DOCUMENT RESUME

ED 230 184 IR 010 705

AUTHOR	Evans, Evan C., III
TITLE	Establishing Data-Exchange Networks Through Data Management & Telecommunications.
INSTITUTION	Naval Ocean Systems Center, Kaneohe Bay, HI.
PUB DATE	15 Jan 83
NOTE	25p.
PUB TYPE	Reports - Descriptive (141) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Conservation (Environment); *Databases; Data Processing; Design Requirements; Environmental Standards; *Information Networks; Information Retrieval; Information Storage; *Man Machine Systems; *Natural Resources; Online Systems; Scientists; Telecommunications
IDENTIFIERS	*Database Management Systems; Navy; Resource Management

ABSTRACT

This paper describes several pilot systems of data management using telecommunications links, which have been tested by the Navy during an 8-year period in which emphasis has been on the development of relational database management systems, exchange protocols, and man-machine interface. An introduction discusses the background of the project, which began as an attempt to computerize natural resource and environmental survey data for Navy-controlled United States land. The three prototype management systems described were developed because of the multidisciplinary character of the data and the diversity of the data uses. The fundamental problems of taxonomy, habitual procedures, and reliability are addressed. Emphasis is on the natural scientist as a computerized-system user, the user interface, and data exchange applications. An expanded database management system currently under development is also briefly described. (LMM)

Establishing Data-Exchange Networks through Data Management & Telecommunications

Dr Evan C Evans III
Naval Ocean Systems Center/Hawaii Lab
Kaneohe Bay, Hawaii 96863

15JAN83

As the custodian of nearly 3½ million acres of land in the United States & as responsible tenant on ¼ million additional acres in many foreign countries, the U.S. Navy has an important requirement for efficient natural resource management & for environmental quality control. The immense geographic spread of these areas & the need for long-term time-series comparisons in both natural resource & environmental management dictate an efficient means of data storage, manipulation, & exchange. In consequence, the U.S. Navy has tested several systems of data management & data exchange using telecommunication links. Special emphasis has been placed on the development of relational database management systems, on exchange protocols, & on the man/machine interface. A thorough understanding of this interface & of the practical applications required by the user are paramount to the success of any data-exchange network. This paper describes several pilot systems which have been tested over the last eight years. The fundamental problems of taxonomy, habitual procedures, & reliability are addressed. The emphasis is on the user interface & on the applications that efficient data-exchange makes possible. An expanded database management system, currently under development, is also briefly described.

Establishing Data-Exchange Networks through
Data Management & Telecommunications

Dr. Evan C. Evans III
Naval Ocean Systems Center/Hawaii Lab
Kaneohe Bay, Hawaii 96863

Introduction

As the operator of ships, submarines, aircraft, and landbased facilities on a global scale, the U.S. Navy clearly has a requirement for sophisticated, efficient data management and for data exchange through telecommunication. The 8-year project described here began as an attempt to computerize natural resources and environmental survey data for the 3½ million acres of land controlled by the Navy within the United States (Hura, 1976; Evans, 1977a). The multidisciplinary character of these data and the extreme diversity (both in operational requirements and in geographic location) of the data users forced the development of a generalized, relational data management system. Since most expertise in the natural sciences is found on university campuses, in museums, or in organizations (both public and private) outside the Navy, the data management system that evolved was expressly tailored to facilitate strong interaction with these "outside" sources. An important aspect of this project has been an overt attempt to entrain individual users and their observations into the system through the excellence and affordability of the data management service provided. The generality of the relational data management systems so far developed has permitted their effective use in many other fields, such as meteorology, microelectronic component properties, technology transfer, conference administration, chemical oceanography. Three prototype data management systems have been developed and tested, the last of which (R*B-2.4) is currently operational. The project is continuing with the development of R*B-3, the first full-function data management system, expected to become operational in 1985.

The Pilot Systems

The current project evolved out of a Navy biological survey of Pearl Harbor, Hawaii (Evans, 1974). Computer analysis of this survey data showed that similar "ship signatures" could be detected in Hawaiian and a number of west coast harbors, and showed further that such analysis applied to the observations of others could reveal biotope patterns not recognized by the original observers (Evans, 1977b). These discoveries led to a search for other harbor survey data that might corroborate these findings. At that time, Navy data was archived in the University of Hawaii's Hawaii Coastal Zone Data Bank (HCZDB), a file management system using PANVALET. While the HCZDB was adequate for those familiar with its contents, its lack of a data management system rendered it quite unsuitable as a generalized database that could be shared. Since 1975 the Naval Ocean Systems Center and the Computer Sciences Corporation (under contract), have collaborated in developing the type of data management system required by the Navy. This effort commenced with an evaluation of data management systems extant in the 1975-77 time frame to find one suitable for the kinds and amounts of data being collected.

Many data management systems (among them ENVIR, TAXIR, Bio-STORET, System-2000, UPGRADE, DMS-1100) were examined. None could accommodate in an adequate and affordable manner the wide range of multidisciplinary measurements characteristic of environmental survey data. At that time, no relational data management system existed, although two (System-R and INGRES) were in the early stages of development. Furthermore, most data obtained from other harbor surveys could not be used because of inadequate supporting information. The latter situation leads to the oft heard statement: other people's data are no damn good. This statement is inaccurate. Verified scientific measurements have lasting value if they can be marshalled for the right application with a full set of supporting information. Usually it is the absence of necessary supporting information that disqualifies otherwise useful data obtained from outside sources. All our findings substantiated a definite need to develop a data management system that could be shared with equal facility by different scientific disciplines. Hierarchical data management systems were obviously inadequate for such multidisciplinary application. Thus, the decision was made in 1977 to follow the relational theory recently advanced by E. F. Codd (Codd, 1970).

At that time, the penalties for selecting a relational approach loomed large. Chief among these were the sequential search requirement and the repetition of ancillary or supporting information in each tuple (record). Proof of high search-rate capability and of effective data compression was paramount to the success of the relational approach. From the beginning our prime goal was the management of very large data bases at a cost that was affordable to universities and museums, the principal sources of verifiable environmental observations. To assure that all necessary supporting information was correctly associated with an observation regardless of the discipline or circumstances under which it was made, we adopted the concept of a data template, see Figure 1. This template, developed in 1976 and still in use, has been tested against many different types of observations (scientific and otherwise). It has proven entirely adequate for our data management applications. The first data management system, called BIODAB for BIOlogical DAta-Base, became operational in APR78 (Key, 1979). It was built to determine three things:

* the adequacy of the data template as a discipline-independent vehicle for scientific observations
* the recovery times for complex searches directly on data stored in relational format
* the data compression obtainable using various coding or linkage techniques.

The results of the BIODAB test were positive on all three scores. As said above the data template proved wholly adequate. Rates of 300,000 records per CPU-second* were obtained for complex searches directly on data. BIODAB tuples could be compressed from 56 words to less than 6, see Figure 2. The high search rates were obtained by means of the masked search instruction available on UNIVAC machines. While instructions emulating the UNIVAC masked search can be written for other mainframe computers, our data management systems continue to be specific for the UNIVAC 1100 series. The philosophy of the project is to run a given system on the machine that is optimal for the processes involved and to bring the user into contact with that machine through telecommunication.

BIODAB was tested for two years and then retired. During that test-period, its better features were incorporated modularly into the first of the RELATABASE or R*B systems, see Figure 3. As indicated in this schematic, the development of all follow-on data management systems was driven by strong interaction with the user community employing an existing prototype in real job situations. This interactive aspect of system development is essential to the success of any data management system. The user community must be created simultaneously with the data management system itself. Note also that the same data management system was used for several quite different data bases (the Oceans '79 Conference database, the Integrated Circuit DataBase, the Natural Resources DataBase). Further discussion of meaningful involvement with the user community, the man/machine interface, and data structure follows in the next section.

Because the R*B systems were developed primarily for archiving, manipulating, and sharing or exchanging numeric data, each included a statistical processor that permitted interactive data analysis in the sense of John Tukey (Tukey, 1977). This processor would permit any user employing the numeric observations of another to probe or shape his newly-acquired file through interactive analysis before applying more sophisticated statistical treatments, like factor analysis. The importance of such probing has been stressed by J. Stuart Hunter (Hunter, 1980). BIODAB contained a partial implementation of Don McNeil's interactive data analysis programs (McNeil, 1977). Such simple displays as stemleafs, boxplots, scatter plots, and regressions to the third power were possible. Follow-on data management systems (R*B-1 and R*B-2) improved or enhanced these interactive capabilities. Futhermore, BIODAB was not a strictly relational system. Its 18-character taxonomic code (see Figure 4), while fully capable of accommodating Latin names, common names, and synonymy for all living organisms, was hierarchical, a format not permitted in relational systems. R*B-1 development involved two major efforts, viz:

* taking all useful features of the BIODAB design and further generalizing them so that they were strictly relational, and
* designing and implementing means whereby any user could create his own database (BIODAB did not have this capability).

---

*CPU-second = Central Processing Unit-second or machine-second.

RELATABASE, version 1 or R*B-1, development included the design and implementation of processors to permit a user to:

* define his own database,
* insert records into the database so defined,
* remove selected records from that database,
* update selected records in that database, and
* unload or move part or all of that database to another file.

In addition, an editing capability was added to the search and report-writer processors. The report-writer was also enhanced by the addition of sorting and listing options. R*B-1 became operational in JUN79.

Several engineering groups were attracted to the R*B-1 system with the result that the R*B project lost its predominantly environmental cast. The initial environmental slant had, however, served a definite purpose. To test a generalized data management system, one must have both complex multidisciplinary problems and a good supply of different but fairly well organized data sets. Many scientific disciplines have complex data management problems, but often available sets of organized data tend to be lacking. Environmental studies and surveys offered taxonomic complexity, convoluted and overlapping geographic and jurisdictional boundaries, and "constants*" that change as a function of location. The engineers soon found certain enhancements to R*B-1 to be highly desirable. They were accommodated by a series of modifications culminating in R*B-1.4, while a full revision, R*B-2, was being implemented. The enhancements available in R*B-2 were:

* optimized search routines to achieve higher search speeds,
* surrogate link values making record insertion easier and cheaper,
* use monitors to collect operating data on various R*B processors and to provide more detailed cost breakdowns,
* a text attribute so that text or long comments could be stored,
* a list directive so that new or intermittent users could refresh their memories on the contents of any relation,
* a menu option to prompt new or intermittent users inputting data,
* real number representation (not implemented in R*B-1.4)
* further improvements to the stats processor, such as adding intrinsic functions and an equation processor.

R*B-2 became fully operational in APR81; the current modification R*B2.4 was released AUG82. Search rates in this version were clocked at between 500,000 and 800,000 records per CPU second. Details of the sys-

---

* For example, the bald eagle is endangered or protected or both depending on its location. Its classification can change as it flies across state or county lines. Classification also depends on whether the bird is considered as a species or as a raptor. This curious sort of variation, resulting from different laws and their interpretation, represents a problem for the Navy as well as a real challenge to the designer of data management systems.

tem are described elsewhere (Key, 1979; NOSC, 1982a; NOSC, 1982b). Briefly, any R*B user can create (define), maintain (update, insert data, remove data), and unload individual databases. R*B also maintains individual user files which can be displayed, described, labeled, or deleted. Any major relation in any master database can be searched for specific values and the material so retrieved stored in a file assigned to the individual user. A report-writer (permitting a wide range of format specifications) and a statistical processor is also available. R*B also has provisions for self-tutorial help and for sending messages or bulletins. All versions of R*B currently operational are considered prototypes. A full-function data management system R*B-3, discussed in the final section, is currently in the definition phase of development.

Since a data management system, of itself, contains no data, an application of R*B-2.4 to the Natural Resources DataBase (NRDB) is briefly described. The NRDB was established to manage the records of the Navy's natural resource managers, who are widely distributed among many Navy facilities throughout the continental United States and Hawaii. Their concerns involve 3½ million acres of land (including 96 thousand acres of ponds, streams, and wetlands, and 80 thousand acres of forest in timber production) and around 2 million civilian guests per year, who hunt or hike or perform scientific studies on Navy land. Their records include, but are not limited to, such disciplines as: agriculture, archaeology, biological survey, chemistry, cultural registration and restoration, endangered species protection, erosion control, forestry, historic preservation, hydrology, geophysics, grazing regulation, land use, management plan development, meteorology, outlease inspection, pollution monitoring and prevention, recreation control and development, resource management, soil analysis, timber surveys, vegetation mapping, well logging, and wildlife management. The many individuals in the work force employ different methodologies, data formats, and filing systems. Certainly, the application is a challenge to any data management system. The NRDB is comprised of four major relations (tables), viz: OBSERVATION, CLASSIFY, USAGE, and EVENT, and of five support relations, viz: SOURCES, CONTACTS, TAXON, METHODS, and GLOSSARY. The details of all these relations and lists of attribute values contained in any of them are stored in the system and may be called for at any time. An overview of data types in the NRDB is given in Figure 5. Currently, the NRDB contains about 1 million records, each containing many items (a mean of 41 for the major relations and of 17 for the support relations). Its size is doubling annually and is expected to approach 6 million records by the time R*B-3 becomes operational.

User Interaction

As mentioned above, strong interaction with a user community is an essential aspect of the development of effective data management systems. Since its inception, user interaction has been an important part of this project; quantitative study of user activity, however, commenced in 1979. Many things are involved, including telecommunications, network protocols, reliability, man/machine interfaces, user behavior, natural language, data structure. Only a few of these subjects can be touched on here. Fortunately, good reports of user interaction exist (Hiltz & Turoff, 1978; Vallee, 1978; Johansen 1978). Hiltz and Turoff's excellent summary of what system designers must expect should always be borne in mind. Users will:

* fail to notice even the most explicit instructions
* do the unexpected, the unanticipated, and the forbidden
* disregard or forget instuctions
* often fail to ask for help when they need it
* form opinions based on inadequate knowledge
* use the system only if it benefits them.

Hiltz & Turoff (p 61) emphasize the crucial importance of a user-oriented monitor, providing in-person or telephone training and serving as a point of contact with system designers or operators. They also describe (pp 46-61) the animosity of established ADP* groups to the development of new computerized systems. This project has had exactly these same experiences. The importance of one (or more) full-time, user-oriented monitors cannot be overemphasized.

Hiltz and Turoff's observations are confined to computerized conferencing systems which do not involve the sophisticated management of scientific data. Our experience overlaps theirs in the areas of user support and in electronic mail, the latter being used in conjunction with but not as part of the R*B development project. The R*B systems interface with ARPANET, a packet switching network implemented by Bolt Beranek & Newman for the Advanced Research Projects Agency in 1969, see Figure 6. The electronic mail and file transfer protocols associated with ARPANET were used extensively. Experience using these systems as well as the R*B systems is here summarized. The emphasis is on the natural scientist as a user of computerized systems. As shown in Figure 7, there is a wide spead in amount of individual use. Of the NRDB user community, about 80% fell into the light-to-occasional category. These users tended to disappear unless they were expressly cultivated by R*B monitoring personnel. The reasons for their disappearance were various, but prominent among them were dislike or fear of computers or failure to appreciate the utility of computerization in their work. The remainder of the R*B community was divided into moderate users (15%) and heavy users (5%). About half the moderate user category tended to move upward into the heavy user category.

Often scientists tend to be curiously ambivalent with respect to their own data in that they regarded them as both worthless and highly proprietary. This behavior is the result of fear of preemption or misuse combined with the fact that data are regarded as the raw material

* ADP = Automatic Data Processing; also EDP = Electronic DP.

which ultimately supports publication. It is difficult therefore to get scientists either to share their data or to store same in a rigorously accountable manner. We estimate that about 90% to 95% of all basic scientific observations are so poorly archived as to be essentially lost. This follows from the fact that scientists are trained to extract information from data, not to husband data after the manner of accountants. The NRDB with its data template can, therefore, be regarded as an educational tool. Monitoring observations showed that continued use of the NRDB improved both field and laboratory procedure in the sense of thorough and more accountable note-taking. Since verified basic measurements (as opposed to the reduced data published) tend to have a high degree of commonality and to retain their value indefinitely, any procedure that archives data in exchangable form is decidedly cost-effective. This is especially true in the environmental sciences where long-term time-series analyses are required to detect subtle changes.

While scientists' customary behavior usually results in massive data loss, other important and unreported data sinks are to be found in the military. The 3-year tour of duty with its associated name/code changes for groups, commands, projects, buildings, bases, &c adds up to a thumping loss of corporate memory. The penchant for acronyms does not help. Often the basic measurements are still on file but the supporting information necessary for their use has been lost. On the basis of our experience in sequestering data from various sources, we estimate that the half-life of basic measurements is less than 3 years in the military, between 7 and 10 years in the private sector, and between 20 and 30 years on university campuses. J. Stuart Hunter quotes a National Bureau of Standards estimate that in 1977 the U.S. government spent $690 million for data gathering (Hunter, 1980). The cash value of these data losses can, therefore, be inferred to be significant. Verifiable basic measurements are in themselves a valuable resource and should be conserved. The applications supported by the R*B systems are expressly designed for that purpose. During the life of this project, the cost of computerized data storage has become far less expensive than any other means. With appropriately designed data management systems, access to data so stored becomes flexible, efficient, and affordable.

The man/machine interface continues to receive insufficient attention. Obviously the person who can compose at a standard QWERTY keyboard has a monumental advantage over those who cannot. The proliferation of non-standard additions to that keyboard displays more of the American penchant for packaging than of a coordinated approach to user needs. These problems, while admittedly beyond the purview of a project to develop generalized data management software, are nonetheless felt as we canvass our users. Data-linking reliability is a second problem in this beyond-our-control category, and one that has been so severe as nearly to cause the demise of the project. As stated above, strong interaction with the user community is paramount, not only as a requirements source for system designers but also as a means of developing a user community while the data management system itself is being developed. With the wide-spread aversion to computers, particularly apparent among the natural scientists, low data-linking reliability or long down times is the primary cause of user loss mentioned above. The probability of a remote user being able to access his data is the product of the reliabilities of at least three systems which have nothing whatsoever to do with the data management system itself. At the least, these are: the telephone link, the ARPANET link, and the host computer. During a six-month monitoring period in 1981, these reliabilities were estimated to be 0.79, 0.91, and 0.58 respectively, for a product of 0.42. In short, the data-linker could be assured of reaching his/her data slightly less than once in two tries. This is an admittedly worst case situation in that we were obliged to use very noisy telephone lines and also our host computer (one that interfaces ARPANET, not the UNIVAC where the data was housed) was a severely overloaded machine. To put these figures in better perspective, the probability of reaching the correct person on the first telephone call should be considered, viz: 0.26 success, 0.10 busy, 0.28 no answer, 0.28 wrong person, 0.07 misdial or other problem (Wedemeyer, 1980). The point here is that most users are very tolerant of the telephone without realizing it, whereas they tend to be extremely intolerant of computerized systems. It should be added that many NRDB users are hardwired into the UNIVAC and therefore enjoy high access reliability.

Our data management philosophy of bypassing "portability" and bringing the user into contact with the mainframe computer that can best do the job desired requires that these data-linking problems be solved. ARPANET's recent (01JAN83) switch from NCP* to TCP* and a significant upgrading of our host computer has greatly improved matters. Current probabilities are estimated at 0.79, 0.95, and 0.95 respectively, for a product of 0.71. This still leaves room for improvement. The quality of telephone lines and the manner in which the itinerant data-linker is handled by the telephone companies also needs improvement. The problems of the itinerant data-linker, the one moving about the country carrying a portable terminal, seem to be largely neglected by the telephone companies. Dialing protocols change with variations from rotary to touchtone instruments and change more confusingly as one moves from regional exchange to regional exchange. Directions cannot be conveniently found in the telephone directories, nor can they be obtained from the operators, who are trained to give only a limited set of responses. These are minor, but nonetheless real, problems which currently cause the itinerant data-linker severe heartburn. The switch from NCP to TCP suggests, however, that interactive data systems, portable terminals, and the like are at last coming into their inheritance. Thus, the difficulties enumerated here should shortly be resolved.

* NCP = Network Control Protocol; TCP = Terminal Control Protocol.

There are, however, more severe problems, deserving more attention than currently seems to be lavished upon them. The business of manual writing and machine-tutorial composition needs all the attention it can get. Arthur Naiman's Introduction to WordStar™ is an example of progress in this direction (Naiman, 1982). There seems to be a need for 3-color printing so manuals can distinguish unequivocally user-input, machine-output, and comments concerning the first two items.

While the project gives careful attention to the preparation of manuals and machine-tutorials, the matter of data structure is more clearly in the province of NRDB support. This is a complex, difficult, and often neglected field which is essential to the establishment of practical and efficient databases. A thoughful inspection of the U.S. Library of Congress' call numbers for botanic monographs preserves in stone, as it were, the pitfalls of insufficient consideration of data structure. We do not claim now to have finalized data structure for the NRDB. A few examples, however, are provided to illustrate the problem and our approach to same. NRDB users frequently consider complexes of actions to be separate entities. Consider the complex:

consultation/conference/meeting/briefing/congress/seminar/workshop

or another such:

inspection/inventory/survey/tour/observation-set/reconnaisance.

Certainly, there are differences between the elements of these complexes, but are the differences sufficient to require separate treatment in defining a relation? In our estimation, there is roughly 80% functional similarity between the elements within each exemplary complex. We have attempted to use the concepts of natural language (Sager, 1981) and of the selection properties of words (Bloomfield, 1933) to assist us with these problems. However, careful study of user work habits and continuous dialogue between the user and the user-oriented monitor appear to be the most efficient means of solution. The situation is part of a larger problem which is central to the success of any database, viz. an efficient and rigorous taxonomy.

The taxonomic codes employed by BIODAB worked beautifully, but they were hierarchical and therefore not admissable into a generalized relational system. All efficient formal taxonomies are hierarchical and the problem of mapping such a system into relational format is not a simple one. The current TAXON relation in NRDB uses the Linnaean binomial/trinomial system since it has withstood the tests and trials of over 200 years. The system, however, is confounded by the fact that botanic usage (Int. Code, 1975) and zoologic usage (Int. Code, 1961) employ the same names for different levels in the hierarchy. Worse, the same discipline will use the same name at two different levels! Obviously, such practice cannot be tolerated in a computerized system. The solution currently employed in TAXON is shown in Figure 8. The use of flags is regretable but necessary. A better solution is still being sought during R*B-3 development. Our current solution is somewhat mollified by the fact that R*B-3 will support customized applications that maintain user-profiles (one user may have several aliases). One or more of these profiles can automatically set the taxonomic flags customarily employed by a given user when he/she logs into the system. Other dictionary or menu solutions are also possible, but thus far the difficulty remains. Non-biologic taxonomies are also hierarchical, thus our TAXON solution can be applied to them as well. An example using a formal taxonomy for man-made objects (Chenhall, 1978) is provided in Figure 9. Please note that cladistic or evolutionary significance is emphatically not implied by these arrangements. They are erected simply for the orderly accommodation of a wide range of entities in a computerized database.

Costs often loom large in administrators eyes when computerized databases are proposed. More often than not, these administrators are still thinking in terms of the industrial age, as opposed to the information age (Giuliano, 1982). They view information handling as essentially non-productive work and data husbandry as a serendipitous pastime rather than as a logical response to a valuable resource. For the last decade, the cost of personnel has risen at about 10% per year while that of computers and their usage continues to fall at about 25% per year. Already the cost of computer storage has fallen to less than 1/100th that of paper; similar savings are realized on document reproduction. The costs of NRDB support using R*B-2.4 are, of course, monitored. Data obtained during the fall of 1982 are as follows:

- electronic mail — $6/hour and falling
- computer usage — $20-$60/hour (depending on the complexity of the task attempted) and falling
- data preparation — $0.25-$10/record (depending on the state of the raw data and on the complexity of verification) - this is largely a personnel cost and is amortizable as more records are entered in the same category
- data entry — $0.02-$0.75/record (depending on how it is done; demand or batch, for instance)
- data storage — $0.10/record per year (essentially zero if data is archived on tape)

These costs may seem large to some, especially that of data preparation. The higher data preparation costs arise when particularly messy data sets are encountered. Great cost reductions in both this and data entry costs can confidently be expected as the user community modifies field and laboratory behavior to become more compatible with computerization. With routine direct-data-entry, costs of less than 2¢/record are certainly achievable. As said above, the project saw significant changes in user behavior as they continued to use the NRDB and the R*B system. Thus far, there have been only a few instances in which the NRDB was used to prepare a special report....the database is, after all, still new. In all those instances, the cost of NRDB preparation was estimated to be about 1/40th that of doing the same job manually.

The Future

The NRDB supported by R*B-2.4 will continue to be maintained until the full-function R*B-3 system becomes operational around JUN85. At that time all records (an estimated 6 million) will be transferred into the new data management system. The cost of this transfer is expected to be minimal since a continuous dialogue is maintained between the NRDB database administrators and the R*B-3 system designers.

As stated above, R*B-3 currently is in the Definition phase. Development will continue in three more phases, Design commencing the summer of 1983, Implementation commencing the spring of 1984, and Demonstration commencing the spring of 1985. In order to obtain Navy approval, the project was required to show superior performance and cost effectiveness for the proposed R*B-3 system. This was done by comparing the existing R*B-2.4 prototype against commercial relational database management systems and database machines available in early 1982 (NOSC, 1982c). R*B-3 will be a generalized, full-function relational system compatible with the management of multidisciplinary scientific measurements. Search rates of at least 2 million records per CPU-second are confidently expected. More explicitly, R*B-3 will have:

* a common query language syntax and grammar for all user functions
* the ability to merge data from different relations into new combinations
* the ability to support customized interfaces to the database. including specialized menu formats, application packages such as statistical and graphical analysis, word processing and document production
* multi-level security to control access on all levels from a relation (primary table) to a single data item (column-row intersection in a table)
* audit trails to monitor access to data and to provide for automatic recovery in the event data are lost or corrupted
* greatly improved efficiency through use of attribute-packing, trigger, and assertion routines currently being defined.

Finally, it should be emphasized that at least the NRDB application of R*B-3 will continue to operate in the public domain as it does now. Participation by non-Navy organizations, particularly universities and museums, is expressly invited on a pay-as-you-go basis. Private as well as other government agencies working in environmental fields are invited as well. The overall intent of this project is to capture valuable environmental data and preserve them for public use.

References

1. L. Bloomfield (1933), Language, Holt Rinehart and Winston
2. R. G. Chenhall (1978), Nomenclature for Museum Cataloging - A System for Classifying Man-Made Objects, Amer. Assoc. for State & Local History

3. E. F. Codd (1970), A Relational Model of Data for Large Shared Data Banks, Communications of ACM 13:6, 377-397
4. E. C. Evans III, ed (1974), Pearl Harbor Biological Survey - Final Report (3 volumes), NUC TN-1128
5. E. C. Evans III (1977a), Microcosm Responses to Environmental Perturbants - An Extension of Baseline Field Survey, Helgoländer wiss. Meeresunters. 30, 178-191
6. E. C. Evans III (1977b), HCZDB/BIODAB Development - An Interlocking Environmental Data Base System, Proc. of IEEE Oceanic Data Base Information Exchange Workshop, pp 84-93
7. V. E. Guiliano (1982), The Mechanism of Office Work, Sci. Amer. 247:3, 148-164 (SEP82)
8. S. R. Hiltz & M. Turoff (1978), The Network Nation - Human Communication via Computer, Addison-Wesley Publ. Co.
9. J. S. Hunter (1980), The National System of Scientific Measurement, Sci. 210, 869-874
10. M. Hura, E. C. Evans III, & F. G. Wood (1976), Coastal Water Protection the Navy Way, Env. Sci. & Tech. 10, 1098-1103
11. International Code of Zoological Nomenclature, London (1961)
12. International Code of Botanical Nomenclature, Leningrad (1975)
13. R. Johansen, R. DeGrasse Jr, T. Wilson (1978), Effects on Working Patterns, The Inst. for the Future, Report R-41, FEB78
14. G. S. Key, E. C. Evans III, & G. G. Gustafson (1979), BIODAB: A Prototype Relational Data Management System for Scientific Applications, presented at the Marine Technology Society Oceans '79 meeting, San Diego, 17-19SEP79
15. D. R. McNeil (1977), Interactive Data Analysis - A Practical Primer, John Wiley & Sons
16. A. Naiman (1982), Introduction to WordStar™, Sybex Inc.
17. Naval Ocean Systems Center (1982a), Natural Resources DataBase Data Entry and Verification Manual, 30SEP
18. Naval Ocean Systems Center (1982b), NRDB System Overview and Users' Manual, OCT82
19. Naval Ocean Systems Center (1982c), System Decision Paper - One, Natural Resources (Ecological) Information System, AUG82
20. N. Sager (1981), Natural Language Information Processing - A Computer Grammer of English and Its Applications, Addison-Wesley Publ. Co.
21. J. W. Tukey (1977), Exploratory Data Analysis, Addison-Wesley Publ. Co.
22. J. Vallee, R. Johansen, H. Lipinski, K. Spangler, & T. Wilson (1978), Social, Managerial, & Economic Issues, The Inst. for the Future, Report R-40, JAN78
23. D. J. Wedemeyer, ed (1980), Marill & Holden in Papers & Proc. of the Pacific Telecommunications Conference, Honolulu

ATTRIBUTES OF A FULLY DOCUMENTED MEASUREMENT

OBSERVATION

MEASUREMENT

- VARIABLE - entity on which measurement was made, including name & classification, state, & components if reported
- QUANTITY - amount of variable measured; any real number, dimensionless
- UNITS - specific magnitude of quantity adopted as a standard for comparison with other quantities of the same kind

DESCRIPTION

- METHOD - technique by which measurement was made
- LATITUDE - first horizontal spatial coordinate of measurement
- LONGITUDE - second horizontal spatial coordinate of measurement
- ELEVATION/DEPTH - vertical spatial coordinate of measurement
- DAY/TIME - temporal coordinate of measurement
- REGION - regional geographic name where measurement was made
- INVESTIGATOR - person making measurement
- AFFILIATION - organization with which investigator was associated at time of measurement
- SPONSOR - organization which underwrote measurement
- HABITAT - environment or surroundings in which measurement was made

Figure 1. BIODAB Data Template

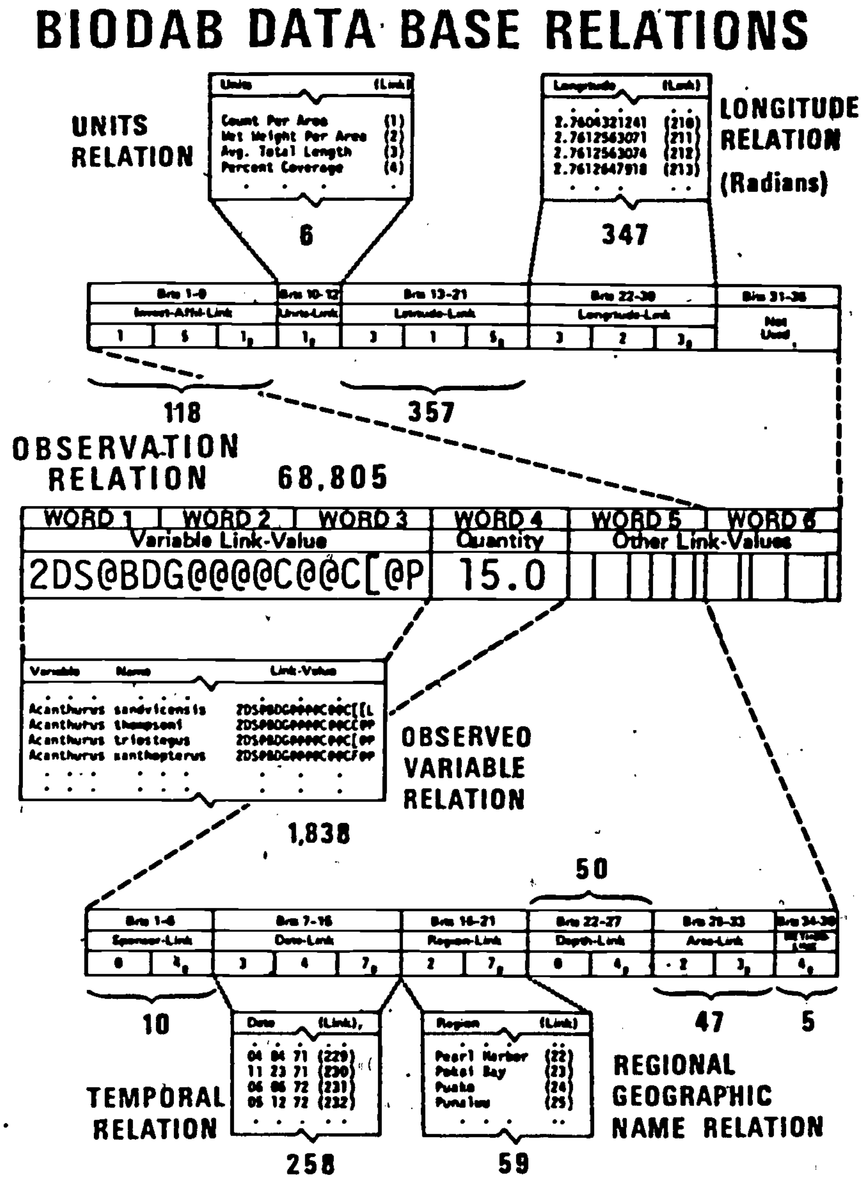


Figure 2. Schematic of Compression Used in BIODAB Tuple

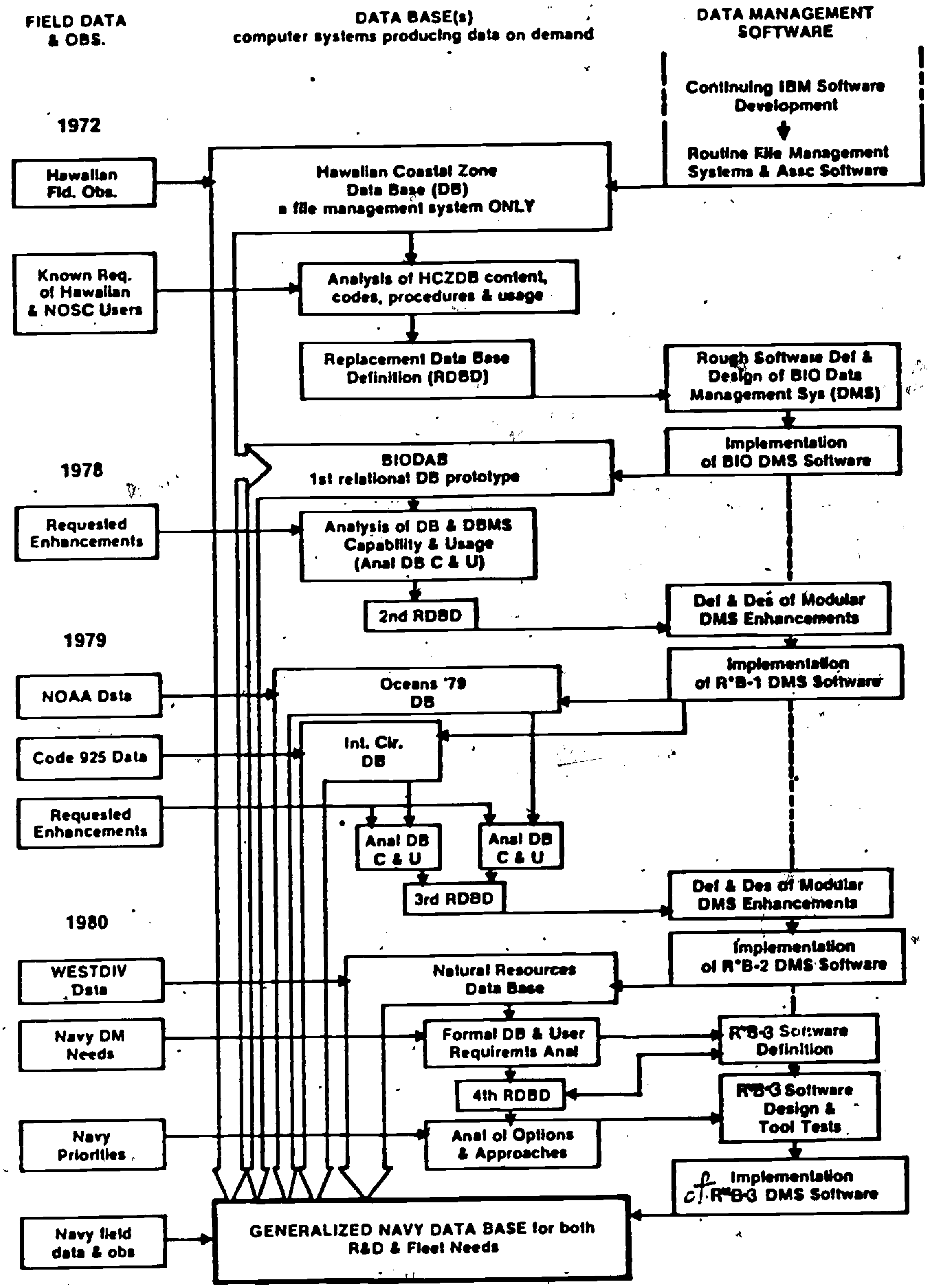


Figure 3. Schematic History of NOSC's Data Management System Development

TAXONOMIC							
Name		Number					
Level	Taxon	TBN†	Dec.	Oct.	Octal Equiv.	Fieldata Character	Binary Representation
Phylum	Chordata	2	50	62	62	2	110010
Subphylum	Vertebrata	1	1	1	11	D	001001
Superclass	Pisces	1	1	1			
Class	Osteichthyes	1	3	3	30	S	011000
Subclass	n. a.	2	00	00	00	@	000000
Infraclass	n. a.	1	0	0			
Series	n. a.	1	0	0	07	B	000111
Superorder	Acanthopterygii	1	7	7			
Order	Perciformes	2	9	11	11	D	001001
Suborder	Acanthuroidei	2	12	14	14	G	001100
Infraorder	n. a.	1	0	0	00	@	000000
Section	n. a.	1	0	0			
Subsection	n. a.	1	0	0	00	@	000000
Superfamily	n. a.	2	00	00	00	@	000000
Subsuperfamily	n. a.	1	0	0	00	@	000000
Family	Acanthuridae	3	001	001	10	C	001000
Subfamily	n. a.	2	00	00	00	@	000000
Tribe	n. a.	1	0	0			
Subtribe	n. a.	1	0	0	00	@	000000
Genus	Acanthurus	2	01	01	10	C	001000
Species	triostegus	3	001	001	01	[	000001
Subspecies	n. a.	1	0	0	00	@	000000
Taxon Level*	Species	3	021	025	25	P	010101

Acanthurus triostegus = 2 D S @ B D G @ @ @ @ C @ @ C [@ P ← (in machine code) ←

* Taxonomic level indicator; 1 = Phylum, 22 = Subspecies; Level Indicator + 30 = Common Name, + 60 = Synonym

† TBN = Three-Bit-Nibble, or three bits read as a byte; half a UNIVAC byte

Figure 4. Taxonomic Code used in BIODAB - How Compression is Accomplished

NRDB Stored Information

DATA TYPE	DATA TYPE

ACQUISITIONS
- Purchase
- Transfer

AGRICULTURE
- Apiculture
- Citrus
- Crop Storage
- Farming
- Fish Farm
- Grazing
- Nursery
- Pasture

ARCHAEOLOGY
- Burial Accompaniments
- Burial Site & Type
- Historic Site
- Prehistoric Site
- Stationary Features

BIOGEOGRAPHY

CONSERVATION
- Cost Avoidance

CONTRACTS
- Agriculture Outlease
- Timber Sales

CORRESPONDENCE

FORESTRY
- Access Roads
- Decade of Origin
- Fire Protection
- Lumber Volume
- Reforestation
- Site Index
- Size Class
- Stocking Density
- Tariff Number

GEOGRAPHIC PLACE NAMES

HABITAT SITES

HYDROLOGY
- Estuarine
- Irrigation
- Lacustrine
- Open Water
- Palustrine

MAINTENANCE
- Outlease
- Water Wells

MANAGEMENT PLANS
- Cooperative Agreements
- Fish & Wildlife
- Forestry
- Land
- Landscape
- Special
- Wildland

MEETINGS & TRAINING
- Agriculture
- Biology
- Computer Networks
- Forestry
- Hydrology

PEDOLOGY
- Flood Deposit Soils
- Lacustrine Terrace Sediments

OBSERVATIONS & SURVEYS
- Agriculture
- Archaeology
- Birds
- Coastal Marine
- Forest Inventory
- Feral Animals
- Hydrology
- Vegetation
- Water Table
- Weather
- Wildlife

RECREATION
- Hunting

REGULATED
- Flora
- Fauna
- Habitat

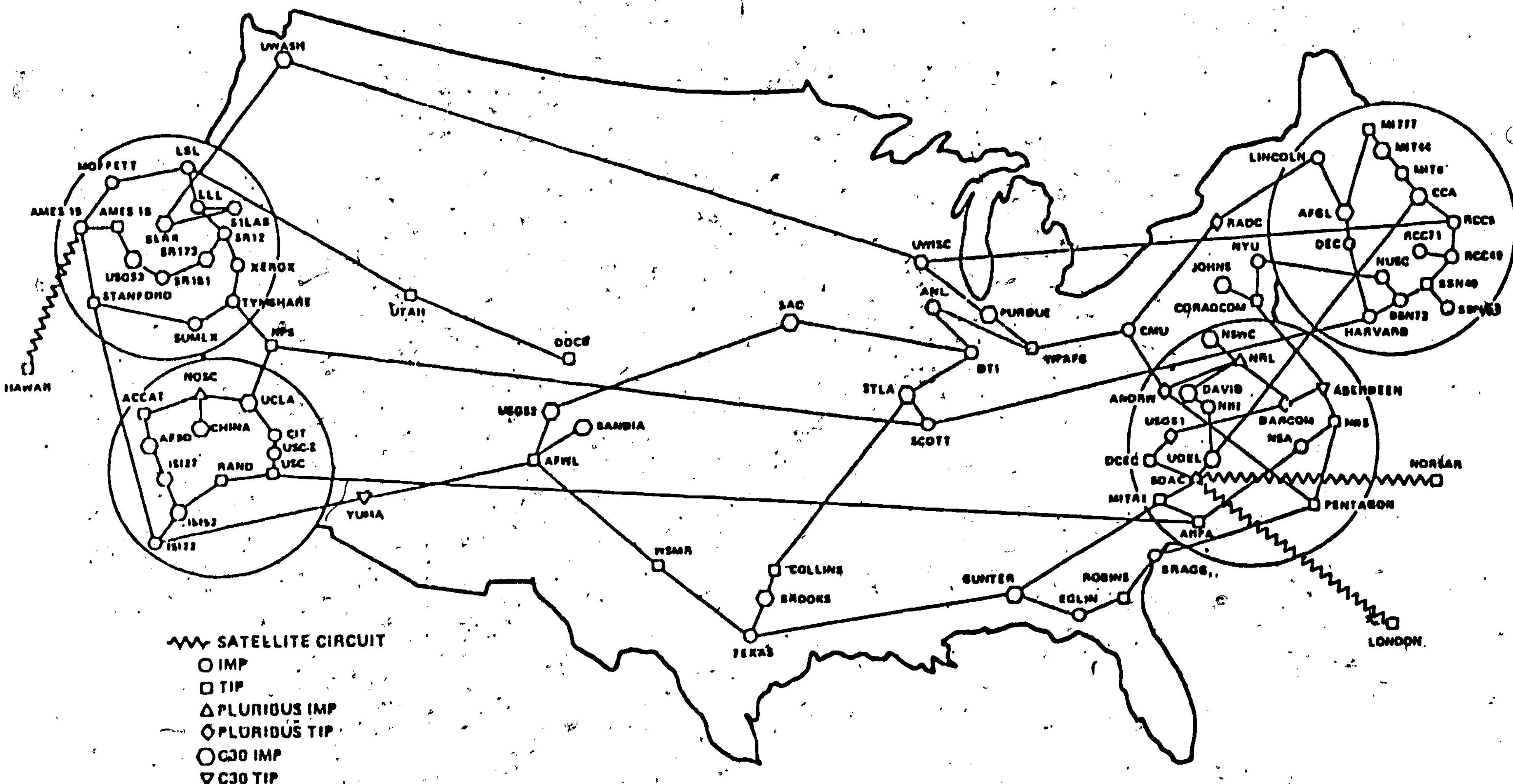
RESOURCE MANIPULATION
- Cows & Sheep
- Rabbits
- Research Natural Area
- Sand Dunes
- Vernal Pools
- Water Wells

SITE IMPACT
- Construction
- Military
- Non-military

TOPOGRAPHY

Figure 5. Types of Data Stored in NRDB, OCT82

ARPANET GEOGRAPHIC MAP, FEBRUARY 1982



(NOTE: THIS MAP DOES NOT SHOW ARPA'S EXPERIMENTAL SATELLITE CONNECTIONS)
NAMES SHOWN ARE IMP NAMES, NOT (NECESSARILY) HOST NAMES

22

Figure 6. ARPANET Geographic Map, FEB82

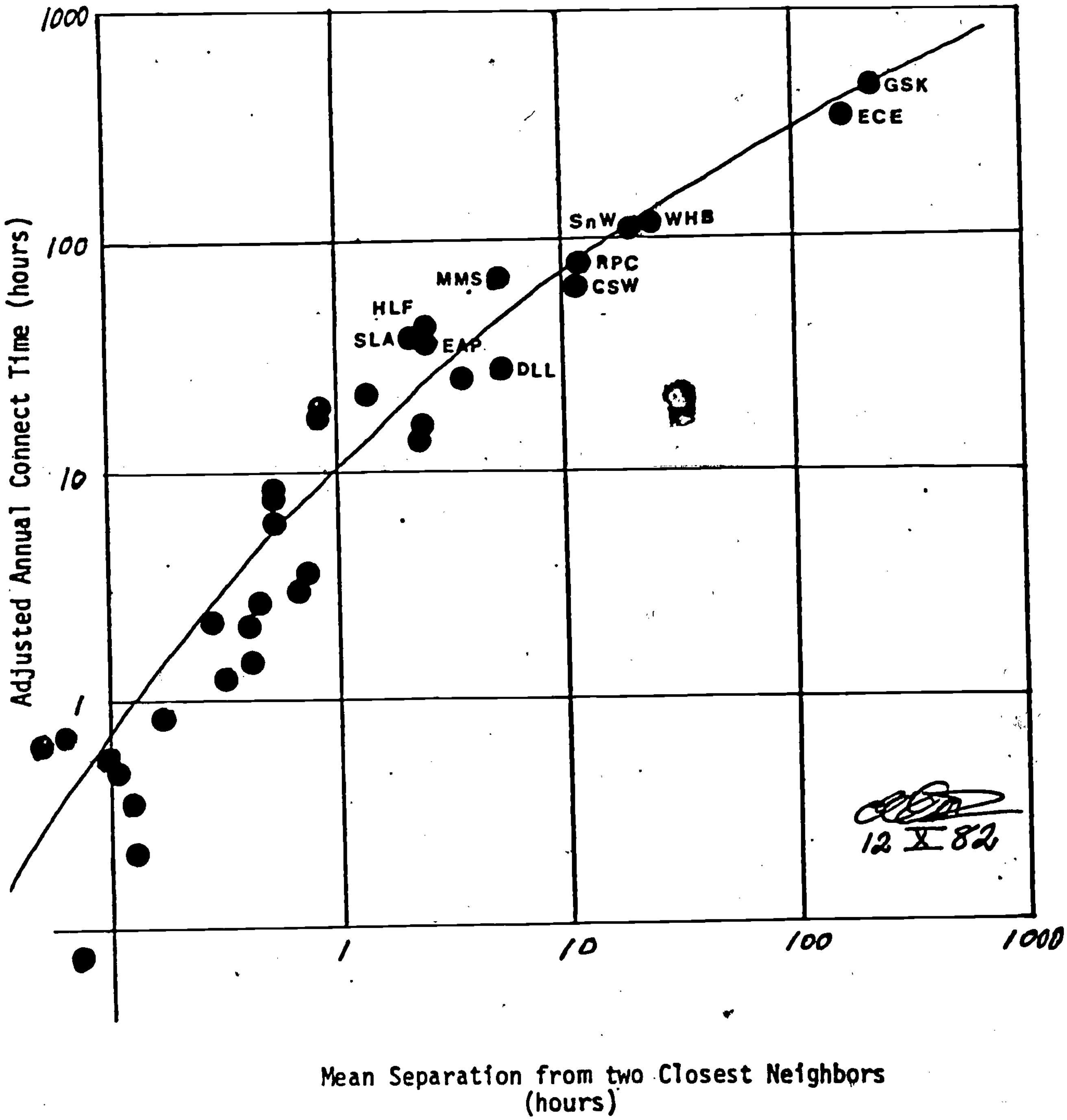


Figure 7. NRDB User Activity, 1981-1982

*1.	KGD	Kingdom
'2.	sKG	Subkingdom/Category
*3.	PHY	Phylum/Division
'4.	sPH	Subphylum/Subdivision
5.	pCL	Superclass
*6.	CLS	Class
'7.	sCL	Subclass
8.	iCL	Infraclass/DivisionF/DivisionI/SeriesC
9.	pOR	Superorder/Cohort/SubdivisionI/SectionI
*10.	ORD	Order
'11.	sOR	Suborder
12.	iOR	Infraorder/DivisionC/SectionC/TribeC/TribeI
13.	pFM	Superfamily/SubdivisionC/SubsectionC/SubtribeC/SubtribeI
*14.	FML	Family
'15.	sFM	Subfamily
16.	iFM	(Infrafamily)/Contribe/DivisionO
17.	TRB	Tribe
18.	sTR	Subtribe/SectionG/SeriesO
19.	pGE	(Supergenus)/SubseriesO
*20.	GEN	Genus
'21.	sGE	Subgenus
22.	STN	Section
23.	SER	Subsection/Series
24.	pSP	(Superspecies)/Subseries
*25.	SPC	Species
'26.	sSP	Subspecies/Variety/Breed
*27.	TAX	Binomial (GEN + SPC) or Trinomial (GEN + SPC + sSP) or Variety/Breed/Form/Race/Cultivar/Cross & their subs, in short the specific entity
*28.	LVL	Taxon Level
*29.	AUT	Authority
*30.	DAT	Date
*31.	STS	Status
*32.	VID	Vide
*33.	CMT	Comments
*34.	UPD	Update

Notes:

attribute types -

* = primary taxonomic level or master attribute status
' = secondary taxonomic level
= sliding taxonomic level or little used taxonomic level

attribute flags -

C = crabs, F = fish, G = grasses, I = insects, O = orchids

Figure 8. Schema for a Generalized Taxonomic Hierarchy (RELATABASE-3)

	Accepted Tuple	Common Tuple	Synonym Tuple	Common Tuple
1 KGD	Animalia	Animals	Ø	Man-Made
2 sKG	Metazoa	Ø	Ø	Ø
3 PHY	Chordata	Vertebrates	Ø	Structures
4 sPH	Gnathostomata	Jawed-vertebrates	Ø	Ø
5 pCL	Pisces	Fishes	Ø	Ø
6 CLS	Osteichthyes	Boney-fishes	Ø	Bldg-Fragment
7 sCL	Neopterygii	Modern-fishes	Ø	Ø
8 iCL	Ø	Ø	Ø	Ø
9 pOR	Acanthoptergyii	Ø	Ø	Ø
10 ORD	Scorpaeniformes	Ø	Ø	Passage
11 sOR	Scorpaenoidei	Ø	Ø	Ø
12 iOR	Ø	Ø	Ø	Ø
13 pFM	Ø	Ø	Ø	Ø
14 FML	Scorpaenidae	Scorpion-fishes	Ø	Door
15 sFM	Ø	Ø	Ø	Ø
16 iFM	Ø	Ø	Ø	Ø
17 TRB	Ø	Ø	Ø	Ø
18 sTR	Ø	Ø	Ø	Ø
19 pGE	Ø	Ø	Ø	Ø
20 GEN	Scorpaenodes	Ø	Scorpaenodes	Door-panel
21 sGE	Ø	Ø	Ø	Ø
22 STN	Ø	Ø	Ø	Ø
23 SER	Ø	Ø	Ø	Ø
24 pSP	Ø	Ø	Ø	Ø
25 SPC	parvipinnis	Ø	guamensis	west panel
26 sSP	Ø	Ø	Ø	Ø
27 TAX	S. parvipinnis	Scorpion-fishes	S. guamensis	west panel
28 LVL	species	family	species	?
29 AUT	Garrett	Ø	Eschmeyer	Raphael
30 DAT	18nn	Ø	19nn	15nn
31 STS	accepted	common-name	synonym	?
32 VID	S. parvipinnis	Scorpaenidae	S. parvipinnis	?
33 CMT	whatever	whatever	whatever	whatever
34 UPD	821115	821115	821115	821115

Figure 9. An Application of the Generalized Taxonomic Schema (RELATABASE-3)