ISSUES IN THE DESIGN OF GIS DATABASES FOR BIODIVERSITY ASSESSMENTS

by

PRISCILLA GAIL MINOTTI

A RESEARCH PAPER

submitted to

THE GEOSCIENCES DEPARTMENT

in partial fulfillment of the

requirements of the degree of

MASTER OF SCIENCE

GEOGRAPHY PROGRAM

May 1995

Directed by

Dr. A. JON KIMERLING

ABSTRACT

National and international organizations are developing spatial databases to assess the distribution of Biodiversity using geographic information systems (GIS). Currently, there are no guidelines or standards for building GIS databases for biodiversity assessment because database design is discipline-dependent and most biodiversity disciplines lack spatial traditions. The present research analyzes the conceptualization of biodiversity information at the spatial level and discusses its relationship to the modeling and structuring of data in GIS databases. The relevant findings of this research are: 1) the degree of data processing in the original sources has to be acknowledged in the design of the spatial model and data structure of the biodiversity database theme, and 2) the spatial uncertainty of biodiversity data is related to the georeferencing method, the classification label (taxonomic, physiognomic,etc) and the ecological context .

ACKNOWLEDGMENTS

This research is dedicated to my beloved daughters Lucy and Ale, that gave away some of their best childhood moments; to my husband Claudio (that was responsible for my going back to school), his patience, love and encouragement; to my dog Toto that traded long walks for window watching, and to my family and friends in Argentina for their encouragement.

I would like to thank Dr. A. Jon Kimerling, my major professor, and Dr. Gordon Matzke for assuming the risks of accepting me (and my problems) to the Geography Master of Science Program. Being a teaching assistant to Mary Lee and Sidney Nolan was a funny and rewarding stage in the beginnings of my program. I experienced Oregon's geo and biodiversity thanks to Charles Rosenfeld and Mary Santelmann. I thrived through graduate courses and T.A.'s evaluation thanks to Jo Anne's and Therese's smiles and friendly words. Karen Anderson, Aileen Buckey and Juan Carlos Torres provided a quality of friendship that allowed me to mix technical discussions with shoulder sobbing. Jeannie Sifneos, Kathy Freemark, Scott Urquhart, Ross Kiester, Blair Csuti, Mary Barzack and Mary Cablk have been great sources of themes and ideas in the workplace.

Very special thanks go to Denis White, who provided a long term friendship that taught me how to think in spatial terms, write GIS programs and encouraged me to use my ecological background as the source of common sense for my project.

TABLE OF CONTENTS

LIST OF FIGURES

ISSUES IN THE DESIGN OF GIS DATABASES FOR BIODIVERSITY ASSESSMENT

I. INTRODUCTION

1. Statement of the Problem

Biodiversity -or biological diversity- is a complex idea that involves the variation and variability among living organisms and the ecological systems in which they occur (US. Congress Office of Technology Assessment 1987). Our survival as a species is closely linked to the maintenance of biodiversity as a source of quality food, water, air, cover, and medicines, as well as aesthetically and ethical pleasure. The increasing awareness of irreducible losses of biodiversity has spurred conservation measurements using all available information to assess the geographical distribution of biodiversity (Groombridge 1992). This assessment is critical to understanding the current status of biodiversity, to locating areas needing priority protection and evaluating management and development alternatives that will maximize biodiversity conservation.

National and international organizations are rapidly developing computerized biodiversity databases. Computerized databases are digital collections of information. As more databases are built and become available by electronic means, data produced by independent organizations and researchers can be combined to assess the distribution of biodiversity on a variety of scales.

To eliminate overlapping efforts and to use all available information efficiently, the trend is to assess the geographic distribution of biodiversity by combining information from databases developed by individual researchers, public agencies and non-governmental organizations. The integration of biodiversity information based on geographic occurrence can be done by means of Geographic Information Systems. Geographic Information Systems (GIS) are computer-based tools for working with geographically referenced data by storing the data in spatially meaningful and logical formats (Thompson and Laurini 1992). GIS databases are structured as spatially indexed layers of information, so information can be overlaid in geographical space for analysis. The implementation of GIS databases involves decisions on how to model, capture, structure and store data that faithfully represent the given geographic phenomena. This decision process is termed GIS database design (Chambers 1989).

The first uses of GIS for biodiversity assessment consisted of translation of already mapped biodiversity information into digital coverages. Biogeographical themes such as species densities by country, or number of extinct species by region, were soon available in digital format. The spatial databases -- called digital maps or coverages- were designed following compilation procedures similar to paper maps (Stoms et al. 1990), even producing drafts manually before inputting the information into the GIS.

Davis et al. (1990) and Stoms et al. (1990) were among the first to address the problem of availability of biodiversity data in spatial format for biodiversity assessment at the state level. They pointed out that spatial information is usually incomplete because spatial databases are produced from sources at different map scales, they are not available in digital format for particular themes, or there is no information at all on some variables. Their work also highlighted the fact that digital maps produced using traditional compilation methods do not allow for changes in the aggregation, classification or inclusion of variables. Standard compilation also fixed the spatial scale of the analysis, posing a different problem to biodiversity researchers, who are concerned with the maintenance of biodiversity across scales.

This short account reflects three important current shortcomings. The first is that GIS is going to be more useful for biodiversity assessment when more data are available as spatial databases. The second is that standard methods or databasing guidelines are needed that would allow researchers to choose a spatially meaningful and logical format to organize their biodiversity information while facilitating the future integration with databases produced by others. The third and most important is that the design of GIS databases is discipline-dependent. The impacts of errors incurred from using unsuitable datasets from badly designed databases can have a profound effect in biodiversity conservation and management, affecting our own survival. Approaches for properly designing biodiversity GIS databases are urgently needed.

2. Objectives

The objective of this research paper is to present some issues in the design of GIS databases for biodiversity assessments. The paper has three sections:

> **Background**, providing an overview on database terminology and GIS database design for geographical databases.
>
> **Spatial issues in the design of biodiversity databases**, analyzing biodiversity concepts and information and their conceptualization at the spatial level for modelling into data structures.
>
> **Discussion**, organizing the previous information into a series of design considerations that should be addressed before implementing or using any biodiversity database.

II. BACKGROUND.

1. Database design terminology.

When we capture information from the real world , we obtain **data** that model a particular aspect of reality. **Databases** are sets of data in digital form representing information about the observable world. The database is a digital description that has to be complete enough to assure that the data adequately reflect the world that is modeled. This representational relationship between data and reality is known as **data integrity**. Databases may hold data with different **degrees of refinement** or **processing**. Different methods and equipment are used to **capture** data, and the term **raw data** is used to describe primary measurements of the real world. Raw data are usually **corrected, calibrated** or **transformed** in some way to assure lack of bias, potential for spatial integration and ease of interpretation. When data are **analyzed** and **interpreted** in the context of a theoretical or **reference frame**, they result in new **information** that represents a **knowledge view** of the real world.

The graphic portrayal of how information is captured from the real world, to its use as knowledge in a decision making process, is known as **the information flow**; it highlights the compartments where information changes format and the flow between compartments (Figure 1).

Database design is the series of steps or procedures that allows one to transform the information of the observable world into a digital description that can be queried or analyzed. The steps in database design are usually presented in a logical order (Figure 2) but this does not represent the actual temporal sequence, as database building loops many times with different refinements at each iteration.

The first step starts with the definition of the reference or theoretical frame, the **scientific or phenomenological model** that sets up the objectives of the databases and analyzes the existing knowledge basis. This phase describes the database **requirements** in terms of **information content** and **intended uses**. It is considered the most critical step in the design, as it defines the frame for assessing errors, integrating disparate data formats and linking different databases (Burrough 1992).

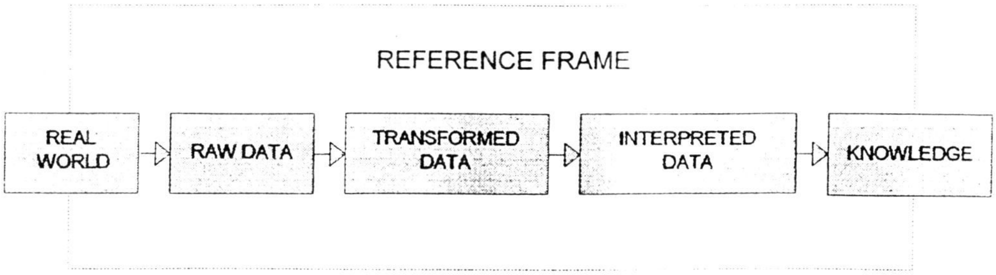


Figure 1. Information flow in the adquisition of knowledge.

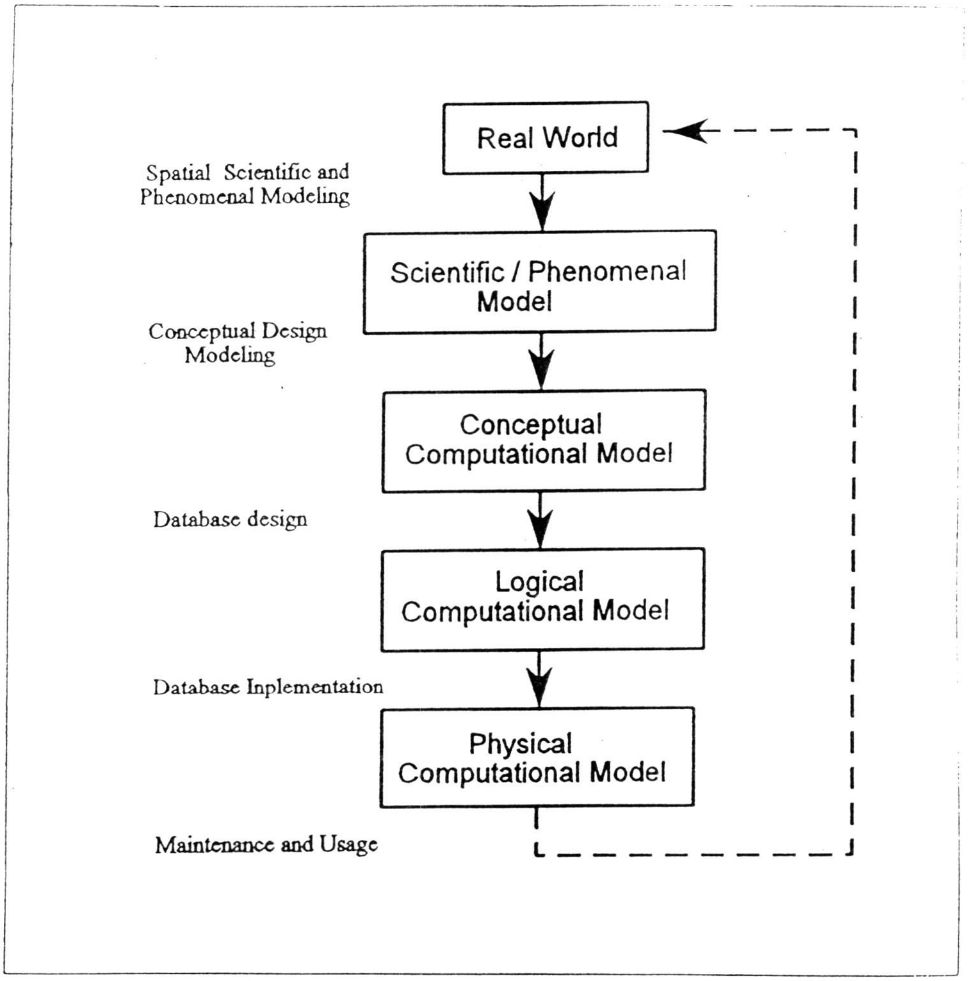


Figure 2. Stages in the design of digital databases.

The next stage is the **conceptual computational model**, which represents the elements or objects of reality and their processes, in forms that are computationally tractable but independent of any specific computer science paradigm. The key component of this step is the definition of what are the **objects** to represent and their **relationship** to the data describing their **attributes**. The conceptual design is a description of data in the database will carry the real world information.

The next design phase builds the **logical computational model**, referred to as the **database structure,** because it uses some of the available **database paradigms** (object-oriented, relational, network, hierarchical) to structure the data according to the conceptual model design.

The last step is the **physical model** that corresponds to the actual implementation of the database within the database management system. The implementation involves the capturing, storing and preprocessing of data and continues with the actual use and maintenance of the database in a database management system. The implementation is based on the documented conceptual and logical designs.

The **database management system** is the computer based record-keeping frame for the integrated and shared repository of databases (Thompson and Laurini 1993). The database system has a declarative component for *describing* the real world, and a operational or functional component to *manipulate* this description. Typical manipulations are 1) **data capture** or **automation** -the entering of data according to a define structure, 2) **modifications** or **updates**, either to refine the model or to track changes that may have occurred in the real world, and 3) **transformations** to derive new descriptions (databases) more suited to a particular application.

2. Database design in GIS.

Geographic information systems differ from other database management systems in that they have special methods and structures to input, handle, analyze and display geographical information. GIS analysts have adapted traditional database design procedures to schemes of their own for designing geographical databases (Figure 3). These adaptations are responses to the spatial nature of geographical data and cartographic traditions for modeling geographic information.

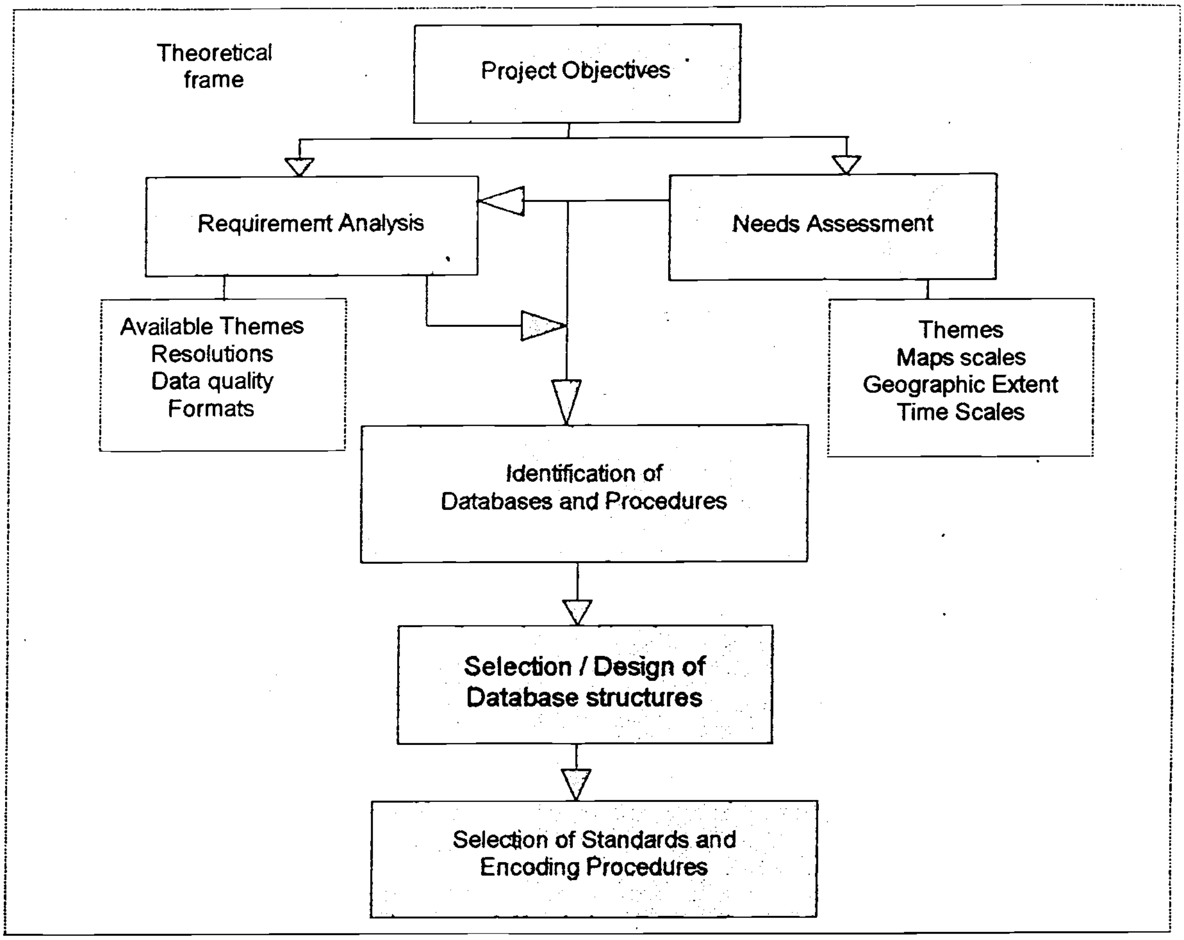


Figure 3. Database design stages in GIS.

Theoretical frame

Geographical data are measured along several dimensions: spatial, temporal, graphical and textual-numerical. Data that present spatial and graphical properties are termed **spatial data**. Spatial properties are location, area, distance, adjacency, connectivity, spatial autocovariation and autocorrelation, volume, containment, incidence and [fractal] extension (Burrough 1986, Burrough 1992). Graphical properties are decomposition into spatial primitives (e.g. point, lines), size dimensions or distances expressed using a metric norm, and visual characteristics such as color, tone, value texture, pattern contrast, and standard symbology representation. Geographical data are spatial data that can identify objects of the real world based on their location. Cartography and spatial theory provide the basis for modeling spatial data from geographical objects into maps.

Spatial data be referenced to a location in different ways (Figure 4). Spatial data may have an **explicit location** or **geocode** that refers to a geographic localization such as the collecting place of a herptile specimen given by latitude and longitude, or the name of a county. Alternatively, spatial data may have **implicit location**, when their positions are given in relative distances to certain referenced points (Peuquet 1988). The location of a pixel in a satellite image is given by row and column, while the location of the whole image is given by the geographic coordinates of its four corners. Spatial data with explicit or implicit location have graphical dimensions, and they can be represented as maps or images using geometric models and graphical data structures such as rasters and vectors. Spatial data can also be georeferenced by the name of a locality, whose coordinates can be obtained from a gazetteer. This type of spatial data has **inferred location** because their spatial distribution is derived from, or attached to other information (e.g. the distribution of a species from the distribution its environmental requirements). Data without graphical properties are usually named **non-spatial or attribute data**. Non-spatial data may have spatial information, but this information is not derived directly from its spatial extension.

Conceptual model

Most GIS database protocols have been developed to translate map and remote sensing imagery information into georeferenced data sets. The theoretical frame is given by cartography. Paper maps and imagery represent models of the earth surface that are easy to translate into computer terms because of their graphical and positional properties.

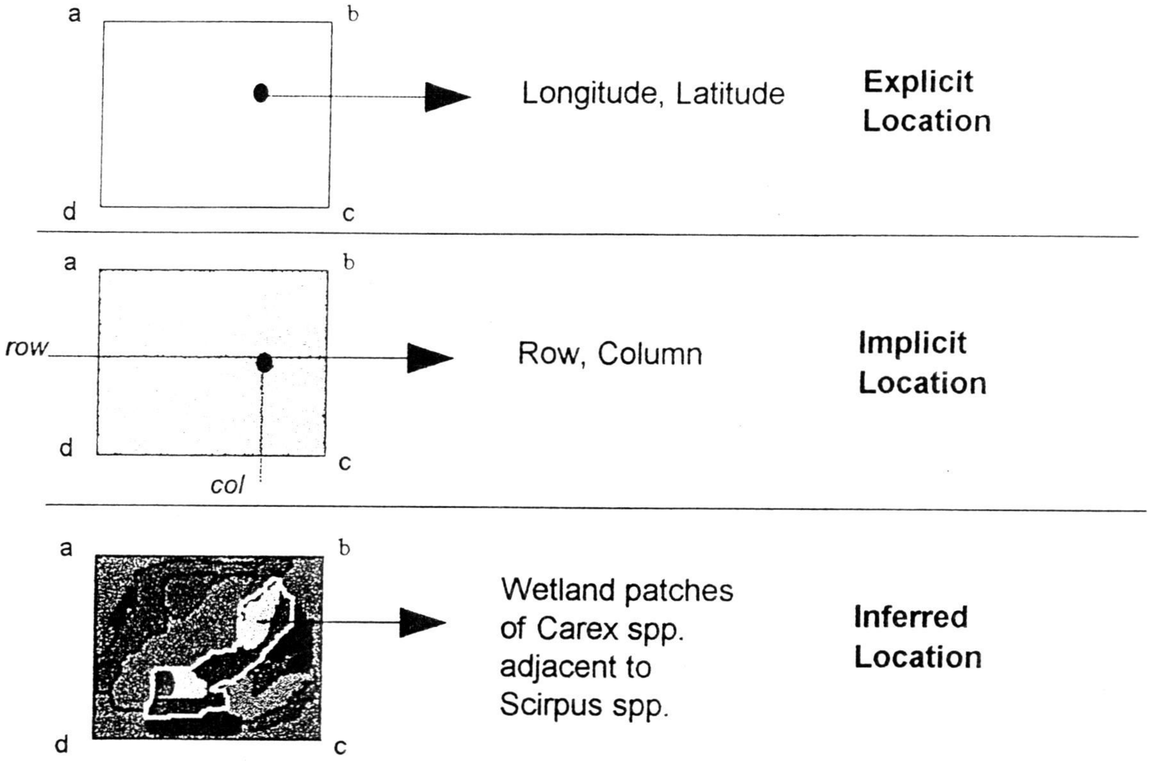


Figure 4. Geo-referencing the location of geographical data.

This stage involves defining:

1) the geographical objects **to be represented**,
2) the **spatial nature of their boundaries** and
3) the **attribute information** needed to **represent** the **behavior** of the geographical phenomena under study.

The list of answers constitutes the **information content** of the database.

The GIS database analyst defines the objects to represent in the database based on the information needs assessment, and/or the statement of requirements. The needs assessment is usually conducted by interviewing or observing the future database users, addressing the information content of the databases to build. It includes what types and themes of maps and images will be required, at what scales, with what accuracy and precision, and for what geographic area and time frame. The requirement analysis focuses on the functionality of the GIS. The requirement analysis evaluates what are the objectives of the GIS database, what is available in terms of resources (human, budget, software and hardware) and data, and how source data can be structured and manipulated in order to meet the objectives.

System analysts use the information content to sketch the relationships between the different objects (entities or elements), to establish how the data would be split or merged into databases, what attributes will each database have, and what attributes will serve as index or pointers to relate the databases. This is called the **entity-relation** approach to database design.

Not always is it easy to define the objects to represent in the database. To represent the hydrology of an area, we can digitize as continuous lines the whole network of a river drainage as sub-basins or we can digitize stream segments as separate entities that can be queried independently. In terms of map output both graphics will look identical; in terms of spatial analysis they behave differently. Some uncertainties can be reduced by building an information model to depict what pre-processing, analytical functions and/or output products are needed to perform a spatial analysis. This is also known as the **business model** or approach that economists use to design their information systems.

Direct **geometric model matching** is another way of conceptualizing geographical objects. A geometric data model is chosen that matches the objects' most relevant spatial

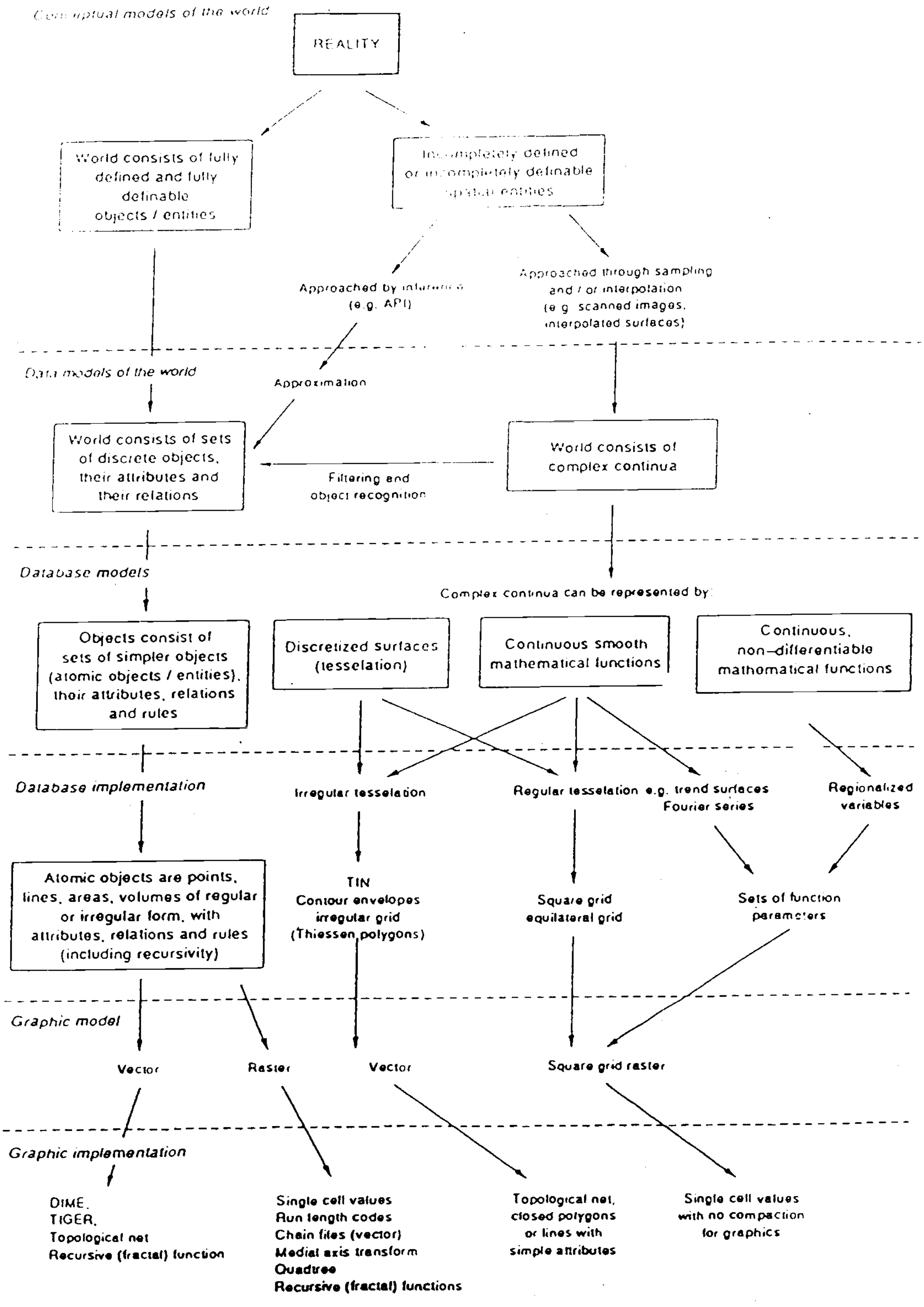


Figure 5. Relation between spatial conceptualization, data model, data structure and database implementation (from Burrough 1992).

aspects. The geometrical representation captures the spatial relations between objects, is easy to implement in the computer, and directly conforms a logical model for structuring the graphic components of the spatial data. The conceptualization of geographical objects presented by Burrough in 1992 (Figure 5) offers a good key for evaluating this matching. The chosen geometric model has to balance the spatial properties and behavior of the object, the ease of retrieval and manipulation with the quality of graphical representation.

Logical Model or Data Structure

The logical model shows how geographical objects with spatial data will be structured into databases. Spatial objects can range from completely discrete and sparse, to continuously covering the whole surface or volume of the Earth . **Discrete objects** have defined spatial bounds; the real world can be seen as made of fully defined or definable entities (Goodchild 1992). These entities can be described digitally by: a) geometric objects that define the spatial boundaries of the entities as lists of coordinate pairs and/or b) by tables with attributes describing explicitly their spatial properties (topological relations, area, perimeter) and other variables indicating internal composition . This approach is also termed **exact** because it has traditionally treated boundaries as lines exactly located on the earth. The graphical description of these bounded objects is usually structured using **vectors** of coordinate pairs representing geometric primitives such as **points, lines, areas** and **volumes**. The non-graphical description can range from a simple table with the identifier for each object and its calculated spatial dimensions (length, size), to complex sets of related tables modeling the topological relations and degrees of spatial interaction, such as those present in network data models.

At the other extreme, geographical objects can be described by some sort of statistical or mathematical function in a complex spatial continuum (Burrough 1992). This approach is termed the **field-based** paradigm, where geographical objects by their nature are incomplete, inexact or incompletely definable. Because they cannot be seen in their entirety, they can only be described by sampling, and they have to be reconstructed by interpolation or by feature recognition methods (Burrough 1992). This sampling may be systematic by completely discretizing or quantizing the field surface or volume using point lattices and areal or volume tessellations or by randomly locating sampling points, lines, areas or volumes, with subsequent interpolation between values. The extraction of the object is based on the pattern analysis of spatial and non-spatial properties. Their graphical representation is usually associated with raster graphical structures such as satellite images, but vector and raster representations may also be used . Attribute information most of the time is stored as stacks of images, each holding a different attribute . At times it is structured as tables or arrays related by cell numerical values, label identification, or cell location.

Physical design

The physical model provides the guidelines to capture data from the source documents and to structure the graphical and non-graphical data, so that it follows the conceptual model. It has to address the contents and quality of the data sources. This step also defines how data will be georeferenced and if databases will be built for the total geographical extent of the study or if criteria for tiling will be used. It also involves data verification. Verification checks should be performed during and after the data have been captured and structured.

The physical design also involves the documentation of how the data are captured and structured. The documentation of the graphical components of the database is a description of the contents, collection and processing methods and quality of the data, that can be attached as a header or as a searchable file. This information is termed metadata or codata (Thompson and Laurini 1992) and many organizations are currently discussing how to provide this description in a standard format. As GIS databases are considered products, they require an assessment of their quality. The physical design should state the quality control procedures to be used for the verification and documentation of the GIS database.

III. SPATIAL ISSUES IN THE CONCEPTUAL DESIGN OF BIODIVERSITY DATABASES.

The conceptual design stage is considered the critical step in the design of a spatial database because it defines the relationship between the identity of the objects we want to model in the database and how we will retrieve their information content based on their spatial representation. In the case of biodiversity assessment, spatial concepts of biodiversity information have never been formalized, so we lack references to assess how good this relationship can be. The following section presents what are considered key areas in the spatial treatment of biodiversity information that can affect the design and quality of GIS databases for biodiversity assessment.

1. Objectives of Biodiversity Assessments.

Biodiversity assessment is the analysis of the geographical distribution of biodiversity at several scales, searching to explain the factors that govern this distribution. This knowledge can be used at local to international levels, to reverse the trend in biodiversity losses. Biodiversity assessment covers three main analytical areas: 1) Biodiversity inventory, 2) Biodiversity factors and 3) Biodiversity conservation.

Biodiversity inventory

Biodiversity inventory is dedicated to the inventory and mapping of biodiversity to answer what is where It involves collating information for different biodiversity data independently. Data on specimen collections and ecological observations are compiled to map species distributions. Ecosystem types and their spatial/temporal distribution are determined from remote sensing documents and field data. This involves characterizing biodiversity by assessing the extent of genetic diversity between and within populations, and determining the dissimilarity in species composition between communities, landscapes or biogeographical regions. It also implies collating information about environmental variation that is related to biodiversity.

Biodiversity factors

The second area analyzes the factors explaining biodiversity distribution. It deals with the generation and testing of hypothesis that explain, at least in statistical terms, what influences the generation, maintenance or loss of biodiversity, and by which mechanisms. Several empirical models relate species richness to latitude, incident energy, geological and climatic history, habitat and environmental diversity. These relationships have been assessed for some taxonomic groups and areas, but more biodiversity indicators should be analyzed , trying to reach global coverage.

Biodiversity conservation

The third area uses all the above information for biodiversity protection and conservation, assessing ways or alternatives to prevent biodiversity losses. The **what if** type of spatial modeling is used to evaluate what will best protect biodiversity, or produce most losses to biodiversity (e.g., the effect of landscape configuration in biodiversity protection) based on current knowledge of biological and ecological systems. This modeling has direct application in the selection of areas to function as biodiversity reserves based on different sets of priorities (Kiester et al. 1993).

2. Biodiversity units.

The biodiversity concept is as complex and ambiguous as the variety of organisms. Its name derives from the union of 'biotic' and 'diversity' (OTA 1987) and became a buzzword before biological scientists started to argue about its validity. Magurran (1988) states that "[bio]Diversity is rather an optical illusion, the more it is looked at, the less clearly defined it appears to be, and viewing it from different angles can lead to different perceptions of what is involved". The root of this assertion lies in the variety of concepts used to represent biological variation. Scientists conceive of living matter as organized in systems that are structured hierarchically in space and time. Biological information can be described as **biological systems** and as **ecological systems**. Each hierarchical level constitutes a biodiversity unit.

Biological systems

Biological systems form a nested hierarchy based on the structure and interchange of genetic information and its products (Cousins 1991). The nested hierarchy implies that they follow principles of spatial containment. Any biological object or unit is an element of the next hierarchical level. A biological object at any given level presents *intrinsic properties* that are derived from its individual components, *emergent properties* pertaining to its condition as a whole, and *extrinsic properties* that are related to its being an element of the next hierarchy.

The smallest systems of interest are **genes** inside cells. Individual or multiple groups of cells form **organisms**. Groups of organisms related by reproduction form **demes**. Single or multiple demes that share the same place at the same time, with free reproductive interchange of organisms, form **populations**. Single or multiple populations of organisms of the same type, that do not present barriers to sharing genetic information by reproduction, constitute **biological species**.

Species can also be organized into taxonomical hierarchies or **taxons**, based on their evolutionary closeness. These relationships are determined by assessing the similarity in gene flow given by the gene structure *(biological species)*, morphological characteristics *(morphological species)* and/or ecological functionality (*ecological species*).

Ecological systems

Ecological systems are based on the structuring of energy flows and interchange of material resources. The hierarchical levels are defined by : 1) the degree of interaction in time and space between biological systems and their environment formed by physio-chemical systems, 2) their spatial extent and 3) the position (scale) of the observer describing the spatial extent, of the interaction (Rowe 1993).

Communities or species assemblages are detected as patterns of co-occurring organisms of different species, sharing space or resources from a physical support system called the **habitat**. Some species in the community may share an evolutionary ancestor while others represent non-related lineages (e.g., steelhead and coho salmon vs. steelhead and caddies flies). Communities may include whole populations of some organism while having only a few individuals of others.

The concept of **habitat** addresses the suite of variables and the range of values of the physical environment where organisms of given populations or communities are found. Distinct portions of the habitat used by individual organisms are called **microhabitats**. For some species, the habitat includes other organisms that form a physical support or host, such as forest for birds, or deer for parasitic nematodes.

Communities and habitats are interchanging matter, energy and information. This pattern of matter and energy exchange coupled in a geographical and biological setting is named **ecosystem**. From a purely biological perspective, the boundaries of an ecosystem are intrinsically fuzzy or uncertain as they are determined by a pattern of energy interchange that is not always possible to detect. From a physical view, ecosystems are bounded by the spatial expression of environmental factors such as geology, climate, relief, air or water movements, and their sizes are a reflection of the combination of processes.

Individual landforms with their communities are **landscape elements.** Repeating combinations of landscape elements form **landscapes**, that reflect a common pattern of geomorphic processes of natural or anthropogenic origin. Patterns of landscapes repeating spatially , with the same climate form **landscape systems**, also called **ecological regions**. When the community of landscape systems presents a dominant physiognomy of vegetation throughout its extent, it is called **biome.** If the landscape system presents the same geologic setting, the term **physiographic region** or province is used. Landscape systems can be single landscapes or can include similar landscape patterns distributed over whole continents, or even have global extent.

3. Information contents for biodiversity assessment

Assuming that all types of biodiversity assessment will be done with a given GIS, is not easy to decide which objects should be represented without a thorough analysis of the information needs of a quite heterogeneous set of scientists.

Molecular biologists measure biodiversity by the **genetic differences between organisms representing populations or species**. The taxonomist or systematic biologist will consider the **number of species** in a given taxonomic group (e.g. family, order) and the **patterns of radiation** of sibling species as his measurement of biodiversity. The biogeographer will stress the **spatial distribution of species**, seeing biodiversity as a property or qualifier of a given area. Biogeographers estimate the **species list** of an area as the basis of their analysis. Areas with high biodiversity may have high counts of species (**species richness**) or high **endemism,** that is, with species restricted to that location.

Community and population ecologists will incorporate the abundance of each species or functional group of organisms at the local level and search for differences between communities. **Alpha diversity** is the species richness of a community at the sampling unit or landscape element level (the local scale). **Beta diversity** is the difference in species between communities. Landscape and regional ecologists include the variety of ecosystems and their environmental processes at different geographical scales. **Gamma diversity** describes the number of species in an ecological region, also known as the **species pool** whereas **delta diversity** describes the change of species between regions. They are also interested in the structural variation within landscapes, called **landscape heterogeneity** and the variety of agents responsible for the generation and maintenance of that variation, commonly referred to as **landscape dynamics**.

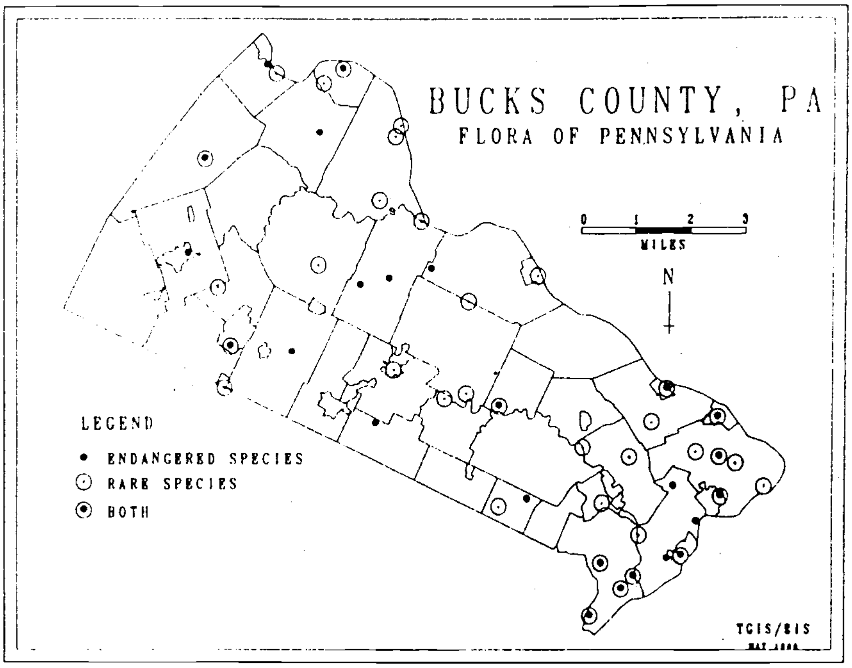
Conservation biologists will weight the species by their population and protection **status**, focusing more on those species that are naturally in low numbers or **rare**, **unprotected** by legislation, and/or those currently in very low numbers with high probabilities of extinction, the **threatened and vulnerable** species. Paleontologists and paleoecologists focus on the **richness of fossil species** and the dynamics of their **paleohabitats**. Human ecologists and anthropologists have their own set of indicators for assessing **cultural diversity,** as local peoples, their cultures and their indigenous knowledge account for the diversity of an area.

Ecosystems carry not only information on biological variation but also on a **variety of physical-chemical processes** that environmental and ecosystem ecologists track. Species in ecosystems represent both a collection of **individual variation accumulated over time** (evolution), as well as **patterns of energy utilization** from the suite of **ecological processes** faced in the past and today, measured by ecophysiologist.

4. Spatial assignement of biodiversity units.

The analysis of biodiversity using GIS requires georeferenced information organized using a data structure model. Burrough (1992) recommended that the choice of data structure models should be governed by an analysis of the spatial nature of the units or elements to be included in the GIS database. Conceptualizing geographical objects using geometrical models forces the selection of a data structure that does not reflect the spatial properties of the data. He stressed that we should concentrate on designing a spatial model (or conceptual design) that faithfully represents the spatial properties of our geographical objects of study. In this way the spatial region of inference for the data is known when the object is captured to the database.

Consider the following example. The spatial extent of a population of granivorous rodents is given by reference to their summer habitat, patches of annual grasses in deciduous forest stands. The spatial extent of the patches cannot be interpreted or extracted from remote sensing photos or images because the crown of the trees form a continuous canopy during warm weather, and during winter all is covered by snow. In this case, grass patches are geocoded by reference to deciduous forest patches, whose spatial extent can be acceptably extracted from the remote sensing sources. The rodent population will be spatially assigned to all deciduous forest stands in the area, whether they contain annual



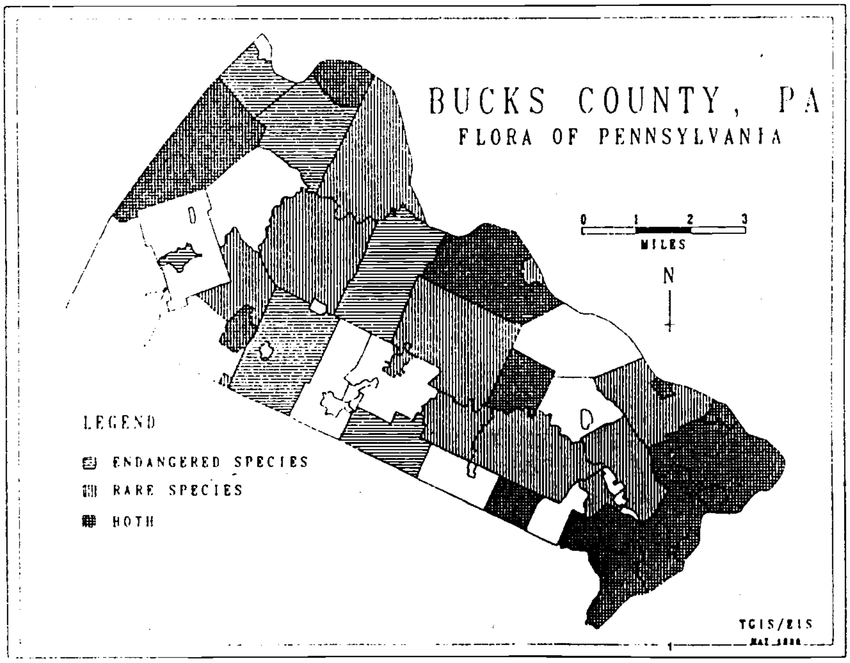


Figure 6. Endangered and rare plants displayed by collection site and township.

patches of grasses or not. The rodent and grass populations will have the same data structure as the forest. If we are assessing the impacts of logging on small mammal populations, our rodents would be at high risk of local extinction due to the disappearance of forest patches, although their numbers actually be increasing due to the expansion of their grass habitat. A similar kind of misinterpretation occurs when researchers generalize points features into discrete areas for map enhancement purposes, for the area of inference of a sampling design or a response model (Figure 6)

The spatial unit data model.

The spatial unit is the minimum unit of geographic space for sampling, collecting, collating or inferring information for a given biodiversity unit. Spatial units can be defined based on: a) their level of processing, and b) the geometric model that describes their extent and topological properties.

Spatial units can be defined based on the degree of numerical and spatial processing into:

1) *Input units* that respect the original spatial model of data collection in terms of geometry , extent, resolution, and density of data.

2) *Analysis units* that optimize the selection of data structures for preparing and processing data according to the spatial scale and resolution of different analyses.

3) *Output or visualization units* that structure the data for maximizing cartographic representation.

Spatial units can be defined based their spatial geometry model, taking into account their Euclidean dimensions, their shape and size, the way to represent their extent in digital format, or their spatial continuity. The following geometric models are used to describe spatially the distribution of biodiversity:

1) *Points* are used to represent the location of collection, observation instances or the geographic reference of sampling units. Points can reflect true point samples (Figure 7) or they can represent aggregated information in which the actual sampling effort may correspond to transect or areal sampling units . They may be area samples that are shown

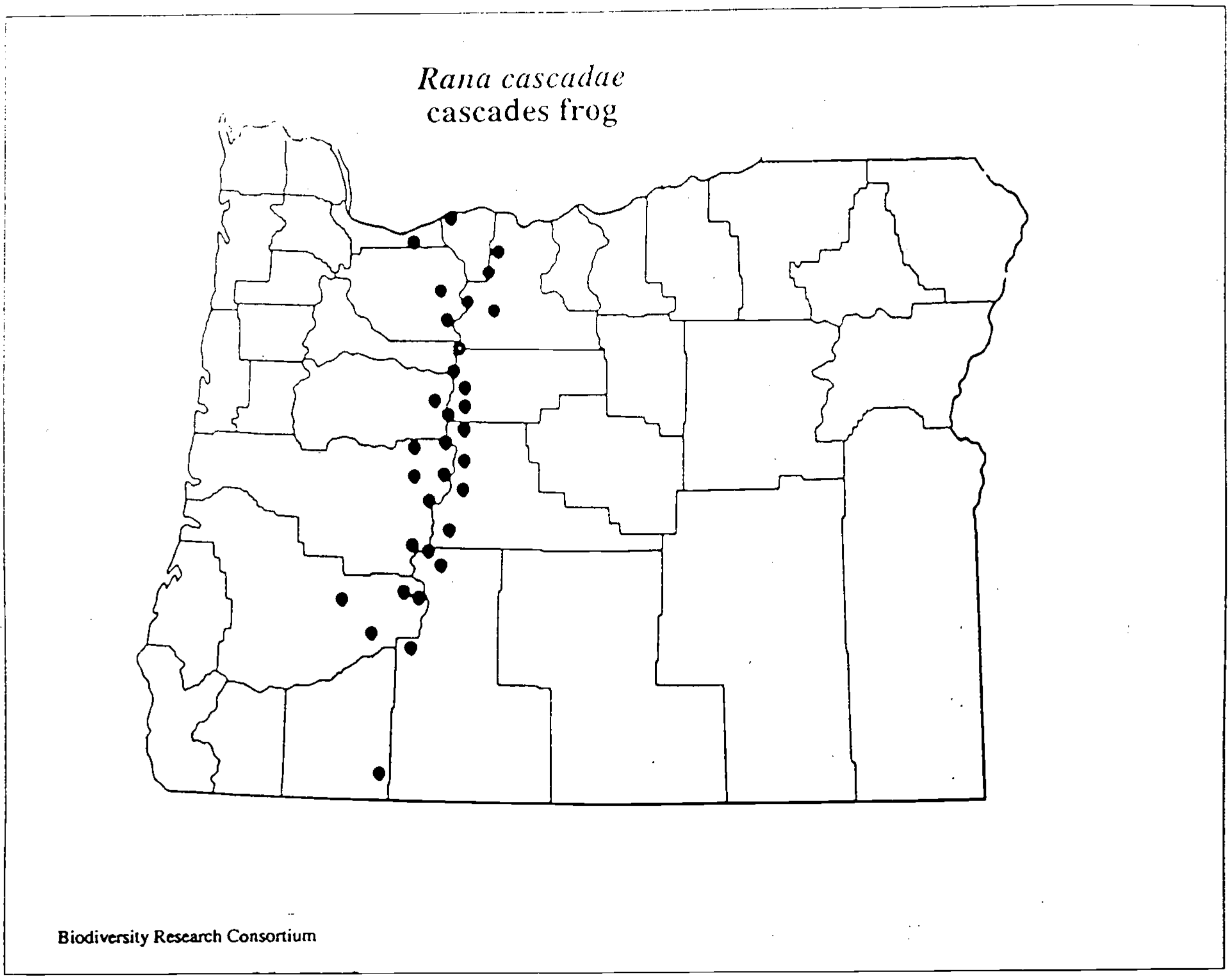


Figure 7. Distribution of Rana cascadae in Oregon by collection sites (from Nussbaum et al. 1983)

as points on a smaller scale map. When a study area is stratified or partitioned for sampling, data can be represented by a point at the centroid location of each partition.

2) *Discrete lines, polygons or blocks* are used to represent the extent of each sampling or observational unit (Figure 8). These units are spatially discrete because no sampling or data collecting was done at other places, so no information is available outside their boundaries.

3) *Continuous layers, response surfaces, or continuous volumes* are used to store data coming from the sampling of a surface or volume field, when the field is spatial units formed by formed by complete partitions or *tessellations* of the studied region. Statistical assumptions are: a) equal sampling effort in terms of areal coverage, so that every partition is sampled, searched or data assigned to; b) equal probability of sampling, if data have been interpolated between partitions; c) equal sampling effort in terms of repetitions or proportion of area sampled, within each partition; and d) internal heterogeneity does not affect the probability of being sampled (searched, scanned), nor the assignment of effort units. The partitions are usually called cells or areal units.

Regular geometric tessellations (Figure 9) are usually derived from sampling or aggregation purposes and provide areal units with fixed boundaries and regular geometric shapes. Cells can have different shapes, such as hexagons or squares. Cells can be of the same size like in a square grid or of different sizes, as in a quadtree. Regular geometric tessellations can be recursive, meaning that they can be repeatedly subdivided into cells of smaller size and exactly the same shape, or they can be non-recursive, when this type of subdivision is not possible. Their metric space can be euclidean, such as square grid on an albers equal area projection map, or can be non-euclidean such as spherical tessellation of latitude-longitude on the globe. Regular geometric tessellations provide a null model to assess or express the distribution of a species, as the placement of the boundary is independent of both the sites of occurrence and the abundance of the species.

Irregular tessellations can have artificial or 'natural' boundaries and can have regular or non regular shapes. Their boundaries are hypotheses of the homogeneous influence of a spatial factor at the scale considered. Examples of *irregular geometric tessellations* are Voronoi tessellations. Voronoi polygons represent similar influence of a factor located at the center of the polygon and have been used to evaluate rodent abundance in briar patches , on the assumption that they provide areas of equal accessibility to cover and food.

Non-geometric irregular tessellations are usually called homogeneous areas or choropleths. Boundaries of homogeneous areas can be considered *artificial* when derived from administrative divisions, such as counties, provinces, countries and ownership boundaries

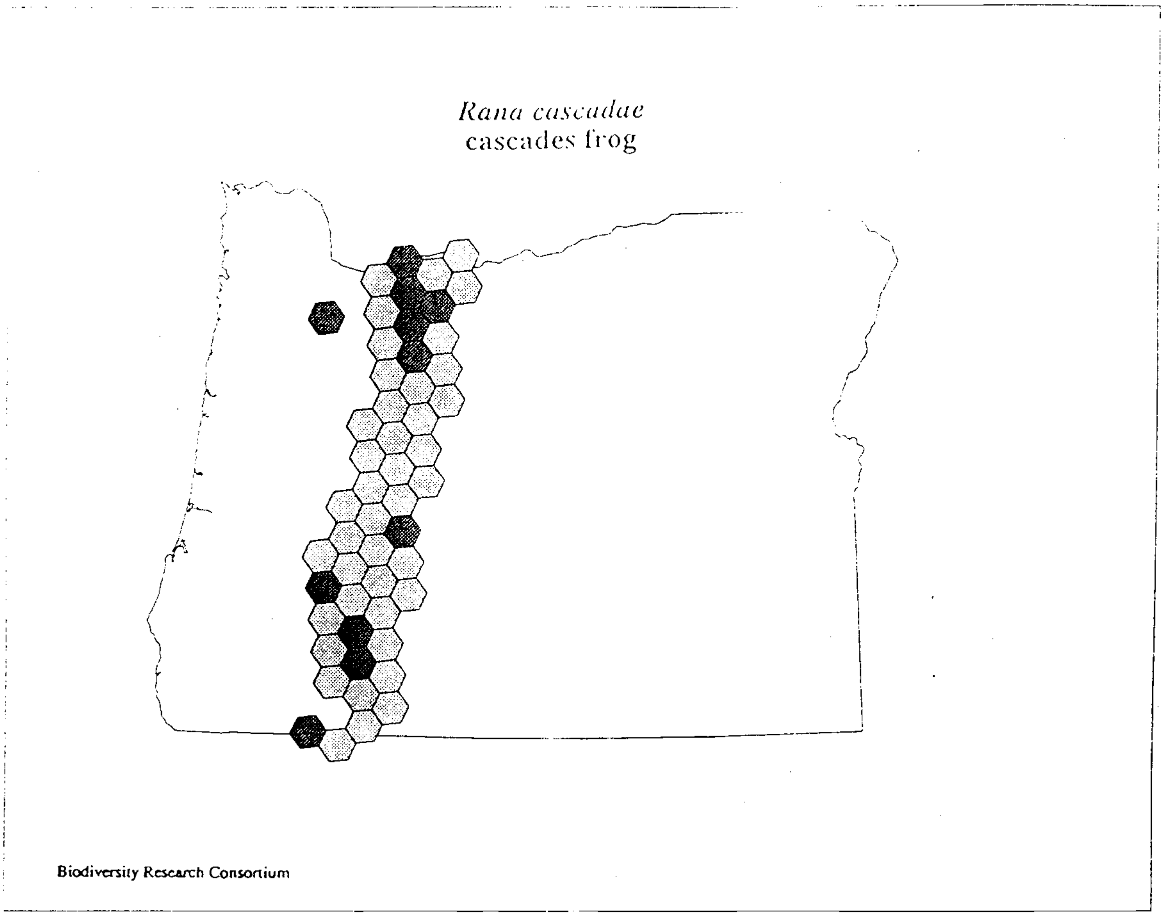


Figure 8. Discrete areas used for collating information on Rana Cascadae.

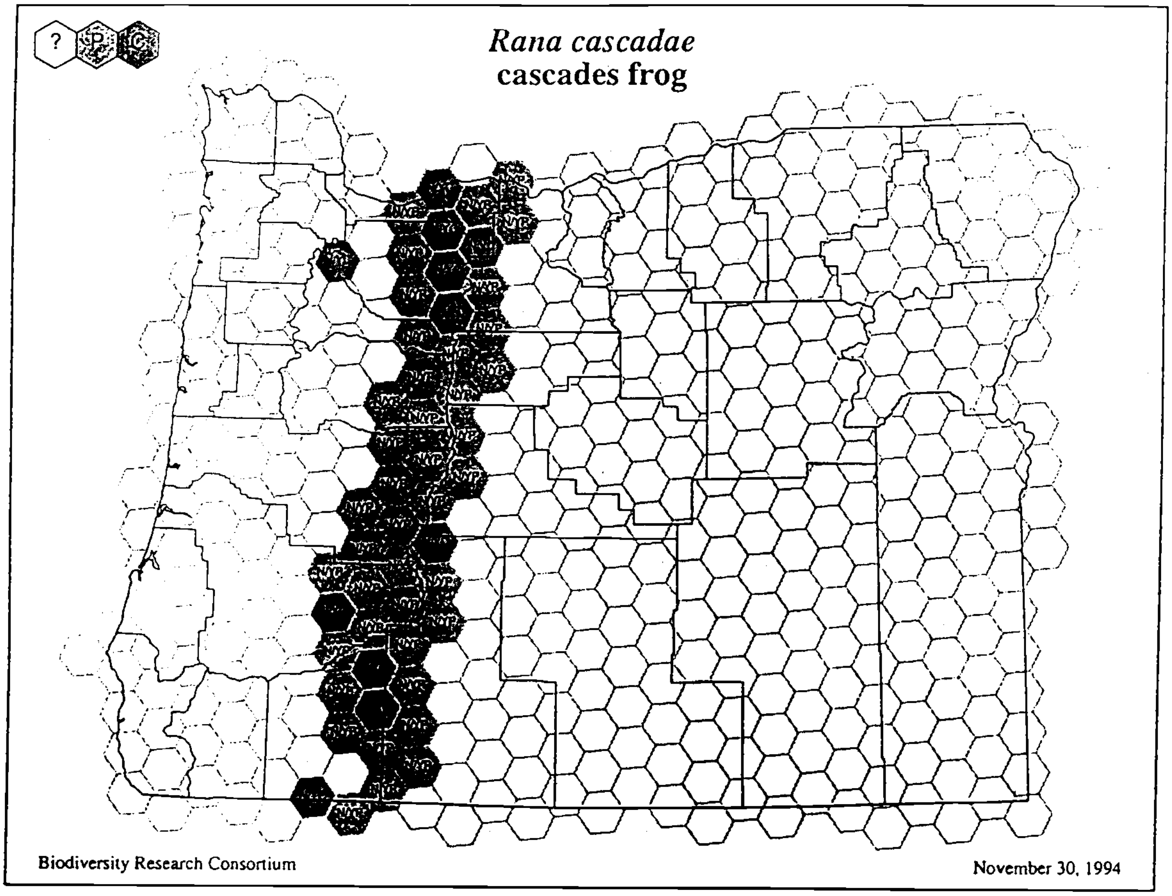


Figure 9. Hexagonal tessellation showing the distribution of *Rana cascadae* based on The Nature Conservancy species database.

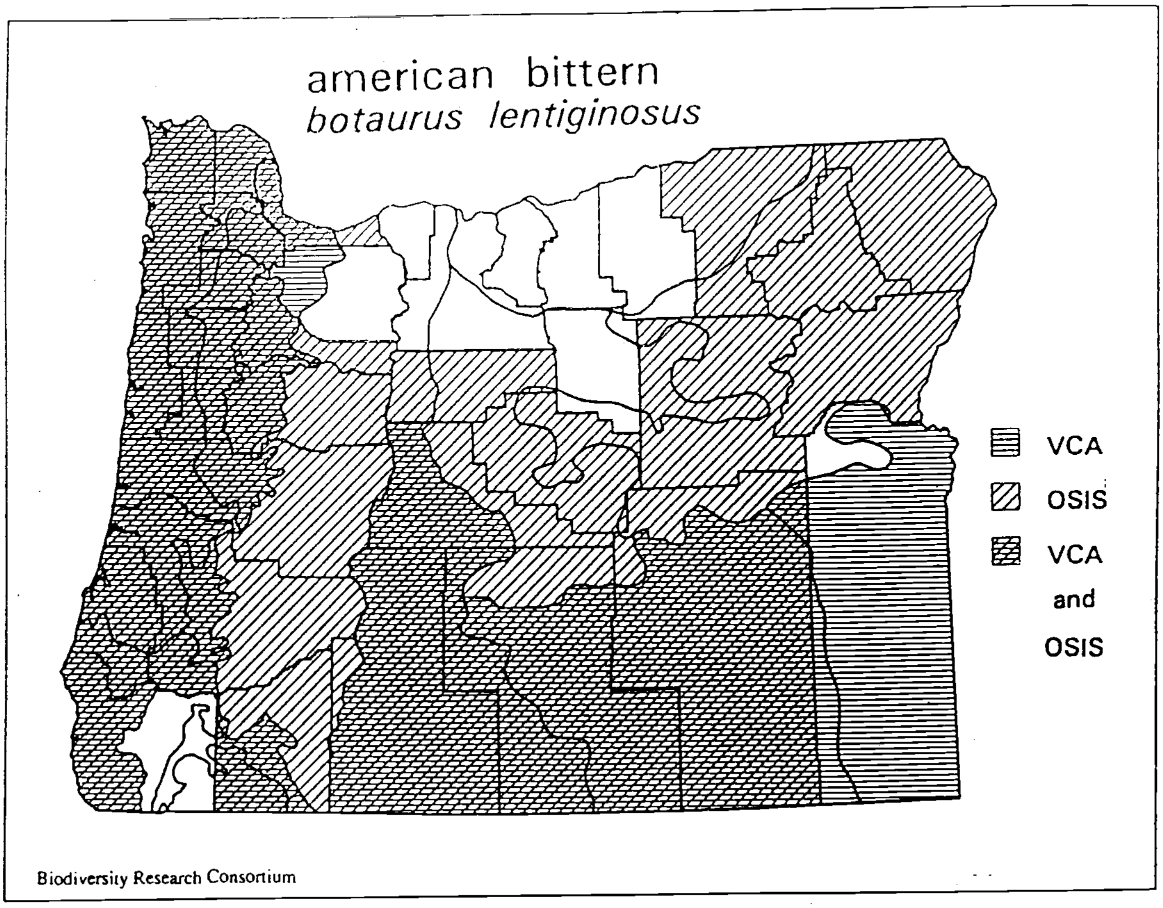
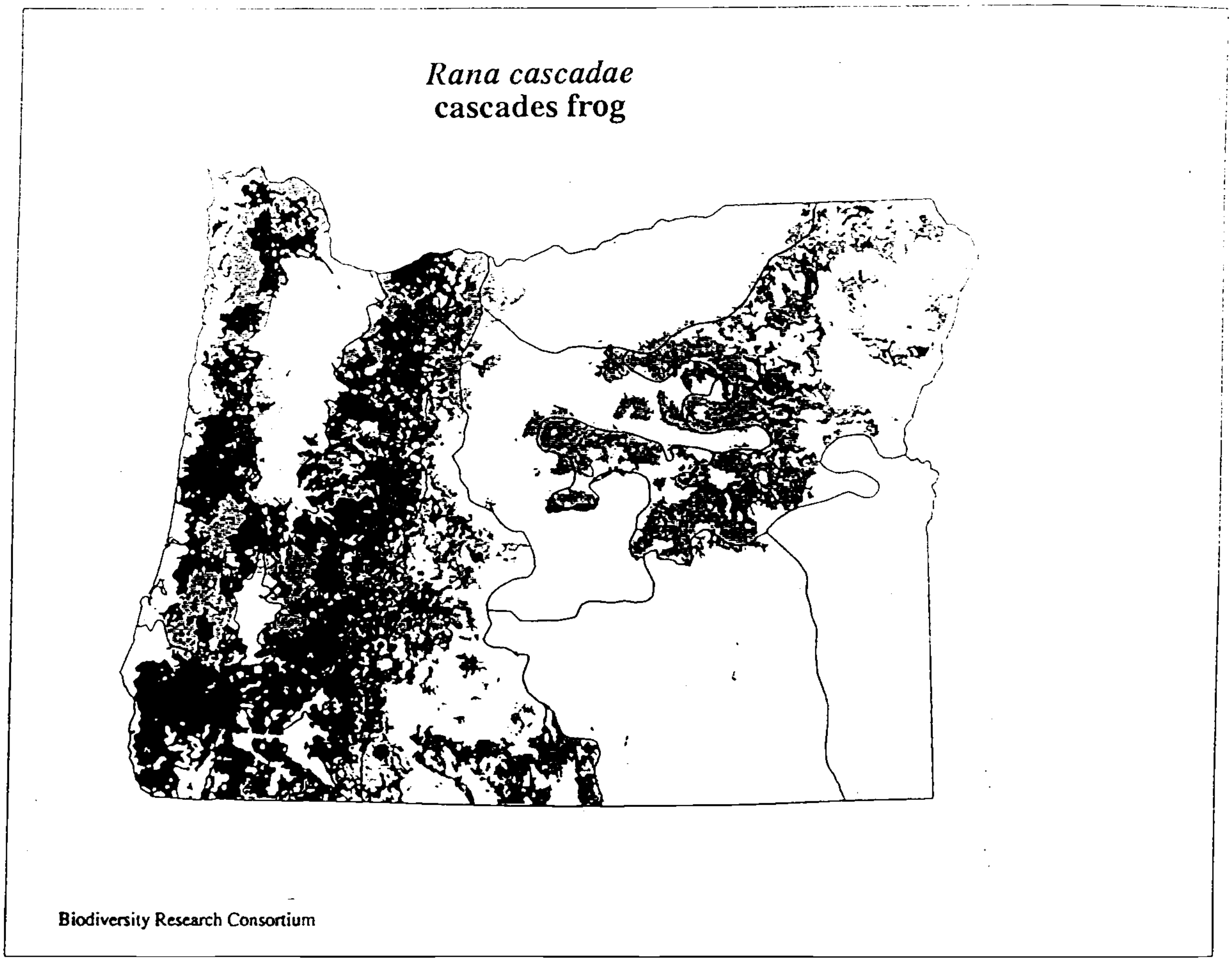


Figure 10. Species data information structured using spatial units defined by administrative boundaries (counties and EPA's ecoregions).



Rana cascadae
cascades frog

Biodiversity Research Consortium

Figure 11. Rana cascadae's habitat distribution based on landscape units.

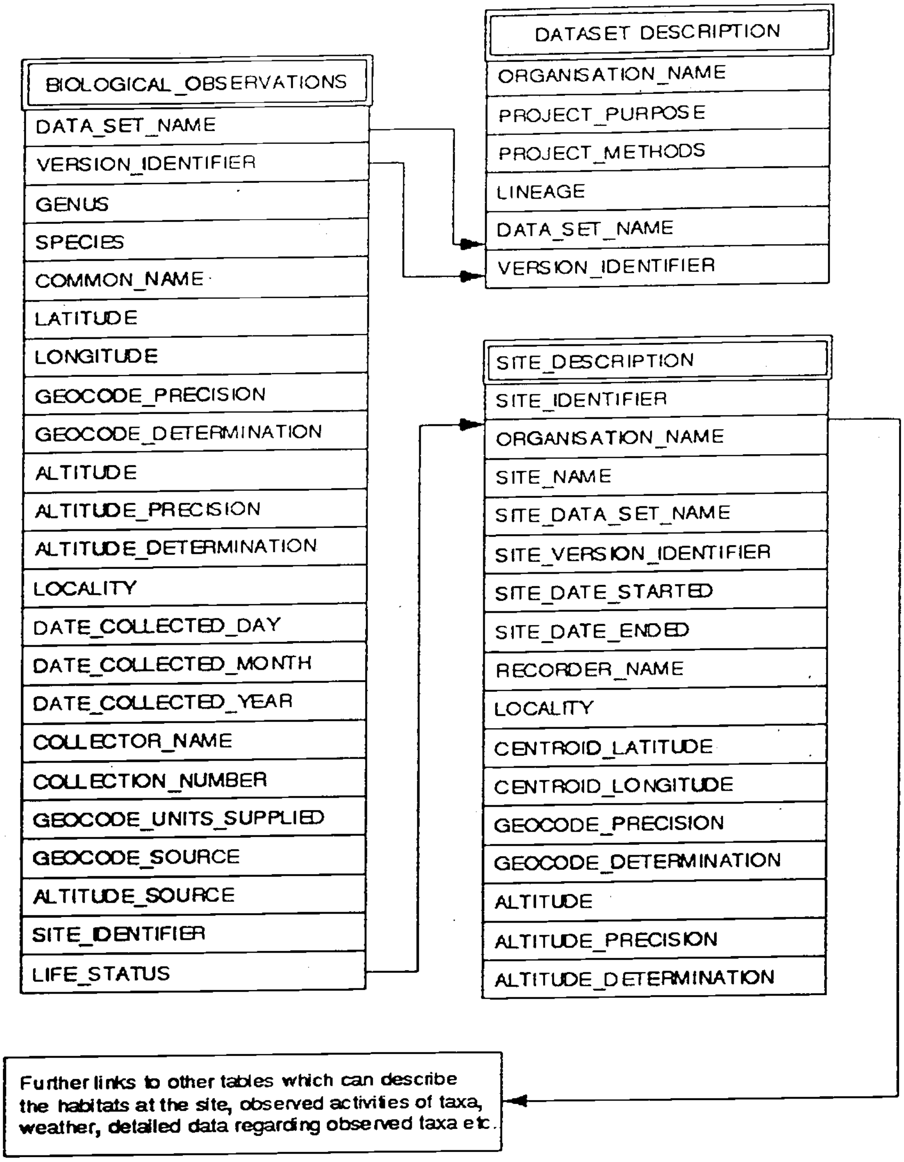
(Figure 10). In this case, the areal units represent an equal policy or management regime. Boundaries are considered *natural* (Figure 11) when the irregular tessellations represent homogeneity of action of ecological factors such as soils, vegetation cover and geomorphology.

5. Modeling biodiversity information into database objects.

By storing biodiversity units and their measurements in spatial units, we could try to design the themes of the databases based on the hierarchies of the biodiversity units. Some biodiversity units can be considered the building blocks of others. Their information can be stored in what could called 'primitive' databases. These "primitives" hold source or primary information that can be used to derive databases for biodiversity units at a different hierarchical level or spatial scale. The following database themes can be considered as primitives for biodiversity analyses :

Specimen databases hold information on localities of collection or observation of individual organisms as independent records. Non-spatial attributes include a taxonomic code for better decomposition into taxonomic levels, the scientific and common names of the species, time of collection, identification tag, habitat, time and date of collection. These are typically stored as a non graphic databases (Figure 12), but the expression of position in latitude and longitude allows one to translate and represent information as points in any spatial scale needed. Unfortunately, position is not always geocoded so precisely. Most records date from more than 30 years. Traditionally, the name of the closest locality was written in the tag, and coordinates were derived more recently from topographic map reading. Records can also contain information on specimens sampled for genetic content , or contaminant levels. Gene data will usually be mapped to the location of the individuals sampled, or to areal sampling units. As genetic data still have limited coverage, it is better to have them as a related non-spatial database linked by species and by specimen tag-ids if they are shared. For individual researchers it is more convenient to store this information as separate databases for each taxon, with exactly the same structures that can be easily updated and also recombined as needed. Range maps are derived from this source of information, using environmental information to delineate species limits.

Species databases hold information at the species level. Two types of independent databases may be developed. The first type is spatial databases holding information on species as discrete spatial units, from scanned or digitized published sources (see Figure



Further links to other tables which can describe the habitats at the site, observed activities of taxa, weather, detailed data regarding observed taxa etc.

Figure 12. Contents of a specimen database.

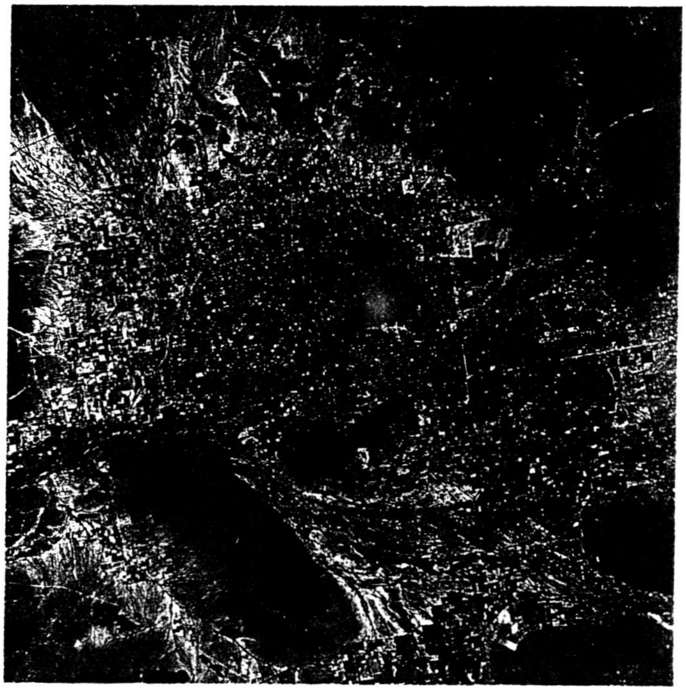


Figure 13. Landscape databases: Satellite imagery as source of landscape information.

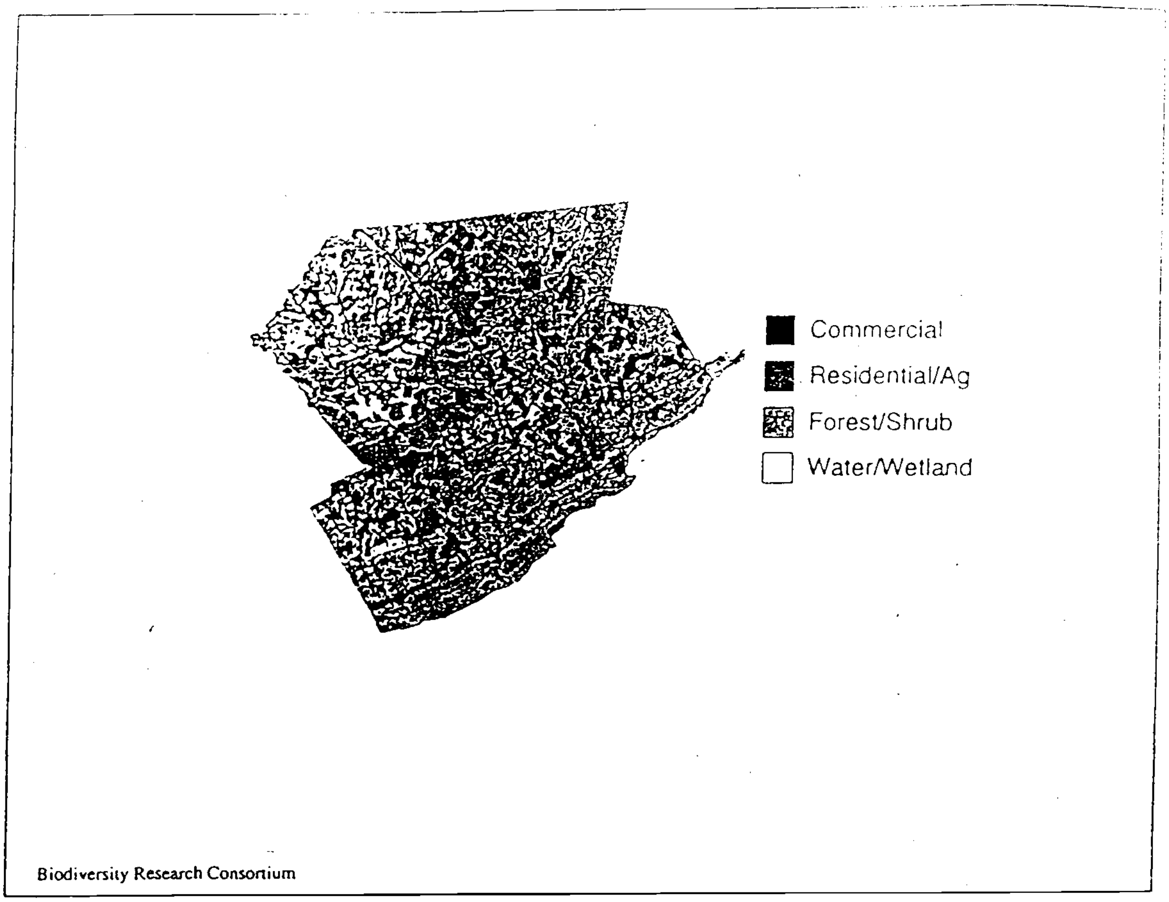


Figure 14. Landscape databases: landcover classes from classified satellite imagery .

7). These databases can be used to derive species occurrences on analytical units at small map scales, to obtain homogeneous areas of species richness by overlay operations at regional scales, or to serve as checks for distributional data input (see Figures 9 and 11). Non-spatial databases hold ecological information on habitat associations and requirements, such as food and home range, landcover/landuse type associations, guilds, official scientific names and synonyms. These allow one to link specimen with landscape information by means of habitat and landcover key.

Landscape databases store information on landscape elements in a complete sampling of territory and are usually expressed as regular or irregular tessellations (Figures 13 and 14). They can be recombined and grouped to delineate landscape patterns at coarser spatial hierarchies. These spatial databases can be described by any typical GIS spatial structure. They derive from visual or supervised and unsupervised analysis of remote sensing images and aerial photography. Attribute information can include landcover/landuse label, habitat type, list of species, area, and perimeter.

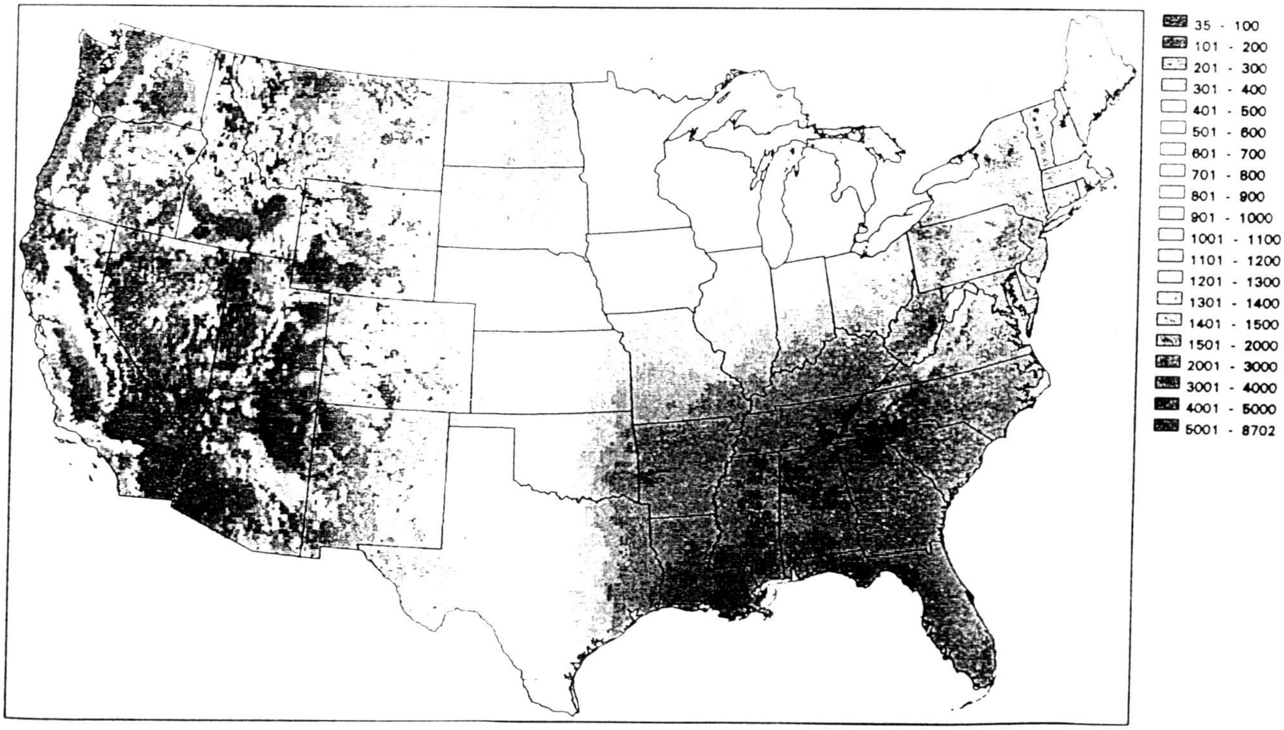
Environmental factors databases include environmental information not expressed by the landscape information (Figures 15) They also represent a complete sampling given by a tessellation or lattices. The information included represents the following groups of factors:
a) Natural factors and covariates such as climate, oceanic currents, elevation, and soils;
b) Stressors induced by human intervention;
c) Natural factors whose values have been affected by anthropic intervention.

Boundaries databases provide referential frames for analysis and presentation of results. They include information on administrative boundaries, spatial sampling frames and spatial units used to integrate or analyze biodiversity information .

The combination of elements in these databases allows one to derive specific databases as needs arise, with the advantage that the process can be modified and repeated as desired before an acceptable result is obtained. Spatial databases derived from compilation of paper maps do not have this flexibility, both in updates and for analysis.

Annual precipitation in millimeters

Based on Daly et al. (1993) PRISM model



Biodiversity Research Consortium Analysis Team

Sifneos,Minotti,Kiester,White Oct 93

Figure 15. Environmental factors databases: Climatic data.

6. Spatial transformations

Biodiversity assessments may require converting the source information into units that are more appropriate for automatic analysis of biodiversity distribution. This conversion can involve converting the units in which biodiversity is expressed, such as the point distribution of collected specimens, into the areal range of their corresponding species label, or its habitat. Changing the biodiversity units may result in a change of spatial unit. Other times, the transformation requires converting information expressed with one type of spatial unit into another, to obtain a biodiversity indicator that is suited for the analysis, for example, the overlay of range maps to estimate species richness on a grid. The results may need to be expressed in a more general way to convey the key points of the analysis, using a more suitable choice of spatial unit. The following list shows the main spatial transformations. The names follow the change in biodiversity information content.

Interpolation is the estimation of values within the spatial extent of the source information. There are different interpolation methods, all based on different models of how closeness between observations affects interpolated value. The simplest are proportional and inverse squared distance; they are estimated by a Delauney triangulation from the original points. Kriging methods allow one to interpolate irregular continuous and discrete lattices to regular lattices or grids, based on the pattern of covariance of the biodiversity or environmental indicator. Home ranges of burrowing animals can be established by calculating Voronoi polygons from the point distribution of burrow entrances.

Extrapolation refers to the prediction of values beyond the spatial extent of the source information. Extrapolation is performed by means of a theoretical or empirical model that is applied to the unknown areas. A common routine is to *extrapolate* species point observations to the geographic range by using a habitat model. The association between specimen localities and the co-occurrence of a certain landscape, or values of an environmental factor, define the habitat model. The occurrence of the species is extended to all spatial units matching the model.

Aggregation condenses different sources or detail of information inside a spatial unit by applying a statistical, mathematical, logical or cartographic operation. Combining species abundance information from four square cells into single congruent larger cell may be done by calculating an alpha diversity species richness, by estimating the mean abundance, or by selecting the maximum abundance as an indicator for the new resolution. When two polygons share the same label , dissolving their common boundary aggregates their species lists. Increasing the number of polygons by a union of layers also represents an aggregation of spatial information from two different sources.

Disaggregation or deconstruction is the application of mathematical, statistical, logical or cartographic operations resulting in the decomposition of spatial information into units of maximum detail, or into original or potential spatial components. The automatic delineation of habitat types from satellite imagery is usually done by aggregating information into spectral classes by means of a clustering algorithm, and then reassigning the pixels to their closest cluster based on their maximum likelihood probability. Each cluster is then related to a habitat type by comparing their mutual spectral signatures. This model of landscape partitioning into non overlapping habitats may be useful for displaying landscape heterogeneity, but is not adequate for modeling the distribution of species that require the spatial co-occurrence of two or more habitat types (Dunning et al 1994). Deriving a probability map for each cluster-habitat from the same source of imagery - a *disaggregated* landscape representation- can result in a more precise and realistic distribution for these species.

Visualization refers to operations on the analysis and resulting spatial units for the sole purpose of reporting and display (See Figure 6). Cartographic generalization can affect the line detail of the spatial units or the color detail of a raster map grouping species richness values, producing a more pleasant graphic document. A three dimensional elevation model can be used to drape landscape polygons, colored to show the degree of genetic similarity of Douglas fir stands, for example.

7. Quality, Validity and Uncertainty in Biodiversity GIS Databases

Whatever the results of a biodiversity assessment are, the outcome will probably be distributed, used and perused, being treated as the latest truth without given proper consideration to the presence of various kinds of uncertainties in these results. Errors can arise from bad original data, encoding operations and computational procedures (Star and Estes 1990), but also from inadequate conceptualization and structuring of the data and their properties (Burrough 1986, Burrough 1992, Goodchild 1992). Measuring instruments, physical or conceptual, can be biased and lead to inaccuracies, that is, the property of not producing the real value. The data may not have a one-to-one correspondence to a geographic location. The input species data used may not come from a valid source, such as the US. Breeding Bird Survey . Specimen location data may be describe by ambiguous geocodes. GIS operations will compound spatial errors with each successive operation (Goodchild and Gopal 1991). The ultimate burden of checking rests on the biodiversity analysts who should document not only how databases and analyses were designed and conducted, but also the quality checks on the data and procedures used. When the assessment involves many modeling steps, a **quality assurance protocol** needs be developed, that combines checking for known sources of errors in data and analytical procedures with checking for results that are unmeaningful or do not make sense, better known as the "gut feeling" or common sense criteria. The control of the quality in spatial databases is based on the duality of a piece of data: as the digital codification (alphanumeric or positional) of a source of information, and as the repository of this information. **Data verification** operates on the codification aspects and revises data coded incorrectly. Involves checking the integrity or consistency of the databases and can be done while the database is created, when completed or when updated. **Data validation** refers to the origin and quality of the data sources. As important decisions are going to be based on the available biodiversity data, the sources, precision and accuracy of the information need to be stated.

Checking for **accuracy or bias** in sources and databases is the most complicated task. To asses **locational accuracy** we need a reference that represents the correct location. Depending on the conceptual model of space for that biodiversity object, locational accuracy may not depend solely on its position given by a spherical coordinate. Accuracy is usually only measured for positional information (how far from the actual location, border, etc.) with no reference to the sources or types of biases that can be more critical for evaluating the quality of the data. Using the species range as an example, locational accuracy can be described as a compound of **positional** (where in coordinates), **taxonomical** (what species is describing this point?) and **ecological or referential** (is this the type of environment where this species can occur?).

Positional accuracy can be described by how far the observation was from the spatial unit that recorded its presence (e.g. fish species given by township while their range map had to depict their distribution in streams and rivers). Many taxonomists extract latitude and longitude positions of collecting points from topographic maps, being unaware of errors introduced by the projection type, its parameters and reference datum, map generalization and/or cartographic license. Positional errors due to displacements in the location of rivers and roads are quite common but hardly ever taking into account .

Taxonomic accuracy tries to evaluate if the recorded species is really such. It falls in the category of label inaccuracies. Taxonomical inaccuracies arise from misclassifications and failure to revise collections. Jennings (1994) presented interesting examples of misclassified amphibian species in the state of California and warned against the misuse of species lists. The probability of misclassification varied from 0 to 100% and was related to Museum source and time since last taxonomic revision of the collection. He blamed the curators for forwarding species lists from collections that have never been revised for consistent classification. This implies that the assessment procedures should be robust against the intrinsic uncertainty in taxonomical assignment. When the probabilities of confounding a species with another are known, fuzzy data models could be used to represent the degree of certainty in the species location based on 1) the presence of a close or similar species within the range and 2) the collection source. When the probabilities are unknown, a Monte-Carlo approach can be used to evaluate the sensitivity of the assessment procedure, by randomly including and excluding a certain percentage of species/taxons from the presence list in a given spatial unit (Kiester et al. 1993).

Ecological accuracy asks if a location really reflects where the species lives. There are hardly any references dealing with this topic that should be the most critical. Species range maps are good to analyze species distributions at continental scales. At landscape scales, part of the range is actually non-habitat. The use of coarse scale range maps in landscape level analysis produces results with uncertainties that cannot be estimated or handled, diminishing the credibility of biodiversity analysis as a tool for decision-making. Ecological accuracy can be estimated by modeling the following question "does this species live here?". A species is potentially in a place if : 1) it is/was consistently observed during at least the last ten years (to account for interannual variability of wet/dry cold/warm climatic cycles); 2) its habitat is present (Csuti 1994); 3) the suite of environmental factors that affect its distribution are within the tolerance limits of the species (BIOCLIM modeling, Bisby 1992); or 4) it is common or abundant.

IV. DISCUSSION

Biodiversity databases need to be a common repository of information on the living world. Biodiversity information should be structured in a way that facilitates and encourages its analysis, so that we can build a knowledge base to diminish threats and enhance opportunities for the recovery of biodiversity. The present research has analyzed some of the problems of representing biodiversity information in a spatial database for conducting biodiversity assessments in a geographic information system environment.

Biodiversity assessment is multipurpose. This implies that databases can have different geographical extents, different structure and spatial resolution and different attribute descriptions depending of the type of analysis involved and the source information.

Biodiversity is a complex geographical object. It can be hierarchically decomposed into constituent units of successively smaller spatial extent according to two approaches. The biological series defines biodiversity units that are hierarchically related by spatial containment but it is not possible to decompose a geographical area into biologically defined biodiversity units. Species contain populations, that in turn contain demes, that contain genes carrying biodiversity information. But to define the spatial extent of a species, we need to sample the area to locate populations, demes or genes belonging to this species, and use these locations to construct the species' spatial extent (species' geographical range). The ecological series defines biodiversity units based on spatial extent and spatial relations. These characteristics allow the partition of a geographical area into biodiversity units for a given hierarchical level. A given partition or classification at one scale does not necessarily translate to other scales, because the property of spatial containment is incomplete between hierarchical levels.

Biodiversity information can have multiple representations, both in terms of attributes and spatial representation. Using the species as a representative biodiversity unit, we find that attributes describing the genetic variety of a species do not serve the purpose of a conservation biologist trying to assess the distribution of endangered species. For assessing the conservation status of a rare species, a point representation of the known localities for its population is preferred to a species range map. For estimating the species abundance, a map of landscape patches containing the species would be more informative. To assess species richness at a state level, the overlay of range maps would be enough.

Biodiversity information is not uniformly georeferenced. Biodiversity units can be georeferenced to a point or pixel area locality or can be geocoded to another biodiversity unit at different hierarchical level. This presents problems for data integration as the cross-referencing is not direct but involves different types of spatial analysis or queries. The georeferencing method used neither guarantees the precision nor the accuracy of the data stored. Both biological and GIS scientist forget that precision is linked to the geographical scale of the source document, the smallest precision will be that obtained directly in the field.

Biodiversity data sources present different levels of processing at the spatial level. We usually consider that data are raw if they have not been processed numerically. GIS analysis requires extending this concept to data that has not been processesed spatially. This degree of spatial processing is important for interpreting data uncertainty and modeling spatial error. The assumptions and criteria used to construct species range maps are rarely documented, as these maps are considered visualizations or conceptualizations of the species actual occurence and not as cartographic documents presenting a valid description of this reality. GIS data capture assumes that biodversity maps are like cartographic documents with defined precision and accuracy in terms of boundaries and label. If the information presented in one type of spatial unit actually refers to a quite different extent, it introduces uncertainty that cannot be accounted for. This means that the best quality source maps for biodiversity assessments are the ones that present biodiversity data in the same spatial units used for sampling (raw spatial units). Data can be generalized or enhanced afterwards according to needs for analysis or modeling. Maps of acceptable quality are those in which the level of spatial processing is fully documented.

The accuracy of biodiversity data is described by at least three components. Positional accuracy describes locational uncertainty in relation to a coordinate georeferencing system. Labeling accuracy describes the uncertainty in the identity (label) of the biodiversity unit. For species, this corresponds to taxonomical accuracy; for communities and landscape units this is known as label accuracy by image analysts and photointerpreters. The third component is named here as ecological context accuracy. It describes the ecological uncertainty in the location of biodiversity data.

The findings just discussed here highlight the importance of a proper database design. Bad conceptualization of biodiversity data at the spatial level can result in the degradation of information. Users -biodiversity researchers- should be actively involved in the conceptual design of biodiversity databases, bringing their discipline's reference frame for assessing inconsistencies and lack of utility, as well as they discover new ways to interact with their data.

V REFERENCES

Bisby, J. 1991. BIOCLIM - A bioclimate analysis and prediction system. In C. Margules and M. Austin (eds) Nature conservation: Cost effective biological surveys and data analysis. CSIRO Publications, East Melbourne, 64-68.

Burrough, P. A. 1986. Principles of geographic information systems for land resources assessment. Clarendon Press, Oxford.

Burrough, P. A. 1992. Are GIS data structures too simple minded? Computer and Geosciences 18: 395-400.

Burrough, P.A. and A.Frank. 1995. Concepts and paradigms in spatial information: are current geographical information systems truly generic? International Journal of Geographical Information Systems 9(2): 101-116.

Chambers, D. 1989. Overview of GIS databse design. ARC News 11(2):1-13.

Chapman, A. 1992. Quality control and validationof environmental resource data. In, D. Chapman (ed.) Data Quality Standards; Commonwealth Land Information Forum, Canberra, Australia.

Cousins, A. 1994. Taxonomy and functional biotic measurement, or, will the Ark work? In P.Forey, C Humphries and R. Vane-Wright (Eds.), Systematics and conservation evaluation. Systematics Association Special Volume 50, Clarendon Press, Oxford: 397-419.

Csuti, B. 1994. Gap Analysis: Identification of priority areas for biodiversity management and conservation. Idaho Cooperative Fish and Wildlife Research Unit , University of Idaho, Mimeo.

Dunning, J.B. B.J. Danielson and H.R. Pulliam. 1992. Ecological processes that affect populations in complex landscapes. Oikos 65(1):169-175.

Getis, A. 1993. GIS and modelling prerequisites. In A. Frank and I.Campari (Eds.): Spatial information theory: A theoretical basis for GIS. Springer -Verlag, Berlin: 322-340.

Goodchild, M. 1992. Geographical data modelling. Computer and Geosciences 18: 400-408.

Goodchild, M. and S. Gopal. 1989. Accuracy of spatial databases. Taylor and Francis, London.

Groombridge, B (Ed.). 1992. Global biodiversity: Status of the earth's living resources. A report compiled by the World Conservation Monitoring Center. Chapman and Hall, London.

Jennings, M. 1994. The use of unverified museum databases for land management decisions: the case of native california frogs. Proceedings of the XI Meeting of the American Asociasion of Ichthyologists and Herpethologists, Los Angeles, California: 108.

Kiester, A. R., D. White, E. Preston, L. Master, T. Loveland, D. Bradford, B. Csuti, R. O'Connor, F. Davis and D. Stoms. 1993. Plan for pilot studies of the Biodiversity Research Consortium. Mimeo.

Laurini, R. and D.Thompson. 1992. Fundamentals of spatial information systems. Academic Press, San Diego CA, 680p.

Peuquet, D. 1988. Representations of geographic space: Toward a conceptual synthesis. Annals of the Association of American Geographers, 78(3): 375-379.

Rhoads, A. and L. Thompson. 1992. Integrating herbarium data into a geographic information system: requirements for spatial analysis. Taxon 41: 43-49.

Rojas, M. 1992. The species problem and conservation: What are we protecting? Conservation Biology, 6(2): 170-178.

Rowe, J. 1992. The integration of ecological studies. Functional Ecology 6: 115-119

Star, J. and J. Estes 1990. Geographic Information Systems: An Introduction. Prentice Hall, Englewood Cliffs, New Jersey, 303 pp.

Stoms, D. F. Davis and A. Hollander. 1993. Hierarchical representation of species distributions for biological survey and monitoring (draft).

Stoms, D. F. Davis, J. Scepan and J. Estes. 1991. Assessing biological diversity: A geographic information system approach. International Conference and Workshop on Global Natural Resource Monitoring and Assessment, Venice Italy, 2: 432-442

U.S. Congress Office of Technology Assessment. 1986. Assessing biological diversity in the United States: Data considerations. Background Paper 2 , OTA-BP-F-39, US Government Printing Office, Washington D.C.

White, D. A.J. Kimerling and W. Scott Overton. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. Cartography and Geographic Information Systems, 19(1):5-22.