D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

DATA INTERCHANGE: HELPING TO SURVIVE EXISTING TRANSFER STANDARDS

David A. Hastings and Carol N. Gerlitz

World Data Center - A, Boulder Centers
National Oceanic and Atmospheric Administration
National Geophysical Data Center
325 Broadway, Boulder CO 80303, USA
Email: dah@ngdc.noaa.gov & cng@ngdc.noaa.gov

KEY WORDS: Spatial Data Access, Interchange, Point Data, Vector Data, Attributes, DBMS storage, GIS access, Spatial Data Transfer Standard, SDTS

ABSTRACT

Exchange of data between different software/hardware systems has traditionally involved porting data directly from System **A** to System **B**. With large numbers of systems, this tradition has created very large numbers of unidirectional translators, which are often incomplete (or even error-inducing) and are difficult to maintain and use well. One must sometimes pass data through several intermediate steps between System **A** and System **B**. Two alternative systems are (1) to develop a broad-based System**A**-System**B** translator with a data description language encompassing many forms of data (such as the National Geophysical Data Center's FreeForm), and (2) to develop a comprehensive transfer standard and format system, thus requiring systems to have only "IN" and "OUT" translators into the common format system (such as the Spatial Data Transfer Standard - SDTS). We offer suggestions for improving the success of your transfers between systems. The new SDTS Point Profile offers an opportunity to facilitate porting point data from data base management systems into geographic information systems.

1 INTRODUCTION

The National Geophysical Data Center (and co-located World Data Center - A, Boulder Centers) distributes several global data integrations, enhanced for use directly in geographic information systems (GISs). These include the Global Ecosystems Database, TerrainBase, Global AVHRR-Derived Climatologies, and Global Land One-km Base Elevation digital elevation model. Most of these compilations have been raster data. We have found that image processing (IP) and raster GISs relatively easily import and export such data if we don't challenge them unfairly with esoteric cartographic projections or similar impediments.

However, point, line, or polygon data are much more difficult to distribute to the general public, as they are harder to pass between hardware/software environments. Several approaches to this challenge have been tried. These include specific direct System**A**-System**B** translators, generalized System**A**-System**B** translators with data description languages, and comprehensive transfer standards and formats.

We currently plan to release the contents of a new vector data base in Digital Line Graph, Arc/INFO ungenerate and export, Spatial Data Transfer Standard (Topological Vector Profile) and rasterized versions, in addition to the original file format of the source GIS. We still anticipate more customer assistance requirements for these vector formats than for our raster data distributions.

This paper discusses these formats, and offers some of our findings on how to improve the completeness and accuracy of your data transfer.

2 DIRECT SYSTEMA-SYSTEMB TRANSLATORS

Direct translators have been developed for many types of data, from word processor proprietary formats to IP and GIS formats.

2.1 Parallel Case: Word Processors

Those with experience with word processors may notice that word processor **A** commonly reads word processor **B**'s files better than **B** writes **A**'s format. This may seem logical, as software developers may be more motivated to import from other systems than to share with those other systems. Yet data distributors must successfully go against this trend by successfully writing to a format that customers can read.

Another observation is that package **A** can usually read older versions of **B**'s files, but not the most recent versions. This is because of the time lag between writing translators (for then-current versions of **B**) and the actual release and use data of **A**. Thus packages that frequently change file formats are often more difficult to transfer to/from than are more stable packages. Thus, in our case, data distributors must be able to write to stable, preferably generic, interchange formats.

2.2 Specific Examples

2.2.1 Digital Line Graph

Digital Line Graphs (DLGs) are base map data produced by the U. S. Geological Survey. Several formats have

D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

been developed for DLGs, though DLG Level 3 Standard and Optional Formats may be the most widely used by GISs. (Levels 1 and 2 do not support topological structuring, while recent formats such as DLG-F have not yet been adopted by many GISs.)

DLGs became so popular among GIS users that many packages included import routines from DLG 3 format to the internal file structure of the GIS. DLG 3 Optional Format was designed strictly for interchange of topologically structured data, and is the format described here.

DLG headers include information about data layer name, date, scale, projection, resolution, number of categories (up to 32) in the DLG file, and bounding coordinates. The header also contains 72 characters of banner headline, as freely developed descriptive text. The remainder of the DLG file contains information about nodes, areas, and lines.

DLGs can handle only limited attribute information. If one's vector data layer contains multiple attributes, it is still possible in many cases to export the attribute table from the DBMS attached to the GIS, then import the DLG and DBMS table into the recipient GIS - then manipulate the attributes in the recipient GIS to get the desired result. To do this, unique attributes (such as unique line and polygon identifiers) need to be in both the DLG and the DBMS transfers. This may require work within the DBMS of the exporting GIS, before the DLG and DBMS exporting is performed.

The DLG format also only handles two decimal places of locational accuracy. With Universal Transverse Mercator projection meters, this may be adequate for all but precision geodetic control. However, if global data are in Latitude-Longitude Degrees, one has precision of about 1 kilometer! Some people convert Lat-Lon Degrees to Lat-Lon Seconds. However, some GISs do not easily handle Seconds. Thus a transfer of high-precision data in Lat-Lon cannot be generalized, but must be designed for the specific exporting and importing GISs involved.

DLGs are not always written as originally documented. The original documentation for DLG Optional 3 format specified full 80-character records (padded at the end with spaces, if necessary). Some current DLG files are not padded. This makes a more compact file, but may confound some translators. In addition, Arc/INFO will not automatically create a correct DLG header without manual intervention in the process (by creating an {in_header_file} and possibly creating major and minor attributes in an INFO file.

2.2.2 Arc/INFO Export Files

The Arc/INFO GIS has more than one internal file format. The EXPORT function may be used, for example, to port between workstation and personal computer versions of Arc/INFO. Some other GISs can read such files.

2.2.3 ASCII X,Y Files

Simple attribute/x,y files are produced by many GISs. These files are similar to files produced by many digitizers. Common features may include a single identifier or class number and a string of X,Y pairs to describe point coordinates. Some, such as Arc/INFO's ungenerate files, terminate X,Y strings with the term "END". Others, such as GRASS and Idrisi, include a counter of the number of X,Y pairs in a header line for that feature. Some versions may allow points, lines, and polygons to co-exist within a single file (as does GRASS), or they may not (for example, Idrisi and Arc/INFO).

Arc/INFO writes its simplified ungenerate format for transfer to other developers' GISs. Ungenerate files permit each line or polygon to have a single identifier, and "any number of" x,y pairs to describe coordinates of component points. Points, lines, and polygons should be UNGENERATEd separately, for separate reimport into the recipient system. Multiple attributes and topological structuring must be recreated in the recipient GIS. This is possible if the attribute tables are carefully prepared and exported from Arc/INFO, and if the attribute tables contain the single attribute allowed for each feature in the ungenerate file.

Because of Arc/INFO's widespread use, many other GISs use Arc/INFO's ungenerate file format as their basic-level interchange format with Arc/INFO. As with the DLG format, this facility may allow one to go between System**A** and System**B** using UNGENERATE, even if neither system is Arc/INFO. However, this by itself is a rudimentary interface.

One can use a DBMS, text editor (if the editor is sufficiently powerful), or utility such as "awk" in UNIX to reformat data between Idrisi, GRASS, and Arc/INFO ungenerate file formats. Separately, functions such as Idrisi's ARCIDRIS and GRASS' v.in.arc (and v.out.arc) can read and write ungenerate files.

Note that some GISs can only handle single precision Arc/INFO coverages via their ungenerate interfaces. This is usually not a practical difficulty, as most GIS data layers are still lower than single precision accuracy. However, as GIS becomes more precise, this single precision limitation will become more of an obstacle.

Idrisi, on the other hand, adds a significant digit to an imported ungenerate file, and populates that digit with a value (thus moving every point location upon import!). Idrisi then removes this digit upon export. Idrisi for Windows may not be able to handle large ungenerate or DLG files, apparently because of a feature of Idrisi's memory management. Idrisi for DOS appears to have no such limitation.

2.2.4 Other Formats

Some other direct SystemA-SystemB translators exist. For example, AutoCAD .dxf format is handled by some GISs. If you have a one-time transfer between specific systems **A** and **B**, it is worth evaluating the levels of sophistication and "pain" involved in each available option -- including possibly passing through a possible system C to make the transfer.

D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

3 GENERALIZED SYSTEM*A*-SYSTEM*B* TRANSLATORS

Several decades ago certain industries developed generalized translators, designed to handle several popular file formats within such industries. Such generalized translators may have read several formats for translation to a single format, or to several formats.

A similar approach was developed several years ago at NGDC to several tabular file formats. This utility, called FreeForm, was a data description language for each understood format to facilitate translations between all listed formats. FreeForm N-Dimensional has been extended to include some additional data formats, including multidimensional raster data bases.

FreeForm can handle some simple ASCII vector file formats, such as Idrisi's and GRASS' formats (which include a counter for the number of X,Y pairs in a feature), but not Arc/INFO (which lacks such a counter in its ungenerate files). A solution to the latter problem might be to develop a script (using perl, awk, or another tool) to add a counter to the Arc/INFO ungenerate file, prior to porting through FreeForm ND.

FreeForm is in the unrestricted public domain, and is described on the Web at http://www.ngdc.noaa.gov/seg/freeform/freeform.shtml

4 COMPREHENSIVE TRANSFER STANDARDS AND FORMATS

Traditional transfer standards are numerous, often weak in functionality and documentation. The Spatial Data Transfer Standard (SDTS) was one attempt to overcome such limitations, by developing a comprehensive environment for data transfers. The SDTS was designed by a committee sponsored by the American Congress of Surveying and Mapping, and including government agencies, universities, and private companies recognizing a requirement for a better way to exchange spatial data. The National Mapping Division of the U. S. Geological Survey is the maintenance agency for the standard (FIPSPUB 173-1). SDTS advocates hope to adapt SDTS for international usability and recognition.

SDTS is subdivided into many parts. For our purposes, perhaps "data content standards" and "profiles" are the most pertinent.

Data content standards are definitions of the philosophy and mechanics of certain types of data. Though these are approved, the process is still sufficiently new that details become overlooked. For example, the original draft content standard for elevation data looked remarkably like a description of the U. S. Geological Survey's flavor of raster digital elevation model -- at least partially overlooking other options for representing elevation data, such as vector contours, points, other raster models, and triangular irregular networks.

Profiles describe a methodology for representing different types of data in a format that can be passed between software/hardware environments.

Until recently SDTS had only the Topological Vector Profile (TVP), which is a stylized impression of vector data. For example, the TVP requires a polygon for successful transfer, even if a data set contains only line work. (The solution is to insert a gratuitous and non-bothersome polygon into the data set before transfer.) In addition, the TVP fails to handle high-precision data, such as GPS-referenced observations or cadastral surveys.

Current TVP translators often fail to successfully transfer data between different hardware/software environments. This appears to be because translators have had little use. As a result, they have not yet had adequate opportunity to mature. In addition, TVP translators have apparently been largely developed and tested only with U. S. Geological Survey Digital Line Graphs (delivered in DLG Level 3 format), not in more general cases of vector data. David Arctur (1996: http://www.ngdc.noaa.gov/seg/tools/sdts/gislis_arctur.shtml) noted that TVP may be misnamed, and also misrepresented if one considers it a general vector data layer translation environment. Arctur also notes that the TVP may have been too ambitious an initial effort (implying that something conceptually simpler such as points might have been a better starting point).

Using existing TVP translators to exchange vector data is still problematic, fraught with undocumented obstacles. For example, a common failure in translations between Arc/INFO and GRASS via the SDTS Topological Vector Profile is that, according to GRASS, "polygon label missing for polygon X." Note that the SDTSEXPORT function in Arc/INFO did not generate error messages when creating the SDTS translation. Nevertheless, the source Arc/INFO coverage was not properly developed, even if it had been in use for years at the source facility. Even if Arc/INFO tolerated data in such a condition, GRASS rejected certain aspects of the data set, with specified unlabeled polygons being rejected. This problem can be solved with the function CREATELABELS in Arc, or by running BUILD in ArcEdit (not in Arc). In addition, INFO tables may not be in good condition. For example, fields that should have values may have incorrect zeros; attribute fields may be missing or mislabeled.

In short, merely pressing "the SDTS out button" on your GIS, and getting some files written, does not guarantee a compliant SDTS transfer. The SDTSEXPORT function in Arc/INFO does not check all of the requirements for a successful transfer. One might complain that (1) the program is immature and buggy; (2) inadequate information is available on requirements for data preparation prior to a successful transfer; or (3) users stop the careful preparation of their data bases once they can work - even if the data bases are still incompletely developed. All of these complaints may be valid. Time and additional experience should solve such problems.

We have been exploring ways to make existing transfer standards easier and more successful to use. We have also helped lead the development of the new SDTS Point Profile, which allows points to be passed between environments without some of the kludges required from

D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

the TVP. The Point Profile also permits high-precision data to be incorporated -- a feature that should probably be added back into the TVP (just as the TVP should handle line files lacking a polygon).

One example of the SDTS Point Profile is the access to U. S. National Geodetic Survey Geodetic Control data. Formerly available only in NGS' traditional "data sheet" format, the data (location and a very large number of attributes) can now be retrieved in the SDTS Point Profile directly from NGS' continuously updated active archive. The Website for this facility is http://www.ngs.noaa.gov (select Products and Services, then select DataSheet).

Initially, it was anticipated that the SDTS facility would not be widely used, as the existing user community was thought to be comfortable with the traditional data sheet format. However, the SDTS option has proven surprisingly popular, suggesting that GIS users not previously able to conveniently work with geodetic control data are evaluating this option. Several developers have expressed interest in developing Point Profile translators for their software. The Environmental Sciences Research
Institute (ESRI) is currently working on a Point Profile translator, scheduled to be released with Version 8 of Arc/INFO.

A generalized SDTS Point Profile encoder is being developed, for free distribution and adaptation by anyone. More information about SDTS and the Point Profile is at http://www.ngdc.noaa.gov/seg/tools/sdts

The Point Profile may become a mechanism to facilitate the porting of tabular data from a DBMS to a GIS. Considering the volume of spatially referenced point data (such as meteorological data, geodetic control, geochemical and precision agricultural field observations), such a capability may help to broaden and deepen the use of GIS.

The Raster Profile has been in draft status for a few years. Review and completion of this profile is currently underway.

Thus current developments in SDTS should help most current distributors of spatial and tabular data to improve their ability to reliably and conveniently distribute their data more robustly than in the past. Creative adaptation of the Topological Vector Profile (by careful data preparation, adding a gratuitous polygon if one is not present in your data layer, and testing your sdtsexport with the import routine of at least one other GIS), and commercial implementation of the Point and Raster Profiles should make this possible.