

# Data Standards Overview

---

Dr. Donna Pequet  
Professor, Geography Department  
The Pennsylvania State University

## Introduction

When producing a map, the cartographer must organize all the descriptive information that goes into the map legend in a particular format. Titles are put in a specific place, tic marks are made a certain way, meters may be used instead of feet, and so forth. A map legend is pure metadata. The legend contains information about the publisher of the map, the publication date, the type of map, identification of the geographic location, the map's scale, its accuracy and many other things.

Metadata standards, like mapping standards, are simply a common set of terms and definitions that describe geospatial data. Besides the gradually increasing dependence upon ever more powerful and accessible computing systems within organizations for handling data, the development of computer networks, such as the Internet, is leading to a revolution with regard to the quantity and diversity of information available for exchange. There now exists a variety of communication systems linking together many thousands of computers and their accompanying data users.

At the same time there is a desire within the GIS community to reduce the high costs of data collection and capture through the sharing of data. Over the thirty year period in which GIS have been in existence, many gigabytes of data have been collected. Much of this is still of use. Modern database management offers a flexible range of tools suitable to the context of the age of digital geographic data in the form of shared digital geographic databases. Most importantly, we are no longer restricted to the old pen and ink conventions of cartographic display.

For the rest of my talk, I will present an overview of the importance of metadata in environmental and other geographical databases, paying specific attention to data dictionaries. First I will give a description of what metadata and data dictionaries are. Next, I will discuss their potential impact, both upon the organization that owns the database, and upon users both inside and outside the organization. Then, I will talk a little about the evolution of geographic database standards as they apply to data dictionaries.

## Why Use Standards?

Thirty years ago, humans landed on the Moon. Data from that space mission are still being used today, and will continue to be used in combination with data coming back from missions like Cassini for many more years to come. Even for a single application, some protocol must apply so that the data are indeed usable. Without some standard, we would have something like this...

<slide here>

As the variety of data gathered at different times for different applications grow, these data will have different definitions for the same things (as well as different formats, and so on) if there isn't some form of standardization. The 'let a thousand flowers bloom' philosophy is not productive or effective when it comes to databases that need to be shared. Thus, a database system that follows one set of rules for defining terms will not be able to communicate with another database using another set of rules. There are two major problems when this occurs:

- 1.) Users tend to lose confidence or simply become frustrated in trying to gain access to needed data
- 2.) Databases tend to be duplicated, and proliferate unnecessarily, because it is often easier to create a new version of the data element than to resolve conflicts between existing ones.

Metadata standards increase the facility of comparing data gathered in different places and at different times. Metadata can help the city planner, the graduate student in geography or the forest manager find and use geospatial data, but they also benefit the primary creator of the data by maintaining the value of the data and assuring their continued use over a span of years. Metadata is simply data about the data. This can include information about the time the data was entered into the database, by whom, and who (or what part of an organization) has responsibility for maintaining specific data elements. A data dictionary is an essential part of this metadata, containing the definitions of data elements and explicit linkages describing how the various elements relate to each other. For example, a data dictionary for hydrological data would contain a list of categories of hydrological features (streams, wetlands, and so on) with the definition of each class and give the list of features within each, with their definitions. A data dictionary would also contain aliases or equivalent terms - for example a bayou = a marsh. One of the primary purposes of the data dictionary is to be able to describe and communicate what specific information elements are in a database to users, or potential users. Another important feature that the dictionary provides is the ability to link or relate two instances of the same feature in different databases. Thus, a common designation of "marsh" with

explicit equivalent terms - such as bayou - would allow the same type of entity from different databases to be compared. A third purpose is to insure the integrity of the data by minimizing the possibility of the same data being entered in slightly different forms (and perhaps maintained in slightly differing vintages). A fourth is to maintain cross-references. These cross-references linking related features within a database, allows the effects of a proposed change in the contents of the database to be recognized before any actual change is made. A data dictionary is thus both a means and an end, a tool and a resource-- a tool for data administration, and a resource for the user.

## **Getting There--Implementing Entity Naming Standards**

One of the major problems that will occur in any effort to introduce entity naming standards - i.e., a data dictionary within an organization - is the multitude of naming systems that inevitably crop up for individual files in the absence of a naming standard. It is a rare organization that has a single naming convention throughout all of its data files. This is because so many files have been built on an individual basis to deal with specific needs, mandates, or projects. Staff members (or outside contractors) have often been able to put their own personal stamp on naming different entities. This is not something that happens by design... People left to their own devices will independently designate -- or create -- whatever name seems descriptive to them. The issue of naming conventions comes loaded with emotional overtones. People who are used to one naming convention are loth to change to somebody else's naming convention for the "greater good". Sometimes the differing application, and therefore the differing perspective, of various units will cause data files to diverge in the focus on their own local needs.

Here you can see discrepancies between parcel data definitions between the King

County Assessor and the building permit review agency (BALD). This came about as the result of two different mandates. Since mandates reflect the social purposes agencies are created for, most mismatches discordances result from different perspectives on the world. The Assessor determines property values for the calculation of property tax and is therefore only interested in already existing parcels. Because BALD decides which areas can be built up through an examination of a site's characteristics and the applicable environmental regulations, it also must consider unincorporated areas that have not yet become parcels for the Assessor, because they are not yet assessable.

This situation is further complicated because BALD depends on the Assessor's parcel data for processing applications. This makes things somewhat dysfunctional for both agencies. Because of the different requirements, the semantics of the parcel databases in BALD and the Assessor are defined quite differently. These differences lead to increasing disparities between the parcel databases. A catalog of contingencies presents an

approach to dealing with the known translations between databases. However, as each agency continues through time in their respective role, dealing with a growing list of contingencies and verifying the integration of data between the agencies becomes more cumbersome and inefficient.

Briefly, BALD used parcels to exhaustively subdivide the county, whereas for the Assessor, parcels represent only the legal division of land. Easements, for example, are not part of the parcel for the Assessor, but they are for BALD. So... how do you implement a standardized data dictionary when two components of the same organization each claim the need to maintain their own definition for "parcel" ? Even in the many cases where there are different names for what are truly equivalent features, people get used to the names THEY use for things, and don't want to change. An outside standard would certainly help to facilitate conflict resolution -- but more importantly, these are definitions deliberately designed for general use, and specific variants that cannot be resolved by nature of specific uses can still be explicitly linked, as variants, to a primary definition. Equivalent terms can be kept as "local" terms that are explicitly linked to the primary term within the data dictionary as aliases. This is part of what a data dictionary does. How do you implement a standardized data dictionary as a data provider to a wide range of outside users ? I like the example of a successful definition standard - that for land use classification. The U.S. Geological Survey came up with what is now THE standard for land use classification.... the Anderson Landuse Classification System, introduced in the early '70's. This immediately caught-on because before such a standard, there was a "tower of babble" situation. The standard allowed people and various government agencies to communicate about real problems that these agencies at multiple levels had to deal with in some coordinated manner- urban sprawl, and so on.

## **Existing Standards**

There is no shortage of standards for the treatment of spatial data. The problem at the moment is choosing among them. Standards for maps had been rather stable in the predigital era, the USGS standard symbology for their topographic maps being a good example of a 'de facto' standard. The US National Map Accuracy Standard (Bureau of the Budget, 1947) was a formal standard that set out a positional tolerance that declared that 90% of "well-defined" points should be within 1/50" of their correct position. Curiously, maps that comply can proclaim "This map complies with NMAS", while maps that do not comply say nothing. In the realm of standards for digital data, most standards efforts within the US at the national level have been focused upon the problem of transferring data, including SDTS - Spatial Data Transfer Standard, NSDI - National Spatial Data Infrastructure, and DIGEST - digital Geographic Exchange Standard, but all of these also include standardized definitions for features.

The SDTS was adopted officially by the federal gov't in 1994. This meant that all

Federal agencies had to use this standard to document newly created data as of January, 1995. This standard (including an open-ended formatting protocol) was developed over a period of over ten years, with a two year external review period.

The NSDI initiative aims to foster enhanced use of spatial data through better management of existing spatial data and through more consistent definition of new data. One of the objectives is the establishment of a National Geospatial Data Clearing House to enable access to all types of spatial data sets, whether generated by government or through partnerships with academia and private companies. A number of prototype distributed data holding sites already exist.

## **Management Commitment**

Management commitment to dictionary implementation is absolutely crucial. This commitment has to be given over an extended period without expecting immediate payoffs. Depending on the staffing of the data dictionary function, payoffs could be visible within one to two years. In most cases, other than a brand new start, the dictionary will have to be retrofitted onto an existing collection of data files. If the sheer number of files is large, the dictionary will play a role in reducing the complexity level by providing documentation and definitional control. The complexity of an organization's data files and their interactions tends to increase, and ability to access the data tends to deteriorate with the addition of new data simply because it is easier to create new data items (with new definitions) without some standard. Thus, the value of a standard data dictionary which records existing data items, their definitions, their explicit interrelationships, as well as such things as aliases and (controlled instances of) exceptions, avoids or can remedy the following situation: