2. Typology Approach and Products

2.1 Typology Database

2.1.1 Database Description

The LOICZ typology database is developed and maintained with the joint support of the partner project, Biogeography of the Hexacorallia. It is based at the Kansas Geological Survey of the University of Kansas, operates in Oracle® database software, and is served to users on the World-Wide-Web through an interface designed in Cold Fusion® software. The features of the database, described below, include the geographic grid and cell structure of the basic design, the present contents of the database in terms of environmental variables (listed in Appendix IV), and the user interface for selecting, modifying, acquiring and analyzing the data.

2.1.1.1 Geographic grid and typology cell structure

The objectives of the LOICZ cell structure are to provide a geographically structured basis for database design that will:

- classify coastal environments in a conceptually useful fashion;
- permit integration and consistent analysis of available global data sets dealing with land, sea, air, and human dimension variables;
- operate at a scale of resolution useful for the data and applications envisioned; and
- provide a manageable number of data points for analysis and global upscaling or extrapolation.

Half-degree resolution has been selected as the most realistically useful compromise between desired resolution and the available data and methods. Of the 259,200 half-degree cells on the earth's surface, a total of 47,057 have been identified as primary typology cells (coastal, terrestrial and ocean-I). The remaining oceanic cells (145,989) are included to provide complete coverage for both the larger LOICZ budget sites and for the NOPP/OBIS partner program, Biogeography of the Hexacorallia. This cell structure also serves the needs of other partner projects.

Land or Terrestrial cells (T) are typically defined as the cells containing only land (or fresh water). Coverage typically extends two cells (one degree) inland from the most landward coastal cell, although they have been extended farther inland in a few areas (e.g., estuaries not well represented by the shoreline data set). These do not have ocean variables.
Coastal cells (C) are defined as those containing a significant length of the World Vector Shoreline. These cover significant areas of both land and (marine or estuarine) water, and are populated with all classes of variables.
Oceanic I cells (O-I) are those extending seaward from the coastal cells the greater of (a) one degree, or (b) the 50 or 100 m isobath in areas of a broad shelf, or (c) to include all of a biogeochemically budgeted area. These will have only oceanic and atmospheric variables.
Oceanic II cells (O-II) are those additional cells needed to complete the in-filling of relatively enclosed water bodies or coastal seas that might be the target of future up-scaled biogeochemical budgeting exercises.
Oceanic III cells (O-III) cover all remaining oceanic areas not included in the other classes.
Inland Cells (I) cover all remaining land areas not included in other classes.

Geographic conventions: NORTH latitude and EAST longitude are <u>positive</u> numbers; SOUTH latitude and WEST longitude are <u>negative</u> numbers.

Cell ID numbers are based on a sequential global grid of half-degree cells (Table 2.1). Numbering (from 1 to 259,200) begins with the cell centered at 89.5 degrees N latitude and 179.5 degrees W latitude (89.5, -179.5) and proceeds from west to east. When a full circle (or row, in a planar projection) of 720 cells (360 degrees/0.5 degrees/cell) is completed, the numbering steps one cell south along the -180 meridian and continues sequentially west to east.

Table 2.1 Statistics on cell types and numbers

Cell Class	Number of Cells
Total primary Typology cells	47,057
Oceanic-III	143,394
Oceanic-II	2,595
Oceanic-I	19,330
Coastal	15,278
Terrestrial	12,449
Inland	66,317
Non-Typology Cells	209,517
Total number of 0.5 degree cells in the world	259,200

2.1.1.2 Data contents

Appendix IV provides a listing of the environmental variables, the type of data, cell assignments, and a summary identification of the source. Biogeochemical budget variables contained in the database are discussed in Section 2.2 and Appendix V of this report. For a complete review and overview of current database contents and form, go to www.kgs.ukans.edu/Hexcoral/Envirodata/envirodata.html. Additional examples and illustrations may be found in the CD-ROM accompanying this report.

The environmental data contained in the database are derived primarily from public-domain sources, and metadata listings are provided so that users may access and refer to the original data source. In order to meet the needs of global synthesis, only datasets that are global or near-global in extent are included in the database at present. The data are divided into categories for ease of searching and reference: these are (in addition to database structure and location references) Atmospheric, Geomorphic, Human Dimension, Oceanic, Terrestrial, Basin and Budget.

The Basin variables are supplied by Charles Vörösmarty, University of New Hampshire, under the aegis of the IGBP-BAHC project. These consist of modeled annual and monthly outflow (runoff) for all world river basins above a size threshold of a few tens of thousands of km^2, plus supporting data in the form of basin population, size, and land cover. The Budget variables, recently included in the database, are the site and biogeochemical budget characteristics assembled by the LOICZ Biogeochemical Budgets activity (see Section 2.2, Appendix V, and http://data.ecology.su.se/mnode/).

2.1.1.3 User interface features

The WWW-based data access site is publicly accessible for review and occasional use by a guest log-in option. More extensive users are issued a password on request; this provision is designed to regulate traffic and avoid electronic conflicts, not to restrict access.

The sequence of normal use proceeds in two phases, first:

- Selection of database version – selected (short form) or full environmental data, with or without budget data. This option was instituted to provide archival access and variety in the full database option, while keeping the list of variables that have to be reviewed by the typical user to a workable number.
- Selection of geographic region – by pre-determined zone or by user input of latitude and longitude.
- Selection of variables.
- Assembly (automatic) of the user dataset.

Second, after the data set is assembled in the Oracle system, the interface offers the user a variety of options for review and modification. These include:

- Reviewing the selection criteria.
- Viewing the statistics (tabular) and distribution (histogram) of any variable; the histogram display is user-controlled for range and number of classes, and offers some data transformation options (e.g., log, square root).
- Excluding cells with null values (no data).
- Modifying the dataset by excluding or resetting numerical values for any variable above, below, or within a user-specified range.
- Modifying the dataset by transforming any variable (e.g., log, ln, absolute value, square root).
- Creating a correlation matrix for the selected variables (with modifications).
- Generating a report specifying the data selected and all operations performed.

When the final dataset is assembled, the user may view or download it as a comma-separated text file, and/or send it directly to the Web-LOICZView clustering tool via an internet link.

2.1.2 Database Access

2.1.2.1 Internet-based

The use of the internet to maintain and serve the data is considered the only cost-effective and reliable way to make up-to-date data readily available to a widespread user community. It is recognized, however, that connection, hardware, and software limitations may make access problematic from some parts of the world. To reduce these problems, all of the interface tools have been designed to run "server-side" to minimize the amount of material that must be actually exchanged between the site and the user. However, since even these measures will not address all problems, alternative approaches to data dissemination are also used and under development.

2.1.2.2 Portable

Distribution of CD-ROM versions of websites, datasets, and products has been used by LOICZ as part of its workshop process, and is available on request to people who have problems with internet access. A particular issue is the provision of remote database equivalents, including the selection functionality. A product under development by Dr L.T. David and existing in prototype form is a database front-end designed to operate in Microsoft Access®, a widely available microcomputer database system that is compatible with the structure of the central Oracle system.

This product would permit distribution of subsets of the database (on CD-ROM), or provision of the entire system on DVD disks, in a form that would support the same basic functionalities and output products as the WWW-based system.

2.2 Biogeochemical Budgets Database

2.2.1 Biogeochemical budgets – background

The LOICZ project set up a "globally applicable" method of estimating fluxes of carbon, nitrogen and phosphorus within the coastal ocean, especially the bays and estuaries of the inner coastal zone (Gordon *et al.* 1995). It was necessary to erect a methodology that could depend largely on secondary data, that had minimal data requirements and that was widely applicable and uniform, in order to allow effective comparison among sites. Finally, the method had to be informative about processes influencing CNP fluxes.

The implementation strategy for developing a biogeochemical budget database was to mount a two-pronged attack on acquainting the scientific community with the budgeting procedures. The first prong has been publication of a web-page (http://data.ecology.su.se/MNODE) that summarizes and updates the budgeting procedures, provides tools for implementing the procedures, provides various forms of teaching materials, and posts existing budgets as they are developed. The second prong has been to hold a series of workshops around the world, in order to teach people how to do the budgets and to get them to prepare budgets that can be used by LOICZ. At time of writing, about 170 budgets have been

developed, largely as products of more than 15 workshops held around the world. The sites budgeted are indicated in Figure 2.1.

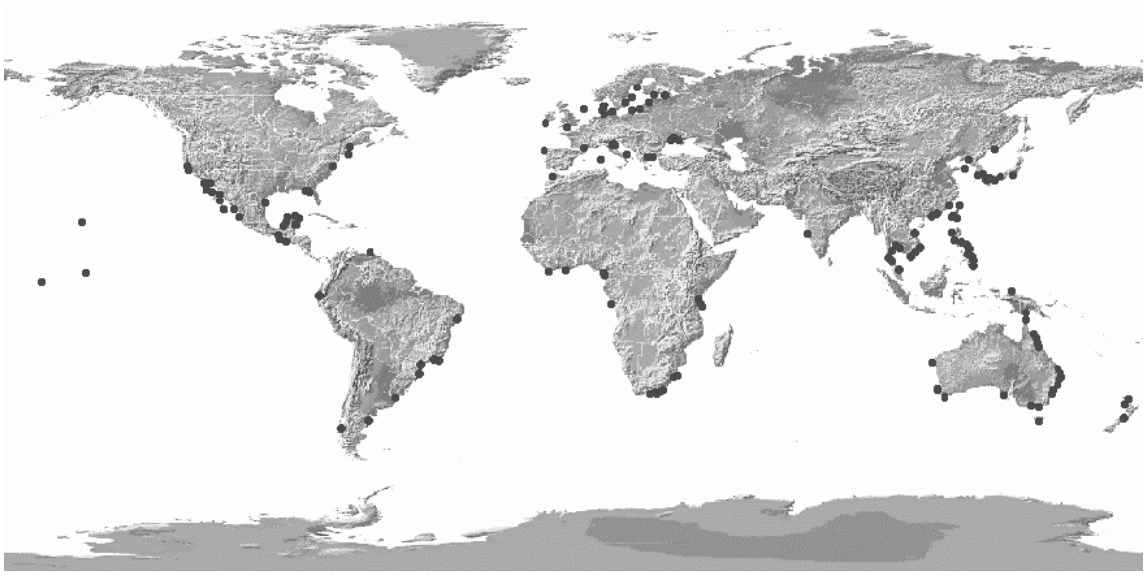

Figure 2.1 Map of LOICZ budget sites, October 2001.

The LOICZ approach is based on one of the most fundamental concepts of the physical sciences: conservation of mass. Details of the approach are given in Gordon *et al.* (1995) and on the LOICZ Modelling web page (http://data.ecology.su.se/MNODE). Steady-state conditions are assumed in which water volume and salt content in the system remain constant over time, as water flows through the system and mixes with adjacent systems. The net flow of water can be described by a water budget. Information about mixing can be deduced from a salt budget of non-reactive materials. The data to establish at least first-order water and salt budgets can be found for many sites around the globe.

Nutrients not only move with the water but also undergo reactions within the system. Nutrient data (especially data on the dissolved inorganic forms of phosphorus and nitrogen, here termed DIP and DIN) can be found for many of these same sites and used to establish nutrient budgets. These nutrient budgets include water flow and mixing, as defined by the water and salt budgets, and an additional term that describes net uptake or release of these nutrients within the system. In the jargon of oceanography, these are termed "nonconservative fluxes," because the nutrients do not follow exactly the flux pathways of water and salt.

The nonconservative flux of DIP can be used as an approximation of net uptake of phosphorus into organic matter during primary production, or release from organic matter by respiration. While it would be desirable to have direct measurement of carbon uptake into organic matter, such data are not available for most locations. Therefore the flux of DIP becomes a proxy for net carbon flux. In the open ocean DIN is often scaled in exactly this manner to carbon. That scaling in general does not work well in the coastal ocean, for a reason that contains a great deal of information itself. Nitrogen fixation and denitrification are important metabolic processes in bottom-dominated systems and can account for most of the observed nonconservative flux of DIN. Therefore calculations are derived from the budgets:

1.) using DIP flux as a proxy to calculate how much net carbon uptake or release has occurred,
2.) scaling DIP flux to estimate the expected nitrogen (DIN) flux (typically using the Redfield N:P ratio of 16:1), and
3.) using the deviation between the observed DIN flux and the expected DIN flux to estimate the net of nitrogen fixation and denitrification.

2.2.2 Interactions with the regional typology workshop process

Development of biogeochemical budgets through workshops has continued in parallel with the regional synthesis workshop process, with budget node personnel heavily involved in the synthesis workshops. This ensured that the latest information on developments in the budget process were available to the synthesis workshop participants, and provided critical opportunities for developing the needed interconnections among the developing biogeochemical budget database, the typology database (Section 2.1, Appendix IV) and the Web-LOICZView (WLV) clustering and visualization tool (Section 2.3, Appendix VI).

The close coordination of synthesis (typology) and budget efforts through the workshop process developed both a cadre of scientists familiar with the developing combined effort, and a series of data integration and analysis trials that led ultimately to the combined database system. This process was augmented by key 'mini-workshops' held among the technical resource personnel in August and October 2001 to work through and test the practical details of the joint database operation and the applications of WLV to the combined data.

2.2.3 The joint "Synthesis Database" product

The budget sites (Figure 2.1) vary dramatically in their characteristics: from lagoons and estuaries of less than 1 km^2 in area, to the 10^6 km^2 East China Sea; from sites that are decimeters deep to sites that are hundreds of meters deep; from sites that are virtually devoid of loading from land to sites that receive heavy loads of inorganic nutrients derived from human wastes, agriculture, and other sources; from sites that are river-dominated estuaries to hypersaline embayments; and from tropical to arctic climatic zones. For some sites, data quality and quantity are both high; other sites suffer in the quality and quantity of information available. For the initial analysis we have identified a "preferred" subset of budget site data that excludes systems for which the basic data are incomplete, open shelf systems, and systems with an average depth >100 m, in order to facilitate comparisons among sites. This preferred dataset includes about 80 systems.

Initial experiments have shown the importance of data transformations and scaling in developing relationships that will be useful for upscaling as well as interpretation. The basic biogeochemical budget dataset has therefore been augmented with alternative presentations of many of the variables, so that users can take advantage of the accumulated experience without having to reformulate the data set.

The entire biogeochemical budget database has been incorporated into the Oracle-based, web-accessible Typology database at http://www/kgs/ukans.edu/Hexacoral/Envirodata/envirodata.html. It can be accessed as either the complete or the preferred budget dataset, and in combination with either the full inventory of environmental variables, or a more concise set of variables selected for their appropriateness to representing the ecosystems processes. The budget database includes descriptive and transformed variables, and all of the budget variables can be manipulated (separately or in combination with typology variables) by the database and WLV features described elsewhere in this report.

Appendix V provides a summary list of the budget variables accessible through the integrated database.

2.3 Geospatial Clustering Tools (Web-LOICZView)

2.3.1 Concept and Capabilities: Web-LOICZView

Web-LOICZView [WLV] is a web-based graphical user interface to a set of data analysis tools. These tools are intended to facilitate analysis and understanding of trends and groupings that exist in a spatially indexed data set. WLV was developed with the support of the Land-Ocean Interactions in the Coastal Zone [LOICZ] program, which is core project of the International Geosphere-Biosphere Program [IGBP]. WLV has been developed mainly for working on geographic data sets with multiple variables defined at each geographic location. It is tightly integrated with the LOICZ/Hexacoral database at the University of Kansas (Section 2.1 and Appendix IV), that contains a wide variety of geographically indexed data sets. WLV is reasonably flexible, however, and can be used with a wide variety of datasets.

The capabilities and use of WLV are addressed in more detail in Appendix VI of this report, and are described in Maxwell and Buddemeier (2001). A more extensively illustrated version of Appendix VI can be found in the CD-ROM portion of this report, as can numerous examples of WLV output. Similar information is also available on the project web sites.

The primary data analysis tool in WLV is a set of clustering routines that group together similar data points into classes. WLV then gives the user a variety of ways to visualize the classes and the data. To provide for ease of experimentation, there is a variety of data management tools that allow users to manipulate and control how the data is treated in each analysis step. For example, you can select subsets of variables and weight them according to their importance.

Data sets may be imported into WLV by transfer from the typology database, or by independent upload according to a set of simple instructions on the website. Similarly, products may be downloaded or captured with browser software.

WLV also includes a tool for executing an eigenvector, or principal components analysis on the data set. This tool permits users to understand the principal causes of variation in their dataset. It also permits effective visualization of high dimensional data sets.

Finally, WLV includes a tool for analyzing the overall complexity of a data set based on the Minimum Description Length principle. This can be used as an aid in discovering the appropriate number of classes for a given set of variables and data points.

2.3.2 Access

Access to Web-LOICZView is offered as a courtesy to the scientific community. As is the case with the Typology database, access is controlled by password not to restrict use, but to manage traffic and prevent overloads. New or occasional users may take advantage of a “guest” log-in, and passwords are issued on request to more frequent users.

WLV runs on a LINUX (or UNIX) operating system, and requires no more than a modest laptop computer to support its operations. It has been transported to, and run locally at, all of the workshops. An early version (LV) was made available for download and local installation on LINUX or UNIX systems. Although this is still available, the rapid pace of development of WLV over the past year has precluded updates of the stand-alone version.

WLV will continue to be served and supported for the remaining lifetime of the LOICZ project, but long-term continuation, and development of updated stand-alone software for distribution, will depend on future institutional and funding arrangements.