Chapter 4

Technical Aspects of Biologcal Data

Chapter 4

Technical Aspects of Biological Data

HIGHLIGHTS

- **Automation and computerization have provided and should continue** to pro**vide many tools to increase the quality and quantity of biological data, to store and manipulate data, and to increase access to data.**
- **These tools also can perpetuate data incompatibility and inaccessibility.**

DATA COLLECTION

Data=Entry Technologies

Developments in automated data-entry systems for use in the field can improve data collection by reducing data-entry errors. These portable systems remove unnecessary intermediate steps, Traditionally, biological data have been recorded on collection forms and carried by hand to central processing facilities to be copied onto computer coding forms. (Figure 1 shows the paths along which data are transferred by conventional methods and by automated data-entry systems.) Developments in the technology of electronic hardware have resulted in small solid-state memory units with increased storage capacity and longer lasting batteries. These advances have permitted the recent development of dependable, portable devices for entering and storing data in the field. When such devices are used, data are immediately recorded in a computer-compatible form, and intermediate steps of transcribing data are eliminated (11]. However, eliminating intermediate steps may mean that not all parts of the field data are entered into the computer, therefore some valuable information (e.g., animal behavior) may be lost. Moreover, these portable computers may not be rugged enough for some kinds of field work where dust and moisture are a problem.

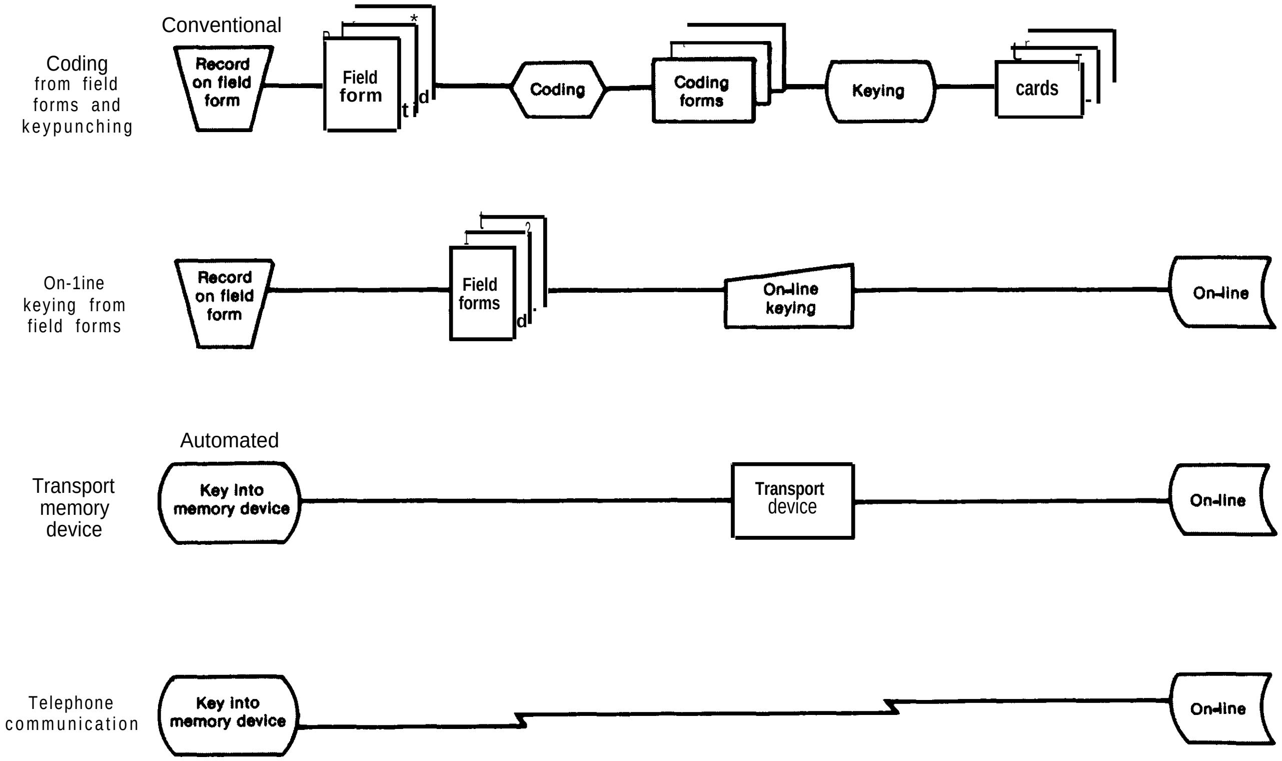
The data-entry process has also been simplified by developments in optical-character-recognition technology, which makes it possible to scan and convert printed materials into digital form. Companies are beginning to market technologies that can scan and digitize almost anything that is printed on a piece of paper. Once digitized, the image is stored on a microcomputer diskette from which it can be accessed and manipulated. Errors occur in copies or field forms but the problems are in the process of being solved. Although this technology provides opportunities for capturing graphic material, the large amount of memory required for storing the images makes the technology relatively expensive (7).

Remote Sensing Devices

Other significant technological advances in data collection have occurred, improving the quality of data and broadening the scope of the data collected. One such technology is remote sensing, which has developed rapidly during the past 25 years as one benefit of space research.

Remote sensing means gaining information without direct contact (6), Remote sensing technology includes aerial photography, radar, infrared imagery, and other devices. With regard to on-site maintenance of biological diversity, remotely sensed information is most commonly used when preparing inventories to establish baseline data and to allow monitoring of changes that occur over time, For *instance,* remote-sensing imagery can detect large animals, such as seals, caribou, and pelicans, thus providing a means to census animal popula-

Figure 1 .–Paths of Data Transfer for Conventional and Automated Data Entry



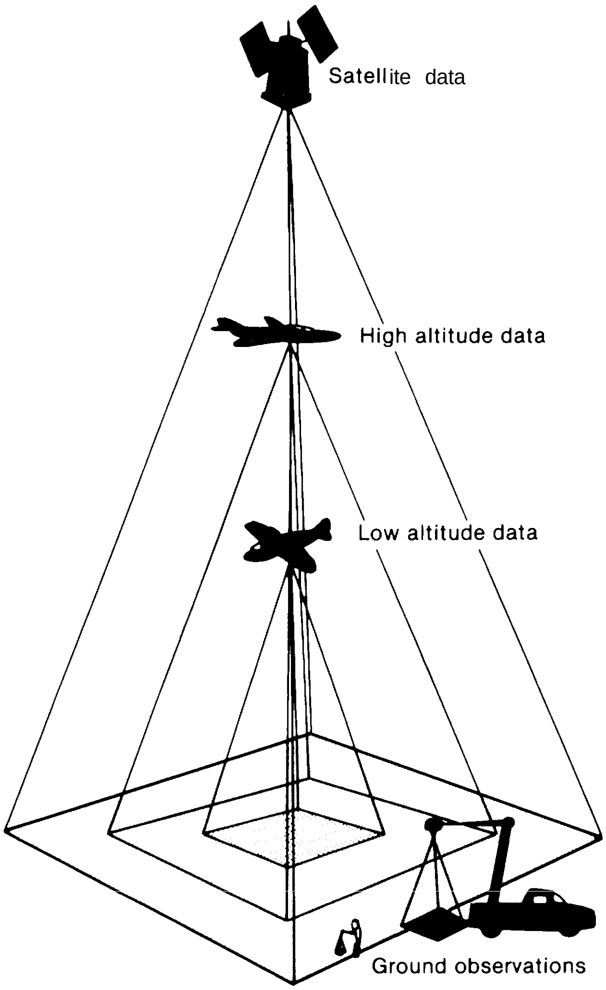
SOURCE, J.A, Newcomer and J. Szajgin, "Evaluation of an Automated Field Data Entry System," */n-P/ace Resource Inventories: Principles and Practices—Proceedings of National Workshop* (Orono, ME: University of Maine, 1981),

tions. However, the utility of this type of data is significantly reduced by several factors, including interactions between the color of the animal and the color of the background, and the obscuring effects of vegetative cover (10). The use of thermal-infrared scanning, which senses temperature differences between animals and their backgrounds, apparently has met with limited success (3).

Remote sensing has been used more as a tool for studying vegetation than for studying animals. Aerial photography is used when detailed habitat data (e.g., information about vegetation cover) are required for relatively small sites. Landsat is used for general reconnaissance surveys of large areas, because it provides multipple views at low cost. Landsat imagery is used to delineate crop production regions or specific forest types. Its repetitive coverage makes the Landsat system useful for temporal monitoring of habitat changes. Low-level aerial photos, especially color infrared, are much better than satellite imagery for showing the dominant genera and species of vegetation. Even so, this type of information is more reliable if supported by field verification (see figure 2) (2).

Although remote sensing has been used routinely by government agencies and other organizations to collect biological data, its costs are often high, especially where time-effective imagery is necessary, making it unaffordable for the average user. Other problems, such as unfavorable weather conditions, frequently prevent aerial photos or satellite images from being available for the season that would provide the best biological information. Special skills are required to interpret and use the information derived with some of the technology. And

Figure 2.— Remote Sensing Devices Increase the Coverage of Area From Which Data Is Collected



SOURCE: R. Best, *Handbook on Remofe Sens/ng In F/sh and W//d//fe* Manage. *menf* (Brooklngs, SD Remote Sensing Institute, South Dakota State Unlverslty, 1983).

in some cases (e. g., monitoring changes in species composition within a small area), the level of resolution may not provide as much detail as field verification does for the kinds of data that are most useful.

The use of remote-sensing devices in coastal areas also presents a variety of problems. The vastness of coastal zones necessitates coverage of extensive areas. Urban areas and near-shore waters are heterogeneous, however, which means that high resolution of the land coverage is needed to identify habitat patches. Information must be provided about the sea's surface as well as its subsurface down through the water column to identify ocean biota. To further complicate the matter, cloud cover in coastal zones requires either more frequent scanning or cloud-penetration capability, raising the cost of the technology (15).

DATA STORAGE

Developments in storage technology over the past **20** years have increased the options for storing data. Information-storage technology probably is the topic of greatest interest to those responsible for managing biological data.

Software

Biological data may be stored in unstructured flat files–two-dimensional files in which data are entered in the order received. But a structured database management system (DBMS) greatly increases the efficiency of data storage and use. The added efficiency results from the fact that most biological databases are dynamic and open-ended: that is, new parameters are added, old ones are dropped, and new ways of examining data are to be expected (4).

A DBMS is a software system that provides access to the database and accommodates a variety of applications using the same data. Most of the generalized DBMSS in use today were developed in business environments and were influenced by the problems with hardware, software, and other factors encountered there. Most of the systems are oriented toward modifying and retrieving formatted data, such as inventories and accounts. Many business-oriented DBMSS are successfully applied to scientific data, particularly for low-level tasks such as keeping structured records and generating reports (14),

Systems designed specifically for managing scientific data began appearing in the early-1970s (9). Most of these systems either lack full sets of data management operations or are specialized. Management of scientific data is most advanced in well-funded areas, such as defense, medicine, and industry-related research, and least advanced in the biological sciences (9),

Some difficulties in designing databases reflect the extreme complexity of the natural environment. A database is an abstraction of the natural system, which cannot be described in all its original detail. Problems with definition of abstract entities occur in such tasks as delimiting entities (from the continuous whole); distinguishing like-entities (e.g., assigning unique identifiers to each pine tree); and identifying changes in the essential natures of entities (l). Some problems with modeling a natural system result from insufficient knowledge of the system itself, Deciding what biological attributes to include for each entity can be a major problem in designing a database, because the types of questions to be analyzed are not yet fully known.

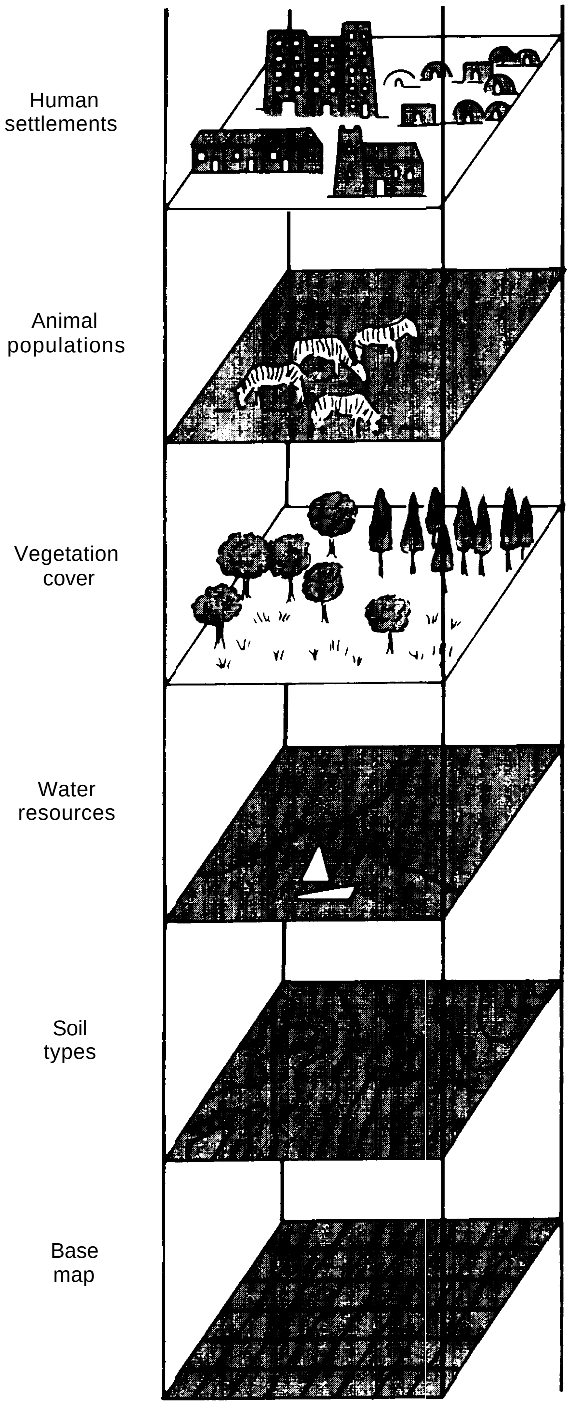
The majority of biological databases available today do not contain the data necessary for spatial analysis (e.g., geographic distribution of a species) and plotting of the spatial data. A special type of data management system, known as *geographic information system* (GIS), has been developed to handle explicit spatial data, make necessary calculations, and plot that data as required. A GIS is particularly useful to biological diversity maintenance because it provides the means to present an integrated view of a geographic area by, for example, overlaying several kinds of data. Overlay mapping allows a view of the distribution of plant and animal species over a large area and reveals the factors (e.g., roads, streams, different habitat types) that might have influenced such a distribution. (Figure 3 illustrates the kinds of information that can be overlaid.) (See the American Farmland Trust 1985 survey of GIS softwares for more information,)

GISS can display any information capable of being mapped on graphics terminals. Most of the several types of GISS can generate overlays and display maps and can change scales readily. The more sophisticated ones can compute areas, distances, peripheries, and intersections; and use shading to enhance understanding. Recent advances allow color coding and color display and printing, Major Federal land managing agencies (e. g., the Bureau of Land Management, the National Park Service, the Fish and Wildlife Service, and the Forest Service) have been using, or are beginning to use, GISS to map natural resources and to conduct project impact assessments.

The large volume of spatial data required to run a GIS restricts its application to small areas, As the level of resolution increases, the volume of spatial data increases, which means that the storage capability must also increase, Unpublished estimates by the Defense Mapping Agency indicate that a world database at the lo-meter level would require a storage capability on an order of magnitude larger than any database known at this time (10). Consequently, many GIS applications in the United States cover only project-specific areas. The largest civil GIS operational today, the Canada Geographic Information System, which is operated by Environment Canada, covers the developed portion of Canada and is still acquiring data nearly two decades after establishment (lo).

A major problem in spatial databases today arises from the desire to represent different levels of spatial generalization within the same database or to change the level of generaliza-

Figure 3.— Representation of a Main Geographic Information System Function of Overlaying Several Types of Environmental Data



SOURCE Untted Nations Environment Programme/Global Environment Monitor. Ing Systems, G/oba/ *Resource* /n70rrnaf(on Databases (Nairob/ GEMS Pub//catlon, 7985)

tion once the database has been created. Where multiple levels of spatial generalization are required, the current solution is to create multiple copies of the database, one for each different scale. Other limitations of GISS include their high maintenance costs, the labor intensiveness of digitizing data, and the need for technical expertise to operate the system. A Forest Service brief estimates the initial installation cost of a GIS to be $50,000 to $100,000.’ Finally, different systems cannot easily share data when the data are stored and handled in different and incompatible ways.

Hardware

Significant hardware developments include 30-fold gains in processing speed, major increases in reliability and storage capacity, dramatic reductions in the sizes and prices of equipment, and improvements in display resolution and graphics capabilities. Data management and analysis have relied primarily on mainframe computers, which are fast, have large memories, and are very expensive. Minicomputers offer many of the characteristics of mainframes at substantially lower prices. Because of storage and processing limitations, however, a minicomputer cannot serve as many users as a mainframe can serve (12)$

Microcomputers have revolutionized the computer industry. They are fast enough for most single-user applications and perform best on tasks requiring quick responses. Recently, however, a new class of computers based on the latest 16- and 32-bit[2] microprocessors has been developed. These supermicrocomputers bridge the gap between the micros and the minis, offering minicomputer capabilities while serving several users at one time. The supermicros now cost between $10,000 and $50,000, but they should become less expensive within the next few years (12).

I This figure is intended only to give a general idea of the magnitude of cost.

‘Generally, the larger the bit size, the greater the amount of memory a computer can manipulate at any given time, and the faster the manipulation. Therefore, *16*- or 32-bit microcomputers are faster and can handle more data simultaneously than the 8-bit micros developed 5 to 10 years ago.

Major developments are also taking place in magnetic storage, optical disks, and a combination of magnetic and optical storage (7). Magnetically stored data is recorded by repeatedly polarizing tiny areas along the surface of the magnetic medium. The size of the polarized areas and how closely together these areas can be packed determine storage capacity. Technological advances continue, allowing more and more bits of data to be packed into given areas.

An optical disk typically stores data as a series of spots on a light- or temperature-sensitive medium. The technology seems ideal for preparing multiple copies of archival data as information is digitized onto a master disk and then replicated for distribution. The lifespan of the optically recorded information is projected at more than 40 years (7), The major disadvantage of optical storage when compared to magnetic storage is that the data cannot be erased.

Efforts to develop erasable optical-storage media currently combine magnetic and optical technologies. Heat from a focused laser beam is used to impose a magnetic orientation. This information then can be read using polarized laser light. As the method of storage is magnetic, information stored in this fashion can be erased to free disk space for the storage of new data.

DATA RETRIEVAL

Until recently stored data had to be retrieved by searching card catalogs and agency file cabinets or by making telephone or mail inquiries —very time-consuming tasks. Today, technological developments have facilitated the ease with which data can be retrieved. Improvements in telecommunications, in particular, have increased the number of options.

Data transmission via telephone is one of the fastest growing fields in the information industry. This growth has been spurred by such developments as fiber-optics technology, which promises improved efficiency and lowered costs. Using laser light and a bundle of glass strands, it is possible to transmit more than 240,000 telephone conversations simultaneously. The major problems with telecommunications for data transmission at present are the comparatively high costs, relatively slow speeds, and fairly high error rates. Satellite technology has great potential to reduce the cost of long-distance transmissions. Because telecommunications depend on the quality of the phone connections, their reliability varies with location and time (8).

The same telecommunication technology that provides access to a large centralized database also facilitates access to smaller databases distributed in different localities. Advances in telecommunication technology make it increasingly practical to build small, local databases that can be remotely accessed and maintained. With the aid of special software to facilitate access, the user can easily access data residing on different computers. Data at available varied locations can be searched, modified, or moved from one computer to another (transferred from a mainframe to a microcomputer, for example) for future use (7).

Telecommunication developments also have facilitated the use of data in printed form. Microcomputers coupled with optical scanning devices can now be used to store printed images, which can be transmitted to remote locations.

OPPORTUNITIES AND CONSTRAINTS

Automation and computerization have provided many tools to increase the quality and quantity of biological data collected, to store and manipulate data in a variety of forms, and to allow more users to have access to data. Computer software and hardware are changing rapidly, are becoming easier to handle, and are less costly to acquire and maintain. The range of new technologies, however, can present some difficulties that require careful planning to overcome.

Many software packages exist, and more are being developed. Software for many operating systems is available in the public domain, but documentation of public domain software tends to be poor, and problems using the programs are common (13). Commercial software packages often are better documented than custom software or software developed in-house. Moreover, the costs of the latter maybe higher than those of commercial software when the hidden costs, such as salaries, are included. In addition, the utility of custom or in-house software may decline when the original developers or users leave, taking away their intimate knowledge of the system (5).

The variety of technological options brings problems as well as opportunities. At present, biological databases are fragmented, and each is designed for its own purpose. (See ch. 3,) This situation creates incompatibilities that may hinder the process of linking databases together or of simply exchanging data between agencies. The diversity of technologies can exacerbate this problem. Caught up in computer enthusiasm, individual database managers may use different software programs or create their own programs, making it difficult to access data on another agency's computer. For example, delineating the goals of data collection before acquiring the software and hardware could minimize the use of several types of software (that may be incompatible) within a particular project or program. Therefore the technical aspects of database development could benefit from careful planning and coordination. To ensure such coordination in planning, designing, implementing, and maintaining databases would require high levels of institutional support,

CHAPTER 4 REFERENCES

1. Bell, J., "Data Modeling of Scientific Simulation Programs, " *Proceedings of the ACM SIGMOD [international Conference on Management of Data,* Orlando, FL, 1982.
2. Best, Robert, *Handbook on Remote Sensing in Fish and Wildlife Management* (Brookings, SD: Remote Sensing Institute, South Dakota State University, 1983).
3. Carneggie, D. M., Schrumpf, B. J., and Mouat, D.A. (eds.),"Rangeland Applications, " *Manual of Remote Sensing,* 2d cd., R.N. Colwell (cd.) (Falls Church, VA: American Society of Photogrammetry, 1983).
4. Gault, F, D., "Database Management Systems for Science and Technology," *Database Management in Science and Technology,* J.R. Rumble, Jr., and V.E. Hampel (eds.) (New York: North-Holland, 1984].

.5 Gurtz, M. E., "Development of a Research Data Management System ,'' *Research Data Managemen t in the Ecological Sciences,* Belle W, Baruch Library in Marine Science No, 16, William K. Michener (cd.] (Columbia, SC: University of South Carolina Press, 1986),

6. Hardy, E., "Remote Sensing for Land and Water Resource Management, " *Remote Sensing for Resource Management, C,* Johannsen, and J. Sanders (eds.) (Ankeny, IA: Soil Conservation Society of America, 1982).
7. Kennedy, E., and Kelly, M., "ADP Technological Perspectives of Biological Survey Systems, " *Proceedings of* Hearings *on a National Biological Survey* (Manhattan, KS: Association of Systematics Collection, 1986).
8. Klopsch, M. W., and Stafford, S. G., "The Status and Promise of Inter-site Computer Communi-

cation," *Research Data Management in the Ecological Sciences,* Belle W. Baruch Library in Marine Science No. 16, William K. Michener (cd,) (Columbia, SC: University of South Carolina Press, 1986).

9. Komarkova, V., and Bell, J. L., "Characteristics of Scientific Databases and Database Management Systems, " *Research Data Management in the Ecological Sciences,* Belle W. Baruch Library in Marine Science No. 16., William K. Michener (cd.) (Columbia, SC: University of South Carolina Press, 1986).
10. Marble, D., "Technical Aspects of Databases in Natural Systems," paper prepared for the Office of Technology Assessment, unpublished, 1985.
11 Newcomer, J. A., and Szajgin, J., "Evaluation of an Automated Field Data Entry System, " *In-Place Resource Inventories: Principles and Practices—Proceedings of National Workshop* (Orono, ME: University of Maine, 1981).
12. Stafford, S. G., Klopsch, M. W., Waddell, K. L., Slagle, R. L., and Alaback, P. B., "Optimizing the Computational Environment for Ecological Research, " *Research Data Management in the Ecological Sciences,* Belle W. Baruch Library in Marine Science No, 16, William K. Michener (cd.) (Columbia, SC: University of South Carolina Press, 1986.)
13. Stafford, S. G., Alaback, P. B., Waddell, K. L., and Alagie, R. L., "Data Management Procedures in Ecological Research," *Research Data Management in the Ecological Sciences,* Belle W. Baruch Library in Marine Science No. 16, William K. Michener (cd,) (Columbia, SC: University of South Carolina Press, 1986),
14, Town, W. G., Powell, J., and Huitson, P. R., "Recent Experience of the Application of a Commercial Data Base Management System (ADABAS) to a Scientific Data Bank (ECDIN),'' *Information Processing & Management* 16:91-108, 1980.
15. Zetka, E., "Coastal Zone Management Information Needs: Potential Landsat Applications, " *Remote Sensing for Resource Management, C.* Johannsen and J. Sanders (eds.) (Ankeny, IA: Soil Conservation Society of America, 1982).