Geographic Information Systems Data:
A Description of the StreamNet GIS Data
By David Graves
StreamNet GIS Specialist
December 21, 2001

I. Issues in developing regional GIS data sets

A. Availability of Data

The primary considerations to be addressed when developing regional GIS data products are if reliable spatial data is available and if so, whether they can be integrated into a regional product. Although a wealth of data may be available, it takes time to seek them out among the many regional data collectors and to ascertain the scope and reliability of the data. Data must then be compiled into a usable regional format that is consistent and meaningful before data analysis can be initiated.

1. Data providers/ Data collection

Identifying data providers can be a challenging task for a regional project. Entities that collect data often do not have any incentive to make this data widely available. Data are often collected for specific local purposes, and may or may not be able to be assimilated into larger data sets. A project that aggregates data must perform the time-consuming task of evaluating data and determining how they can be aggregated at a regional level. A particular concern with GIS data is scale. A data set that has been collected at a rough scale (example: 1:250,000) should not be recompiled at a more precise scale (example: 1:24,000) because the accuracy of the data would be misrepresented.

Many types of GIS data are currently available on a regional level (land use, digital elevation, hydrography, etc.) but fisheries data have traditionally been collected at the state or tribal level and thus must be pieced together to make a consistent regional product. Fisheries and habitat data in the Columbia Basin have principally been collected by the state fish and wildlife agencies, the tribes, and by the federal agencies (U.S Fish & Wildlife Service, U.S Forest Service, Bureau of Land Management, etc.). With an increased emphasis on watershed-level planning, however, it is anticipated that in the future there may be many more local sources of spatial data.

2. Data collection and compilation

Once a data source has been identified, data must be collected and compiled into a consistent format. Spatial data may already exist as a GIS data set, but more commonly spatial data are actually tabular data that have spatial references. A GIS coverage may be created with georeferenced tabular data, but scale and the referencing system used is important. Stream-based data must often be converted to a standard format (example: the LLID stream-based routing system) in order to make them meaningful for comparison to

other data. Although two sets of data may be collected on the same stream, if they are not georeferenced in the same format, then geographic analysis is made much more difficult. The spatial data collection and compilation process thus seeks to bring in data from disparate sources and georeferencing systems and combine them into a consistent GIS data set.

3. Data exchange formats

Data exchange formats detail specifically how data should be exchanged. They should include tables and item names and types, required information, and file types. Data exchange formats are very useful when a data provider has an incentive to deliver their data to the regional level in a specific format and has the capability to do so. Both spatial and tabular data are more useful when they are exchanged in a pre-determined format.

B. Creation of GIS data

Spatial data which arrive in a tabular format are converted into a GIS format for use in GIS analysis and display. An example of conversion of tabular data works as follows: A record of data identifies that something (fish distribution, water pollution, habitat quality, etc.) is located on Smith Creek from river mile 4.5 to 5.4. To create GIS data from this tabular data, this data record is "placed" on a hydrographic GIS data set (a set of streams and waterbodies). Smith Creek is identified on the hydrography and the record is attached to this creek from 4.5 to 5.4 miles along its course. The GIS data record is thus composed of the original tabular data (fish distribution, habitat quality, etc.) and the physical topography of the stream over this extent. The compiled GIS data offers the advantages of display and geographic analysis which cannot be accomplished with the data in the original tabular format.

1. Georeferencing of data (i.e. which system to use? LLIDs, etc.)

Spatial data may be georeferenced with a variety of different methods and systems. Georeferencing may be crude, with general geographic descriptions like "headwaters of Alsea river" or precise "i.e. Alsea river, river id #5473, from 5250.3 feet to 5796.8 feet). Precise georeferencing offers the advantages of easier conversion to GIS format and more accurate data, allowing increased confidence in subsequent analysis. A number of different systems exist which may be used to georeference data. For stream-based data, the "LLID" system has gained wide acceptance in the Pacific Northwest region. LLID (Longitude Latitude Identifier) is a unique stream identifier assigned at the mouth of a stream. Routes have been developed at the 1:100,000 scale along each identified stream which allow precise data measurements on a river. Multiple data sets that are standardized in this format may be easily compared and analyzed.

While StreamNet uses the LLID system to georeference stream data, other hydrographic systems also exist for this purpose, including the National Hydrography Dataset (NHD). The NHD was recently completed for the entire nation through a joint effort of the US Environmental Protection Agency (USEPA) and the US Geological Survey (USGS). In contrast to the LLID system, which uses stream-based routes, the NHD principally uses reach-based routes, created when routes are assigned to a stream reach (the area on a stream between confluences).   In November 2001, with the cooperation of the USEPA

and the USGS, StreamNet completed a database application for converting stream data between the LLID and the NHD systems. To download this application or learn more about it, please visit http://www.streamnet.org/nhdllid/. For more information about the differences between stream based and reach based routing systems, please see http://www.streamnet.org/nhdllid/Help_Documentation.html#faq.

2. Metadata and data accuracy issues

With the recent proliferation of available data, metadata (referential data about data sets) have become increasingly important in distinguishing the quality and utility of data sets. Researching and developing metadata can be a time-consuming task but it is a fundamental one for regional data collection. Spatial data is commonly documented through FGDC (Federal Geographic Data Committee) methods or related variations, which are thoroughly explained on their website at http://www.fgdc.gov. Metadata about spatial data sets should contain information about scale, sources, geographic extent, development method, and contact information.

A "data dictionary" is sometimes included under the scope of metadata, and is a necessity for most data sets. A data dictionary is a detailed description of data tables and items which allow the user of a data set to understand all of the information that is contained therein.

II. Issues in distribution of regional GIS data sets

Regional products should be distributed in an easily accessible manner in order to realize their greatest potential use. Internet distribution is the easiest way to distribute data, although there are differing philosophies on how to present the data. Depending on the resources and expertise of a user, varying approaches for distribution of GIS data may include any or all of the following:

* Premade maps
* Online query systems through which users may select and view GIS data
* GIS data files made available for download
* Interactive online mapping

A. Premade maps/ Online Query systems

Maps have long been a powerful medium for displaying spatial data. Many users respond better to maps than to spatial data because maps quickly convey information and trends without requiring a user to disseminate large amounts of data. The disadvantage of maps is that they may only display a limited amount of information at a time, and it is time-consuming to anticipate and produce maps which answer the variety of questions that may be asked about a spatial data set.

Online query systems offer a more flexible method for displaying spatial data. A user selects the information and parameters of interest, and then may choose to display them in a variety of formats (maps, graphs, tables, etc.). Another advantage of online query systems is that they typically do not require expensive GIS software for an end user, only a web browser and a sufficient computer. The disadvantages of online query systems are that they require time and

expense to program and it is difficult to tailor a system that works well for both novice users and more technologically savvy users. New interactive online mapping technology may simplify the delivery of GIS data, but the underlying GIS database will still have to resolve all the issues previously discussed.

B. GIS file distribution (ESRI formats, SDTS, etc.)

The most straightforward method to distribute GIS data is to provide GIS information for download. This assumes that a user will have the necessary time, expertise, and software to utilize this data. GIS data may be distributed in a few different file formats. Since ESRI (Environmental Systems Research Institute) software has gained widespread acceptance in the GIS world, ESRI formats are the defacto standard for distribution. These formats may be converted into most other GIS platforms for use with different software. The two standard ESRI formats are arc export format and shape file format. Arc Export format offers the advantage of storing a GIS coverage in a single file. Shape files require three files for one coverage but are more convenient for ArcView users. SDTS (Spatial Data Transfer Standard) is a standard that was developed by the USGS for transferring spatial data. The SDTS is useful for transfer between dissimilar systems and is advantageous in that it captures data, metadata, and data dictionary information in one format. It can be cumbersome to use, though, and is used less frequently than the other standards.

C. Benefits of standardized data (i.e. local vs. regional level distribution, metadata)

Standardized GIS data offer many advantages of local distribution. To use regional data, the steps of data collection, data compilation, and data standardization must be performed. If these tasks are organized and completed by a single entity and then adequately distributed, it alleviates this workload for each end user that is able to utilize the data. Metadata lists are popular means of locating spatial data but they still require the laborious tasks of data compilation and standardization that must be performed to aggregate meaningful data sets.

III. Spatial data storage (tabular and spatial)

Spatial data is commonly stored in both tabular and GIS formats. A tabular format offers the advantages of lower disk space use and easier integration with other tabular data sets. Tabular data which have adequate georeferencing may then be converted to a GIS format when this is necessary for display and spatial analysis. Stream-based data are commonly stored in this method because linear geographic information is readily compatible with tabular formats. This format is also referred to as "event data" because records are stored as data events (tabular records that precisely describe the location of the data on a GIS route system – for example, data on stream #7432 from 2000 to 3000 feet) but may also be displayed as GIS data on a stream coverage or used to create new GIS data coverages. This method utilizes the efficiency and simplicity of tabular data storage while offering the flexibility of use with powerful GIS systems.

Point data, which are used to describe objects that fall at a specific location (dams, hatcheries, barriers, etc.) are also amenable to tabular data storage. If a point falls along a stream, it may work well with existing hydrography (example: a point is on stream #7432, 2543 feet up river).

Non-stream points are commonly georeferenced with longitude/ latitude measurements, which may also be stored easily in tabular formats.

Polygon, or area information, is less amenable to tabular storage because it requires specific geographic referencing. It is usually easiest to transfer and store polygon information in a GIS format unless the geographic extent can be easily described with a pre defined location (i.e. Roosevelt Lake, King County, etc.)

III. Conclusion

Geographic Information Systems offer powerful new tools to collect, manage, analyze, and deliver information that is spatial in nature. StreamNet utilizes these tools to improve our service in these areas and will likely continue to do so in the future. As with other types of data, issues of accuracy and compatibility are imperative to the use of GIS data and are a primary focus of the StreamNet project. Once data is compiled to a standardized format, management of the data is necessary, and it is possible and sometimes advantageous to manage spatial information in a tabular format rather than a GIS format. This spatial data may be then be delivered through GIS, where some of the most exciting advances of recent years have occurred with data delivery techniques, including new capability to deliver spatial information over the internet via interactive mapping technologies.