GALVESTON BAY DATA INVENTORY

Principal Investigators:

Neal E. Armstrong
George H. Ward
The University of Texas at Austin

EXECUTIVE SUMMARY

In the past, a wealth of data has been collected from the Galveston Bay system, relating to the movement and quality of water, the biology of the bay, navigation, socioeconomics and fisheries, some of this information dating back more than a century. Therefore one of the early tasks in the Galveston Bay National Estuary Program (GBNEP) was to locate and inventory these data. The specific objectives of this project were to: (1) survey local, state, and federal agencies and other organizations for data sets; (2) prepare in a standard format written descriptions of data sets; and (3) compile the data inventory in an electronic, searchable, microcomputer-based information base. The product of the project should enable a researcher with a specific data requirement to determine where (and whether) the historical data exist and how to access them.

At the outset of this work, a data management system (DMS) had to be chosen to form the software basis for the inventory. There is a phenomenal selection in DMS products presently on the market. The most recent buyer's review available in 1989 listed 80 PC-based relational database management software packages. To sharpen the choice, thirteen criteria were formulated that the GBNEP DMS must satisfy (five of which had been specified by the GBNEP management and were required by the contract), dictated by the anticipated characteristics of the entries and the probable requirements of the users. Ultimately, dBase IV was adopted as the data inventory software.

The data resources for the Galveston Bay system take many forms, including point observations (such as grab samples), time series (streamflow records, tide scrolls), line series (cross-sectional profiles, scanner imagery), areal delineations (maps, aerial photography), anecdotals (event descriptions), regional statistics (bird rookeries, population profiles, economic activity). This variety of forms demanded a considerable flexibility--and therefore complexity--in the electronic Galveston Bay Data Inventory System (GBDIS). One of the major features of the GBDIS data-base structure is the use of multiple files, elements of which are "related" (i.e., logically identified for access and retrieval purposes). Because different types of data have different properties, their logic structures are different. Retrieval is accomplished 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.

One of the more important retrieval fields is that specifying the locations within the system at which the data observations/measurements were made. We anticipate one use for the data inventory to be retrieval of data of a specified type pertaining to a specific region of Galveston Bay. 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 have 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 we believe justifies its employment.

There are many sources for data on Galveston Bay, including open literature, grey literature, file documents, transient literature, formalized data tabulations, organized data archives, and raw data. The task of location proceeded simultaneously on several fronts: review of bibliographies and indexes; *direct* review of journals and reports; visits and contacts with likely sources. All of this work was carried out by the project principal investigators, personally; no student help was employed. It is important to differentiate this project, whose objective was to inventory extant data from Galveston Bay, with the companion GBNEP project, to compile information on the Bay. The former (i.e., this project) focuses upon raw measurements, while the latter focuses upon the technical literature.

One of the major classes of data sources is the unpublished holdings of agencies and individual researchers. The approach to this class of data was stepwise, starting with inquiry letters and proceeding to direct contact; visits by the PI's to inspect and assess holdings; completion of the inventory, assessment of data perishability and acquisition of copies where appropriate. For the key state and federal agencies (most of which are participants in the GBNEP), the strategy (proposed by the GBNEP) was to identify a point-of-contact in that agency who would facilitate the location of data holdings and make the necessary internal arrangements for the PI's to visit and inventory the data. Individual researchers posed a greater problem, in that there were many more of them, individually with smaller data sets, difficult to locate and contact, and frequently uncooperative.

In summary, the project proved to be far more complicated and time-consuming than originally envisioned. Several factors contributed to this:

(1) There proved to be a large number of data sources for Galveston Bay, but only a minority could be described as major projects (e.g., the TWC Statewide Monitoring Network, the Galveston Bay Project, the TWDB Bays & Estuaries Program, etc.), i.e. the data resource can be described as a few large projects and a great many small projects, which served to multiply contact time and logistics;

(2) The point-of-contact approach failed, requiring much greater time and effort of the PI's to find and gain access to agency data holdings;

(3) In general, the response of the data sources to our inquiries has been poor, necessitating multiple letters or calls, and requiring months (at best)

to finally gain access to data. In 1991, after the project was technically over, many were only then responding.

However, the dominant reason is that the management of older data--and by this we mean any data taken prior to 1980--is by-and-large a shambles.

The principal conclusions regarding the data resource for Galveston Bay drawn from the experience of this project are:

(1) Most of the data sets for Galveston Bay taken prior to 1980 are presently unavailable. The majority of this data appears to be irrevocably lost.

(2) When one considers that the data prior to 1980 comprises the vast majority of data taken in Galveston Bay ever, this implies that most of the data resource has vanished.

(3) The factors which have led to this loss of data are still operating today.

These conclusions apply primarily to data on the biological, water quality and hydrographic features of the system, which are the most important insofar as the GBNEP objectives are concerned, however they probably apply to other categories of data as well.

These conclusions of course must be qualified for specificity. For example, sediment quality data is of more recent concern, and has benefited from advances in analytical technology, so is in relatively good shape. Also, specific data collections with national archival procedures are well-managed, e.g. the historical mapping of the National Ocean Service and its predecessor agencies, and the data collection efforts of the U.S. Geological Survey. On the other hand, for many major and fairly recent data collection projects implemented by federal, state and regional agencies, the data are *totally missing*. Additionally, data sets, which had been entered on digital media, now only exist as one or a few *hard-copy* tabulations. Thus the utility of the data is severely truncated, and the effort invested in putting the data in a utilitarian format is lost. The situation is worse for research data of individuals.

The factors that contribute to this data loss include:

(1) problem-specific operation of most agencies, and the valuation of older information as "obsolete";

(2) low priority assigned to archiving and preservation of older data, and the general perception of archiving of information as an unwarranted expense;

(3) personnel turnover in the agencies combined with little or no documentation;

(4) agency instability, i.e. dissolution, merging and reorganization of an agency, as well as frequent displacement and relocation;

(5) natural calamities (fires, floods, hurricanes) in poorly protected housing;

(6) changes in data management technology, without upgrading of historical files;

(7) proprietary attitude toward data by individual investigators.

All of these are mutually exacerbating. All of these are continuing to operate and permit continued loss of data. In our view, the problem is critical.