

Database and Design Considerations for an Online Urban Atlas

Yanlin Ye and Robert Cromley

Abstract: *The U.S. Bureau of the Census Urban Atlas project in the 1970s was an attempt to provide the public access to maps of selected census statistics. The technology of the time, however, only permitted a very limited number of census characteristics to be displayed in an atlas and at only one census geography—census tracts. The atlases also had limited availability and accessibility. Today the World Wide Web (the Web) provides a platform for an online version of the Urban Atlas concept that can be available to a global community for accessing the full release of census characteristics and geographies. This paper presents the system and database design and implementation of such an urban atlas for the state of Connecticut. Using a full database approach for both census statistics as well as geographic units, differing study areas, geographic units of display, and original or derived census variables, can be easily defined and mapped. The database platform also permits a future expansion of data services such as downloading as well as analytical services, including statistical analysis that may be of interest to the user community.*

Introduction

The U.S. Bureau of the Census Urban Atlas project in the 1970s was based on selected computer-generated maps for different metropolitan areas published in document form. Of the 269 standard metropolitan statistical areas (SMSAs) that existed in the United States in 1970, only the largest 65 SMSAs were included in this series of atlases. Each atlas was based on the 1970 census data and provided a graphic presentation of selected statistics at the census tract level for an individual SMSA. Only 12 choropleth maps with the same mapping variables were repeated for each atlas.

Although the Urban Atlas project prepared maps from digital databases of census statistics and census tract boundaries, the technology of the time only permitted an output strip of microfilm, from which press negatives were made for printing the atlases (U.S. Bureau of the Census and Manpower Administration 1974). The printed atlases were limited then in terms of the number of census characteristics displayed and the number of census geographies that could be displayed (only census tracts). In addition, as static displays, the maps were not able to be reused and were only one possible presentation of the data. More importantly, the atlases had limited availability and accessibility. Most atlases resided in federal depository libraries, and although available, the general public would need to plan a special trip in order to use an atlas. Finally, updating such urban atlases is energy-, time-, and money-consuming and the urban atlas series was not repeated for the 1980, 1990, or 2000 census.

However, there is still a need for atlases of census information, and printed atlases are still being produced such as the recent *Mapping Census 2000: The Geography of U.S. Diversity* (Brewer and Suchan 2001). Such printed atlases still are limited in terms making available to the public the full range of information collected and processed by the Bureau of the Census. Today though, the Web enables an

online version of the Urban Atlas concept to be available to a global community that can access the full release of census characteristics and geographies. This paper discusses the design and implementation of such an urban atlas for the state of Connecticut.

Web-based Map Dissemination

The development of the Internet and the Web in the 1990s presented a new media for information dissemination allowing different methods for communicating map information. In the general model of cartographic communication, static maps are being replaced by interactive ones that allow the map user more control over how and what information is depicted (Peterson 1995) as well as multimedia presentations of that information (see Cartwright, Peterson, and Gartner 1999). All forms of mapping products are being redefined. For example, Slocum (1999) has defined an electronic atlas as “a collection of maps (and database) that is available in a digital environment. The more sophisticated electronic atlases enable users to take advantage of the digital environment through a variety of means, such as Internet access, data exploration, map animation, and multimedia. Electronic atlases may also permit users to create their own maps and analyze spatial data.” (Slocum 1999, 233.)

Publishing these atlases and maps on the Web can take many forms. Kraak (2000) classifies maps published on the Web as being either static or dynamic. Static maps are bitmap images that are primarily used for display but that also can be used as a spatial index interface to other information (Kraak 2000). Dynamic maps are animations of spatial processes that also can be used in a view only environment or in an interactive one. Static maps typically are viewed by a Web browser using HTML tags, and can be used as a spatial index using the coordinates of polygons to define a “clickable region” (Kobben 2000).

By using the Common Gateway Interface (CGI) and integrating HTML code with scripting languages, server software can be developed to allow information transfer from the user to a server application. In this manner, users can query server databases and retrieve information “on-demand” either in the form of map or data table. This is the approach taken in the construction of the Online Connecticut Urban Atlas (OCUA), whose system design and database design are described in the next two sections.

System Design

The OCUA is designed as a server-side application, which means that most services are on the server side; users at the client side access data and mapping services provided by the OCUA via the Internet using a Web browser. This server-side-application architecture has some advantages over a typical client-side-application architecture. First, the OCUA system processes data locally and sends results to the client side, while a client side application downloads data from a data server and processes the data locally at the client side. However, processing user’s mapping requests may involve a large set of spatial and/or attribute data, and transferring the large set of data over the Internet could be very slow, especially for those who use dial-up connections. The OCUA system does not transfer raw data from the server side to the client side. Instead, it sends a minimum amount of information to the client side, receives a user’s requests, and then sends back to the client side the requested maps, which are in a compact image format such as GIF. Second, because most service procedures are on the server side, no software or plug-in installation is needed on the client side. The minimum requirement for the client-side system is access to the Internet. This makes it possible for those who are using public computers, such as college students, to make use of the system. The requirement of installing software or plug-in could prevent users without administrator privileges to use such software or plug-ins. Also, updating the system or its components is relatively simple, because it does not involve the client side. Finally, user’s interaction is achieved by using forms and client-side scripts in dynamic HTML pages. The server-side-application architecture was once considered less capable of interacting with users for a long time. However, with the development and adoption of new Internet techniques, especially the Dynamic HTML technique, which includes the Document Object Model (DOM) and scripting languages such as ECMA Script and VB Script along with some existing techniques such as HTML forms, user interaction can be achieved by embedding these techniques into HTML pages.

The system architecture of the OCUA makes the client side as simple as a Web browser, while the server side holds most of the function modules. The server side is further divided into three layers, including the CGI layer, the service layer, and the DBMS layer.

CGI Layer

On the server side, the requests from a user are directly handled by the Web server and then forwarded to corresponding scripts

on the CGI layer. CGI scripts interpret the user’s requests and call some specific service modules, which are on the service layer, to perform data processing and to generate maps, and organize results by generating HTML pages on the fly. Most CGI scripts are written in Perl (<http://www.perl.com/>), which is an open source programming language with useful third-party modules freely available, and is very suitable for text processing and Web development (Christiansen and Torkington 2002). Because Perl is an interpreting language and runs on a large number of platforms, including most UNIX variants and other systems like VMS, DOS, OS/2, and Windows, scripts written in Perl are platform-independent and highly portable.

Service Layer

In the CGI layer, scripts interpret the user’s requests for data and mapping and call services on the service layer. Some simple services such as data catalog queries are tied with CGI scripts on the CGI layer. For example, there is a list of census data sets available on the server, and after the user selects the data set of interest, the list of all census tables in the selected data set are dynamically generated by CGI scripts. At the same time, most complex and efficiency-demanding tasks, especially data compilation and map generation, are handled by some specific, stand-alone procedures on the service layer.

Service procedures, unlike CGI scripts, are developed using C/C++, instead of Perl. Perl is good for high-level application development, especially light-weighted CGI scripts. However, it is not as good as C/C++ for processing large amounts of binary data. Furthermore, Perl scripts are less efficient than programs written in C/C++. GNU’s Compiler Collection (GCC, <http://gcc.gnu.org/>) is used for compiling C/C++ source code. Due to GCC’s multiple platform support, including support for most operating systems such as UNIX and its variants, DOS/Windows, etc., the migration of programs developed in C/C++ between different operating systems often means only re-compilation and some minor modification of the original source codes.

Some open source software from third parties is used in the service layer as middleware. Among them the most important one is ImageMagick (<http://www.imagemagick.org/>), which serves in the server layer as a graphic server and provides map-drawing functionalities.

DBMS Layer

The data, as well as the CGI scripts and the service modules, are the most important components of the Online Connecticut Urban Atlas. All census data for the year of 1990 and 2000 and the geographic data for the state of Connecticut are stored in one central database. This database design, which is discussed in detail next, provides the convenience of data management and compilation and takes full use of the power of modern Database Management Systems (DBMS).

MySQL is selected for the implementation of the OCUA. MySQL is an open source and free software package (<http://www.mysql.com/>). It has the ANSI SQL syntax support and also the

cross-platform support. A complete set of online documentation is readily available to program developers. Besides data management, the DBMS also provides to the service layer data access interfaces, through which the upper level service procedures access the census data and the geographic data.

Database Design

The database design involves the census data, the geographic data, and the coding tables, which link the geographic data and the census data together.

Census Data

As previously mentioned, all of the census data, as well as the geographic data, are stored in one central database. Each set of census data for a particular census year, such as 1990 Summary Tape File 1, 2000 Summary File 1, 2000 Summary File 3, etc, is organized as a conceptual sub-database. Individual census tables in each sub-database are combined as several larger tables, which are subsets of the census data set. For example, all census tables of 2000 Summary File 1 are stored in eight large tables, and most large tables have more than 1,200 data fields (Table 1). By combining smaller individual census tables into larger ones, the total number of tables is dramatically reduced and the efficiency of data queries is normally improved, for less JOIN operations are necessary when user queries involve data from different census tables. The combined census tables, the individual census tables, and the individual data fields are registered in several indexing tables.

Geographic Data

The coordinate values for state boundaries, MSA boundaries, and county subdivision (or town) boundaries were derived from a 1:24000 database of town boundaries in the Connecticut State Plane Coordinate System NAD-83 (Connecticut Department of Environmental Protection 1996). The boundaries were simplified so that each census unit is represented by a simple polygon. The coordinate values for census tracts and block groups were downloaded from the census Website, www.census.gov.

The boundary files for MSAs, county subdivisions, census tracts, and block groups, are also stored in MySQL tables. Each boundary file is organized as one table. Because the geographic boundaries are stored in a relational format, a fixed-length car-

tographic data structure is used. As choropleth mapping is the only current operation of the system, a non-topological, entity-by-entity data structure is sufficient. Each polygon representing a census unit is stored in its boundary table as a series of ordered records. Each record includes an x,y coordinate and its polygon or Geographic ID. This follows the data structure used by the SAS/GRAPH mapping system (SAS Institute 1988). This organization of the boundary files is also similar to the OpenGIS simple features specification for SQL. In this standard, up to five x,y coordinate pairs can be stored per record and more complex geometric features than simple polygons can be represented (Open GIS Consortium, Inc. 1999). The simpler structure is used here to improve processing time.

The Geographic ID is used in conjunction with coding tables to enable the dynamic definition of study areas.

Coding Tables

The coding tables are used to link the census data and the corresponding geographic data, as well as the geographic units at a higher level. The links are implemented through foreign keys. The census data are linked through a Logical Record ID, which is the common primary key of all combined census tables for a given set of census data, such as 2000 Summary File 1. The higher level geographic units are linked by referring to their Geographic ID, which are the primary keys of geographic tables. For example, census tracts are linked to MSAs and towns through MSA ID and town ID, respectively. This allows the user to select a study area for higher level geographic units. For example, selecting census tracts of certain counties or towns within a certain study area definition only requires an SQL SELECT statement. All census tracts within the towns of Hartford and East Hartford can be selected using the following SQL statement:

```
SELECT * FROM TRACT_CODE WHERE TOWN_ID
IN ('09043', '09064')
```

in which TRACT_CODE is the lookup table for the 1990 census tracts, TOWN_ID is the Geographic ID for the towns, and "09043" and "09064" are the IDs for the towns of East Hartford and Hartford, respectively.

The geographic look-up table for the towns includes fields for the town name, a town identification code, the identification code

Table 1. Segmentation of Census 2000 Summary File 1

Table Name	Census File Segmentation	Starting Table Name	Ending Table Name	Number of Data Fields
SF1_2000_A	01-06	P1	P16I	1331
SF1_2000_B	07-11	P17A	P35I	1134
SF1_2000_C	12-18	PCT1	PCT12C	1312
SF1_2000_D	19-24	PCT12D	PCT12I	1254
SF1_2000_E	25-30	PCT12J	PCT12O	1254
SF1_2000_F	31-36	PCT13A	PCT17I	1230
SF1_2000_H	37-39	H1	H16I	595

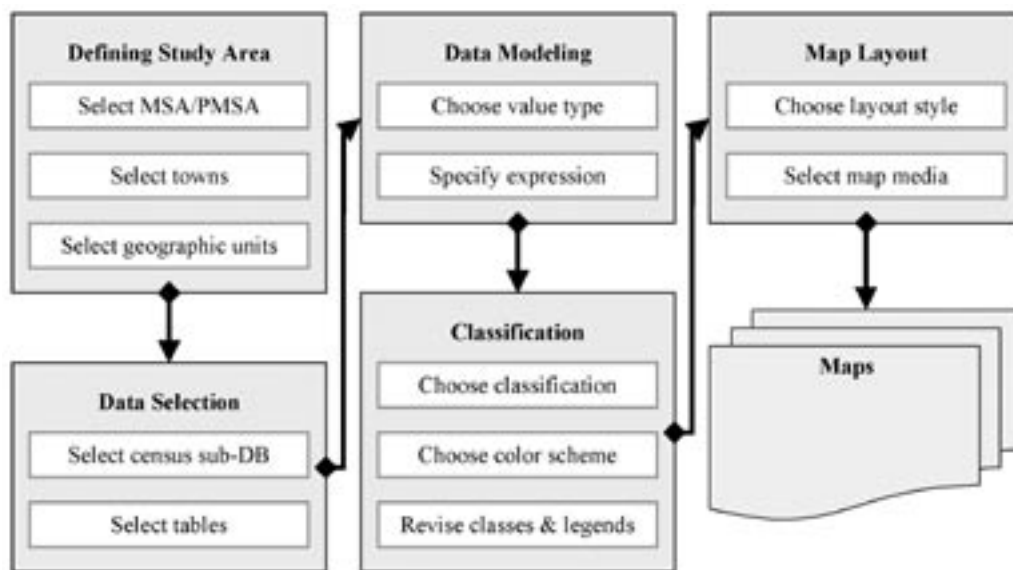


Figure 1. Flowchart of the application design

for the county in which the town is located, and the identification code for the state in which the town resides. The geographic look-up table for counties includes fields for the county name, the county identification code, and the identification code for the state in which the county resides. Although the state identification field in both tables includes only one value for Connecticut, this field is important in order to select objects at the state level. The presence of this field also permits the system to be scalable to a larger region, such as New England, at a future date.

Application Design

The application design organizes how information is collected from a user as he or she decides what map to produce (Figure 1). It involves taking the user through the steps of defining the overall study area and the geographic units of inquiry, and retrieving the potential census characteristics for display. The user can then define the specific statistic to be mapped and the classification scheme for presenting its spatial distribution. Finally, the user has some control over the elements of the final map layout.

Study Area and Units of Analysis

For now the user can select one or more MSAs or PMSAs in the state of Connecticut. The list of all MSAs or PMSAs in Connecticut is dynamically generated from the MSA/PMSA look-up table on the server. This makes the scripts adaptive to future data updates. When newly updated MSA/PMSA data are available, only the look-up tables for the new MSAs/PMSAs, but no script modifications, are necessary.

Once the user submits his or her MSA/PMSA selection, a list of towns within the selected MSA/PMSA is displayed on the next page. These towns are ordered either by town names only or by MSA/PMSA names and then town names, depending on the

user's choice. These towns are all selected by default. However, the user can narrow down the study area to several towns by selecting only those towns of interest, thus increasing the flexibility of defining a study area.

After defining a study area, the user selects the geographic units of analysis. Three options, including the towns, the census tracts, and the census block groups, are available, while the Census Bureau's Urban Atlases have maps of several statistics only at the census tract level. The information about the user-defined study area and the selected units of analysis is passed on to the next step, data selection.

Data Selection

A list of all available census data subsets is generated, according to the previously defined study area and geographic units of analysis. The user can select one set of census data and then all census tables in that data set are listed on the next page. The census tables are the minimum data selection units and the user cannot further select individual data fields within a census table. Otherwise, the list of individual data fields could be too lengthy. One or more census tables can be selected at one time.

Value Definition

Based on the selected data, the user then can define the value to be mapped. In order to simplify the process of value definition, value expressions are grouped into the following four categories: single field, field as percentage of its universe, ratio of two fields, and a generic algebraic expression.

The first category is the simplest format for value expressions. Some census statistics, such as mean and median values, are ready to be mapped without further calculations. These statistics fall into the first category and the value can be directly mapped by

choosing a single field from the list of all selected fields. However, most of the census statistics are total values, such as race, sex by age, household type by household size, etc. In choropleth mapping, it is common to calculate these totals as a percentage of some larger universes before mapping. This instance falls into the second category. The system will automatically search the metadata for the universe of the selected field and then calculate associated percent values.

Ratios, such as the ratio of males to females, fall under the third category and can be modeled by selecting one field (male) as the numerator and another (female) as the denominator. The system has a tool that can test any field to determine if that field is suitable to be designated as a denominator.

These three categories are special cases of a generic algebraic expression. Each of the three situations could be modeled by specifying them as an algebraic expression. For example, the sex ratio in the year of 2000 can be calculated using the expression $(P012002/P012026)$, where $P012002$ is the total population of males and $P012026$ is that of females. Some mathematical functions, such as $\text{power}()$, $\text{exp}()$, $\text{log}()$, etc., can be used to construct complex expression.

Classification and Legends

The selection of appropriate classification methods and legends are of key importance for demonstrating the output of the user-defined models. As noted earlier, the maps presented in the original urban atlases were only one portrayal of the data. As in most desktop mapping packages, the Online Connecticut Urban Atlas system has support for a variety of classification algorithms and legend options.

The OCUA Website currently supports four classification methods, including equal intervals, natural breaks, quantile, and standard deviation, that offer capabilities for the user to experiment before choosing a final thematic map. The first three classification methods require preferred number of classes as parameter, while the last one requires the standard deviation interval. As an alternative to classification, the OCUA also allows the user to create a classless choropleth map (see Kennedy 1994 for a discussion of the advantages of this approach).

The user can select a sequence of colors for the classes to be identified. For choropleth map without classification, the colors will apply to the legends of sampled value sequence. Three kinds of color schemes are available, including monochromatic colors, dichromatic colors, and continuous colors. Monochromatic color sequence is generated using one single color (red, green, blue, etc.) with variant brightness. Dichromatic color sequence is generated using two colors, one with brightness varying from dark to bright and the other with brightness varying from bright to dark. Continuous color scheme uses two or more colors and is generated by linear interpolation in the RGB (red, green, and blue) color space. The sequence of colors also can be inverted. The OCUA dynamically renders a preview of the color sequence while the user makes his or her choice on colors.

When the selections of classification and color scheme are

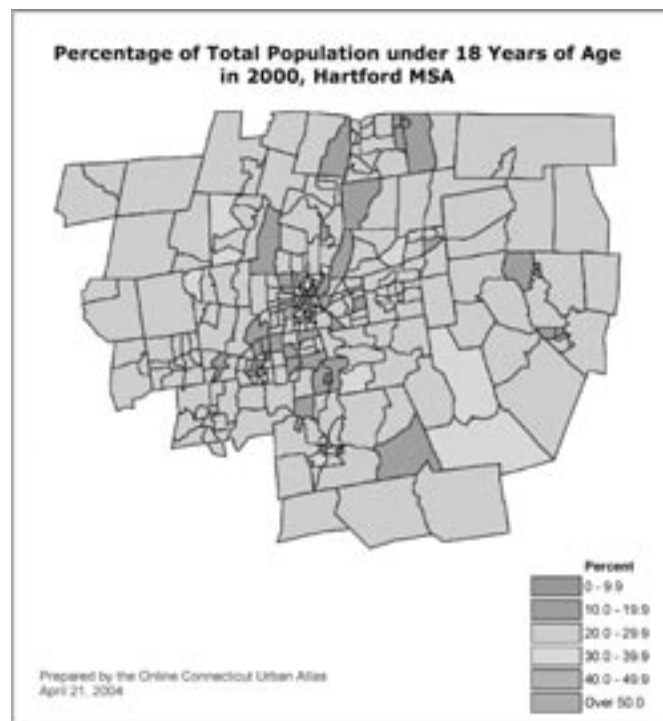


Figure 2. Percentage of total population under 18 years of age in 2000, Hartford MSA

submitted, the OCUA generates a series of classes as well as their class labels and legends, based on the data model specified on an earlier step. Although the classes are automatically generated by the system, all information about the classes, including classification rules, class labels, fill colors, etc., can be modified by the user for his (or her) own purpose. If the user manually changes the classification intervals, though, some mapping values may fall into none of the classes. These values will be displayed with a white fill color by default. In this case, the user may want to include the "Other" class to hold these non-classified values. The user also can modify the color of boundaries of geographic units, which will be displayed with a black color by default.

Layout

The last step of creating a choropleth map is defining the layout of the map. This includes the creation of some other map elements, such as map title, scale, north arrow, and positions of the elements. The current version of the OCUA system only has limited support for map element definition and positioning, but more options will be included in a future version.

The user can select different types of media in which the map will be displayed. Two media types are supported now, including screen and printer. When the user chooses to prepare the map for displaying on screen, all functional or look-and-feel elements of the Web pages, such as the logo of the UCCGIA, hyperlinks, and background colors, will be rendered along with the user-defined map. When printer is selected as the desired media, these elements will be excluded and only map elements, such as map

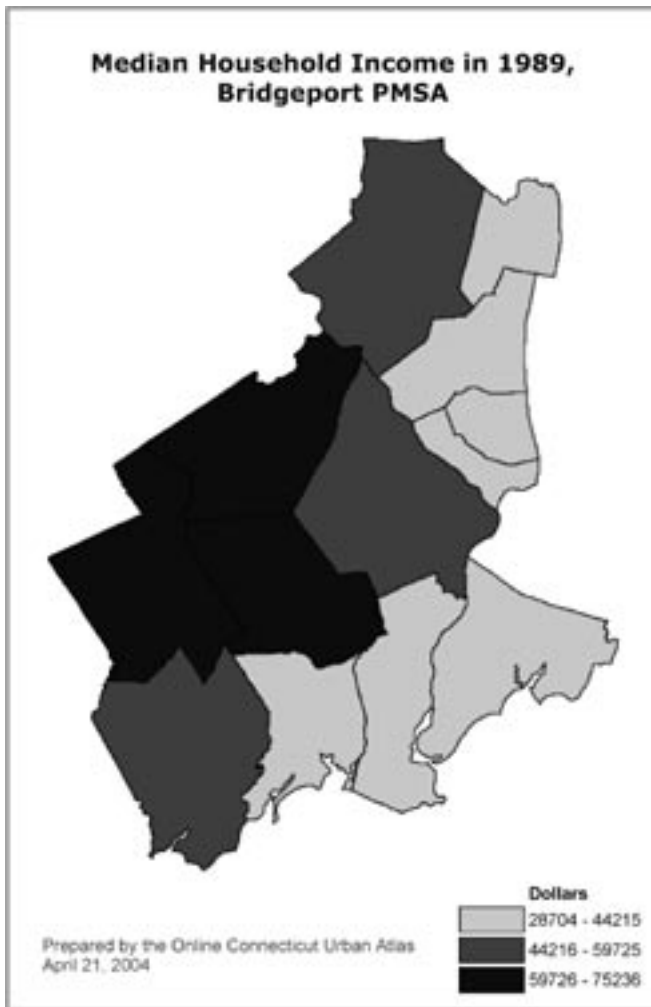


Figure 3. Median household income in 1989, Bridgeport PMSA



Figure 4. Median household income in 1989, Bridgeport.

title, the user-defined map, and the map legends, will be rendered for printing on a paper page.

It is likely that one will refine the map he or she creates if the mapping result is of interest. One can always use the “Back” function of the Web browser to return to a previous page to adjust the parameters for map creation.

Mapping Examples

The first mapping example is a reconstruction of a statistic presented in the original Urban Atlas for the 1970 Hartford SMSA—percentage of the total population that is under 18 years of age (Figure 2). This map is produced by selecting the entire Hartford MSA as the study area with the census tract as the resolution unit. The mapped value is prepared by the algebraic expression $(P001001 - P005001) / P001001 * 100.00$ (total population minus population aged 18 and over divided by total population).

The next example is a map of median household income in 1989 for the Bridgeport PMSA mapped at the town level. In this example, data from the 1990 decennial census is displayed and the mapped value is a single field, P080A001 (median household

income), taken directly from the census.

The final example is the same data value displayed for the inner city of Bridgeport mapped at the block group level. Five classes are chosen for this map because of the greater number of the observation units. By contrasting Figures 3 and 4, it can be seen that the median income within Bridgeport has as much spatial variation as the median income within the entire PMSA.

Conclusion

Compared to the U.S. Bureau of the Census Urban Atlases, the Online Connecticut Urban Atlas is much more flexible in different ways. The study area and geographic units can be easily defined. Not only the whole MSA/PMSA, but also partial sections of the MSA/PMSA, such as one or more towns, can be designated as the study area. The theme of a user-defined map can be any statistical variable of census data or a value derived from one or more census fields. The user also has the options for how the theme is presented by applying different classification methods and different fill colors.

The OCUA integrates census data and the geographic data with mapping functionalities. This design nearly eliminates the

cost of data collection, conversion, and compilation, because the data are already there ready for analysis. Both professionals interested in urban analysis and the general public could benefit from the system. If an urban planner wanted to examine some social factors within an urban area, the OCUA can be used to conduct a pilot study without the planner having to collect any census statistics or geographic outlines. For the general public, creating maps of their own interest is relatively simple and requires no training. The main necessary background is some understanding of data classification systems for mapping and the user always can accept the default mapping parameters if in doubt. The simple, easy to use capability presented by the system is a necessary early step in the education of society to the benefits of the routine use of mapping. In addition, the general public has direct access to census data regarding their locality without the need to travel to a census repository.

The Online Connecticut Urban Atlas, which is available at the Website of the University of Connecticut Center for Geographic Information and Analysis (<http://www.uconn.edu/ocua.html>), is more accessible than the Bureau of Census' Urban Atlases, due to the accessibility of the Internet. The OCUA is easy to update, not only the system itself, but also the data on the server. Because updates of the system have nothing to do with the client side, the process of updating is fairly transparent to the users of the system, except for noticing an increased functionality or data availability.

Although the OCUA has been initially developed for the state of Connecticut, the issues of portability are fully addressed in the stage of system design. Its capabilities can be easily packed and made available to other states, which share the same database structure if the Census statistics are the major source of mapping data. With some minor modifications to the database design, the system also can meet the mapping needs of agencies with data from other sources.

Forward Thinking

In a world of new cartographic possibilities, more possibilities exist for information stored in the spatial databases underlying atlases and interactive maps than the visual world of maps and exploratory data analysis. Statistical packages, spreadsheet programs, and database management systems, as well as word-processing systems make the new world multi-analytical as well as multimedia. Map users may be geographic data users who need maps for answers to certain questions and statistical analyses for answers to others. The growth in geographic information system (GIS) use also means that many geographic data users would have to have available software to view cartographic databases if these databases were available to them. In DiBiase's (1990) model of private visual thinking versus public visual communication, it is the spatial database that provides the scientist or cartographer with the most flexibility in the private world of exploration and analysis. Without direct access to the database, the user is always restricted to the viewing and analytical capabilities of the Internet software. Future data services will include data downloading so

that the researcher is unrestricted by the current functionality of the system.

For the public without access to specialized analytical software packages, some analytical services, including basic descriptive statistics and spatial measurement, that are useful to municipal officials and urban planners will be added. Other databases besides the decennial census of population and housing also will be added. The goal eventually is to transform the Online Connecticut Urban Atlas into an Online Urban Information and Analysis Server for Connecticut.

About the Authors

Yanlin Ye is a PhD candidate in the Department of Geography, at the University of Connecticut.

Robert Cromley is the Director of the Center for Geographic Information and Analysis and a professor in the Department of Geography, at the University of Connecticut.

References

- Boutell, T. 1996. CGI programming in C & Perl. Reading, MA: Addison-Wesley.
- Brewer, C., and T. Suchan. 2001. Mapping Census 2000: The geography of U.S. diversity. Redlands, California: ESRI Press.
- Cartwright, W., M.P. Peterson, and G. Gartner, eds. 1999. Multimedia cartography. New York: Springer-Verlag.
- Christiansen, T., and N. Torkington. 2002. General questions about Perl, <http://www.perldoc.com/perl5.8.0/pod/perlfaq1.html>.
- Clarke, K. 1995. Analytical and computer cartography. 2d ed. Englewood Cliffs, N.J.: Prentice Hall.
- Connecticut Department of Environmental Protection. 1996. 1996 town boundaries. [Online]. Map and Geographic Information Center. Available at <http://magic.lib.uconn.edu> [October 1, 2000].
- DiBiase, D. 1990. Visualization in earth sciences. Earth & Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences 59(2): 13-18.
- Kennedy, S. 1994. Unclassed choropleth maps revisited/Some guidelines for the construction of unclassified and classed choropleth maps. Cartographica 31(1): 16-25.
- Kobben, B. 2000. Publishing maps on the Web. Chapter 6 in Kraak, M-J., and A. Brown, eds., Web cartography. London: Taylor & Francis.
- Kraak, M-J. 2000. Settings and needs for Web cartography. Chapter 1 in Kraak, M-J., and A. Brown, eds., Web cartography. London: Taylor & Francis.
- Marble, D. 1987. The computer and cartography. The American Cartographer 14(2): 101-103.

- Open GIS Consortium Inc. 1999. OpenGIS simple features specification for SQL, Revision 1.1. OpenGIS Project Document 99-049.
- Peterson, M. 1995. Interactive and animated cartography. Englewood Cliffs, N.J.: Prentice Hall.
- SAS Institute Inc. 1988. SAS/GRAPH user's guide, Release 6.03 Ed. Cary, N.C.: SAS Institute.
- Slocum, T. 1999. Thematic cartography and visualization. Upper Saddle River, N.J.: Prentice Hall.
- U.S. Bureau of the Census and Manpower Administration. 1974. Urban Atlas, Tract data for standard metropolitan statistical areas: Hartford, Connecticut. Washington, D.C.: U.S. Government Printing Office.