

## **Improving Uncertainty Management Through Readily Accessible Metadata**

Valerie L. Carney, University of Maine and U.S. Army Engineer Research and Development Center ([Valerie.L.Carney@usace.army.mil](mailto:Valerie.L.Carney@usace.army.mil))

### **Abstract**

The importance of metadata is becoming more apparent with the high cost of data generation and the increased use of geospatial data clearinghouses and warehouses. Metadata allows data users to determine the fitness of a data set for their particular needs. However, current implementations of metadata typically consist of text files that are separate from the geospatial data. When a question is raised about the quality of the geospatial data, a user must find the metadata file and search it manually. This approach interrupts the workflow and is inefficient for specific quality-related questions. This paper proposes an integrated environment where links between geospatial data and their quality-related metadata will be established. In this environment, quality-related information about the geospatial data will be readily accessible to the user. Additionally, this paper presents a set of metadata operations that correspond to related geospatial data operations (i.e., query, display, update). These operations will allow quality-related metadata to be queried by interaction with a geospatial data unit, updated as geospatial data are transformed, and displayed graphically as attributes of the geospatial data.

### **Introduction**

With the emergence of geospatial data warehouses and clearinghouses and the high cost of data generation, data sharing has become more common. With the increase in data sharing, data users need to evaluate the fitness of the geospatial data in question – they need to determine whether the data are good enough for their needs. Data users can search clearinghouse nodes for available geospatial data by querying the metadata stored in the clearinghouse. After evaluating the metadata, a user can decide if the data appear suitable to their needs, then identify how the data producer provides access to the data. The user typically receives both the geospatial data and the associated metadata text file. Having a separate metadata text file is acceptable for the use described. However, when a data user is working with the geospatial data and questions arise regarding the quality of the data, the user must find the metadata file and search it manually. This approach interrupts the workflow and is inefficient for specific quality-related questions. Additionally, if the user modifies the geospatial data, the metadata must also be updated in order to maintain the integrity of the data. However, there is a good chance that these updates are not made since the user must actually make these changes manually in a separate text file.

This paper proposes an integrated environment where links or relationships between geospatial data and their quality-related metadata will be established. These links will enable quality-related information about the geospatial data to be readily accessible to the user. This information becomes more important as time passes. As personnel change within an organization, undocumented data may lose their value. New personnel will

have little understanding of the quality of the data and may not want to use the data for analysis (FGDC Secretariat, 1998). In turn, this could lead to an organization spending more money to either generate or purchase a new data set of known quality.

This paper also proposes a set of metadata operations that correspond to related geospatial data operations, in addition to links between the geospatial data operations and metadata operations. With operations and links established, users would be able to perform operations such as query, display, and update on the metadata. These operations and links would enable users to query quality-related metadata by interaction with a geospatial data unit, update metadata as geospatial data are transformed, and display metadata graphically as attributes of the geospatial data.

### **What is Metadata?**

Metadata is data about data. It provides information about the producer, content, quality, and other characteristics of the data. Why is metadata important? Firstly, metadata help insure an organization's investment in data. As personnel change or time passes, information about an organization's data will be lost and the data may lose their value. Collecting and updating metadata will insure its value. Secondly, by making metadata available through data catalogs and clearinghouses, organizations can find data to use, partners to share data collection and maintenance efforts, and customers for their data. And finally, the metadata will help the users that receive the data to process and interpret data, incorporate data into its holding, and update internal catalogs describing its data holdings (FGDC 2000). Therefore, metadata should accompany the transfer of a data set.

Most of the current work with metadata in the field today includes reviewing and developing metadata standards to determine where improvements can be made (i.e., Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM), Open GIS Consortium (OGC) Abstract Specification for Metadata, and the International Organization for Standardization (ISO) Metadata Standard), in addition to developing techniques for the creation and implementation of metadata.

Since generating metadata can be a laborious task, many organizations have developed metadata generation tools to ease the task of creating metadata. According to Hart and Phillips, metadata tools may be separated into four categories based on their operating characteristics and function: 1) intelligent, 2) forms-based, 3) ASCII and word processor templates, and 4) utilities. Intelligent tools automatically extract information, such as bounds, projection information, attributes, and vector feature count, from the geospatial data sets. They do not perform all documentation, however. The user needs to provide the descriptive information such as the abstract, contact and distribution information, and explanation of attributes. Forms-based tools provide a user interface that assists the user with the documentation process. The user interface typically consists of a series of forms with fill in boxes or drop down lists. Some of these tools also indicate which elements are optional and which are mandatory. ASCII and word processor templates are not actually metadata tools. They are strictly text editors and word processors that are used to edit template documents that contain the metadata elements. The last category, utilities,

includes tools and services that are used to process metadata in some form instead of actually producing it. Examples include tools to find data sets, to pre-process metadata into consistent format, and to validate metadata (Hart and Phillips 1998).

Metadata standards in use today focus on documenting metadata at the data set level. This is not sufficient for those interested in metadata at the feature level. With the development of web servers for geospatial information, the need for more detailed metadata is increasing. The Open GIS Consortium has developed a Web Feature Server where data users can insert, delete, update, query, or retrieve specific features of interest of multiple sources (OGC 2001). This capability will require feature level metadata in order for data users to understand the quality of the geospatial data that they are using for analysis.

As a short-term solution, Hart and Phillips suggest that a limited amount of feature level metadata could be stored as the attributes of each feature (Hart and Phillips 1998). But this will not be sufficient as the exploitation of multi-source data continues to increase and Web Feature Servers grow in popularity. This paper suggests a metadata model that would allow for feature level metadata in an integrated environment.

### **Conceptual model of integrated environment**

The proposed integrated environment lends itself well to object-orientation because of the characteristics that object-oriented models have: structural characteristics, such as classification, generalization, and aggregation; and behavioral characteristics, such as inheritance and propagation. In this paper, I will only address classification.

A class is a set of objects (or instances) that are all of the same “type”. For example, a common class found in a geospatial database is a road class, which consists of several instances of roads. The class provides the blueprint for all of the instances within that class (R.C.M. Consulting, et. al 1993). Using the road example, the road class would define the variables required for each road object. Figure 1 shows the road class with instance variables for storing the road type, surface material, width, number of lanes, and road name. There are many road objects in the database, however there is only one road class.

<b>ROAD</b>
Road Type
Surface Material
Width
Number of Lanes
Road Name

*Figure 1: Road class with instance variables*

Each instance of the road class has the same kinds of instance variables. They each have the same structure, however, the content within the variables is different (R.C.M. Consulting, et. al 1993). Each instance contains the attributes that pertain to the particular

road that the object represents. Figure 2 shows an example of three road objects with the values of their respective instance variables.

Major Road	Divided Highway	Residential
Asphalt	Concrete	Asphalt
48	60	32
4	4	2
Hayfield Rd.	S. Kings Hwy.	Duddington Dr.

Figure 2: Three road objects and their values

Before mapping this concept to metadata, it is beneficial to look at a high-level conceptual model that integrates geospatial data with their metadata (see figure 3). In the object-oriented environment, everything is an object. Not only is every instance of geospatial data and metadata an object (represented by the circles), but the collections of geospatial data, metadata, and classes are also objects. In this environment, even the relationships between the geospatial data and metadata are objects (Goldberg and Robson 1989).

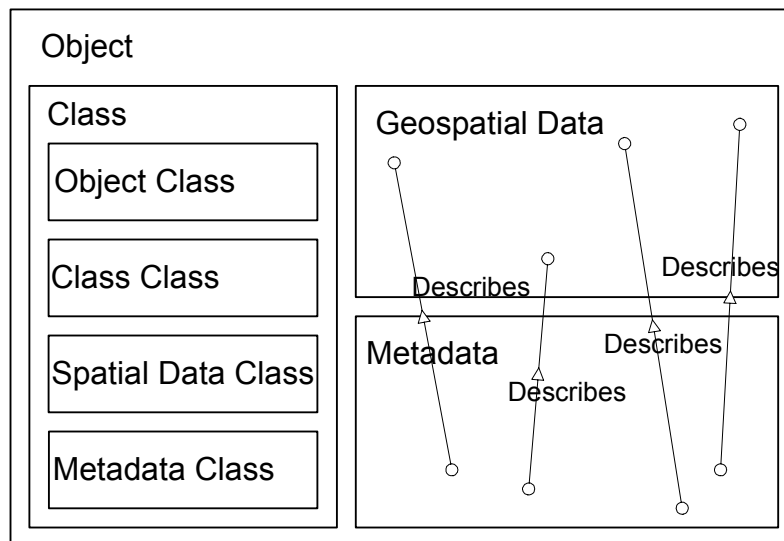


Figure 3: Integrated model in an object-oriented environment

This conceptual model shows that each instance of metadata is related to an instance of geospatial data by the “describes” relationship (i.e. one instance of metadata describes one instance of geospatial data). With this high-level model in mind, we can now narrow the focus to the metadata model. For the purpose of this paper, the metadata standard used is the FGDC’s CSDGM. The CSDGM is divided up into seven main sections and three supporting sections as shown in figure 4 (FGDC 2000).

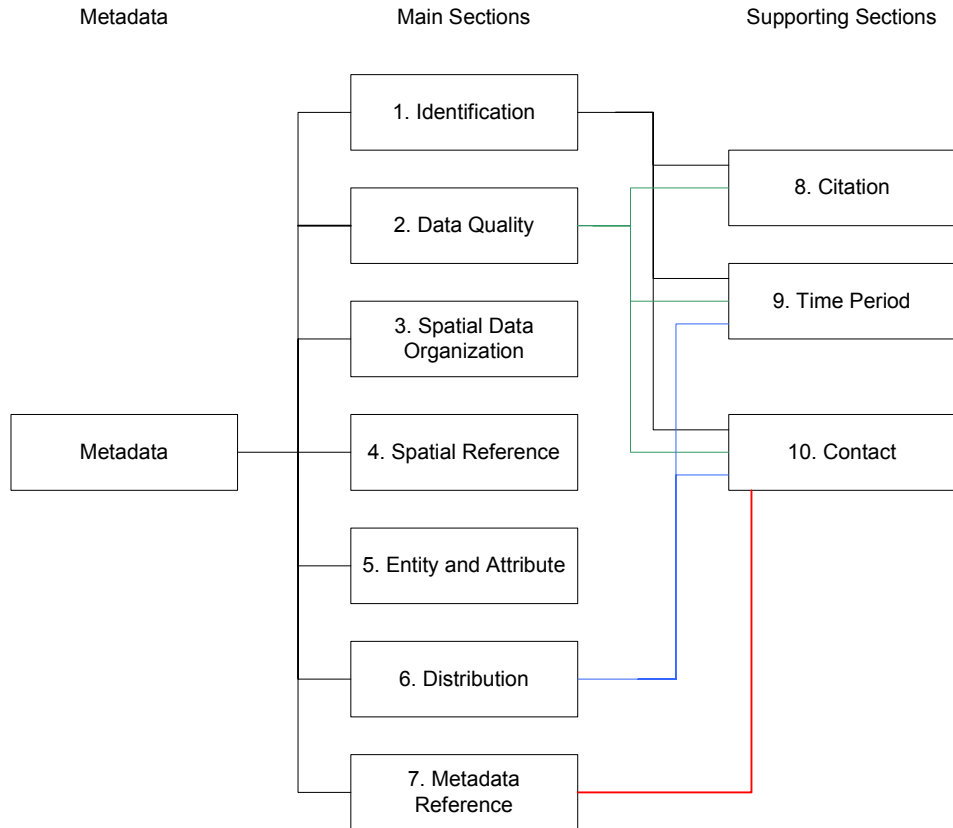


Figure 4: Sections of the FGDC CSDGM (FGDC 2000)

This paper will focus on the data quality portion of the metadata, which itself is comprised of six sections with a number of subsections as shown in figure 5. The three-dimensional boxes in the graphic indicate data entry fields (FGDC 2000). Just as with the road class, we now have a metadata class that provides the blueprint for all of the instances within the metadata class. Therefore, each instance of the metadata class will have the same kinds of variables. Once again, the structure of each instance will be the same, however, the content will differ (R.C.M. Consulting, et. al 1993). For the sake of simplicity, we will collapse the data quality information into just the six sections without the subsections. In this example, we will assume that the road features were collected from different source data and at different levels of accuracy. Therefore, each instance of the road class has different metadata associated with it. Figure 6 illustrates how the data quality information is related to the road objects by the “describes” relationship. Other classes of geospatial data in the GIS would also have their related metadata linked to their geospatial data. With these relationships established between the geospatial data and the metadata, we can now start thinking about operations on the metadata.

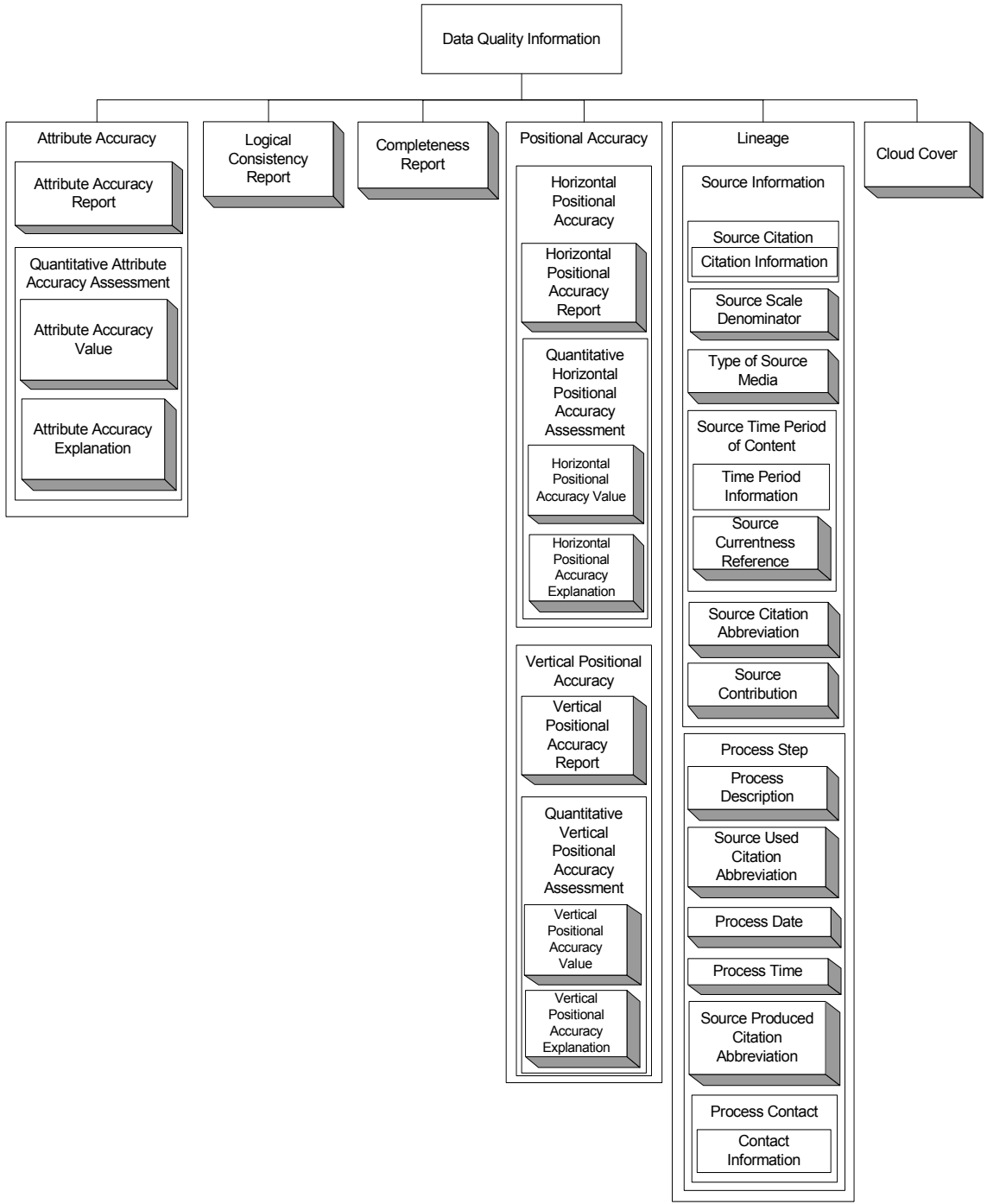


Figure 5: A graphic representation of the data quality information from the FGDC CSDGM (FGDC 2000)

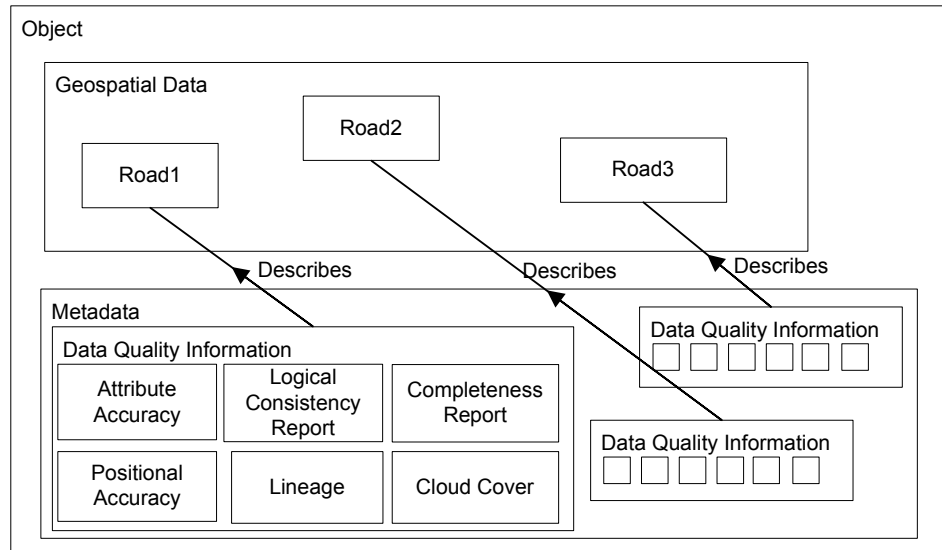


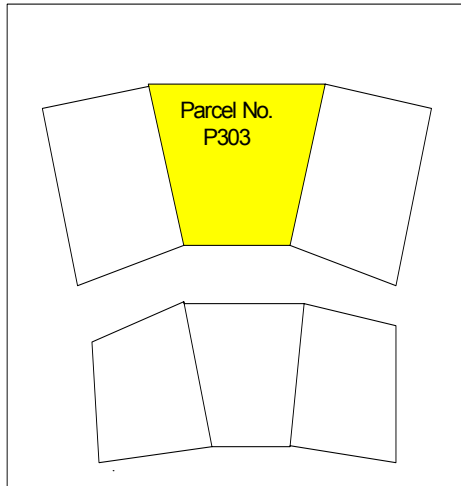
Figure 6: Integrated environment linking the data quality information to the road objects

## Metadata operations

As discussed earlier, the technology in use today focuses on metadata generation of which there are a lot of tools available to assist data producers with this task. However, there are no metadata *operations* to manipulate the metadata, and there are no *links* to the geospatial data. Therefore, when any operation is performed on the geospatial data, nothing is done to the metadata. For example, if the geospatial data undergoes a coordinate transformation, the metadata will only be updated if it is done manually. Capabilities to automatically perform these operations on metadata whenever the geospatial data is changed need to exist.

This section describes a set of metadata operations, such as query, display, and update, to manipulate metadata in the database. The first operation is query. For purposes of discussion, GIS county parcel data will be used as an example. A user can query the geospatial data to ask for information about a particular parcel. The query checks the parcel attribute table and returns information such as the owner, parcel size, assessed tax value, tax assessor, number buildings on parcel, geographic coordinates, etc. However, the user might also want to know the reliability or currency of this information. In this case, the user would want to query the metadata and ask such questions as: what is the positional accuracy of the parcel? How old is the tax information? Who was the tax assessor? By linking the metadata to the geospatial data, the user will have immediate access to this information.

The second operation is display. With a display metadata operation, a user could conceivably toggle between the display of the geospatial data with their attributes and the display of their metadata in order to examine the quality of the geospatial data being used. Figure 7 shows a display of parcels in a database and subset of the parcel attribute table. Figure 8 then shows a display of the positional accuracies using error ellipses.



Parcel ID	Owner	Size	Tax Value
P303	R. Smith	.55	\$150,000

Figure 7: Geospatial data and their attribute table

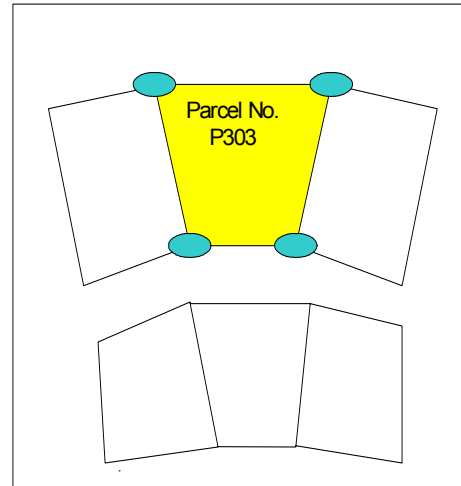


Figure 8: Error ellipses showing positional accuracies of the parcel boundaries

The last set of operations is the set of update operations. Update is a general term that can include operations such as edit, insert, delete, reclassify, and overlay. Each of these geospatial operations would require a specific type of update operation on the metadata. For the sake of discussion, this paper uses the reclassification operation as an example. Reclassification is simply replacing existing values with new values. One of the most common reasons to reclassify data is to generalize or simplify the data, which in turn, reduces the number of classes (McCoy and Johnston 2001). For this operation, vegetation data will be used as an example. For instance, an analyst may want to group 13 various types of vegetation in a database into seven vegetation classes using a reclassification operation. With a metadata reclassification operation and links established, the instances of metadata for the vegetation classes that were grouped together would have to be merged. Furthermore, the lineage would need to be updated to include information describing what operations had been done to the geospatial data, who made the changes, and when were they made.

### Discussion and Future Work

This paper proposed an integrated environment using an object-oriented model where the geospatial data and their metadata are linked, in addition to proposing a set of metadata operations to manipulate the metadata. The end result of this research, therefore, will be an integrated model that will provide data quality information with minimal interruptions to workflow. Additionally, users will be able to perform operations on both geospatial data and metadata simultaneously, which in turn, keeps consistency in the geospatial database. Without this capability, metadata updates are extremely tedious and, in all likelihood, not performed at all.

Benefits of this integrated model will be seen by any organization using multi-source data where they would need to have this capability in order to understand the quality of their geospatially-referenced products. Furthermore, with the development of Web Feature Servers, this capability would be beneficial for users that retrieve features from the web since they would have metadata at the feature level.

Future work includes further developing the conceptual model of the integrated environment, in addition to expanding the set of metadata operations that correspond to other geospatial data operations, such as overlay, rotation, translation, and scale transformation. Once the conceptual model and the set of metadata operations are finalized, I will develop a physical model of the integrated environment. The physical model will then be implemented as a prototype in a commercial GIS software package, where it will be tested and evaluated for effectiveness.

### **Acknowledgements**

The author would like to thank Dr. Kate Beard of the University of Maine and Mr. Brian Graff and Mr. Richard Joy of the U.S. Army Engineer Research and Development Center for reviewing and providing comments on this paper.

### **References**

FGDC Secretariat (1998). The Value of Metadata. Reston, VA, U.S. Geological Survey; National Spatial Data Infrastructure: 1-3.

FGDC (2000). Content Standard for Digital Geospatial Metadata Workbook Version 2.0. Federal Geographic Data Committee. Washington, DC.

Goldberg, A. and D. Robson (1989). Smalltalk-80: The Language, Addison-Wesley Publishing Company.

Hart, D. and H. Phillips (1998). Metadata Primer – A “How To” Guide on Metadata Implementation, NSGIC, <http://www.lic.wisc.edu/metadata/metaprim.htm>.

McCoy, J. and K. Johnston (2001). Using ArcGIS Spatial Analyst. Redlands, CA, ESRI.

OGC (2001). OpenGIS Implementation Specification – Topic Web Feature Server Implementation Specification, Version 0.0.14, p. 84.

R.C.M. Consulting, Multimedia Productions, Inc., and Open Strategies, Inc. (1993). The Open Systems Video Library - Object Oriented Design.