Open GIS and On-Line Environmental Libraries

Kenn Gardels

Center for Environmental Design Research
University of California
Berkeley, California 94720-1839

gardels@regis.berkeley.edu

ABSTRACT

An essential component of an Environmental Information System is geographic or geospatial data coupled with geoprocessing functions. Traditional Geographic Information Systems (GIS) do not address the requirements of complex digital environmental libraries, but are now incorporating strategies for geodatabase federation, catalogs, and data mining. These strategies, however, depend on increased interoperability among diverse data stores, formats, and models. Open GIS™ is an abstration of geodata and a specification for methods on geographic features and coverages that enables compliant applications to exchange information and processing services. For EIS, Open GIS provides an architecture for selecting geodata at its most atomic level, fusing those data into structured information frameworks, analysing information using spatial operators, and viewing the results in informative, decision-supporting ways.

1. Environmental Information Libraries

Environmental Information Systems are comprised of complex collections of a broad range of data types, including textual and graphic documents, digital geospatial map and imagery data, real-time acquired observations, legacy databases of tabular historical records, multimedia components such as audio and video, and scientific algorithms. On-line digital libraries of environmental information provide higher-order classification, cataloging, and human interface functionality for intelligent access to systems.

1.1. Objectives

The objectives of the users of environmental information systems are as diverse as their individual applications, but essentially distill to the intelligent access to and utilization of complex, distributed, heterogeneous information about the world.

Providers of environmental information want to ensure that users can learn of the existence of data stores, how they are constructed, and specific methods for accessing them. As importantly, they want users to know the limitations of the data, and its suitability for use in various application contexts.

Interests of data consumers are symmetric to providers' goals: they want to be able to locate, access, and interpret complex information, and have the tools that make this straightforward and error-free. Consumers also need information to make reasoned comparisons among different data sources supplying information on the same topic or area, and to validate that any data used are correct, current, and unambiguous.

1.2. Geographic Information Systems

A significant, arguably essential, component of any EIS comprises geodata and geoprocessing (geomatics) technology. Historically, GISs have been closed, monolithic systems that mitigated against integration into larger systems, whether those were more encompassing information or decision support systems, distributed systems, or ad hoc network-based collections or aggregations of geodata. Similarly, the functionality of GIS has not been amenable to interoperation with other data processing environments such as database management, statistical manipulation, or even desktop office applications. Many GIS vendors now recognize these limitations, and have begun to address them, albeit with mixed results.

1.3. Organizational Principles

As with any information system, there are degrees of regularization of how disparate datasets (whether centralized or distributed) are managed. No matter the degree of complexity, it is theoretically possible that a single overarching design can be implemented or maintained across different collections, sites, or user groups. In practice, however, systems arise and evolve independently. Geographic information systems in particular have historically been limited by software architectures that were not intended to support distributed data or enterprise-wide usage. As a result, data models and dictionaries tended toward discrete, customized, non-extensible solutions, virtually precluding the interoperability required for integration into larger information systems.

With requirements for integration of geodata into net-accessible libraries, three technical strategies evolved: federation, cataloging, and data mining.

1.3.1. Federation

The concept of GIS federation has reached its greatest success in the context of *frameworks* — collections of synoptic geodata mapped and classified to maximize utility to the broadest possible group of users [FGDC 1996a]. In essence, federation implies that a single dictionary can be applied to multiple datasets addressing the same theme. For example, the same overall classification of soil types could be applied to maps developed and used in different regions of the continent. Of course, there are numerous instances where a dictionary must be extended to address local conditions, but these can typically be accommodated using hierarchical classification systems.

Additionally, federation implies an organizational structure in which multiple data themes can be used concurrently, since this requirement is at the heart of GIS applications. Typically this manifests itself in the choice of a single software environment for data development, though this is neither necessary nor sufficient for effective integration. What is necessary is a common data model; this can be operationalized in the actual data collection, or only in data views generated for different users. Data model here means the conceptual view of a set of information, for example map layers, discrete features and objects, images, observations, or numeric or algorithmic descriptions. It is virtually impossible to cross-compare information compiled according to fundamentally different models without significant structural transformation and semantic translation.

Also necessary is a well-defined spatial reference system. Data do not all have to be stored using the same coordinate system, projection, datum, and so on, but this information most be available to facilitate conversion from one reference system to another where data are managed in different environments.

1.3.2. Catalogs

Cataloging approaches to distributed data collections and libraries utilize complete, well-structured metadata to enable use of disparate geodata. Whereas a federation may best be described by a shared dictionary, cataloged geodata is referenced via a thesaurus of commonly used terms, descriptors, data types, references, and structures. Metadata are used to accomplish at least three objectives in managing and using geographic information effectively: at a catalog level, metadata help to *identify* data that may be useful to a particular problem or application; at a more descriptive level, metadata help users *evaluate* the suitability of a dataset to their problem; and at a disclosure level, metadata provide the detailed information needed to *interpret* the data correctly.

Catalogs describe the thematic domain associated with specific datasets, and additionally enable browsing and searching via structured clearinghouse mechanisms. In effect, the metadata act as a wrapper around geodata, and inform the software mechanism used to access the data as to the correct protocols for returning information back to the original client. Catalogs may be implemented as a single directory, as is the case with the California Environmental Resource Evaluation System [CRA 1996], or as a network of registered servers, such as that being constructed for the National Spatial Data Clearinghouse [FGDC 1996b].

1.3.3. Data Mining

Data mining is perhaps an over-used term, but in the context of geodata libraries, it means using the spatial and non-spatial attributes of the data to search for and retrieve relevant information, absent any *a priori* knowledge of data set organization or content. A now well-established metaphor for this is web searching. Currently geodata mining is primarily a process of text (or web-page) searching for keywords describing geodata; these may or may not be housed in a metadata catalog as described above. There are limited capabilities for spatial searching now being developed, based on comparing a supplied string of coordinate values with the spatial extent of a dataset.

The assumption is that geodata thus located has been autonomously developed and maintained, and that it is essentially up to the user to determine applicability and meaning. Although intelligent browsers or user agents can help guide a search, and data transformation services can make information usable at some basic level, the semantic content must be inferred from multiple unrelated data characteristics, such as source, original scale, classification methodology, mapping standards, and so on. Effective exploitation of unknown, heterogeneous geodata collections remains an important research area.

2. Distributed Heterogeneous Environments

The physical landscape which GIS attempts to describe is in many ways a metaphor for the "data landscape" that users see. To some, it may be a rich, diverse environment supporting a large number of constituents; to others it may be an impenetrable jungle of unknown dangers; and to others it may simply be a vast wasteland. Successful utilization of the resources out there requires knowledge plus the tools necessary to apply information resources to solving problems.

2.1. Interoperability

Interoperability is the key to the use of heterogeneous data and processing resources throughout a networked environment in a single user session or workflow. Interoperability does not mean data standards or conversion tools, nor does it mean complex application suites. It does mean specified mechanisms for data exchange and software interaction, along with accepted protocols or authorities for defining them.

The fundamental requirements of GIS interoperability are:

- shared data space - a generic data model supporting a variety of analytical and cartographic functions
- compatible applications - a user workbench that is configurable to utilize the specific tools and data necessary to solve a problem
- heterogeneous resource browser - a method for exploring and accessing the spatial information and analytical resources available on a network

These three broad requirements must all be linked into an overall system architecture. Although each may be seen as a distinct set of capabilities, they all must coexist in a common framework that defines how system components interact. Of course, these components are themselves complex and have multiple levels and modes of interaction with each other.

The traditional mechanism for achieving the objectives of interoperability is format conversion — software tools for translating the structure of one system to that of another, typically using an intermediate, neutral format. Considerable effort has been invested in these tools, to better deal with issues of divergent data models. Although such strategies are rooted in an earlier generation of independent information systems, batch processing, and bulk data transfer, the file conversion paradigm persists in the new age of networked systems and online data access. Data standards like DIGEST [DGIWG 1996a] or SDTS [USGS 1993] are very limited in terms of retrieving individual geodata objects from a collection. Newer object-oriented protocols such as SAIF [SAIF 1994] or SQL3/MM(spatial) [ISO/IEC-JTC1/SC21 1996] facilitate on-demand data transfer, but still fail to address all the interoperability requirements listed above.

Commercial and research software developers are likewise addressing issues of interoperability, by extending traditional programming interfaces (APIs) to support distributed spatial query via standardized interfaces and geodata modeling languages. Database developers in particular (both relational and object-oriented) have recognized the capabilities for complex data management that their systems can bring to problems of GIS interoperability, not to mention the potential market of integrated geospatial and other data types in enterprise information systems.

2.2. Open GIS

True interoperability within the geomatics discipline is fundamentally driven by an open systems approach to geodata and geoprocessing. The open systems model is an approach to software engineering and system design that enables and encourages sharing of data, resources, tools, and so forth between different users or applications. When applied to the domain of geographic information systems, the intent is to move away from the current paradigm in which specific GIS applications and capabilities are tightly coupled to their internal data models and structures. Open GIS™ [OGC 1996] facilitates exchange of information not only between individual GISs but also to other systems, such as statistical analysis, image processing, document management, or visualization.

In 1994, the Open GIS Consortium was created to develop a consensus solution to issues of GIS interoperability. The long-range vision is the full integration of geospatial data and geoprocessing resources into enterprise information technology, and the widespread use of interoperable GIS software and data throughout the entire information infrastructure. OGC has defined its mission to ensure the collaborative development of interoperable geoprocessing technology specifications and to promote the delivery of certifiably interoperable products.

2.2.1. Architecture

The concept of open GIS is embodied in the Open GIS Abstract Specification [OGC 1996c], comprising the essential model, the open geodata model, and the open GIS services model. The essential model of open GIS describes the processes of associating the "real world" (as viewed by a particular community of geographic information users) with a formalized "project world" (world view), and then representing the project world in the formalisms of geographic data types, schema, and services [Kottman 1996]. Nine layers of abstraction have been defined, as shown in Figure 1. Five of these model the abstraction from real world to project world, and the interfaces between them: the essence of the real world is captured in the names and descriptions of the conceptual world; the conceptual world is modeled in the simplified or generalized constructs of geomatics in the geospatial world; the geospatial world is more rigorously defined in the metrics and spatial objects of the dimensional world; and finally the dimensional world is codified in the project world of a particular domain, along with its definitions of space, vocabulary, methods of observation, and classifications of entities and phenomena.

The remaining four abstractions define in increasing detail the mechanisms of encoding the elements of the project world in terms of the data and services models of Open GIS. In order, these are: the points world of coordinate geometry and spatial reference; the geometry world of well-known geodata types; the feature

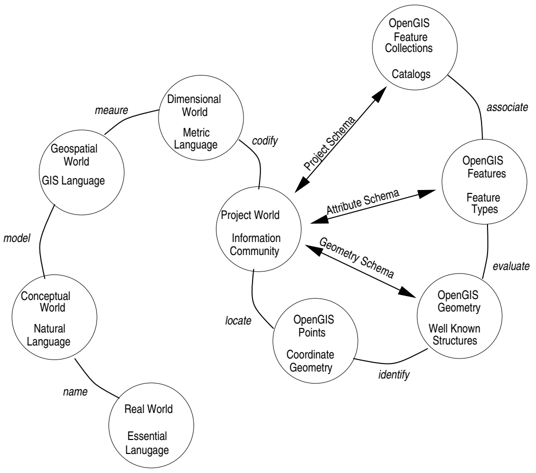

Figure 1. Open GIS abstract model, interfaces, and schema.

world which combines spatial and non-spatial attributes of geographic objects according to schemas of properties and their values; and the feature collection world which provides an overall structure and taxonomy to all of the elements within the underlying project world.

The Open GIS Geodata Model (OGM) comprises the language for a complete, consistent, coherent description of geography - that is, entities and phenomena which have a location and possibly an extent in the real world [OGC 1996b]. Entities are recognizable, discrete objects that have relatively well-defined boundaries or spatial extent. Typically, the spatial position of entities is perceived as secondary to the description of the entities. Phenomena, on the other hand, vary continuously over space and have no specific extent. A value or description of a phenomena is only meaningful at a particular point in space (and possibly time). Their mathematical representation consists of geometrical models of space in one, two, or three dimensions plus a temporal dimension as well as the spatial and temporal reference systems in which they are embedded.

The geographic elements of OGM are used to abstract or represent earth features. The Open GIS Geodata Model may be viewed as a type hierarchy, as in Figure 2. It shows that the generalized notions of dataset and feature may be decomposed into their constituent spatial, semantic, and metadata components.

The fundamental purpose of the Open Geodata Model is to enable interoperability among multiple data collections in satisfaction of a user's information requirements. As a model of geographic information, OGM provides a consistent, unambiguous, and comprehensive set of geodata behaviors/methods for geographic entities (features) and phenomena (coverages), along with dictionaries or thesauri for metadata, spatial reference, and naming systems. The focus is to model the earth, not to model maps and charts, which has been a limiting factor in some traditional GIS implementations. At the same time, OGM must accommodate the wide range of existing stores of digital geographic information, many of which have been implemented as maps in an automated form.

Open GIS services are used by system developers to build interoperable geospatial applications, based on standard Open GIS interfaces. In general, they provide a mechanism for collecting OGM datatypes to form complex representations of spatial phenomena, query individual elements or collections, and document the contents of geographic information repositories. The Open GIS Services Model provides the functions by which individual objects and their associated interfaces may be assembled into complex queries, transformations, analytical functions, and presentation directives. Critically, it enables the construction and promulgation of catalogs that enable users to identify, evaluate, and interpret complex geographic information dispersed throughout the net.

2.2.2. Interfaces and Well Known Structures

It is important to understand that the Open GIS Specification is an operational model, not a data standard. As such, the geodata model elements or features are defined in terms of interfaces, not structures. This is consistent with object modeling and encapsulation of data and methods generally. From the standpoint of geographic interoperability, it means that individual implementations can be developed based on profoundly different fundamental data organizations but those differences are transparent to the user. For example, a client application may have use for the identity of defined subareas within a specified region; this would use intersect and select operations against the abstraction of OGM, rather than directly invoke a vector overlay and point-in-polygon procedure on a known data structure.

At the same time, analytical functions on an ad hoc set of information obtained from multiple data repositories may in fact require conversion of a selected set of information for purposes of data integration. Thus each of the data types in OGM has an interface to return an Open GIS Well Known Structure [OGC 1996b], which is a rigorously defined datatype or structure such as a sequence of *x,y* coordinates for a line string or a two-dimensional array for a grid.

3. Open EIS

For online environmental information systems to satisfy the data requirements of their users, technologies for spatial data exchange and functional interaction must inform the deployment of the next generation of digital libraries. Although conceived of primarily as mechanisms for interoperability among legacy information systems and conventional software environments, the models for Open GIS can be brought directly to bear on issues of environmental library design and implementation.

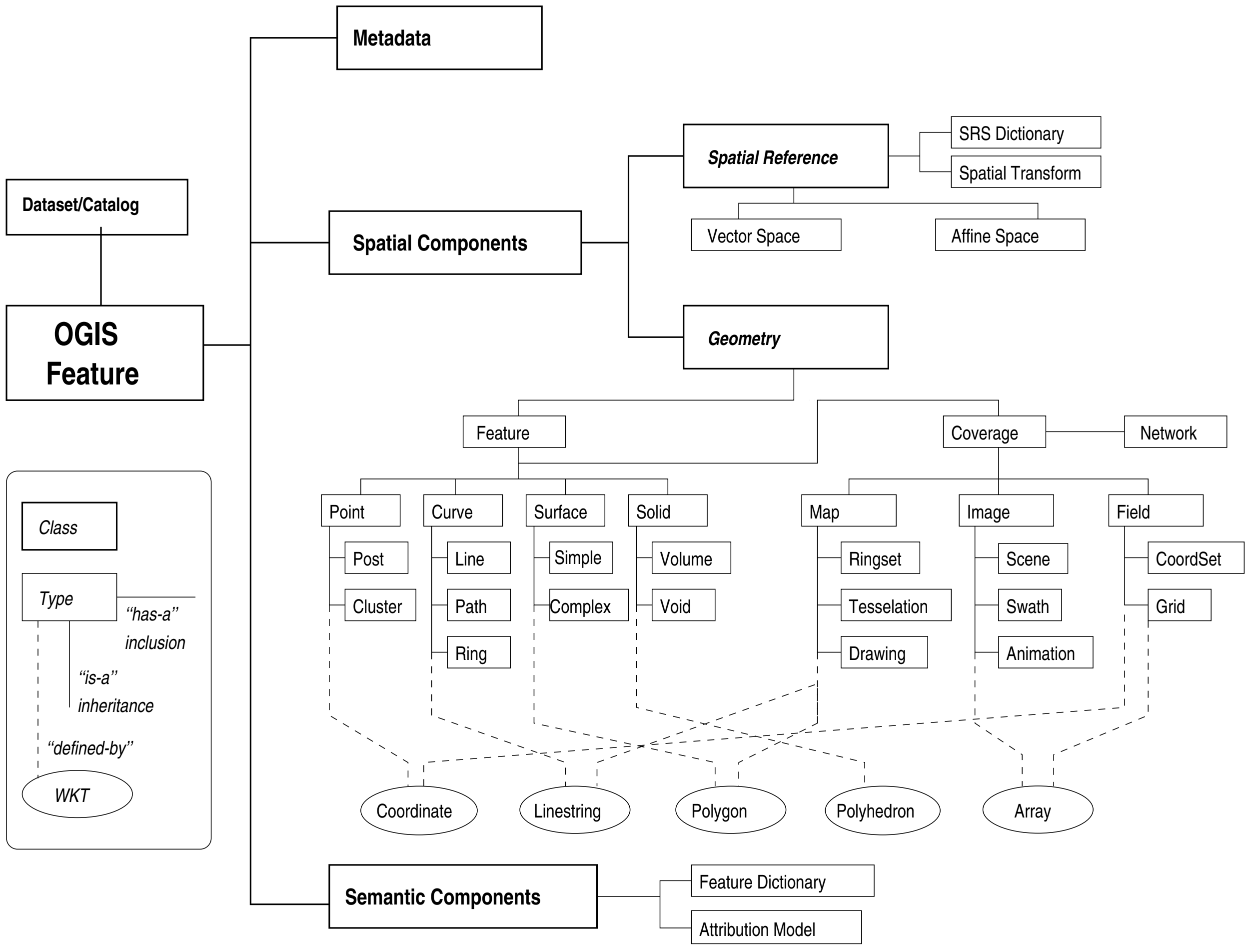

Figure 2. Open GIS Geodata Model class hierarchy.

3.1. System Integration

In this context, system integration is founded on concepts of autonomous subsystems, interacting individually according to prescribed methods. Although a simple metaphor is a layered architecture of separate components, connected possibly via a network "bus," a more accurate, if somewhat less informative, picture is the Web itself, populated with ad hoc links and special purpose tools for various uses.

The flow of information in an online environmental information system is not the same as in a GIS repository. Data elements are provided in terms of specific queries, not arbitrary datasets. Since there is not a single kernel or engine managing all the data in the collection, issues of conformity of specific data structures are largely irrelevant to the end user. Geographic concepts are expressed in terms of the modeling language, not the terms of the GIS software environment. In this regard, the feature interface orientation of the Open GIS geodata model is a close match to the data access requirements of EIS.

3.1.1. Process flow

Similarly, the software environment of an EIS is not a monolithic geoprocessing system, but rather specific tools, operations, functions or "applets" addressing a particular data management need. Through standardized interfaces and backbone distributed database and distributed object technologies, components can interact in a dynamic actor/agent setting in which each may be a client or a server in any individual transaction. New Web-based software capabilities, such as Java applets, allow a more fine-grained approach to crafting customized tools for specialized geodata retrieval, analysis, and interactive display requirements.

3.2. Components

Research is underway at several locations at this writing into the best mix of server and client functions for online access to geospatial data. While some work appears oriented primarily to geospatial browser widgets (for example, BADGER [BASIC 1996], shpclient (http://www.gis.umn.edu/fornet/java/shpclient/), MapQuest (http://www.mapquest.com), the various digital library projects[1] are investigating issues surrounding geodata modeling, information selection and transformation, geoprocessing, and user interface. In particular, the digital environmental library project at UC Berkeley is developing information access methods that integrate text searching and parsing, image recognition, geodata modeling and analysis, and other multimedia tools relevant to components of environmental documents [UCB 1996]. The Alexandria project at UC Santa Barbara is focussed on cataloguing, searching, and visualization functions for extensive map and imagery collections [UCSB 1996]. Common to virtually all of the projects is an abstracted view of a heterogeneous data collection, accessed by special purpose operators, and utilized by generic but context-sensitive viewers, all of which is facilitated by cataloging systems and indices. Interoperability per se is seen as a function of converters lying beneath the data access level, coupled with multi-catalog searching.

However, infusion of Open GIS constructs, interfaces, and services into this model creates an online environment that is more robust, extensible, and comprehensive than current approaches. Individual system components can work with specific data and process elements while still interacting on behalf of an end-user. The relationships among these components is shown in Figure 3.

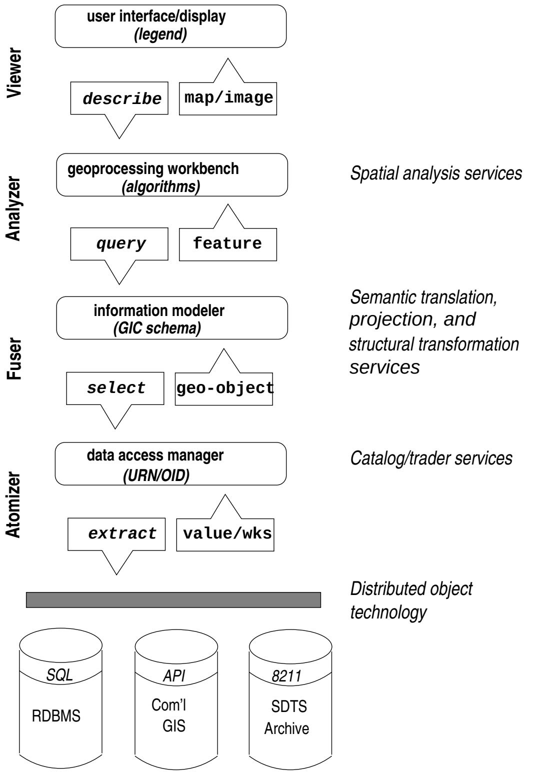

Figure 3. Components of the open environmental information system.

3.2.1. Atomizer

Most geographic information is stored in hierarchically organized datasets comprised of themes, tiles or mapsheets, map or image series, and georelated database tables. Modern GIS assembles datasets from individual elements such as polygons or grids, though such elements may be intertwined in a complex data structure. Too often, though, when such information is introduced into a repository, the most atomic level of the information is beyond reach; applications are forced to retrieve entire datasets and then infer meaning about dataset components absent their initial definition. In contrast, users must be empowered to selectively extract or query the data of importance. Multi-tier interfaces as defined in Open GIS perform this function in the logical sense. In the case of large, deep repositories, an information server should perform the selection operation at least to a first approximation; small datasets may be more effectively queried by interaction with a locally-cached copy. A critical factor in performance is intelligent anticipation of subsequent queries, so that requests for higher resolution data or spatially adjoining data do not require a complete re-initialization of the selection process.

[1] This term is applied here generally to the Digital Library Initiative projects sponsored, at six locations, by the National Science Foundation with support from DARPA and NASA, and to the various activities in digital libraries sponsored by NASA under the Cooperative Agreement Notice.

3.2.2. Fuser

Except in the case of simple *identify* queries against a single datafile, applications require that multiple data sources be integrated into a common framework. This may require conversion of one data structure or model into another; transformation from one spatial reference system to an alternative projection, coordinate system, or datum; translation of the semantic content associated with a theme into the nomenclature and definitions of another; and other direct or indirect mappings from source to target. These services may be supplied by the data server, and intermediary processor, or the client application. In any event, the required transformations may be invoked using regular interfaces; the additional complication is that many of them, especially semantic translations, may require additional thesaurus databases to accurately instantiate the mapping.

3.2.3. Analyzer

Analytical operations on geodata, generally described as geoprocessing, are the single- or multi-source higher order functions to interpret and understand geographic information. These can be generally identified as map algebra — arithmetic operations, spatial searches, boolean combinations, geometric transformations, buffering, statistical summarizations, and so on. The nature of geoprocessing requires that, for the most part, all of the source data and the algorithms themselves operate in the same memory space, implying that data are shipped to the function, which is probably resident at the application client. However, in comprehensive online libraries, all of the required data for an analysis may reside within a single (virtual) repository, and so may be subject to pre-defined analyses on the server. Although it is clear to the Open GIS development community that geoprocessing services must be subjected to the same rigor applied to basic geodata access, less progress has been made in the area of defining the actual analytical interfaces.

3.2.4. Viewer

The user's viewport into all of the data and process described above necessarily determines the extent of his or her ability to interact with the data. Intelligent viewers must expose all of the functionality and geodata resources potentially available throughout a network of environmental information systems. To do this, a viewer must be designed to translate the user's directives into the structured queries or interfaces understandable by geographic servers, and correctly manage the returned values and data structures. From a utility point of view, the viewer must include such functions as render, pan, zoom, point/click (query), select by value, and legend display. It must also provide mechanisms for constructing, modifying, and reusing procedures or recipes for information requirements.

3.3. Conclusions

Properly designed, geodata access and analysis tools, combined with open environmental information systems, can provide sophisticated decision support to the users of geographic information. Traditional centralized GIS approaches are evolving toward geodata mining in a web of distributed, heterogeneous information, pulled by the need for better, more timely environmental information and pushed by the emergence of new technologies for networked systems. Be they professional planners, policy makers, or the general citizens, open technology can reveal environmental information in hitherto unknown ways.

References

Bay Area Shared Information Consortium (BASIC) (1996), *Bay Area Digital GeoResource (BADGER)*, **http://www.basic.org**.

California Resources Agency (CRA) (1996), *California Environmental Resources Evaluation System (CERES)*, **http://ceres.ca.gov**.

Digital Geographic Information Working Group (DGIWG) (1995a), *Digital Geographic Information Exchange Standard, Version 1.2a*, **http://132.156.33.161/Engineer/DIGEST_1.2a/cover.htm**.

Digital Geographic Information Working Group (DGIWG) (1995b), *Feature Attribute and Coding Catalogue, Version 1.2a*, **http://132.156.33.161/Engineer/DIGEST_1.2a/covers/part4.htm**

Federal Geographic Data Committee (FGDC) (1994), *Content Standards for Digital Geospatial Metadata*, **http://www.fgdc.gov/metadata**.

Federal Geographic Data Committee (FGDC) (1996a), *FGDC Activity Areas - Frameworks*, **http://www.fgdc.gov/Fram**.

Federal Geographic Data Committee (FGDC) (1996b), *National Spatial Data Infrastructure*, **http://www.fgdc.gov/nsdi2.html**.

ISO/IEC JTC 1/SC 21 (1996), (ISO Committee Draft) *SQL Multimedia and Application Packages (SQL/MM) - Part 3: Spatial*, ISO/IEC JTC 1/SC 21 N 10441, ISO/IEC CD 13249-3:199x (E) SQL/MM:MAD005. **ftp://speckle.ncsl.nist.gov/isowg3/sqlmm/MADdocs/**.

ISO/TC 211 Secretariat (1996), *ISO/TC 211 Geographic information/Geomatic*, **http://www.statkart.no/isotc211/**.

Kottman, Clifford (1996), *Geographic information communities, an object-oriented approach using open GIS*, **http://ogis.org/members/project/96-007.html**

Mackay, D. S., and K. Gardels. *Interoperability of Geographic Information*, Geographic Information Interoperability Working Group, **University Consortium on Geographic Information Science**. In press.

Open GIS Consortium (OGC) (1996a), *Vision and Mission Statements*, **http://www.opengis.org/vision.html**.

Open GIS Consortium (OGC) (1996b), *The OpenGIS Guide - A Guide to Interoperable Geoprocessing*, **http://ogis.org/guide/guide1.html.**

Open GIS Consortium (OGC) (1996c), *The OpenGIS(tm) Abstract Specification*, **http://www.opengis.org/public/abstract.html**.

Spatial Archive and Interchange Format (SAIF): Formal Definition (rel 3.1, 1994), Survey and Resource Mapping Branch, Ministry of Environment, Lands and Parks, Province of British Columbia, Canada.

University of California Santa Barbara (UCSB) (1996), *The Alexandria Project*, **http://alexandria.sdc.ucsb.edu**.

University of California Berkeley (UCB) (1996), *UC Berkeley Digital Library Project*, **http://elib.cs.berkeley.edu**.

U.S. Geological Survey (USGS) (1993), *Spatial Data Transfer Standard*, **http://mcmcweb.er.usgs.gov/sdts/standard.html**.