Sharing the Data You Have—Getting the Data You Need

by William J. Craig and Nancy von Meyer, Ph.D.

The statements made or opinions expressed by authors in **Fair & Equitable** *do not necessarily represent a policy position of the International Association of Assessing Officers.*

This article provides a classification of data-sharing issues originally identified by Will Craig in 1992, with some slight expansion to accommodate changes in technology. This classification is important because it provides a common language for discussing data-sharing issues and helps formulate solutions by grouping common problems into topic areas.

In 1992, most digital data were captured by digitizing, and most parcel maps were still hard copy. Data sharing was important for two reasons: to reduce the cost of data capture and to gain the benefits of synergy—the result of combining data sets. In the 16 years since then, there has been an exponential expansion in the availability and consumption of digital spatial information. Computers routinely handle large file sizes and process large data sets more quickly. Nevertheless, the importance of data sharing continues to be to reduce duplicative data collection and hence the cost associated with data capture and to combine various data themes.

The biggest changes and impacts on data sharing since 1992 have been the proliferation of Web sites; online search, query, and view; and, most recently, data as a Web service.

- Data access provides the use of data through view, search, and query, and data presentation is typically through an Internet site or a presentation Web service. Information can be discovered, viewed, analyzed, and printed, but the underlying spatial data are not provided to the consumer. For example, millions of users get directions and navigation from MapQuest, print maps, and embed directions to their businesses in other applications without taking possession of the underlying street, road, and feature data sets.
- Data sharing and data distribution provide the data set to consumers who take possession of the data and incorporate the data directly into their applications or products. Data sharing may be limited to government-to-government transfer for specific applications; data may be readily available for download from Web or FTP download sites; or data may be packaged onto media for distribution.
- Data as a Web service allows consumers to add data to their applications directly from producers, gaining the benefit of the complete data set without having to take possession or ownership of the data. This service makes it much easier for users to have access to the most current data and minimizes the amount of data that has to physically change hands, but it also increases the need for data standards and may even expand the demand for digital spatial data.

The taxonomy, or classification, of data-sharing problems discussed here applies to data sharing or distribution and to data as a Web service.

Unusable Material, Format Problems, and Standardized Formats

How many computers can read a floppy disk or even a tape or zip drive? With the Internet and the downloading of files from FTP and Web sites, hard format issues are less of a problem now than they were in 1992, but data-sharing problems related to material, format, and standardized format still exist.

Applications that need parcel data but do not understand the finer points of parcel information cannot distinguish between taxable value and market value, much less decipher which value in the local data set contains the necessary value information.

Versions and Other Formats

With spatial data, the format and version of the geographic information software can make good digital data unusable. Spatial data format seems to change with every version, which seems to change every year. Last year's software can't read this year's geographic data. There have been attempts to build universal spatial data formats, but some of these formats have unexpected consequences on data structure, format, and accuracy, and many data producers provide data in *native* formats and leave getting the version right to the consumer.

Standardized Formats

This was not identified as a specific data-sharing issue in the 1992 article, but certainly business applications that need data

from many jurisdictions benefit if those data are presented in a standardized format. For example, collecting data to support the Wildland Fire response was easier when the local data were translated to a standard format. Applications that need parcel data but do not understand the finer points of parcel information cannot distinguish between taxable value and market value, much less decipher which value in the local data set contains the necessary value information.

Translation Problems

In 1992 translators were identified as a separate issue, but in 2009 the file translation from one format to another (from CAD to GIS or from an AS400 to SQL Server Database) are format problems. Older mainframe or AS400 platforms often require specialized or one-at-a-time export programs or scripts to convert data into a structure that can be shared.

Uncommon Classification

The organization creating the data file frequently has a perspective on important and classified objects that differs from what the consumer expects. With assessment and parcel data classification problems often come from parcel use or tax classification codes. If property is not taxable, it is often not included in the tax mapping data set, yet land use managers and open space analysts want to include those *non-taxable* areas as some of the most important space in their applications.

A similar data-sharing problem is coded data without an explanation of the codes. Zoning and land use are two common examples in parcel data for which codes are often used. Without the key to translate the codes, downstream users cannot use the data.

Cost and Sustainable Funding

In his 1992 article Craig interviewed officials in Australia from a state department of transportation, from a state natural resource department, and from a city government who could not afford to acquire the state-created parcel map to use as a base for their mapping. The city government had two options: (1) drop GIS altogether or (2) build its own base map. In either case, the state loses the benefits of establishing a multipurpose synergistic database, and if the city moves ahead with re-collecting data, there is duplicative data capture.

On the other side of the cost problem is a jurisdiction that does not receive some recoverable fees for its data; it can be left without a funding stream to maintain the data. Can data sales generate a sustainable source of funding? Does charging for data limit its use and cause third parties to re-create an existing data set? What other funding sources can ensure a sustainable system and be justified to elected officials every one, two, or four years. Balancing the fair cost, government-to-government data sharing, and private citizen needs is difficult and tricky for many local jurisdictions.

Balancing the fair cost, government-to-government data sharing, and private citizen needs is difficult and tricky for many local jurisdictions.

Excessive File Size

Even with faster computers, higher network speeds, more memory, and larger disk space, a big file is a big file. Six-inch color ortho-imagery, raw lidar data, or a terrain model data set can overwhelm many data downloads and systems. One approach to sharing larger data sets is to identify a core data set that is a subset of the complete parcel data so consumers can determine where they need more detail. Between government agencies data size may be less of an issue than with the public, whose computers are optimized for home use and are not GIS-ready quad-core processing behemoths.

Data Quality

Data quality has several components, not only spatial accuracy. Quality is perhaps one of the most important considerations in data sharing.

Incompleteness

Many GIS operations optimize their data in ways that do not affect their own day-to-day needs but create problems in data sharing with consumers who are not familiar with local conditions. One example of incomplete data is constant values such as the state or county name, which could be added to all the attributes in a jurisdiction when the data are shared. Another example is missing data, such as having only the site address for some of the parcels, not having complete information on values in an assessment data set, or having coded values without an explanation of the meaning of the codes.

Geographic Positioning and Data Resolution

Geographic positioning is a general term for data quality related to the accuracy, precision, and vertical registration of spatial data. Data resolution, a similar issue, is related to the degree of generalization and/or the intended scale of presentation of the data. Figure 1 presents two views of the Washington-Oregon coast. The shaded (crosshatched) green is the coast as shown from a county boundary file. The more angular solid green is the coast from a state boundary file. The state boundaries have been generalized and probably collected from a less accurate source. The coast is less accurate form the state boundary file but may have an intended use at a smaller scale than the county boundary file.

Lack of Topology

In the topological vector world, lines run continuously, polygons close, and a lake holds its water. Snapping is the digitizing technique that forces this connectivity; lines are connected mathematically, not just visually. In a GIS world this is topology (the relationships of objects to each other). If parcel polygons overlap, have large unexplained gaps, or do not close as polygons, their usability is diminished. Figure 2 shows four Public Land Survey System (PLSS) townships downloaded from a federal data Web site. The red shaded areas are PLSS townships that overlap, and the white areas are gaps. In Township 22S R17E, the line doesn't close across the top. Such an error in

topology creates havoc with the analysis. The usability of these data is diminished because of these topology errors.

Figure 1. Washington-Oregon coast from two sources of data

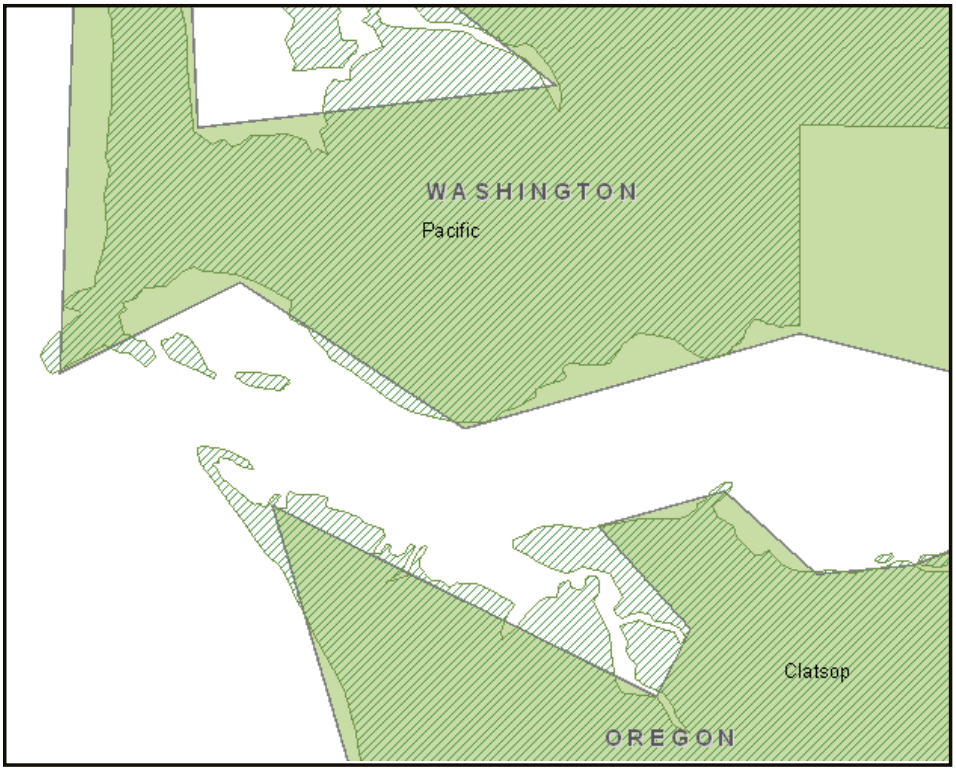
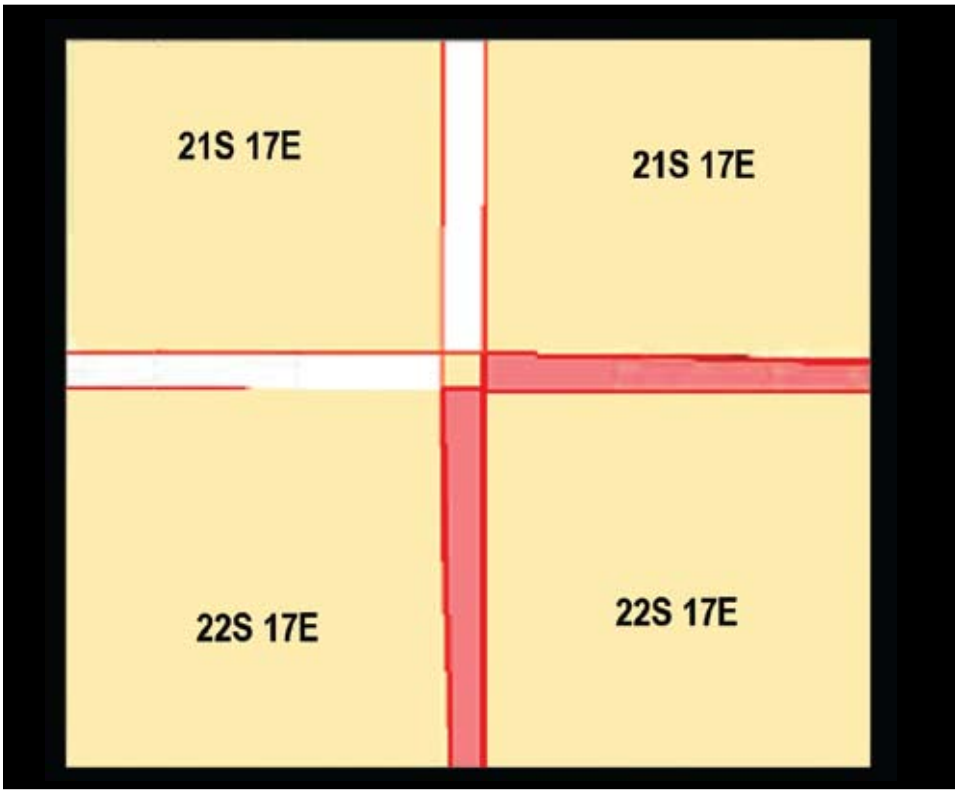


Figure 2. Topology errors



Source: Geocommunicator (http://www.geocommunicator.gov/GeoComm/index.shtm). Data for the State of Utah accessed October 14, 2008.

Label Problems

Much GIS power comes from having data attached to geographic features. The relationship between the mapped feature and the data file is forged through an object ID or label. If the feature does not have a *handle*, or label, then a GIS cannot operate on it. Similarly, if mapped features have multiple or non-unique labels, they are just as unusable.

Currency

Knowing how current the data set is or when it was last updated can affect its usefulness. Not knowing the currency of the data set can limit its usefulness. In general, the most current data are in the greatest demand. In some jurisdictions parcel data are published annually, and updates throughout the year are not shared until the end of the year.

Whether you are a data producer who has data to share or a data consumer who needs to find data for your applications, defining and understanding any problems or issues is a step toward a solution.

Missing Metadata

Metadata contains the definitions of the data fields, codes or domains, spatial coordinate system, currency, and data steward contacts. Capturing and maintaining metadata are often viewed as tedious and unnecessary tasks, but without metadata, data users may use data incorrectly or have to track down the data producer.

Authoritative and Trusted Sources

Not all data are created equal. Because many sites replicate data or provide repackaged digital data, it may be difficult to weed through various data sets to find the most current and most dependable. Data from an authoritative source (a source that can certify the content) or from a trusted source (a source that is sanctified by the data producer) is more

in demand and effective for data sharing. Another article in this issue addresses and defines the role of authoritative sources for parcel data.

Whether you are a data producer who has data to share or a data consumer who needs to find data for your applications, defining and understanding any problems or issues is a step toward a solution. At the root of data sharing is which lesson is at play—*All I Really Need to Know I learned in Kindergarten* (Fulghum 2004) or *The Little Red Hen* (Williams 2006).

The top ten lessons from *All I Really Need to Know I Learned in Kindergarten* are:

- Share everything.
- Play fair.
- Don't hit people.
- Put things back where you found them.
- Clean up your own mess.
- Don't take things that aren't yours.
- Say you're sorry when you hurt somebody.
- Wash your hands before you eat.
- Flush.
- Warm cookies and cold milk are good for you.

The lesson of *The Little Red Hen* is:

Then, probably because she had acquired the habit, the Red Hen called: "Who will eat the Bread?" All the animals in the barnyard were watching hungrily and smacking their lips in anticipation, and the Pig said, "I will," the Cat said, "I will," the Rat said, "I will." But the Little Red Hen said, "No, you won't. I will." And she did.

References

Craig, William J. 1992. Why we couldn't get the data we wanted. *URISA Journal* 4 (2): 71–78.

Fulghum, R. 2004. *All I really need to know I learned in kindergarten,* 15th anniv. ed. Ballantine Books. http://robertfulghum.com/index.php/fulghumweb/booksentry/all_i_really_need_to_know_i_learned_in_kindergarten/ (accessed January 6, 2009).

Williams, Florence White. 2006. *The little red hen.* www.gutenberg.org/etext/18735 (accessed January 6, 2009).

Nancy von Meyer, PhD, PE, RLS, GISP, is vice president of Fairview Industries and works with counties, states, and federal agencies on parcel and land records projects. Nancy has worked on land records projects from Oregon to Florida. Focusing primarily on counties and federal land agencies, finding efficient and economical ways to build, sustain, and publish land records information.

Nancy has worked with land records and parcel related system design, research on new technologies and development of new approaches to parcel issues for nearly 20 years. She has served as the President of the Wisconsin Land Information Association, in 1994 on the Urban and Regional Information Systems Association, URISA Board of Directors (1992-1997), on several National Academy of Sciences panels and has authored numerous articles and publications.

Will Craig is the associate director of the University of Minnesota's Center for Urban & Regional Affairs. He was an early pioneer in GIS, working with state and local government since the late 1960s. His 1980 PhD dissertation in Geography was based on his work with building permits in Wright County Minnesota. He is president-elect of NSGIC, the National States Geographic Information Council.