CREATING DATA FOR USE IN ARCVIEW

INTRO

ArcView is a very powerful software and comes to schools and libraries bundled with a lot of data. However, it cannot possibly come with all the data that anyone might ever need. New information is being created all the time, and users develop new desires constantly. The challenge for teachers and librarians is to help users find and access, or generate, this new data. Sometimes the process is very straightforward; sometimes it involves transforming data from one format to another.

1. DATA VS. PROJECTS

ArcView uses different kinds of files for different things. Remember that a project file is just a recipe for the data that is being displayed; it is not the actual data itself. Thus, when you share a project with someone, that person needs to have the data sets on which the project relies. It is not sufficient simply to have the project, just as it would not satisfy one's craving just to read a recipe for chocolate chip cookies.

2. FEATURE VS. ATTRIBUTE DATA

The first piece to clarify is the nature of the information. Does the information describe a feature, or is it the location of the feature? That is, is this the newly discovered island of Atlantis, with a specific shape and an actual position on the planet, or is it characteristics of the island, such as population and percent forested?

"Feature data" (the shape and location of the entity) tends to be fairly consistent, though political boundaries do indeed change, and rivers carve new courses on occasion, and we build new streets all the time. "Attribute data" (characteristics of an entity) are very numerous and may change quite often, such as the population of a school district, the air quality at a specific point, or the speed limit along a stretch of road.

Creating new feature data is not difficult, if the need for precision is low. Ensuring absolute precision, though, can be a time-consuming process. Making sure that a line zigs and zags in just the right spots on the planet often demands intense study.

Creating new attribute data, on the other hand, is extremely easy, and often what the user truly seeks anyway. For instance, students and library patrons may be less concerned about the exactly perfect outline of each state, and more concerned with representing the results of the latest election in each state.

This document will focus on the creation of ATTRIBUTE data, with only brief attention to creating new FEATURE data.

3. RELATIONAL DATABASES

ArcView is a relational database. That is, given a set of features which have unique identifiers (such as states with unique 2-character postal codes) and a data table which contains information about those features and includes the same unique identifiers (i.e. the table of states includes a column with postal codes), ArcView can "relate" the information, or join the features and attributes together.

This capacity is extremely important. It means that, for any given feature set, as long as there is some way of identifying features in a unique fashion, any external data table with parallel identification can be joined to the feature set. For instance, as long as you have a way to identify unique counties in the U.S., you can create and join tables which describe population patterns, economic data, historical attributes, ad infinitum.

Clearly, the key is having these unique identifiers in the geographic features and the same ones in the attribute data. As long as these "twins" exist on each side, ArcView can join together any number of tables with their appropriate features.

4. NAMING SCHEMES

Fortunately, there are standards for identifying many different geographic features. States, for instance, have 2-character postal codes. This is handy because the code is clear, precise, and universally understood. Other abbreviations of states might not work; I might abbreviate Missouri or Mississippi or North Dakota in a fashion that might differ from yours. And ArcView is extremely particular -- twins must be absolutely identical, not just close. "Minn" is not the same as "Minn." because only one name includes a period.

So it is extremely helpful to use standard identifying codes when they are available, both in the feature data and in the external attribute data. For schools and libraries accessing data from the ESRI bundle, these identifying codes are present for almost all feature types. All you need to do is identify what coding scheme is used and make sure that you can match your data table with it.

States have a 2-character postal code, but what about counties? How do I distinguish between Boone County in Missouri and Boone County in Kentucky? The U.S. government has devised a system (Federal Information Processing Standards, or FIPS) which identifies each state with a 2-digit numerical code, and each county in each state with a 3-digit numerical code. Merging these gives each county a unique 5-digit numerical code. There are, for example, many counties known as "121", but only one known as "13121" -- Fulton County in Georgia. Georgia is state "13" and Fulton County is "121" inside Georgia. (Incidentally, there are eight "Fulton" counties in the U.S.) There are FIPS codes for countries, states, counties, cities, census tracts, block groups, and blocks.

However, it is also important to recognize that, to a computer, the number "6" is very different from the alphanumeric sequence "06". ArcView sees these as not even close to being twins. One is a true number, the other is an alphabetic look-alike, referred to as a "string" ... not a match at all. When attempting to join data, it is crucial to ensure that the type of identifier is the same. Otherwise, you may end up trying to match the number "36" to the string "36" or to the string "036", and none of these would match to either of the others.

5. FIELD NAMES & DATA DICTIONARIES

ArcView relies on standard styles of data. Attribute information can be accessed from tables saved in a "dBASE" or "delimited ASCII" (text) format, or in "INFO" format from ArcView's parent software ARC/INFO. Data prepared in other software packages must be converted into one of these formats.

These formats impose certain restrictions. Field names (column headers) can only be 10 characters long in the native table. ArcView allows you to give fields different names once inside the program, but their native restriction applies when creating the table in the first place: 10 characters, max.

It can be challenging to translate some detailed information into only 10 characters. For instance, how do you distinguish these typical choices: (a) total individuals divided into 10 age groups, (b) percent of population divided into those same 10 age groups, (c) data for all categories broken down by race/ethnicity, and (d) data for all of the above at three different times? One might use "p_1_api_90" to represent "percent of population in age group #1 who were Asian or Pacific Islanders, in 1990."

Naming conventions vary. Typically, fields start with an alphabetic rather than numeric character, and no punctuation marks except underscores, dashes, dollar signs, or percent signs are used.

For any such complex coding scheme, obviously, it is important to have a "data dictionary" which defines each field. This is a separate document which the user could access at any time that would provide assistance in understanding what a field is, or which field name to use when trying to find "absolute numbers of children aged 0-4 in 1992"

(Additionally, while Macintosh computers can use long file names, other computers are not so flexible. In order to assure greatest use by the most people, it is important for Mac users to name files according to the DOS standard of "eight-dot-three" -- filenames of no more than eight characters, followed by a period, followed by a three character extension which identifies the type of file.)

6. DOCUMENTATION

Along with providing translation between computer names and common language equivalents, it is important to document the source of the data. Are these numbers which were produced out of thin air as guesses, or are they based on careful scientific analysis by an organization noted for their skills? What procedures were used to determine if something fit into one category or another? If a user is confused by some data, where can s/he turn for assistance?

All these issues are important for users to make sense of data which they have not produced. Even the data producers themselves can, over time, forget key details of production. Making sure that the sources of information and procedures for processing are carefully catalogued may mean the difference between users conducting powerful analyses or simply ignoring the data.

7. SCRATCH VS. TEMPLATES

When thinking about a new project, it may be tempting to start just creating tabular data from scratch. However, this can actually mean doing more revisions and spending more time than might be necessary. Starting to work from a template may be a much easier route because it can prevent problems and help you think of small details from the start.

Fortunately, templates are readily available. Often the easiest route is to rely on a table that is already well-known. Make a copy of this file, strip away from the copy any unnecessary data fields, and just start adding new fields and data elements as needed. Be careful when deleting data fields. It takes only an instant to delete a field, and it can take many minutes to replace it, if it later appears valuable. Experienced data crunchers often leave the deleting of questionable fields to the end.

Most important in using existing tables is the capacity to rely on feature identification codes that are known, understood, and tested. Using a template can help ensure proper coding. If you can avoid typing in even a few FIPS codes, or if you can ensure that a string field of numeric characters doesn't accidentally get stored as numbers instead of text characters, you will be able to avoid headaches.

8. TXT OR DBF

It is pretty easy to create data files in a variety of software packages. You can create the file inside ArcView, but often it is easier for people to use more familiar word processing, database, or spreadsheet applications. Even software as basic as Windows Notepad, or SimpleText on a Mac, or even AppleWorks on an AppleII, can be used to create attribute tables which can be imported later into ArcView projects.

The package used will control the options available for saving the file. ArcView will add tables in either DBF or TXT format, but there are advantages to each. TXT files are simpler, more commonly available, more importable in other software packages, and they require less space than DBF files for a given data set. DBF files are "smarter", they draw faster in ArcView, and they can be directly modified on the fly inside ArcView.

This last point bears repeating. TXT files cannot be modified within ArcView. In order to change a spelling or add a missing number in a TXT file, you must return to a text editor, or use ArcView first to convert the data set from TXT into a DBF file for editing inside ArcView. This is not a problem, but it is another set of steps to accomplish.

Regardless of application used, the principles are similar. Create a table consisting of rows (records) and columns (fields). In a word processor, the columns may not line up as anticipated because of word wrapping. Key is to ensure that the columns stay consistent. If the number of fields is small, this can be easily accomplished, using a comma or tab to demarcate movement from one field to the next, and the ENTER/RETURN key to note change from one record to the next. For this reason, when constructing tables using a word processor, it is often helpful to use the "hanging indent" feature, if available.

Use of both commas and tabs can lead to unpredictable results. It is best to ensure that there are no commas or tabs WITHIN cells, and only commas OR tabs are used between cells. Then, when the data is exported and imported, it will be much easier to modify, as needed.

9. CONTENTS

Of course, each record must have the same number of columns and a unique identifier. If particular cells have no data, it is best to fill them with an obviously impossible value which nonetheless is formatted properly. Empty cells can lead to unpredictable results, so it is best to have some indicator of the appropriate type (i.e. number or string) in every cell. (ESRI uses "-99" or some such obvious indicator as a "null value.") Thus, you will need to make sure that a data dictionary accompanies your data set.

Applications in which you are preparing data will vary in terms of ability to identify and set the format of a field. Sometimes, you will want to indicate that numerals are a string rather than a number. Different packages accomplish this in different ways. If any of the numeric strings will begin with "0", it is especially important to explore the process of creating and saving the data properly. Remember, "36" and "036" are not identical twins.

Recall, too, that ArcView has some extremely powerful table/spreadsheet functions. Data sets which have been saved in an imperfect fashion may perhaps be changed and even enhanced more easily inside ArcView than in the original application. For instance, you might have data in one column that you would like to modify according to data in another column. ArcView can accomplish this and much more.

ArcView's robust database and spreadsheet tools provide tempting options. It is important to remember the limitations of data accuracy. Generally, "processed information" can only be as specific as the weakest piece of data. For instance, if you know there are exactly one million people in the state and exactly 10,000 square miles, it would not be accurate to say that each of the state's five counties contains 200,000 people. If you create data about your community, it is important to inform the user about the nature and accuracy of the data, in order to avoid improper decisions, such as where a homeowner should plant a garden.

10. CREATING FEATURE DATA

So far, we have referred mostly to ATTRIBUTE data. ArcView allows you to create new FEATURE data as well. You can create points, lines, and polygons. (Spatial Analyst in ArcView 3 for Windows also allows creation of grids.) Using "heads-up digitizing," you can draw these elements inside an ArcView view. Naturally, you should include some indication of how these features were created.

You can also create point data in other applications such as word processors or databases. Creating these points involves adding the external data table into the project and then creating an "event theme" in a view.

ArcView can read a data file which contains street addresses and match these against a street network to create points. In order to accomplish this, it is important to know how the street addresses are constructed in the existing network, because there are many possibilities. However, once formats match, it can be very easy for students and teachers to construct extremely powerful databases.

Even easier still is creating an "X-Y" point theme from a table that includes longitude and latitude fields. For instance, this table could be brought into ArcView and mapped very easily:

```
SITE,LAT,LONG,COMMENT
"1",45.0,-93.2,St.Paul MN
"2",38.8,-77.0,Washington DC
"3",0,0,Equator and Prime Meridian
"4",71,-157,Barrow AK
```

(Notice in this example that the fourth field, "COMMENT," contains no commas in the cells, since commas would be interpreted as indicating a new field.)

If the table uses "LAT" and "LONG," or "Latitude" and "Longitude," as field names, ArcView will automatically identify these fields and use them. If the Lat/Long fields carry different names (such as "DegrLat" or "Internal Point Longitude"), you can direct ArcView to use particular fields for Lat/Long references. Once the fields are set, ArcView will read the data table in the view and be able to draw the points, which can then be saved as a shape file.

CONCLUSION

Creating data for use with ArcView is an extremely easy and powerful way to enhance ArcView's existing data sets. Local information and up-to-date statistics can be incorporated with ease by understanding and following these basic principles. Computers not powerful enough to run ArcView can nonetheless become vital links in a class project through being put to work as data creation stations. This is one of the easiest, most inclusive, and most effective ways for ensuring that more students at any one time can be actively engaged in working with spatial information.

Environmental Systems Research Institute, Inc.
380 New York Street
Redlands, CA 92373-8100 USA
voice: 909/793-2853
fax: 909/793-5953
e-mail: info@esri.com
http://www.esri.com

ESRI Schools and Libraries
1305 Corporate Center Drive, Suite 250
St.Paul, MN 55121-1204 USA
voice: 612/454-0600
fax: 612/454-0705
e-mail: k12-lib@esri.com
http://www.esri.com/base/markets/k-12/k-12.html