

3. DATA MANAGEMENT SYSTEM

To implement the Galveston Bay Data Inventory System (GBDIS) portion of this study, it was necessary: to select software that matched the requirements of the system to be developed; to develop the structure of the database to accomplish the various types of retrievals desired; to write the code to permit data entry into the database not only by the Principal Investigators and their staff, but by the Galveston Bay National Estuary Program staff or others as well; and to write the code to permit data retrieval from the database using menus easily understandable to the lay public. These steps are described below.

3.1 Software

One product of this project is an extensive listing of available information on the Galveston Bay system. The character and treatment of this "information" base are discussed in the following section. It was formatted in a digitized data base, most efficiently accessed and manipulated with a data management system (DMS). At the outset of this work, it was necessary to select a DMS to form the software basis for the Data Inventory System.

There is a phenomenal selection in DMS products presently on the market. Further, the available software is reviewed regularly by various periodicals and ranked according to generally desirable properties, including speed of execution of basic operations such as reads and sorts, simplicity of execution, mathematical function capability, and data fields accessible. The most recent buyer's review available at the outset of the project (Badgett et al, 1989) lists 80 relational PC-based database management software packages. Many of the DMS product features are not relevant to the intended use in the Galveston Bay Data Inventory System (such as speed, since we do not anticipate the GBDIS as being time-bound), while others are very relevant, such as ease of learning and documentation, implicit in (9) above.

The approach to evaluation followed here was to enumerate the properties needed or desired for the specific application of the GBDIS. Thirteen criteria were formulated that the DMS must satisfy, some of which were required by the contract (1-5 below), and others were dictated by the anticipated characteristics of the entries and the probable requirements of the users.

1. The DMS must be electronic, microcomputer-based software, and retail below \$950.
2. The DMS must allow searching of the data base, based on key descriptors and/or fields related to the content of the data set descriptions.
3. The DMS must be sufficiently flexible to allow future updates with descriptions for new agency and project data, and future long-term monitoring data.

4. The DMS software should be standard and generally available for data base management.
5. It is desirable that the software be suitable for IBM-compatible equipment, available throughout state agencies. However, consideration should be given to the possibility of integration of the data base with COMPAS or related Macintosh systems.
6. The selected software should allow several (3-5, say) files to be accessed simultaneously.
7. The DMS software should be relational, i.e. permitting cross-comparisons between different files of information.
8. In addition to standardized retrievals, the DMS software should allow the construction of special-purpose retrievals, e.g., logical conjunction and disjunction of different key descriptors. Therefore, the DMS should possess a programmable capability. This should include the ability to construct custom menus.
9. The DMS should be user-oriented and have a high level of acceptance among PC users.
10. The DMS should allow variable-length records (to conserve storage space) and permit multiple index fields (for efficient retrieval).
11. The DMS should be capable of employment for other applications, and should allow easy interfacing with standard, readily available data processing software, e.g. spread sheets and statistical packages.
12. The information-retrieval system is expected to be a permanent, sustained entity, to be continuously updated and provided to various entities of the state and public requiring its use. Therefore, the selected DMS software should evidence a potential for longevity, i.e., to be supported and supplied by its manufacturer into the foreseeable future.
13. While not an immediate requirement of the data inventory system, the ability for networking will prove important in maintaining an updated data base at a central location that is accessible to users at other locations and in different agencies. Therefore, the DMS should include provision for networking.

Some discussion of these criteria is warranted. The first five, of course, are requirements delineated by the Management Conference and made a requirement of this contract. (The cost limitation in the first criterion was been added by the Principal Investigators.) Criteria (3) and (4) together imply the need for longevity, but for emphasis this is stated separately as Criterion (12). The rapid changes in PC

capabilities and available software will render any software choice "obsolete" within a matter of months. However, so long as the basic software structure is maintained by the manufacturer and is "upward compatible," the effort in structuring the data base and its retrieval logic will not be lost. The quality of longevity is not trivial. Of the relational DMS software reviewed by Jacobsen (1984) five years ago, 45% are no longer marketed. Of the five packages given detailed consideration in the 1989 *Personal Computing* review (Badgett et al., 1989), two did not exist five years ago. Probably the best measure of longevity is the demonstration of the manufacturer to upgrade and maintain support for its basic DMS product. This clearly introduces a bias into the selection against newer companies (or companies newly entering the DMS market), but the importance of the Galveston Bay Data Inventory System is too great to risk on future viability of an unproved company.

Cost, it will be noted, was not a criterion *per se*, apart from the retail ceiling in (1) of \$ 950 (which was set because there seemed to be no logic in paying several thousand dollars for capabilities that could be acquired much less expensively). This is because the costs of various software packages satisfying the other criteria were in the range of \$ 300-900. Differentials in this range become miniscule compared to the expense of personnel time, and therefore do not comprise a decision variable (except in the case of two otherwise identical packages--which, in any event, did not apparently exist). However, economy of investment in a broader sense *is* a criterion and represents the motivation for (11), in that a DMS that is used solely for manipulation of the GBNEP Data Inventory is a poor investment if another, more flexible system could be used for other purposes as well.

One potential additional use of concern to the GBNEP is the manipulation of the digitized data sets themselves, which in many instances will be the next step of an investigator studying the Galveston Bay system. The requirement (11) that the DMS permit interfacing with spread-sheet software addresses this concern specifically. (There are DMS packages with graphics capability, but this is only a part of the numerical manipulation that a researcher may desire.) To a certain extent the ability to export and import ASCII files will satisfy this requirement, but the anticipated application would be greatly facilitated by direct export/import of a standard spread-sheet format, specifically Lotus 1-2-3. As we became involved with the input of information, this criterion proved to be additionally important to simplify keyboard entry of large data files.

Criterion (5) regarding transferability to the Macintosh environment can be accommodated by export/import of ASCII files, so this did not prove to be a discriminating property. The reference to COMPAS, the NOAA Coastal Ocean Management Planning and Assessment System, is somewhat misleading, since COMPAS is a system for the assimilation and display of estuarine data *per se*, not entries in a data inventory, which is the present objective. However, a researcher's next step upon locating data sources will be to manipulate this data, which might include importing into COMPAS. Moreover, some of the information in the GBDIS can be of value in COMPAS, e.g. the file of sampling station coordinates. The flexibility implicit in criteria (8), (9) and (11) will permit incorporation in a Macintosh-based system, including COMPAS. (Indeed, in the development of data base files, we

frequently performed data entry in a Macintosh environment then exported to dBase on the PC. The reverse is easily accomplished.)

These criteria were applied to the selection of the DMS for the Galveston Bay DIS. The first broad criterion applied is (5) that the system be PC-based, given the wide usage of IBM-type equipment in the Texas Water Commission and other concerned state agencies, and (7) that the system be relational (i.e., permitting access to more than one file simultaneously, and allowing logical linking, comparison and sorting of elements from separate files). The eighty products listed by Badgett et al. (1989) satisfy these constraints, and, while others may exist, we considered this list sufficiently exhaustive for the purposes of this selection. Note that criteria (2), (4), (6) and (trivially) (1) are immediately satisfied as well. Next, in order to pare this list to something more manageable, we applied (8), that the system be programmable, and (11), that the system specifically export/import Lotus 1-2-3 formatted files. The software systems satisfying these criteria are listed in Table 6. In this table are presented the extent to which the remaining criteria are satisfied by each DMS. It turned out that all candidates satisfy (10), so this criterion did not form a basis for discrimination. Application of the criterion of longevity (12) was simple but perhaps a bit brutal: was the software of sufficient significance five years ago to appear in the review of Jacobsen (1984)?

Two candidate systems emerged from this screening, viz. dBase IV and DataEase 4.01. The basic differences between these systems is that DataEase emphasizes user simplicity at the sacrifice of programming power, while dBase, though providing through its sophisticated programming capability more user-custom flexibility, is considered more difficult to learn and apply (see Sander, 1988, and Blackford et al., 1988, as well as Badgett et al., 1989). This seems to have been reputation of dBase products residual from the previous versions, e.g. dBase III and dBase III+. To a large extent, this weakness of dBase has been remedied in dBase IV through the addition of user-friendly menus.

It was our recommendation that, of these two, dBase IV be adopted as the Data Inventory software. This decision was based upon the additional observations that dBase IV seems to be more readily available, both in local retailers and in mail-order houses, and (more significantly), few, if any, of the TWC staff that rely upon database-manager software, use DataEase while about half of this staff use dBase. Additional advantages accruing to the choice of dBase were (1) dBase seems to be the preference for EPA data base systems, as typified by the Clean Lakes Clearinghouse DMS, (2) the contractor for the Gulf Initiative data inventory, Sverdrup Technology Inc. (1989), has selected dBase III+ for that system.

One of the final choices in Table 6, Paradox, has become quite popular with some of the TWC staff, equal or a close second to dBase. Therefore, some additional comment is warranted as to the elimination of this software from consideration. Examination of Table 6 shows that Paradox was screened out for failing the

Table 6: Dms Software Considered For Galveston Bay Data Inventory System

<i>Software</i>	<i>Criterion</i>			
	(3) flexible	(9) user- friendly	(12) longevity	(13) net- working
dBase IV	Y	Y	Y	Y
dBase III+	Y	N	Y	Y
Paradox 3.0	?	N	N	Y
DataEase 4.01	Y	Y	Y	Y
Informix-SQL	?	Y	N	Y
Knowledge-Man/2	?	Y	Y	Y
R:Base DOS	Y	N	Y	Y
Ramis/PC 3.0	?	?	N	Y
Oracle	Y	?	N	Y
Super-base 4	?	?	N	Y
Key/500	?	?	N	N
XDB-SQL	?	?	N	Y

criteria for longevity and for user-simplicity. As noted above, the former can certainly be faulted for bias against newer software; certainly the present prominence of Paradox in the market would suggest that this product has achieved as much claim to permanence as its competitors. However, we believe the requirement for longevity to be valid, even if difficult to formulate as an objective criterion; further, once an objective criterion has been stated, it should be applied uniformly and unilaterally. The fact is that Paradox fails that criterion. The failure of Paradox in user-simplicity is based upon the large suite of functional options, which are intimidating to a first user, and the complex of manuals that must be penetrated in order to apply Paradox. Certainly, this deficiency is eliminated if the user has access to instructional seminars, as is the case for the TWC at Austin. Clearly, however, this cannot be generally assumed for Data Inventory users.

Finally, we must observe that this entire selection process is specious. This is because the immediate anticipated requirements of the GBNEP Data Inventory System are modest and will not tax or exhaust the capabilities of any of these data management systems. Indeed, any of the systems of Table 6, and many more, would accommodate our immediate needs. Therefore, the choice of dBase IV is in fact rather arbitrary, being based upon extraneous concerns, such as networking capability, or surmountable "deficiencies," such as user-friendliness. This arbitrariness is compensated by the fact that this decision is not really locked in. Because of the import/export capability of dBase, as well as our recognition of the need for generality in structure, the GB Data Inventory System data base should be capable of later transfer to another DMS, should the GBNEP or TWC deem appropriate. Therefore, the use of dBase IV in this project can be viewed as an expedient to act as a basis for retrieval structuring and data input, but not an irreversible software commitment.

3.2 Data Inventory Structure

The design of the data inventory database structure was driven by several constraints. First, the contract called for certain items to be included in the database, namely:

1. Agency/institution name
2. Data file name
3. Agency data base manager or contact, with telephone number and mailing address
4. Data description paragraph including name of the program resulting in the data collection, and objectives and use of the data collection
5. Period of duration for the collection of the data
6. Description of any technical data collection procedures utilized, including:
 - a. data collection locations
 - b. sample frequency
 - c. methods and materials
 - d. sample preparation/preservation
 - e. laboratory procedures and methods
 - f. results (data) entry and editing methods

- g. data scrubbing/error trapping procedures
 - h. period of record, by parameter
 - i. parameters/information collected and units of measure
7. Complete technical specification for any computer storage media utilized for the data, including file formats with field layouts, software applications, and accessibility; for spatial data, aerial coverage, scale or resolution, digital vs. other forms of storage, units stored, and methods and coordinate types for location determinations.
 8. Citations of any publications which have used or reported the data
 9. Complete description of quality assurance and control for data collection, editing, and storage
 10. Other information specific to data set.

Second, even though the general user of the database would not need to be aware of the structure of the database, the types of user searches envisioned dictated some of the structure. For the person unfamiliar with the GBDIS system, searches were designed to include the following:

by agency performing the work;

by data file name;

by principal investigator(s);

by keywords descriptive of the data set or present in the title of papers or reports written using the data;

by duration (year or range of years);

by location (latitude and longitude, water quality segment, segment name, and other identifiers);

by parameter (physical, chemical, or biological);

by priority problem; and

by combinations of two or more descriptors in fields.

Third, the GBNEP or group that would update the database would need to be able to input data to the database without difficulty. And fourth, the more sophisticated user of the system may want to do more detailed searches of the database at the "dot prompt" and would need to be able to search the proper portion of the database without difficulty.

All of these items except Items 6b through 6i could be easily accommodated in a single database, which is characterized as the "general information" collected about each data set. Item 6a was included initially in the general database as it was anticipated that many searches would be on location of data. However, because the number of sampling stations used in some sampling efforts became extremely large,

it became necessary to create a separate file linking station numbers and locations to individual data sets. As the types of data to be gathered began to be examined in detail, it was clear that Items 6b through 6i would have to be subdivided into classes of data type and databases created for each type with a connecting reference to link each of these data type databases. This was a marked departure from what had been envisioned in the Work Plan, but resulted in the only workable solution to accomplish the goals of the GBNEP. For Item 6, the subdivisions of data type used were:

- Morphology
- Hydrography
- Hydrology
- Water Quality
- Sediment Quality
- Biological
- Public Health
- Pollutant Loading
- Sociologic
- Economic

and for each of these data types, two databases were created: an information database to include Items 6b through 6g and a data database to include Items 6h and 6i. These are described in more detail in Armstrong and Ward (1991).

One of the prime features of the GBNEP data-base structure is the use of multiple files, elements of which are "related" (i.e., logically identified for access and retrieval purposes). This is because different types of data have different properties (called "fields" in the data-base management patois), so their logic structure must be different. An element (i.e. entry) of the GBNEP data base is a "project", referring to a uniform, systematic, autonomous data-collection enterprise. A "project" might be a one-time collection of sediment samples, a one-year study of shrimp communities, or a routine collection of water quality data at regular intervals over many years. A project might concentrate upon one geographical region of the bay, or might involve samples throughout the bay. Retrieval is implemented by searching on field variables, perhaps constrained by user-specified relations, and by keyword textual searches of the title and abstract information in the data entries. This dual approach to retrieval allows both quantitative sorting of the information, as well as qualitative searching.

The information contained in the databases allow all the information called for in the contract list above to be entered as well as to allow for the various types of data to be entered, data such as:

point observations/measurements,
analog time series,
analog line series,
continuous or discrete areal delineations,
anecdotal.

Likewise, the sources for these data took several forms as follows (with examples of each):

open literature (books, journals)
grey literature (technical reports, project studies)
file documents (unpublished manuscripts, internal memoranda)
transient literature (newspapers, diaries, historical collections)
formalized data tabulations (publications, computer-readable media)
organized data archives (indexed maps, aerial photos)
raw data (field sheets, strip charts, cassette logs)

and could also be entered. For those entries from the open literature, a citation in the usual scientific format sufficed to uniquely identify the source of data and permitted a researcher to access the source. Thus, field elements obviously included author, title, and journal or book bibliographic identifiers. Any of the above properties will constitute a retrieval parameter (or field), e.g., "author", "aerial photograph" or "tide record". Other retrieval fields would include types of information or measurement, such as "salinity", "*Callinectes* sp", or "water depth". Additional retrieval parameters, incorporated as fields for direct retrievals or descriptors for textual searches, include: specific chemical measurements; geographic location; date (of sample or of fundamental information in citation); date of publication

A GBDIS response query to the specific field parameters activated during the data entry operation has been developed, offering a selection of qualifying information to the entry clerk. This is the means of entering information on the type of measurement or analysis employed, the Q/A procedures, and so on. As an example, a retrieval on "salinity" should produce all of the parameters conventionally used as measures of salinity, including conductivity, chlorinity, and density (hydrometer). Fortunately, there are a small number of potential methodologies applicable for a given variable, so it was feasible to build up a file (or files) of these as part of the data system structure. Like the ADS files described below, these files will be "transparent" to the user. Their contents would be accumulated during the initial data base formulation and entries; we anticipate that after a short period, these files would become essentially static, and would serve from then on to prompt the data entry technician for more detail. The philosophy is that at this point in the process--as the data entry is made from the primary source--it is easiest to search out and input such relevant details as Q/A and analytical methodologies.

The linkage among all the database files is a unique reference number assigned to each data set. Thus, once a data set is identified as the one from which information is desired, then information from any of the associated database files may be retrieved.

3.3 Data Entry

Data entry into the inventory database is facilitated by a program written in dBase with menus and data forms to allow the user to select the type of data to be entered (i.e., morphological, water quality, etc.) then to enter that data via a form on the screen into the database. Some error checking is done during the data entry process (to avoid duplication of reference numbers for example) and helps are provided in terms of lists of parameter names corresponding to those used by the TWC and EPA. Again, the data entry program and procedures are described in Armstrong and Ward (1991).

3.4 Data Retrieval

Data retrieval from the GBNEP Data Inventory System is achieved through a program written in dBase using menus and help screens so that the lay user as well as the experienced user may search the database in a number of ways. As noted above, the types of retrievals now possible using the system are:

1. by federal, state, or local government agency or private corporation performing the work using an acronym for the agency or corporation;
2. by data file or program name (i.e., maintenance dredging);
3. by principal investigator(s);
4. by keywords descriptive of the data set or present in the title of papers or reports written using the data;
5. by duration (month/day/year or period from one month/day/year to another);
6. by location (latitude and longitude, water quality segment, segment name, and other identifiers);
7. by parameter (physical, chemical, or biological);
8. by GBNEP priority problem; and
9. by combinations of the above.

With such search capabilities, for example, one can determine all the studies inventoried which were conducted by a federal, state, or local governmental agency or other groups (search 1) or those studies carried out by that agency in a specific period of time and/or location (search 9). It will also be possible to locate specific data sets, for example, the previous Galveston Bay Study data set (search 2). If it is desired to know the studies performed by a particular report author (search 3) conducted in particular parts of the bay system at particular points in time (search 9), that can be done. Searches for studies in which particular types of water quality constituents and biological components and processes were sampled again in space and time can be done. Finally, searches for studies with information pertaining to particular GBNEP priority problems are possible.

One of the more important retrieval fields is that specifying the locations within the system at which the data observations/measurements were made. One anticipated use for the data inventory was retrieval of data of a specified type pertaining to a specific geographical subarea of Galveston Bay. After much consideration and review, latitude-longitude coordinates were adopted as the basic position specification. This decision entailed a considerable effort in the data entry process; because relatively few data sets had the measurement positions specified by latitude and longitude, it was necessary to map these points and determine the coordinates ourselves. However, the generality and flexibility of this approach justified its employment. Further, latitude-longitude coordinates and geographical descriptors can be cross-referenced, thereby facilitating searches given only the geographical name of a feature in the Bay.

3.5 Computer Hardware Required

The Galveston Bay Data Inventory System at the University of Texas is implemented on a dedicated 386-based microcomputer of PC architecture, operating at 20 kHz, and equipped with high-density disk drives and an 80 MB hard drive. The actual system, in its present form, requires some 10 MB of hard drive storage including the 2.5 MB needed for dBase IV software, so the system can be accommodated on a more modest machine. The size of the data files and the complexity of logical searching do, we believe, mandate a short cycle time. We recommend therefore that the system be installed on at least a 286-based machine (i.e., AT equivalent) with at least 20 MB hard drive. Clearly, if the machine is to be used for any other purposes requiring hard drive access, then a larger capacity drive may be necessary.

The original project scope assumed that any user of the Galveston Bay Data Inventory System would separately purchase dBase IV software. Upon reconsideration, we have invested in the Developers Version of dBase IV, which allows the production of compiled, executable codes that obviate separate software (and will free some of the hard drive storage as well). Therefore, a potential user no longer needs a separate purchase of dBase IV.