features with 'STREAM' as their definition. This yielded several thousand names with primary and secondary coordinates for each name. In a separate effort, certain stream names were captured for the 1:2,000,000 scale National Atlas sheets. This file contained 36 names.

The areal names file was also an extensive file with several thousand names. All names with feature class definition of "LAKE", "TANK", "PARK", "SUMMIT", "FOREST", "BENCH", "CEM", "CIVIL", or "AREA" were retrieved. This file contained state and county FIPS codes along with the latitude/longitude of the point.

The processing done to generate the two data bases is described below.

POINT NAMES DATA BASE

Name placement is straightforward if one does not worry about the number of names and the clutter on a given map. However, if a decision has to be made as to which names to place at a certain scale then a selection methodology must be devised. For the present research, the method outlined by Kadmon(5) was utilized. This method uses rank values and weighting. Rank values are assigned to the variables chosen to describe a point location. A final weighted rank is computed by multiplying each rank value by a weighted vector for a given case. The equation for n points with m rank values would be

$$R_i = \sum_{j=1}^{m} r_{ij} w_j \quad (i = 1,2...n) \qquad (1)$$

where R_i is the final weighted rank vector, r_{ij} is the matrix of rank values and w_j is the weight assigned. Once the weighted ranks are computed, the selection consists of sorting these values and placing the top p values selected, where the value of p is assigned by the user or calculated using the scale of the source and output maps and the number of names contained on the source map. A selection criteria of this type is described in a paper by Töpfer-Pillewizer(6).

The point names data base created for the present research has seven variables associated with each name in addition to the place name and latitude /longitude values. These are: state FIPS codes, population, administrative, post office, bank, and trade ranks, and a remoteness factor. The names, latitude/longitude values and the state FIPS codes were taken from the file received from the GNIS. The place name consists of up to 24 characters stored in upper case because of the restrictions of the plotter utilized for testing. Population rank values were computed from the GNIS file of population values using the following rank assignments:

Rank	Population
0	None
1	1-1,000
2	1,001-5,000
3	5,001-10,000
4	10,001-50,000
5	50,001-100,000
6	100,001-250,000
7	250,001-500,000
8	500,001-1,000,000
9	Over 1,000,000

The administrative rank was assigned depending on the status of the place: 3 if National Capital, 2 if State Capital, 1 if County Seat and 0 otherwise.

The data for the post office, bank, and trade ranks was obtained from the Rand McNally Commercial Atlas. A value of 1 was assigned in the appropriate fields if a post office or bank exists at a given place. The trade rank is based on the total retail trade for a place in millions of dollars: 5 for over $1,000, 4 for $500-$1,000, 3 for $250-$500, 2 for $100-$250, 1 for less than $100 and 0 for none.

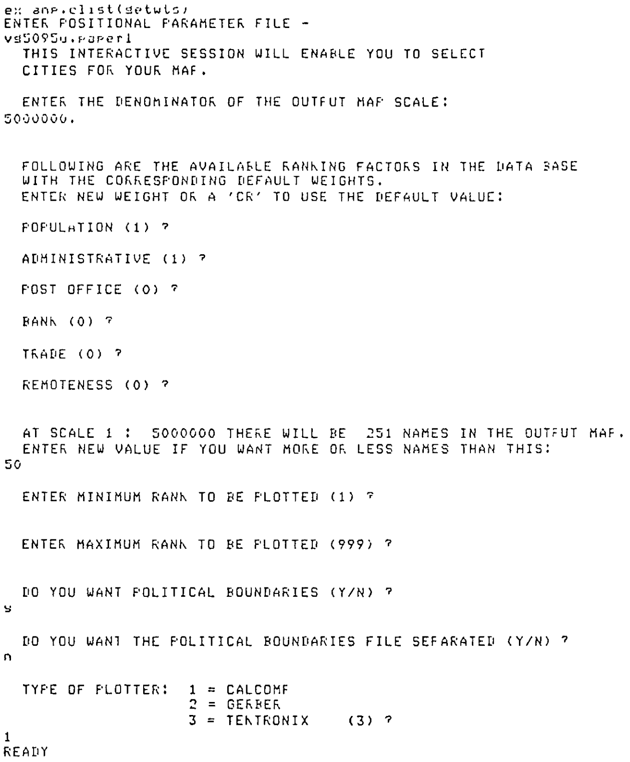
The remoteness factor in the data base was introduced to make it possible for seasonal, historic, remote and other significant locations to appear in certain maps. National Mapping Division personnel selected locations in the test area having these properties. A remoteness factor of 1 was entered for these place names.

Two programs were written to retrieve and verify the point names data. The first is an interactive program. It converses with the user to determine output scale, number of names to be placed on the map, the rank weights, minimum and maximum computed rank weights, output plotter type, political boundaries output and whether to file separate the output overlays. A file is created which is used by the second program to generate the plot. If the output medium is the Tektronix 4014, then a plot is generated automatically. Otherwise, the user must submit a job to be run using batch processing. Figure 1 shows an interactive session, Figure 2 shows the corresponding plot generated by the plot program.

LINEAR NAMES PROCESSING

The processing involved in creating the data base for the linear names was much more complicated. The GNIS portion of the processing consisted of examining the two files mentioned earlier. The file which was digitized from the 1:2,000,000 scale map was unsuitable because only one coordinate at the mouth of each stream was digitized. Therefore, it was virtually impossible to match any of these names with their line segments in the DLG files. To solve this problem, we decided to use the names from the 1:2,000,000 scale file to extract coordinates from the more detailed 1:24,000 scale GNIS files. Retrieving all coordinates associated with the 1:2,000,000 scale names from the 1:24,000 scale file gave us sufficient data for use in the matching process.

In a parallel effort, the streams overlay of the DLG file was processed. This file contained ramdomly digitized line segments which were topo-



```
ex anp.clist(setwts)
ENTER POSITIONAL PARAMETER FILE -
vs50950.paper1
  THIS INTERACTIVE SESSION WILL ENABLE YOU TO SELECT
  CITIES FOR YOUR MAP.

  ENTER THE DENOMINATOR OF THE OUTPUT MAP SCALE:
5000000.


  FOLLOWING ARE THE AVAILABLE RANKING FACTORS IN THE DATA BASE
  WITH THE CORRESPONDING DEFAULT WEIGHTS.
  ENTER NEW WEIGHT OR A 'CR' TO USE THE DEFAULT VALUE:

  POPULATION (1) ?

  ADMINISTRATIVE (1) ?

  POST OFFICE (0) ?

  BANK (0) ?

  TRADE (0) ?

  REMOTENESS (0) ?


  AT SCALE 1 :  5000000 THERE WILL BE  251 NAMES IN THE OUTPUT MAP.
  ENTER NEW VALUE IF YOU WANT MORE OR LESS NAMES THAN THIS:
50

  ENTER MINIMUM RANK TO BE PLOTTED (1) ?


  ENTER MAXIMUM RANK TO BE PLOTTED (999) ?


  DO YOU WANT POLITICAL BOUNDARIES (Y/N) ?
y

  DO YOU WANT THE POLITICAL BOUNDARIES FILE SEPARATED (Y/N) ?
n

  TYPE OF PLOTTER:  1 = CALCOMP
                    2 = GERBER
                    3 = TEKTRONIX    (3) ?
1
READY
```

Figure 1

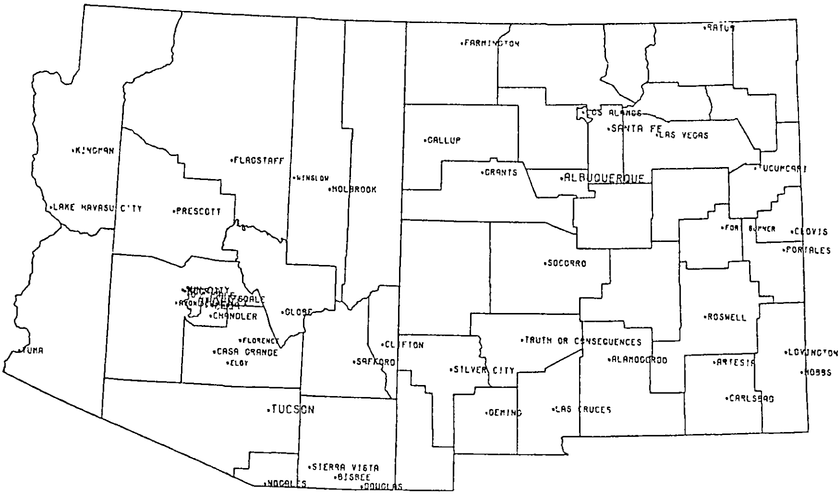


Figure 2

logically correct; that is, end points of each line segment matched the end point of the next line segment. As a result of this, it was possible to chain line segments to produce single lines representing a river or stream. This involved a computer intensive process of searching through the line segments to match the end points. Considerable knowledge of the DLG file structure and attribute coding scheme was required.

The final step of the linear names processing before building the data base, was the matching of the points from the GNIS files (those marked as '+' on the plot) with the coordinates from the DLG files (the line segments on the plot). The software written for this step checks the coordinates for each name against the coordinates of each line segment. The name is assigned to the line segment which contains the most matches within a given tolerance. Twenty-one names were matched correctly with their corresponding line segments within a 15 mil tolerance. These names were entered into the final data base.

AREAL NAMES PROCESSING

The processing for areal names again involved several different pieces of software. The first was a selection program which retrieved the areal names from the GNIS file depending on the specified feature class definition. In the first execution of this program a file for the water bodies overlay was created by selecting "LAKE" and "TANK" feature classes and the second execution created a file for the administrative boundaries overlay by selecting "PARK" and "FOREST" feature classes. Both of these files contained the name and the latitude/longitude coordinates as digitized from the 1:24,000 scale maps. It should again be noted here that the present effort is to test the software developed and any of these files could be expanded by selecting other feature classes or through numerous other techniques.

The administrative boundaries and the water bodies overlays of the DLG files are also topological files which make it relatively easy to create individual closed polygons of the areal features. This was done because closed areas will be required so that names can be placed within them if the scale permits and to identify the name for the present processing. Software was written to create these individual polygons and put the coordinates in a file by utilizing segment chaining operations.

Final processing of areal features involved matching the names from the GNIS file with the polygons created from the DLG file. Using point-in-polygon techniques, these comparisons were made and the names for which the identification point fell within the polygon were inserted into the data base.

The DLG polygon creation yielded 113 polygons for the water bodies. Point-in-polygon processing found 30 points from the GNIS file that fell within these polygons. Two of these polygons had multiple names within the area. This is due to the vast differences in scale of the two files. Coordinates for another 15 names were within a small distance of the closed polygon and these could have easily been included in the data base.

There were 282 polygons created for the administrative boundaries overlay. Only 33 polygons had points from the GNIS file within them and most of these had multiple names. This was followed by a manual checking process which showed only 10 of these to be correct. More research needs to be done in this area if administrative boundaries are to be included in the automated processing.

LINEAR AND AREAL NAMES DATA BASE

Following the above processing, a two level data base was created for linear and areal names. The first level contains the descriptive features: the name of the feature, a code indicating whether it is a linear or areal feature, the length of the line forming the feature (this value is in inches), the area of the feature for areal features (this value is in square inches), state and county FIPS codes where applicable, the coordinates of the minimum bounding rectangle, the number of coordinates forming the line segment, and a pointer to the second level. The second level is a "cartographic" data base; that is, it only contains the coordinates that form the line. Presently, the second level file is a sequential file. Figure 3 shows contents of this data base.

PROBLEMS IN CREATING THE DATA BASES

Attempting a research of this magnitude is a major undertaking. The data bases that were available are quite comprehensive. However, their contents and capacities were somewhat inadequate for this type of processing. Both data bases are sophisticated enough to handle many other requirements but unfortunately are lacking in attributes required if they are to be utilized by a name placement system.

Starting with the GNIS files, for the point names files, the population value alone is not sufficient as a selection criteria. Although it is difficult to include different types of data in these files, a value such as administrative status could easily be stored. Secondly, we found that several places with populations over 30,000 were missing from the files. This reflects the "age" of the file even though the population values were from the 1980 census.

The linear and areal names files had not yet gone through the final edit process and there were quite a few points well beyond the error tolerance. However, the overall quality of these files, especially the linear features, was quite good. The file for linear names that was digitized for the 1:2,000,000 scale map as a separate effort is almost useless for matching purposes. Identifying a single point at the mouth of a river and trying to match that point to a line segment in the DLG file is an impossible task. This file is useable only if manual matching operations are used between the two files.

The DLG files have other problems, the most significant of which is the lack of documentation. The processing on the linear features would not have been possible without several meetings with the USGS personnel on these files. Also, lack of vertical control between overlays in these files may be a problem later on in this research.

The problem that is unavoidable is the one resulting from the differences in scale. The over abundance of data from the GNIS files was most obvious in the processing of linear and areal features. Even after selection by feature class, there was still too much data.

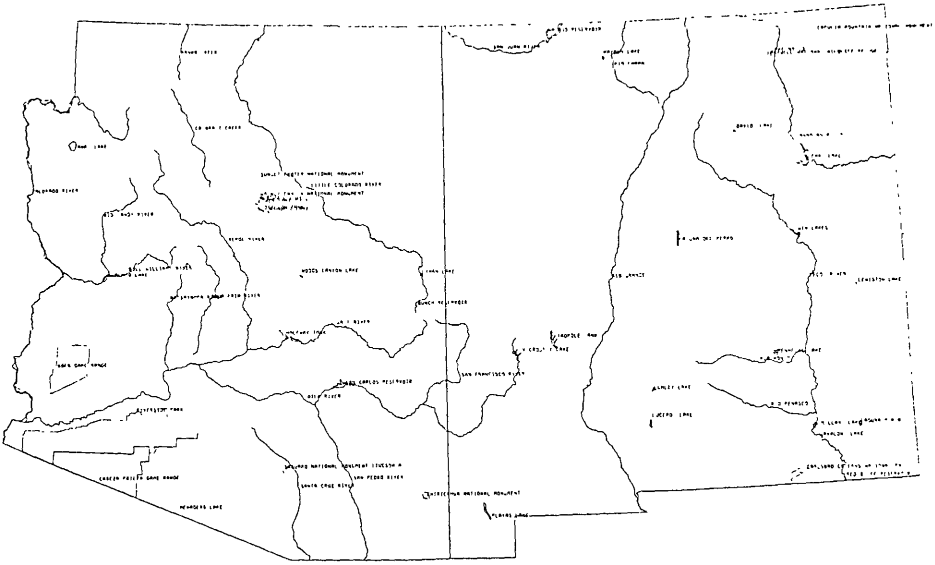


Figure 3

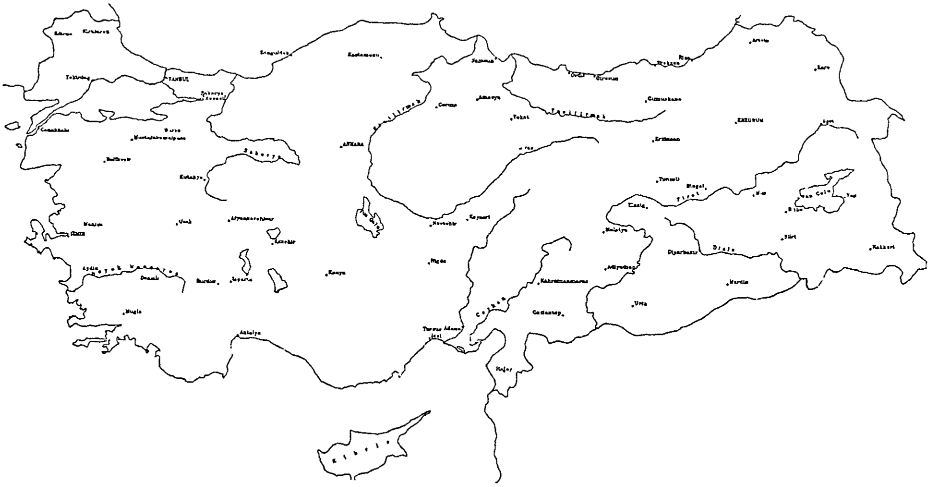


Figure 4

Identification of several names within a given area during the administrative boundary processing also points out this fact.

FUTURE OF THE RESEARCH

The creation of names data bases was the first phase of this research. The subsequent phases will deal with the actual placement problem. Prevention of overplotting of point names and features within an overlay will be the next step of the project. The algorithms developed during the author's Ph.D. research will be expanded to determine whether they are feasible in a large scale operation. The phase following this will be the development of software to place names within areal features and along linear features. The algorithms for these were also developed during the author's Ph.D. research. Figure 4 shows one of the maps produced utilizing those placement algorithms. The goal here will be to duplicate those results in a larger scale effort.

The future of the names data bases is quite promising. The effort so far has shown that two different data bases such as the GNIS and DLG can be manipulated successfully. Matching of names and lines for linear features was the most successful of these processes. More work needs to be done to produce improved results for areal features.

The data base attributes for point names were manually entered. Since Rand McNally stores its Commercial Atlas data in digital form, it would be feasible to acquire this data and write software to generate these files using automated techniques. The variable "remoteness factor" requires clearer definition and understanding. It is very difficult to quantify this variable and therefore to automate its generation process.

The topic of type size and style selection has been left out of this discussion because other research activities are addressing it and it is well beyond the scope of our present research. We feel that, for the present time, development of techniques for automated name placement outweighs the presentation of the type. This aspect of name placement will be improved by the time the placement techniques are developed.

CONCLUSIONS

In this short presentation an attempt was made to describe what has been done during nine months of research. By definition, research many times is a trial and error process. We have had our share of trials in this research so far. The diverse nature of the two data bases made this initial phase even more challenging and we believe this effort shows that even with the vast differences in scale of the two data bases, successful results have been achieved. This is especially true for the linear names. More linear features could have been identified if we had chosen to include more names from the GNIS files or used lower level classification attributes from the DLG files. Areal names processing unfortunately was less successful than we had anticipated for the reasons mentioned earlier.

As the present session of Auto Carto V shows, the interest in automated name placement is growing and the future looks quite promising.

REFERENCES

1. Pinhas Yoeli, "The Logic of Automated Map Lettering", The Cartographic Journal, Vol. 9, No. 2, December 1972, pp. 99-108.

2. W. T. Wilkie, "Computerized Cartographic Name Processing", M.Sc. Thesis, University of Saskatchewan, 1973.

3. Stephen A. Hirsch, "Algorithms for Automatic Name Placement of Point Data", M.Sc. Thesis, State University of New York at Buffalo, 1980.

4. Roger Payne, "Geographic Names Information System", Auto Carto V, August 1982.

5. Naftali Kadmon, "Automated Selection of Settlements in Map Generalization", The Cartographic Journal, Vol. 9, No. 2, December 1972, pp. 93-98.

6. F. Töpfer and W. Pillewizer, "The Principles of Selection", The Cartographic Journal, Vol. 3, No. 1, June 1966, pp. 10-16.