DEVELOPING AN INFRASTRUCTURE FOR SHARING ENVIRONMENTAL MODELS

Scott J. Crosier, Michael F. Goodchild[1], Linda L. Hill, and Terence R. Smith, National Center for Geographic Information and Analysis, Alexandria Digital Library, and Department of Geography, University of California, Santa Barbara, CA 93106-4060.

ABSTRACT

The Internet and the World Wide Web offer a new solution to the problem of sharing scientific knowledge. Unlike traditional libraries based on print media, these new technologies facilitate the sharing of any information that can be expressed in a binary alphabet. Environmental models expressed as computer codes are instances of such information objects, and codes are the last of a four-stage process of model formulation. The transition to digital technologies changes the relative importance of the four stages. We present a six-stage model of the process of searching for information in distributed digital libraries. Search is made possible by

[1] Corresponding author, Department of Geography, University of California, Ellison Hall 3611, Santa Barbara, CA 93106-4060, USA. Phone 1 805 893 8049, FAX 1 805 893 3146, Email good@geog.ucsb.edu

metadata, which serve several distinct purposes, including description of the contents of information objects and their fitness for specified use, and the details needed to make use of information objects once accessed and retrieved. We present a strawman structure for model metadata, explain the process used in its development, and invite its evaluation.

INTRODUCTION

The past decade has seen explosive growth in the use of digital information technology. Early computer applications stressed numerical processing and large-scale data management – applications in which the computer acted as a human assistant, performing functions that humans found difficult or tedious to perform by hand. Today, however, digital computer technology is seen more as the means by which humans communicate information, and geographic information systems (GIS) in particular as a set of tools for communicating geographic information (Goodchild, 2000; Sui, in press; Sui and Goodchild, 2001). A complex set of standards allows virtually all kinds of information to be expressed in commonly understood binary formats, including the text standard ASCII and UNICODE, the image standards TIFF and JPEG, the music standards MP3 and MIDI, and the U.S. Spatial Data Transfer Standard for geographic

information (ANSI, 1998; Morrison, 1987). These standards in turn allow massive economies of scale to be achieved, since all information can be stored, processed, and shared using computers and networks built from the same digital components and operating on a simple binary alphabet.

But the benefits of this success are not yet ubiquitous. Consider the case of scientific knowledge, which is expressed in many forms. Scientists have traditionally communicated their findings in the form of text, augmented with equations and figures, and a massive investment has been made in the associated infrastructure of libraries, publishers, mail and package delivery services, standards and protocols, and personal collections. Much effort in recent years has gone into improving this infrastructure by exploiting electronic media, allowing much more rapid and effective processes of search, browse, retrieval, and sharing of the information that traditionally was transmitted through printed books and journals. Many journals are now published in electronic form, and indexed in online directories. The role of the research library as a means of access to the storehouse of printed knowledge is changing, as more and more of its services are offered online, and the technology already exists to store all of the text contents of a major research library in digital form, and to allow

tens of thousands of users to read this collection independently and simultaneously through network connections.

Similarly much progress has been made in the electronic sharing of data. The traditional library model of cataloging documents has been extended to the structured description of data through such metadata standards as the Content Standard for Digital Geospatial Metadata (CSDGM) of the U.S. Federal Geographic Data Committee (1998), the metadata specification for educational materials developed by IMS Global Learning Consortium, Inc. (IMS, 2001) which has been adapted for use by the Digital Library for Earth System Education (DLESE, 2001), and the more general Dublin Core Metadata Initiative (Dublin Core Metadata Initiative, 2001). All of these standards assist the process of search for suitable data and other materials, by informing potential users of the properties that are likely to be important in determining a data set's appropriateness and fitness for a given use. In effect, metadata standards perform the digital equivalent of the library's cataloging function, by capturing a succinct and searchable description of the contents of a data set, just as a catalog indexes books by author, title, subject, and location in the library, and supports search by presenting each index in alphabetically sorted order. Metadata standards extend that function, however, by

capturing a description of the properties that are likely to be important in handling and using the data set, such as details of formats; and properties that are likely to be important in helping the potential user assess the data set's fitness for a given use. Major searchable archives and warehouses of digital geographic data can now be found on the Web, including the FGDC's National Geospatial Data Clearinghouse (U.S. Federal Geographic Data Committee, 2001), ESRI's Geography Network (ESRI, 2001), and the Alexandria Digital Library (Alexandria Digital Library Project, 2001). Each of these collections also provides search tools based on common metadata standards. Additional organizations such as UNIDATA (University Corporation for Atmospheric Research, 2001) and the network of National Geophysical Data Centers (U.S. Department of Commerce, 2001) exist primarily to provide the infrastructure needed to support sharing of data by scientists.

But in comparison with data, other forms of scientific knowledge have received much less attention, and little infrastructure yet exists for sharing them. This paper addresses one such form, the dynamic environmental model. It begins in the next section with a discussion of the nature of environmental modeling as an activity, and the increasing importance of sharing the resulting models beyond the immediate research

group. The following section examines the sharing of environmental models within the context of the infrastructure that already exists to support sharing of more favored forms of scientific knowledge. The paper then focuses specifically on metadata, and reports on work aimed at the development of a metadata standard for dynamic environmental models, as a necessary foundation for an infrastructure. The final sections of the paper describe our proposed standard, which we term the Content Standard for Computational Models (CSCM). Our intent in proposing a standard is not to argue that it represents a final solution, but to draw attention to the need for a standard, and to present a strawman structure to stimulate further research and discussion.

For the purposes of discussion two contrasting examples of models are used in the paper. One, known as SLEUTH (Clarke et al., 1997; Clarke, 2001), models urban growth, simulating the transition of land use based on factors of slope, proximity to roads, proximity to existing urban areas, etc., using the techniques of cellular automata (Toffoli, 1987). The other models the development of slopes and the formation of drainage channels during the process of erosion (Smith et al., 1997a, 1997b; Smith, 2001), through a computational model based on two partial differential equations (PDEs).

We have used both models as examples in the development of the proposed standard.

DYNAMIC ENVIRONMENTAL MODELS

DEFINITIONS

Dynamic models exist in many forms, but the domain of this paper is somewhat narrower, and some basic definitions are introduced in this section. First, *geographic information* is defined as information about the state of the Earth's surface and near-surface. An item of geographic information has the form of the tuple <**x**,**z**>, where **x** denotes a location in a space-time frame that is limited to the Earth's surface and near-surface, its history and its future, and **z** denotes one or more properties associated with that location, such as temperature or surface elevation, or country name. Geographic databases are constructed from large numbers of such tuples, through the processes of data modeling defined below. Geographic information is limited to a range of spatial and temporal resolutions; few applications appear to require spatial resolutions outside the range from roughly 1 m to 10 km.

Dynamic models attempt to predict future states, in other words, to predict geographic information at future times (models are also sometimes

run into the past, largely for purposes of validation by comparing outputs to known prior conditions, but also to fill gaps in past knowledge, for example in archaeology). A dynamic *environmental* model predicts properties that are primarily associated with the physical environment, though human dimensions are also relevant. The dynamic models addressed in this paper are also *spatially explicit*, a term that is used in a variety of ways in the literature. Here, it means that the model has a finite geographic *domain*, and that predictions of the model vary across that domain, in other words that the output of the model is geographic information in the sense defined above. It follows that the predictions of the model are not invariant under relocation. Moreover, the number of locations for which predictions are made is large and possibly infinite; models that address a geographic world divided into a small number of domains (e.g., two) are not spatially explicit in the sense of this paper, though they may be termed so in the literature. Besides the outputs, it is likely that one or more of the inputs of the model is also spatially explicit.

This definition of spatially explicit explains the concern with coupling such models with geographic information systems (GIS), which are designed as repositories and processors of geographic information, including the inputs and outputs of models. The basic objective of this paper

is to investigate the infrastructure that would be needed to facilitate the sharing of such models, while recognizing the progress that has been made in recent years in building a comparable infrastructure for sharing GIS data.

The focus of this paper is on *computational* models, in other words, computer codes that emulate the operation of some real environmental process. A dynamic environmental model must be *executable*. In the context of sharable digital information, a dynamic environmental model is a collection of bits, just like a data set or an image, but with the difference that it can be executed in a computer in combination with inputs to produce predictions. Thus many dynamic models will have distinct operating system file extensions, such as .exe in the Windows world, that serve to differentiate them from collections of bits representing documents, images, or data sets, all of which are readily communicated using present technology.

The term *model* has many meanings, even in the context of dynamic environmental modeling. Modeling in the sense of *data modeling* is not within the domain of this paper. Data modeling addresses issues of representation of data, and distinguishes between vector, raster, relational, object-oriented, and other options. Data modeling may be static, if what is to be represented does not change through time, but it may also be dynamic

(Frank et al., 2001; Langran, 1992; Peuquet, 1994). However data modeling includes no possibility of *prediction*, or the creation of new states at future times; and it focuses entirely on knowledge of form, or how the geographic world *looks*, whereas dynamic environmental models focus on how it *works* by representing knowledge of process. Hereafter the term *model* will be used in the sense of dynamic environmental model as defined above.

In this paper we also address those models that compute properties of an environmental system from other properties measured at the same point in time, such as models that compute evapotranspiration from measurements of precipitation, temperature, humidity, and other variables; or models that compute solar insolation. Although not strictly dynamic, such models are also expressed in the form of executable computer codes, and we anticipate that they can also be described effectively using the metadata approach proposed in this paper.

Any model is a form of abstraction or generalization, since perfect description of the geographic world is impossible. The scientific community traditionally values dynamic environmental models highly because they suggest a high level of generalization about the world; much higher, for example, than a single measurement, or a data item from a GIS database. In the loose hierarchy that ranges from data through information, knowledge,

understanding, and wisdom (and for an alternative typology see Checkland and Holwell, 1998), we suggest that dynamic environmental models rate higher than many other collections of bits, including measurements of environmental phenomena or digitized maps.

The *inputs* to a model include a representation of the initial state of the system, often termed the *boundary conditions*. In the case of dynamic environmental models these include GIS layers representing such varied phenomena as land cover type, topographic elevation, water depth, soil pH, or atmospheric temperature. Such layers may also serve to parametrize models, if geographic variation in a parameter such as slope plays a role in the operation of the model. We define a *parameter* of a model as a value that remains constant throughout the execution of the model; a parameter may take a single value over the spatiotemporal application domain of the model, or may vary through space or time, or both. We define a *variable* of a model as a value that changes during the execution of the model; a variable may also be constant or varying over the spatial domain of the model, and the *outputs* of a model are variables.

Many model inputs are unique to the spatiotemporal domain of a model's application. For example, a model of tides in an estuary requires unique input data representing the configuration of the coastline and the

estuary's bathymetry. Other inputs may be constant over many different applications of the model, and may have been calibrated when the model was initially developed and evaluated. Such inputs clearly need to be bundled with the model for sharing. In other cases it may be important to share the input data associated with a specific application, and again such inputs may need to be bundled with the model for sharing, and described in a single form of metadata associated with the bundle. We have allowed for this situation in the proposed standard for computational-model metadata.

STAGES OF MODELING

Many efforts have been made to create taxonomies of models (e.g., Chorley and Haggett, 1967; Goodchild et al., 1993; Sklar and Costanza, 1990). For the purposes of this paper it is useful to adopt a typology that distinguishes four stages of modeling, at increasingly specific levels of representation, in both syntactic and semantic terms. These are the conceptual, symbolic, algorithmic, and coded representations of the model (Figure 1). The proposed metadata standard provides a structure for these levels of description through both narrative elements, and elements for specific details of input and output variables, parameters, data sets, and processing flow.

[Figure 1 about here]

The conceptual representation describes the model at the highest level. For the erosion model, for example, it would characterize the model in terms of land and water surfaces, the conservation of water flowing over a surface, and the conservation of sediment eroded from the surface and transported by the water. The symbolic representation is typically, but not always, in terms of some mathematical or logical language, with an interpretation of the symbols in terms of real-world phenomena. In the case of the erosion model, this representation takes the form of two PDEs. The algorithmic representation provides a high-level view of how the symbolic representation is converted into a set of computations, while the coding representation of these algorithms provides codes that are, or can be compiled into, executables in some specific computing environment. The erosion model, for example, is specified at the algorithmic level by indicating that the water flow equation is transformed into finite difference form using an upwind scheme, and that the land surface erosion equation is transformed into finite difference form using a Crank-Nicholson scheme (Smith et al., 1997b). At the coded level, the model is specified by a set of C-language programs and the environment in which they would run. Hence we may view the information represented in these four categories as moving

from a high-level description of the model and its applicability to the details needed to execute it in a specific computational environment.

The urban growth model, on the other hand, is specified conceptually through the rules that are applied in every iteration of the cellular automaton. Rules are used to assign probabilities of transition to cells, based on the parameters input for each cell, and on the current state of each cell and its immediate neighbors. There is no equivalent in this model of the PDEs of the erosion model's symbolic stage, and the cells do not function as the units of a finite difference approximation. Instead, the algorithmic stage follows directly as a computational formalization of the conceptual stage.

The models that are the focus of this paper became feasible in the 1960s with the widespread availability of digital computers. Since the invention of the printing press, methods for sharing scientific information have been dominated by print media, including books and journals. In this medium it is comparatively easy to express the conceptual and symbolic stages of modeling, through appropriate descriptions, rules, and equations. But while some more specialized books and journals include algorithms, often in semi-formal languages described as pseudo-code, and may also include printed computer codes, despite the obvious cost of transcribing

printed code to machine-readable media, clearly print media are far more conducive to sharing the results of the first two stages of modeling than the latter two.

More recently it has become feasible to include machine-readable media, notably the diskette and compact disk, as inserts into books and journals. But the development of the Internet and the World Wide Web (WWW) has now made it feasible to share the results of modeling at all four stages, and greatly reduced the effort required to transfer a modeling capability from one scientist to another. It has also radically altered the economies of scale associated with sharing scientific information. Print media are much more economical per unit to produce when production runs are large, and as a result favor the sharing of materials that are of widespread and general interest over materials that are comparatively specialized. The Internet on the other hand is virtually free to its scientific users. Even after all of the necessary preparation by the scientist is complete, it can take months or years to publish and disseminate a model using print media, but only minutes using the Internet.

These changes suggest a shift in the relative priorities associated with the four stages. The symbolic stage is perhaps the most abstract and succinct, but the coding stage has become the most easily communicated

and shared. Moreover the symbolic stage has little meaning for cellular automata and other models that are written directly as algorithms. Although the symbolic stage of PDEs is inherently independent of the details of implementation, such details are explicit in models based on cellular automata, where the size of the cell is a basic parameter of the model, and where it is difficult to generalize from one cell size to another.

The popularity of the WWW has undoubtedly changed the habits of researchers and students, because it is far easier to search for material or a reference using a WWW browser than to visit the local research library, and more and more scientific knowledge appears in the first instance on the WWW, whether or not it eventually appears in print (see, for example, the e-Print archive pioneered by Paul Ginsparg; Los Alamos National Laboratory, 2001). It also may have changed the relative importance of print and digital media in the sharing of models, and may eventually allow more researchers to execute models than read about them in print. Of course easy access to models carries its own dangers, if it invites use without adequate understanding or critical evaluation.

SHARING INFORMATION OBJECTS

THE PROCESS OF SHARING

The dominance and limitations of print media are consistent with a particular model of science, in which the individual investigator is responsible for every stage from problem formulation and experimental design through data collection, analysis, interpretation, and generalization. In this model only the results of the project are shared, in the form of summary journal articles that present the methods and the major results in succinct text form. Methods are traditionally presented in sufficient detail to allow another investigator to reach the same conclusions, but many relatively unimportant details, such as the raw data, the computer code used to perform numerical analyses, or the coded stage of any dynamic model, are often not communicated. In such situations the reader's ability to rerun the experimental process, with the same or new data, can be impaired.

But this model no longer presents an accurate description of the way science is done in many disciplines. Re-analysis of raw data is increasingly important in such fields as medicine, where *meta-analysis* is emerging as a paradigm for comparative analysis of many potentially conflicting prior studies. Individual scientific expertise is now so specialized that the solution to a problem often requires the collaboration of many investigators with different disciplinary backgrounds. Most importantly for this paper, the output of one person's science may be in increasing demand as the input to

the science of others; for example, a model of part of the Earth system developed by one investigator may be a valuable component of an integrated model of several parts being developed by another team, provided integration is both technically feasible and scientifically valid. In such situations it is obviously most efficient to share models at the coding stage, together with sufficient documentation.

We conceptualize the process of sharing scientific information in six stages, independently of the type of information being shared (Figure 2). First, the potential recipient of information attempts a *definition of need*. Depending on the type of information, this need is expressed in the appropriate form of metadata. For example, a need for geospatial data can be formalized as a CSDGM metadata record. Second, a mechanism of *search* is instituted across some set of collections. In the case of geospatial data this stage might be formalized as a search across the collections that are components of the National Geospatial Data Clearinghouse (U.S. Federal Geographic Data Committee, 2001). Third, the search process results in the *discovery* of information objects with metadata similar to the defined need. Fourth, a mechanism must exist to make an *assessment* of these discovered objects and their *fitness for use*. In the case of data sets this might occur either formally through a measurement of the similarity

between each data set's metadata and the defined need, or informally through a user-directed process of browse and evaluation. Fifth, the user must be able to initiate a *retrieval* of the information object, unless the capability exists to work with the object in its current location (for example, some geospatial data archives offer the ability to perform certain common GIS operations on archived data sets). Finally, the capability must exist to perform an *opening* of the contents of the information object, perhaps using information about the object that is formalized in its metadata, such as details of a data set's internal format.

[Figure 2 about here]

Many aspects of this scheme are difficult to formalize in practice, but some degree of formalization is essential if metadata are to be processed automatically. In the case of data sets, the CSDGM and other metadata standards include many items that defy straightforward coding, and must instead be provided in the form of text. For example, the items defining data quality will almost always be supplied as descriptive narrative, rather than in any formalized way. In such cases it is virtually impossible to devise metrics of similarity between the items in a data set and the items in a specified need. An obvious exception is the geographic coverage of a data set, which can be defined precisely. For example, the Alexandria Digital

Library defines geographic coverage as a sequence of latitude and longitude coordinates, and allows similarity of coverage to be assessed in various ways that are under the user's control. In such situations the needed coverage and the discovered data set's coverage can be compared to produce a simple binary result (data set *overlaps*, data set *lies within*, data set *contains* the needed coverage), or a metric of similarity of coverage that can be used to rank results, such as the ratio of the square of the area of intersection to the product of the areas of the two coverages. But other properties of data sets defy such simple formalization.

Instead, many search processes resort to a simple expedient. Only those properties of information objects that can be precisely defined are identified by the user in the specification of need, and these are assessed either by requiring an exact match, or by metrics such as the one just described. All other properties are left undefined in the definition of need. Discovered information objects can be ranked by the metrics, but all other issues must be resolved informally by browsing through the discovered objects. Clearly this process can be highly inefficient, but it represents a pragmatic solution to a real technical problem. In effect, it is virtually impossible to formalize the process of assessement completely.

The need to rely on informal browse also results from another general issue associated with the description of information objects. Objects such as geospatial data sets can differ on a vast number of dimensions. Although it is desirable for the metadata description to be very much shorter than the data set itself, only the data provide a complete description of themselves; and metadata often include additional properties that are abstracted, interpreted, or generalized from the data and thus not derivable from them through any simple procedure. Thus the potential exists for metadata to be more voluminous than the data themselves. But in practice the task of describing information objects is expected to be substantially less than the earlier task of creating the objects, and there is little incentive to undertake precise and massive description. Instead, assessment is usefully seen as an iterative process, in which the metadata associated with an information object are used to discover an initial set of candidate objects; and the metadata include pointers to additional sources of information, such as telephone numbers or bibliographic references, that could provide the additional metadata needed for a subsequent stage of resolving among the discovered candidates. We have adopted that principle in the design of the metadata standard proposed in this paper.

THE ROLES OF METADATA

It should be apparent that metadata play several roles in the search process, and in this section we attempt to enumerate them. First and most obviously, metadata provide a description of the contents of an information object. We argued that such descriptions are essentially hierarchical; short descriptions are needed to support the initial process of discovery, but longer and more detailed descriptions are needed to support assessment.

A major aspect of this description concerns quality, and characterizations of quality play an essential role in supporting assessment. We understand quality to refer to a comparison between the actual contents of the information object, and the real-world phenomena that the contents are intended to represent (this definition applies both to data sets and to models). Quality defined in this way implies a knowledge of intent, which we interpret to mean the potential user's knowledge of the producer's intent, and this understanding of quality has won widespread acceptance with respect to geospatial data (see for example Longley et al., 2001). But it raises a difficult issue, because the user's knowledge must also be obtained from the metadata. For example, suppose that a metadata record describing a geospatial data set contains no specification of the geodetic datum used to define coordinates. The user is left uncertain about the datum, which can

imply an uncertainty regarding positions of as much as several hundred meters. We use the term *specification* to refer to the producer's description of intent, and regard it as an essential component in the assessment of quality from metadata.

The quality of models can be assessed in many ways. The developer of the model may have validated it using standard procedures, by comparing model predictions with observations, or by incorporating submodels that have already been calibrated and have known quality. Information may be available documenting prior efforts by others to use the model, with formal or informal reports on the results. Demonstration runs of the model might be linked to the model's metadata, together with reviews by peers and other more casual assessments. Formal comparisons of competing models of the same processes are also sometimes available. All of these are potentially useful ways of addressing model quality.

Besides defining need and supporting assessment, metadata are also required to play essential roles in retrieval and opening. To the Internet all information objects are simply collections of bits, and are communicated identically. But geospatial data sets are often very large, so the method of retrieval may need to be conditioned on the data set volume. Opening also clearly requires knowledge of formats, data structures, programming

languages, hardware requirements, and other details of the internal organization of the information object. This is particularly true in the case of models, because of the large number of coding languages in use, and wide variation in system requirements.

In summary, metadata play numerous roles that go well beyond the summary description of an object's contents. The standard proposed in this paper attempts to address all of these roles, and to allow sufficient flexibility to accommodate different priorities and requirements.

COMMONALITIES IN DESCRIPTION

The focus of this paper is on metadata for models, within the domain defined earlier. It is possible in principle to think of domains at any scale, and to devise metadata standards for any of them. The Dublin Core standard (Dublin Core Metadata Initiative, 2001) was devised to describe information objects over a very broad domain, much broader than the domains of the CSDGM (U.S. Federal Geographic Data Committee, 2001) or our proposed standard. There are of course economies associated with very broad domains, but also diseconomies. Particularly, the broader the domain the more difficult it is for a metadata standard to support precise discovery. For example, the Dublin Core standard does not include

geographic coverage as a core formalized element, making it difficult to use this standard to discover data sets based on geographic overlap with the area of need. On the other hand this is a formalized element of the CSDGM standard, and of the derivative metadata standard of the Alexandria Digital Library (Alexandria Digital Library Project, 2001).

Our approach attempts to defuse the scale issue somewhat, by adopting design principles that are similar to those of other standards, and by referencing other standards wherever possible. For example, a metadata standard that requires reference to people should adopt standardized ways of identifying people, rather than invent its own. If different classes of metadata follow similar design principles, it is possible to build generic digital library technologies that are able to operate across many different domains. For example, the similarities between our standard and the CSDGM allows digital library search engines to use the same approach to comparing geographic coverages. Our approach is based on the ISO metadata standard for geographic information (International Organization for Standardization, 2001), which is based in turn on the CSDGM (U.S. Federal Geographic Data Committee, 1998). Our approach also is compatible with the generic *bucket* middleware feature of the Alexandria Digital Library, which was designed to support simultaneous search over

collections that use different forms of metadata. In other words, describing different types of information objects using similarly designed but domain-specific metadata standards can lead to greater interoperability between digital collections.

MAKING MODELS EASIER TO SHARE

Many attempts have been made to create sharable collections of models or otherwise to improve the infrastructure of model sharing, and such efforts date in some cases from well before the advent of the WWW (see, for example, Benz and Knorrenschild, 1997; Mowrer, 1997; Squire, 1990). Web-based collections and directories are maintained by, for example, Old Dominion University's Department of Civil and Environmental Engineering (ODU, 2001), the U.S. Army Corps of Engineers Construction Engineering Research Laboratory (CERL, 2001), and the ECOBAS project of the University of Kassel (ECOBAS, 2001). Few if any of these efforts took place within the framework of print-dominated libraries, and in general it seems that libraries have not paid much attention to the sharing of this particular form of scientific knowledge. The impediments to doing so have already been identified: their comparative recency, the difficulty of handling digital media in a print-

dominated institution, and the specialized nature of models. However, these problems are largely absent in digital libraries, and we believe that our effort is the first to consider the problem of sharing models within the broad framework that they provide, and to develop methods that are compatible with digital library technology.

The complexity of computing environments is reflected in the vast range of forms of coded models. Model codes vary by source language (e.g., C vs C++), specifications of the host computing system (e.g., conventional vs parallel), type of operating system (e.g., Unix vs PC), formats and structures of data inputs and outputs (e.g., raster vs vector), and many other factors. Clearly the task of sharing would be easier if such aspects could be standardized. This might be done, for example, by requiring that all models be coded as scripts in some standard high-level language, and this approach has been used successfully by several GIS, including PCRaster (Karssenberg et al., 2001; Netherlands Centre for Geo-Ecological Research, 2001), which is based on the modeling language devised by van Deursen (1995). Efforts have been made to define sets of elementary components that can be reassembled as models (e.g., Bennett, 1997; HPS, 2001; Villa, 1999), and this approach is consistent with recent trends in software engineering towards reusable software components, as

reflected for example in Microsoft's COM standard. We assume in this paper that the modeling field is not likely to converge around such scripting and component standards in the near future, and that an infrastructure for sharing models must be designed with a broader perspective.

Our focus is on the six-stage process defined above, as the basis for search across distributed collections of models assembled in digital libraries. The key to this process is metadata, and we now discuss several approaches to metadata definition.

Just as the only complete description of a data set is the data themselves, similarly the only complete description of a model is the model itself, in one of its four forms (conceptual, symbolic, algorithmic, or coded). But we have already argued that metadata must serve purposes beyond description, and that even description can involve abstractions, interpretations, and generalizations that are not themselves contained in the model.

Another approach might be based on the nature of models, which transform input data to output data. In principle, then, a model can be described fully through a description of its inputs and its outputs, using standard ways of describing data sets. But there are several arguments against this approach. First, a model is a set of *general* rules for

transforming a class of inputs to a class of outputs, rather than a *specific* transformation of one set of inputs to one set of outputs. The task of abstracting the general rules from specific instances is fundamental to science, and the basis on which many models were initially created. Thus this approach to metadata would require us to automate the same process of generalization, and it is difficult to imagine that this could be done simply and without the involvement of human intelligence. Moreover, this approach like the previous one would yield only a subset of the properties that are important in the search process.

A third approach might be based on a study of the actual practices of scientists, because the existing infrastructure for sharing models is entirely embedded in human communication and discourse. Scientists currently find models by asking each other about them, by hearing rumors and incomplete descriptions, and by reading about them. The priorities a scientist gives to different aspects of a model when describing it to others may be a useful guide to the design of a metadata standard. We have relied on this user-centered approach in designing our standard, and we suggest that it also forms a reasonable basis for evaluation of our standard.

AN OVERVIEW OF THE PROPOSED STANDARD

In this section we briefly review the contents of the proposed standard. As noted earlier, our purpose in doing so is not to promulgate a standard, or even to argue for its adoption by some authority such as ISO (International Organization for Standardization, 2001), ANSI (American National Standards Institute, 2001), or ASTM (American Society for Testing and Materials, 2001). Rather, our purpose is stimulate greater research interest in the topic through the presentation of a strawman.

The CSCM consists of approximately 165 elements divided into ten sections:

1. Identification information
2. Intended use
3. Description
4. Access or availability
5. System requirements
6. Input data requirements
7. Data processing
8. Model output
9. Calibration efforts and validation
10. Metadata source

The following overview of the purpose of each section appears in a previous paper (Hill et al., in press). The full CSCM and example metadata records for the two models included in our discussion here can be found at our Web site (Alexandria Digital Library, 2001).

CSCM DESCRIPTIVE DESIGN

Identification (CSCM Sections 1 and 3)

Identification elements supply the basic citation information, such as title, responsible parties, version, date, and identification numbers. Description elements include conceptual and symbolic-algorithmic descriptions of the model, model typology, topic or field of study, geographic and temporal coverage, and links to related models and to additional information about the model being described.

Functionality (CSCM Sections 6, 7, and 8)

Functionality is described in terms of input data sets, modeling constructs (parameters and variables), data processing steps, and the characteristics of the output data. This also includes any post-processing procedures required on the data. A potential user should be able to use this information to evaluate the model for fitness of use; a current user should be able to use this information to understand how to link data to the model and

how the model uses the data. Eventually, a computer service should be able to use these data to evaluate candidate data sets for their suitability for use with a model.

Fitness for use (CSCM Sections 2 and 9)

Creators of models have information about the intended use of the model, in terms of the intended application, including the geographic areas and time periods for which the model is believed valid, and, if designed for an education purpose, the intended educational level. This information is useful, along with the description of the conceptual, symbolic, algorithmic, and processing details, to determine if a particular model is suited for a particular use.

One of the key purposes of model documentation is to provide information about the calibration and validation tests that have been used, the experiments that have been run, the peer reviews that have been published, and the current known uses. Particularly useful is a citation to a data set that can be used to test the model. Some of this information will accumulate through time as the model is used and may exist independently of the metadata description of the model itself. However, to the extent possible, having citations to external sources containing reviews and experiments will be very valuable for evaluation of fitness for use.

Access and constraints (CSCM Section 4)

Metadata need to clearly explain how to obtain the model and all administrative and legal considerations that might limit its use. Possible constraints include cost and ownership issues. Access information includes email and mail addresses for the access person or organization, ordering procedures, and the URL for direct download, if possible. Related access and use considerations are described in the Environment elements.

Environment (CSCM Section 5)

Both human and computational environments for the model need to be explained to a potential user. Human requirements include the expertise needed to obtain, install, and run (access and open) the model, and to interpret the results. System requirements include the hardware and operating system for which the model was designed, and any auxiliary software required.

Metadata documentation (CSCM Section 10)

The source of the metadata must also be documented to record the creator of the metadata, the creation and modification dates, the sources of information providing the metadata content, and the metadata standard and version that was used. It is important to know if the metadata were created by the model creator or by someone else, and how to contact this person in

case there are corrections to make or questions to ask about the metadata themselves.

DISCUSSION AND CONCLUSION

We have identified methods for sharing models as a critical missing piece of the current infrastructure for sharing scientific models. Models are clearly most easily shared at the coding stage, and we have proposed a six-stage process for sharing coded models, from the definition of need to the opening of a model for execution. Appropriately designed metadata are the key to each of the stages of this process, and we have proposed a strawman to stimulate discussion of suitable approaches to a metadata standard.

The proposed framework places much of the onus for metadata creation on the developer of a model, as the person who is best able to populate its elements. Additional valuable input can be provided by other researchers whose experiences in using the model are a potentially important component of the assessment of a model's fitness for use. A successful infrastructure should provide the means for registering such additional inputs, and making them accessible to users. This process of accumulating and disseminating test results and comments has the effect of incrementally extending the model's domain and usefulness, and is the

means whereby models become widely accepted in the research community. In this way the infrastructure for sharing becomes the mechanism for building trust in a model.

Fundamental to the concept of a digital library is the notion that many of the activities and services of the traditional library are possible candidates for automation. A user of a digital library conducts an automated search through a store of metadata, replacing the tedious task of searching by hand through card catalogs, and extending its effectiveness in many ways. On the other hand not every aspect of library use is suitable for automation. In our model of the search process the human user is involved at several stages, notably in the definition of need and in browsing. We suspect that it will never be possible to automate fully the assessment of fitness for use, or the determination of whether a given model is capable of accepting a given data set as input. Nevertheless, automation can greatly reduce the number of options open to the human, including the size of the set that must be browsed, and the number of candidate data sets for input to a given model. Our proposed standard attempts to narrow options in several respects, notably with respect to those elements such as geographic coverage that can be defined precisely and formalized. Other elements must

be left to human judgment, and presented in a user interface that is designed to make the task as easy as possible.

These arguments suggest a number of possible ways of assessing the success of an infrastructure for sharing models. First, the infrastructure should minimize the magnitude of the human tasks associated with the search process, including the initial definition of the need, and the numbers of models that must be assessed individually by the user during the browse process. Second, it should support the accumulation of information from multiple sources that can aid the user in assessing fitness for use, including access to comments and experiences in using the model. Third, it should encourage honesty in modeling by making it possible to elicit, capture, and disseminate well-defined and objective measures of model success. Finally, the infrastructure will be successful to the extent that it results in a greater degree of sharing of models.

Our proposed metadata standard should not be seen as a replacement for model documentation, or for the tradition of describing model development and assessment in the scientific literature. Instead it is designed as a tool to assist in the search process; in the spirit of the previous paragraph, it is designed to automate as much of that process as is reasonable, to support initial assessment of fitness for use, and to point the

researcher to sources of additional information. Interested readers are encouraged to examine the more detailed information at the Alexandria Digital Library Web site (www.alexandria.ucsb.edu; additional background information is at http://www.geog.ucsb.edu/~scott/metadata/index.html), and to undertake their own evaluations of the proposed standard.

ACKNOWLEDGMENT

The Alexandria Digital Library's Digital Earth Prototype (ADEPT) is supported by the National Science Foundation under Cooperative Agreement 9817432. This research is also supported by the Center for Spatially Integrated Social Science (CSISS.org), which is supported by the National Science Foundation under Award 9978058.

REFERENCES

Alexandria Digital Library Project, 2001 *Homepage*, http://www.alexandria.ucsb.edu

American National Standards Institute, 1998 *Information Technology - Spatial Data Transfer Standard* (ANSI NCITS 320-1998), http://webstore.ansi.org

American Society for Testing and Materials, 2001 *Homepage*, http://www.astm.org.

Bennett D A, 1997, "A framework for the integration of geographical information systems and modelbase management" *International Journal of Geographical Information Science* **11**(4) 337 - 357

Benz J, Knorrenschild M, 1997, "Call for a common model documentation etiquette" *Ecological Modeling* **97** 141 - 143.

Checkland P, Holwell S, 1998 *Information, Systems and Information Systems: Making Sense of the Field* (Wiley, Chichester)

Chorley R J, Haggett P, Editors, 1967 *Models in Geography* (Methuen, London)

Clarke K C, 2001 *SLEUTH Urban Growth Model* (Version 2.1), http://www.ncgia.ucsb.edu/projects/gig/project_gig.htm

Clarke K C, Hoppen S, Gaydos L, 1997, "A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area" *Environment and Planning B: Planning and Design* **24**(2) 247 - 261

Construction Engineering Research Laboratory, *2001 On-Line Catalog of Conservation Technologies*, http://owww.cecer.army.mil/pl/catalog/index.cfm

Digital Library for Earth System Education, 2001 *DLESE Metadata Working Group Homepage*, http://www.dlese.org/Metadata/index.htm

Dublin Core Metadata Initiative, 2001 *Homepage*, http://www.dublincore.org

ECOBAS, 2001 *WWW-Server for Ecological Modelling*, http://dino.wiz.uni-kassel.de/ecobas.html

Environmental Systems Research Institute, 2001 *Geography Network Homepage*, http://www.geographynetwork.com

Frank A U, Raper J, Cheylan J-P, 2001 *Life and Motion of Socio-Economic Units* (Taylor and Francis, London)

Goodchild M F, 2000, "Communicating geographic information in a digital age" *Annals of the Association of American Geographers* **90**(2) 344 - 355

Goodchild M F, Parks B O, Steyaert L T, Editors, 1993 *Environmental Modeling with GIS* (Oxford University Press, New York)

High Performance Systems, Inc., 2001 *STELLA*, http://www.hps-inc.com

Hill L L, Crosier S C, Smith T R, Goodchild M F, in press, "Developing an infrastructure for sharing environmental models" *D-Lib Magazine*

IMS Global Learning Consortium, Inc., 2001 *IMS Meta-data Specification Version 1.2*, http://www.imsproject.org/metadata/index.html

International Organization for Standardization, 2001 *Homepage*, http://www.iso.ch/iso/en/ISOOnline.frontpage.

Karssenberg D, Burrough P A, Sluiter R, de Jong K, 2001, "The PCRaster software and course materials for teaching numerical modelling in the environmental sciences", *Transactions in GIS* **5**(2) 99 - 110

Langran G, 1992 *Time in Geographic Information Systems* (Taylor and Francis, London)

Longley P A, Goodchild M F, Maguire D J, Rhind D W, 2001 *Geographic Information Systems and Science* (Wiley, New York)

Los Alamos National Laboratory, 2001 *e-Print archive*, http://xxx.lanl.gov

Morrison J, Editor, 1987, "A draft proposed standard for digital cartographic data transfer", *The American Cartographer* **15**(1) 1 - 140

Mowrer H T, 1997 *Decision Support Systems for Ecosystem Management: An Evaluation of Existing Systems*, (U.S. Forest Service, Fort Collins, Colo.)

Netherlands Centre for Geo-Ecological Research, 2001 *PCRaster*, http://www.geog.uu.nl/pcraster/tekst.html

Old Dominion University, 2001 *Civil/Environmental Model Library (CEML)*, http://www.cee.odu.edu/model/

Peuquet D J, 1994, "It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems", *Annals of the Association of American Geographers* **84**(3) 441 - 462

Sklar F H, Costanza R, 1990, "The development of dynamic geographic models for landscape ecology: a review and prognosis", in *Quantitative Methods in Landscape Ecology: The Analysis and Interpretation of Landscape Heterogeneity* Eds M G Turner, R H Gardner (Springer-Verlag, Berlin) pp 239 - 288

Smith T R, 2001 *Smith/Bretherton Erosion Model*, www.alexandria.ucsb.edu/doc

Smith T R, Birnir B, Merchant G E, 1997a, "Towards an elementary theory of drainage basin evolution: I. The theoretical basis" *Computers and Geosciences* **23**(8) 811 - 822

Smith T R, Merchant G E, Birnir B, 1997b, "Towards an elementary theory of drainage basin evolution: II. A computational evaluation" *Computers and Geosciences* **23**(8) 823 - 849

Squire G R, Hamer P J C, 1990 *United Kingdom Registry of Agricultural Models: 1990* (AFRC Institute of Engineering Research, Bedford, UK)

Sui D Z, in press, "GIS as media? Or how media theories can help us understand GIS and society", in *GIS and Society: An International Perspective*, E Sheppard, R McMaster, Eds (Taylor and Francis, London)

Sui D Z, Goodchild M F, 2001, "Guest editorial: GIS as media?", *International Journal of Geographical Information Science* **15** (5) 387 - 389

Toffoli T, 1987 *Cellular Automata Machines: A New Environment for Modeling* (MIT Press, Cambridge, Mass.)

U.S. Department of Commerce, 2001 *National Geophysical Data Center Homepage*, http://www.ngdc.noaa.gov/ngdc.html

U.S. Federal Geographic Data Committee, 1998 *Content Standard for Digital Geospatial Metadata*, http://fgdc.er.usgs.gov/metadata/contstan.html

U.S. Federal Geographic Data Committee, 2001 *National Geospatial Data Clearinghouse*, http://www.fgdc.gov/clearinghouse/clearinghouse.html

University Corporation for Atmospheric Research, 2001 *UNIDATA Homepage*, http://www.unidata.ucar.edu

van Deursen W P A, 1995 *Geographical Information Systems and Dynamic Models: Development and Application of a Prototype Spatial Modelling Language* (Koninklijk Nederlands Aardrijkskundig Genntschap/Faculteit Ruimtelijke Wetenschappen Universiteit Utrecht, Utrecht, Netherlands)

Villa F, 1999 *Integrated Modeling Architecture*,

http://iee.umces.edu/~villa/IMA

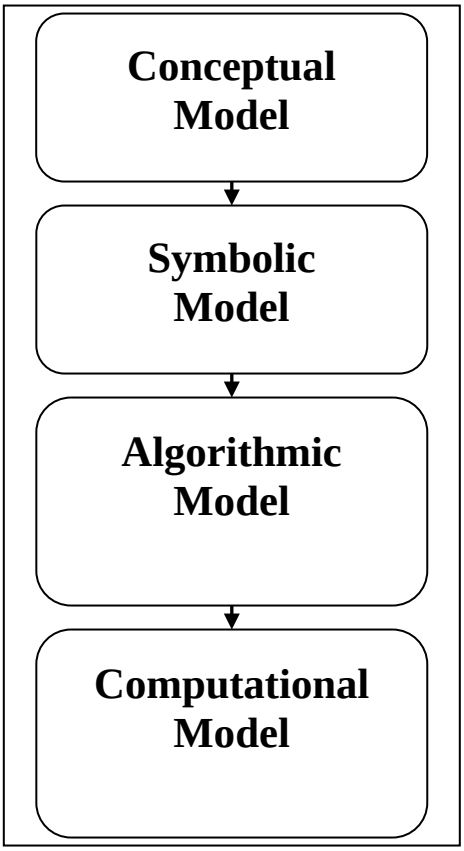

Figure 1: The four stages of modeling

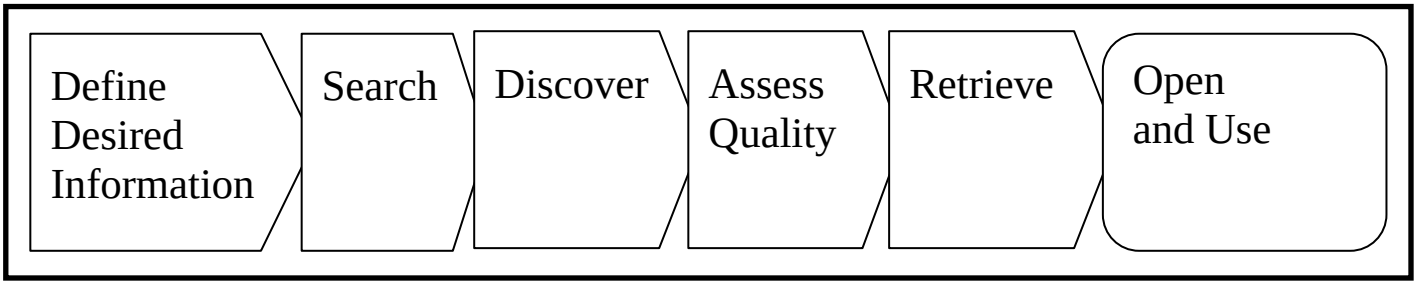

Figure 2: Six stages in the process of sharing scientific information