

Using Digital Spatial Archives Effectively*

DOUGLAS M. FLEWELLING

National Center for Geographic Information and Analysis, University of Maine, Orono, ME 04469-5711, dougf@spatial.maine.edu

and

MAX J. EGENHOFER

National Center for Geographic Information and Analysis, Department of Spatial Information Science and Engineering, and Department of Computer Science, University of Maine, Orono, ME 04469-5711, max@spatial.maine.edu

Abstract

The size and complexity of modern geographic datasets continue to grow and with that growth comes an increased difficulty in assessing the usefulness of a particular dataset for a particular problem. The task of browsing through a large database is rapidly becoming impractical because of the sheer volume of the data, therefore, the information potential of the database is lost because the scope of the database is beyond comprehension. This editorial promotes new methodologies beyond the use of descriptive metadata to support a user's desire to find a dataset suitable for his or her analysis.

1. Introduction

Very large spatial datasets are becoming commonly available through digital libraries (Smith 1996), data warehouses (Garcia-Molina *et al.* 1995), and research archives (Levy and Marshall 1996). Plans to increase the number and size of these *digital spatial archives* through such efforts as the National Spatial Data Infrastructure (FGDC 1994) are being implemented at the Federal and local level. The availability of data and their suitability for a desired purpose do not necessarily go hand-in-hand (Widom 1995). The current state of digital archives is a "buyer beware" market with few controls on what goes into a digital archive or assurances of fitness for use. Efforts such as the Content Standard for Digital Geospatial Metadata (FGDC 1997) and Dublin Core (Weibel *et al.* 1997) are attempting to standardize dataset labeling, but their use is voluntary in many cases. The current euphoria at having and providing access often overlooks the risks and costs associated with using the data. Without knowledge about the suitability or fitness of a dataset for a particular task the user cannot draw meaningful or defensible conclusions. This editorial addresses the need for a means to evaluate the suitability of very large spatial datasets for a given task.

There are two key elements necessary for scientists to determine that a dataset is the best one for their needs. First they must have some knowledge of the dataset's contents, either through experience with the dataset or through a detailed description. In the latter case, metadata—data about data—can provide valuable statistics and data history, which can show that the dataset should have the proper contents to address the problem domain. It is possible, however, that the data collected are incomplete or unevenly distributed through space, or may have any number of hidden problems. For instance, even though the metadata show that each station in a weather

* This work was partially supported by the National Science Foundation under NSF grant SBR-8810917. Max Egenhofer's work is further supported by NSF grants IRI-9309230, IRI-9613646, SBR-9600465, and BDI-9723873, by Rome Laboratory under grant number F30602-95-1-0042, by the National Imagery and Mapping Agency under grant number NMA202-97-1-1023, by the National Aeronautics and Space Administration under grant number COE/97-0015, and by a Massive Digital Data Systems contract sponsored by the Advanced Research and Development Committee of the Community Management Staff.

dataset for 1988 has rainfall as a data field and 86% of the stations reported rainfall, the dataset is of limited use for a scientist studying the 1988 drought if the missing 14% are in the American Midwest. Second, the domain scientist must consider the specific relationships among data elements in a spatial dataset. The relevance of a specific data item in a scientific inquiry is an important factor since there are complex inter-relationships present in spatial data. These specific relationships are, in fact, of greatest concern to spatial scientists and must be preserved in any dataset used.

Scientists who have had experience with a particular dataset or its provider may have developed an understanding of the peculiarities involved. They develop a level of trust—high or low—which effects their future choices of datasets. Inexperienced users of digital archives, however, do not have this advantage in picking their datasets and may have to download several seemingly useful datasets before they find one that meets their needs. In an environment with a charge for network time and data access, finding the right dataset can become expensive. In a traditional market, imagine having to pay a fee to take a shirt off the shelf and try it on, or even having to buy it based on a description like “green plaid, long sleeve, large.” In order to make digital spatial archives useful, users need a mechanism to try out datasets before investing time and money in exploiting the entire dataset. The scientist must be able to assess the fitness of a particular dataset for her tasks and to do so efficiently. The task of *finding* the right data is different from *using* the data in scientific analyses and requires different supporting tools (Flewelling and Egenhofer 1993). The purpose of this editorial is to analyze the options for finding the right dataset in digital spatial archives.

2. Options for Selecting Datasets

The final choice of which dataset to use in an analysis is constrained by a dataset’s fitness for the task, the local computing resources, and the cost of obtaining the dataset, both in time and money. Most important among these constraints is the need for the data to be well suited for the task at hand. The fitness for a particular purpose is defined differently for every type of scientific inquiry, but it can be assumed that domain experts can define fitness, or at least recognize it when they see it. Various domains have defined standards to which all data must adhere. These data interchange formats rely on a common application-level model upon which to base the data types and data contents. These interchange specifications are limited in their domain semantics. By their very nature, digital archives provide access for people who may want to use datasets in non-standard applications with the potential for widely varying semantics (Beard 1987). This means that the fitness for use of a particular dataset cannot be taken for granted just because it is in an archive.

2.1 Download-all paradigm

Within current digital archive architectures the user has a limited number of options available when selecting a dataset for further analysis (Levy and Marshall 1996). At the lowest level of sophistication, the user knows nothing about the dataset except its subject area and some simple description of the content provided by its file name and directory hierarchy. Such a system describes a large number of the digital archives currently available, but is of limited use. Since so little information is available about the dataset, users must download each dataset to their local computer (Figure 1). Once the data are accessible locally they may be processed and manipulated to determine if their content matches the user’s application. Evaluation techniques may include visual inspection of the raw records or loading them into a database management system for querying or into a GIS for spatial analysis. This method of selecting a dataset has the distinct advantage of not requiring an access mechanism beyond a file directory at the data archive end. In this “as is, where is” exchange there are no promises made, nor expected, about the content of the data. The ease with which the data are obtained is quickly offset by the cost of using the data. The user must have a local means of evaluating the dataset and visual inspection of all of the data records is impractical on very large spatial datasets. The number of interrelating factors that are inherently part of spatial

data quickly overwhelms a user on even small spatial datasets, which means some form of geographic information system is necessary to begin to evaluate fitness.

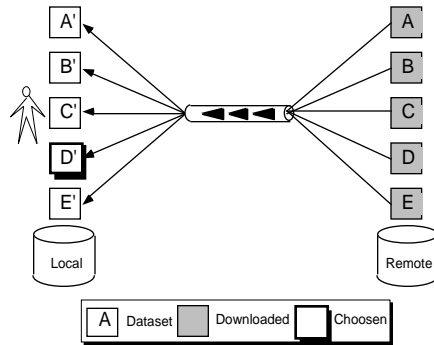


Figure 1: Choosing datasets in a simple data archive.

2.2 *A priori knowledge-based choice*

A significant improvement over examining all datasets is to rely on the advice of domain experts who can recommend a dataset that would meet the user's needs. This professional advice quickly leads to a few datasets of interest; however, due to the dynamic nature of on-line sources, the users cannot be certain that they have the most appropriate dataset. Despite its limitations, this is currently the primary method used by most on-line researchers and digital librarians to locate new data sources. *A priori* knowledge-based selection (Figure 2) has many factors in its favor in terms of resource expenses. Since only a few datasets are transferred, the load on the network is much lower. Although it may have a fee associated with it, it can be more cost effective than analyzing multiple datasets. Processing costs on the local computer and total dataset fees are also lower because of the smaller number of datasets.

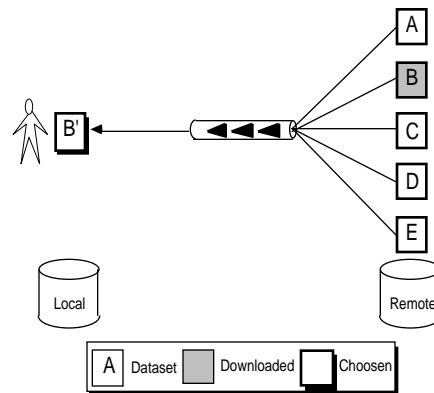


Figure 2: *A priori* knowledge-based choice.

Two issues remain unresolved if *a priori* knowledge-based access is the only method of choice: (1) comprehending the information in the dataset and (2) confidence in the dataset. The scientist's *comprehension* of the dataset's information content can be as difficult as it was with multiple downloads. The dataset must still be loaded into a geographic information system before meaningful analysis can begin and the process of loading and processing such a large dataset can be time consuming. If the topic being studied requires a portion of the data, such operations can most easily be done on spatial data after the load is complete.

Confidence (i.e., trust) in a dataset selected with *a priori* knowledge is heavily dependent upon one's confidence in the source of the knowledge. Users have no means to confirm that the recommended dataset is better than other available datasets without downloading those other datasets. In a situation where the validity of the dataset might be contested (i.e., a courtroom) a narrowly focused analysis may be suspect. Likewise, scientists usually strive to test their hypotheses with the best data available. What is needed is a method to get a feel for a dataset's content and validity before downloading the entire file.

2.3 Metadata-supported choice

Dataset providers and large users are addressing this requirement through metadata and metadata content standards, such as the Content Standard for Digital Geospatial Metadata (CSDGM)(FGDC 1997). CSDGM requires data providers to include a core minimum set of dataset characteristics. Such metadata assist the potential user of the dataset to understand the formats of the enclosed data, their history, and accuracy. In digital spatial archives that support CSDGM it will be possible to download a dataset's metadata. The user may then review them before investing time and resources in any particular dataset (Figure 3).

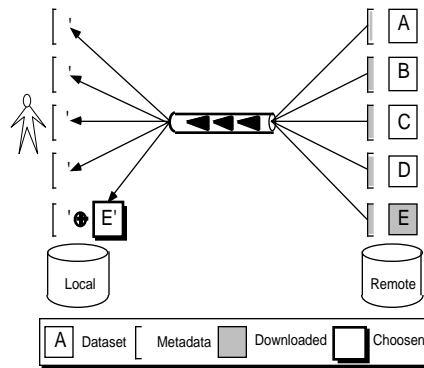


Figure 3: Metadata-supported choice.

Metadata-supported choice clearly has advantages over either of the previous two methods used to finding suitable datasets. While it creates a slightly higher load on the network—metadata must be moved in addition to the dataset of choice—it improves access significantly, particularly when the datasets are very large. The scientist also has the advantage of being able to make an informed choice with more relevant facts without relying on another expert. As a result of this informed choice, confidence that the dataset is appropriate—in a very subjective sense—for the task should increase. There are, however, loopholes in current spatial data standards that could affect the utility of the data quality reports. While the standard provides a means to report the data quality, its information content is currently not standardized, but rather relies on comment fields from the publishing agency. If the publishing agency is not thorough in its description of completeness then the dataset may still turn out to be useless for a particular inquiry. Worse yet, if the metadata description is incomplete important datasets may be overlooked when searched by metadata. The metadata approach provides information and understanding about the dataset, but not necessarily comprehension of the dataset's imbedded concepts, which constrain the utility of the dataset.

2.4 Subset-supported choice

Comprehension of the dataset comes from exploratory analysis of the data (Bresnahan *et al.* 1994). Since this approach can be computationally intensive for very large spatial datasets, analysis of samples of the data is a routine method used by scientists to speed understanding of the limits of

the dataset. Ideally the subset should be data that are representative of the full set, where the term *representative* means that a subset can stand in the place of the superset for a given application or decision-making process and generate highly similar results. Therefore, the subset should have the broad characteristics needed to make an assessment. For instance, if a scientist was looking for a high-resolution image of a city surrounded by agricultural land and near the ocean, it should be possible to determine that a certain image is not going to meet the needs without downloading all of it (Clementini *et al.* 1990). In digital image archives, thumbnail images are often used for exactly this purpose and permit rapid visual confirmation of an image's suitability (Figure 4). As a subset of the image, the thumbnail may contain 1% of the data and nearly 100% of the information needed to make a decision about further use. Subsets for other types of spatial datasets could be used in a similar manner (Flewelling 1997).

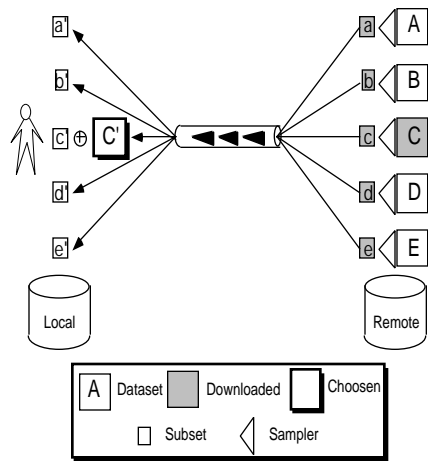


Figure 4: Subset-supported choice.

2.5 Choice with “unlimited” resources

Proponents of a digital society (Negroponte 1995) argue that it is pointless to assume that data users will have the same limits of computing, storage, and network bandwidth that we experience in the late-1990s. They envision a world in which data virtually pours into the user's lap at nearly unlimited speed (Figure 5). However, such a resource-rich environment still would not significantly address the issue of data comprehension, and in fact might only serve to exacerbate the situation. With an ever increasing ability to bring vast amounts of data to the user, the complexity of data analysis over these digital worlds slowly approaches the complexity of the “real” world. A larger dataset does not imply better information.

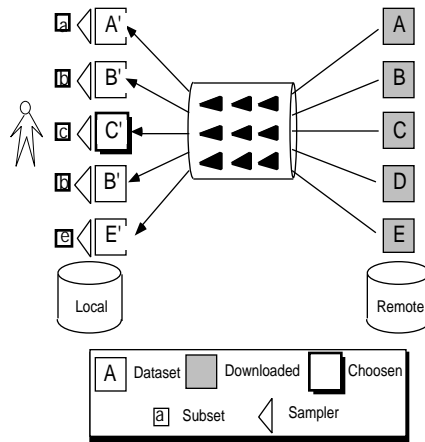


Figure 5: Choice with “unlimited” resources.

The availability of large bandwidth and local storage are appealing and suggest that it would be technically feasible to pursue the previously discussed model of simply downloading everything that looks interesting (Figure 1), since this style of using data could become more economical than paying an expert for advice. However, having data is not the same as having information. Local users need appropriate mechanisms that enable them to extract information from those data accessed. Subsetting still plays a significant role in such resource-rich environments, bridging the gap between the abundance of data and the relatively small amount of information people are capable of processing.

3. Conclusions

Studies in cognitive science have shown that people have difficulty comprehending large amounts of detail (Bruner *et al.* 1956; Miller 1956). Therefore, abstraction methods are necessary to extract the essence of large amounts of data, and to use these abstractions in decision-making and reasoning (Neisser 1976; Norman 1993). Abstraction methods may transform the objects into higher-level objects (aggregation), assign typical characteristics to a single object (prototyping), or may filter out, or select, objects that represent a concept found in the original set.

Metadata as they are currently being implemented in a large number of data collections are seen as a means to describe these higher level concepts. As such, the generation of metadata is primarily a data provider issue; however, metadata are often limited to the uses of the datasets as foreseen by the data provider and encoded into metadata standards. Searching with subsets may overcome these limitations. The issues surrounding the generation of subsets are less those of standardization, but related to the technological advances necessary to support it. Such needs include, but are not limited to:

- defining access mechanisms to do subsetting,
- extensions to information systems to support subsetting,
- research of what spatial concepts should provided by subsetting systems, and
- research on which spatial measures capture high-level spatial concepts.

When users are searching for a dataset that meets their analysis needs it would be helpful if they could use a small subset of the dataset in combination with other metadata elements to make preliminary decisions about suitability and fitness. Since the similarity is user-defined the subset would need to be generated on the fly or would need to be assessed with regard to its similarity to the user’s criteria.

5. References

- BEARD, K., 1987, How to survive on a single detailed database. In *Proceedings of Auto-Carto 8*, edited by N. R. Chrisman (Baltimore, MA: ASPRS & ACSM), pp. 211-220.
- BRESNAHAN, P. J., COWEN, D. J., EHLER, G. B., KING, E., SHIRLEY, W. L., and WHITE, T., 1994, Using Geographical Data Browsers in a Networked Environment. In *Proceedings of Sixth International Symposium on Spatial Data Handling*, edited by T. C. Waugh and R. G. Healey (Edinburgh, Scotland, UK: Taylor and Francis) pp. 921-932.
- BRUNER, J. S., GOODNOW, J. J., and AUSTIN, G. A., 1956, *A Study of Thinking*, (New York: Wiley).
- CLEMENTINI, E., D'ATRI, A., and FELICE, P. D., 1990, Browsing in Geographic Databases: An Object-Oriented Approach. In *Proceedings of Workshop on Visual Languages*, (Skokie, IL: IEEE Computer Society) pp. 125-131.
- FEDERAL GEOGRAPHIC DATA COMMITTEE (FGDC), 1994, *The 1994 Plan for the National Spatial Data Infrastructure*. Federal Geographic Data Committee. Washington, D.C.
- FEDERAL GEOGRAPHIC DATA COMMITTEE (FGDC), 1997, *Content standard for digital geospatial metadata (revised April, 1997)*. Standard Federal Geographic Data Committee. Washington, D.C.
- FLEWELLING, D. M., 1997, *Comparing Subsets from Digital Spatial Archives: Point Set Similarity*. Ph.D., University of Maine,
- FLEWELLING, D. M., and EGENHOFER, M. J., 1993, Formalizing Importance: Parameters for Settlement Selection. In *Proceedings of 11th International Conference on Automated Cartography*, edited by R. McMaster (Minneapolis, MN: American Congress on Surveying and Mapping), pp. 167-175.
- GARCIA-MOLINA, H., WIDOM, J., WIENER, J., LABIO, W., LENT, B., and ZHUGE, Y., 1995, *A Warehousing Approach to Data and Knowledge Integration*. Briefing: Stanford University.
- LEVY, D. M., and MARSHALL, C. C., 1996, Going Digital: A Look at Assumptions Underlying Digital Libraries. *Communications of the ACM*, **38**, 77-84.
- MILLER, G. A., 1956, The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81-97.
- NEGROPONTE, N., 1995, *Being Digital*, (New York: Knopf).

NEISSER, U., 1976, *Cognition and Reality: Principles and Implications of Cognitive Psychology*, (San Francisco: W. H. Freeman).

NORMAN, D. A., 1993, *Things that Make Us Smart: Defending Human Attributes in the Age of the Machine*, (Reading, Massachusetts: Addison-Wesley Publishing Company).

SMITH, T., 1996, Alexandria Digital Library. *Communications of the ACM*, **38**, 61-62.

WEIBEL, S., CATHRO, W., and IANNELLA, R., 1997, The 4th Dublin Core Metadata Workshop Report. URL <http://www.dlib.org/dlib/june97/metadata/06weibel.html> (Accessed October 10, 1997).

WIDOM, J., 1995, Research Problems in Data Warehousing. In *Proceedings of 4th International Conference on Information and Knowledge Management (CIKM)*, edited by N. Pissinou, A. Silberschatz, E. Park, and K. Makki (ACM Press), pp. 25-30.