GIÁO TRÌNH

Quản lý dữ liệu trong nghiên cứu môi trường

Nguyễn Hồng Phương

Chương 1

NHẬP MÔN VỀ QUẢN LÝ DỮ LIỆU

I. MỞ ĐẦU

Nhu cầu tích lũy và xử lý các dữ liệu đã nảy sinh trong mọi công việc, trong mọi hoạt động của con người. Một cá nhân hay một tổ chức có thể đã mặc nhiên có một hệ thống xử lý dữ liệu, cho dù cơ chế hoạt động của nó là thủ công và chưa tự động hóa.

Một bài toán nhỏ cũng cần đến dữ liệu, nhưng không nhất thiết phải quản lý các dữ liệu này theo các phương pháp khoa học. Do khả năng tổng hợp của người xử lý, các dữ liệu được lấy ra, được xử lý mà không vấp phải khó khăn nào. Tuy nhiên khi bài toán có kích thước lớn hơn hẳn và số lượng dữ liệu cần phải xử lý tăng lên nhanh thì khả năng bao quát và quản lý của một người bình thường sẽ trở nên khó khăn. Đó là chưa kể đến một số loại dữ liệu đặc biệt, đòi hỏi được quản lý tốt không phải vì kích thước mà vì sự phức tạp của bản thân chúng.

Lúc bắt đầu công tác tự động hoá xử lý dữ liệu, người ta sử dụng các tệp dữ liệu là nơi chứa thông tin và dùng các chương trình để tìm kiếm, thao tác trên các dữ liệu của tệp đó. Đó là tiền thân của các hệ thống cơ sở dữ liệu. Tuy nhiên một vài người hiểu chưa chính xác về cơ sở dữ liệu; họ coi các hệ quản trị tệp là cơ sở dữ liệu. Việc coi các "tệp dữ liệu" là cơ sở dữ liệu hoặc coi một phần mềm nào cho phép xử lý dữ liệu như hệ quản trị cơ sở dữ liệu...là nhìn nhận không chính xác. Để hiểu đầy đủ các khía cạnh về hệ quản trị cơ sở dữ liệu, người ta cần được trang bị các khái niệm cơ bản.

II. KHÁI NIỆM VỀ CƠ SỞ DỮ LIỆU VÀ HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU

II.1. Cơ sở dữ liệu

Trong kỷ nguyên của cách mạng khoa học kỹ thuật và bùng nổ thông tin, máy tính được coi là một công cụ đắc lực của con người trong việc quản lý những lượng thông tin khổng lồ.

Nhận thức về tầm quan trọng của máy tính điện tử trong việc quản lý dữ liệu đã có từ lâu, nhưng nhận thức này chỉ thực sự được khẳng định từ sau sự ra đời của các máy tính thế hệ 3, điển hình là IBM 360. Một trong những ưu điểm nổi trội của loạt máy tính IBM 360 là ở chỗ, trong hệ điều hành của chúng tồn tại một hệ thống kiểm tra dữ liệu, với một số chức năng quản lý dữ liệu chính cho phép:

- Lưu trữ thông tin về dữ liệu như vị trí, loại, trạng thái, v.v..thông qua hệ thống tổ chức file;
- Quyết định quyền hạn sử dụng dữ liệu, tăng cường các đòi hỏi về bảo mật, cung cấp các quy trình truy nhập;
- Tìm kiếm và cất giữ dữ liệu, chẳng hạn mở hay đóng một file.

Quá trình quản lý dữ liệu dần dần đã vượt xa ra ngoài khuôn khổ của những ứng dụng trong hệ điều hành máy tính. Năm 1959, tại một hội nghị quốc tế về ngôn ngữ cho hệ thống dữ liệu (CODASYL), lần đầu tiên những nền tảng cơ sở cho việc phát triển các công nghệ và ngôn ngữ sử dụng cho việc phân tích hệ thống các dữ liệu, thiết kế và ứng dụng các cơ sở dữ liệu đã được thiết lập. Cho đến nay, lý thuyết về cơ sở dữ liệu đã phát triển tới một mức độ cao và tồn tại độc lập như một lĩnh vực nghiên cứu, thu hút sự chú ý và công trình của đông đảo các nhà khoa học, các chuyên gia tin học và các nhà quản lý dữ liệu trên toàn thế giới.

Một cơ sở dữ liệu được định nghĩa là một tập hợp các dữ liệu về các đối tượng cần được quản lý và lưu trữ theo một cơ chế thống nhất, nhằm thực hiện các chức năng sau đây một cách tối ưu:

- Mô tả dữ liệu;
- Cập nhật dữ liệu;
- Tìm kiếm dữ liệu;
- Trao đổi dữ liệu

II.2. Hệ quản trị cơ sở dữ liệu

Hệ quản trị cơ sở dữ liệu là một công cụ tổng hợp dùng để thực hiện các thao tác đối với một hay nhiều cơ sở dữ liệu lớn. Thông thường, hệ quản trị cơ sở dữ liệu bao gồm một phần mềm hay một hệ chương trình đặc biệt, giúp người sử dụng thực hiện có hiệu quả các quá trình tra vấn, sửa đổi hay phân tích, xử lý dữ liệu.

Một hệ quản trị cơ sở dữ liệu được xây dựng nhằm hướng tới các mục tiêu sau:

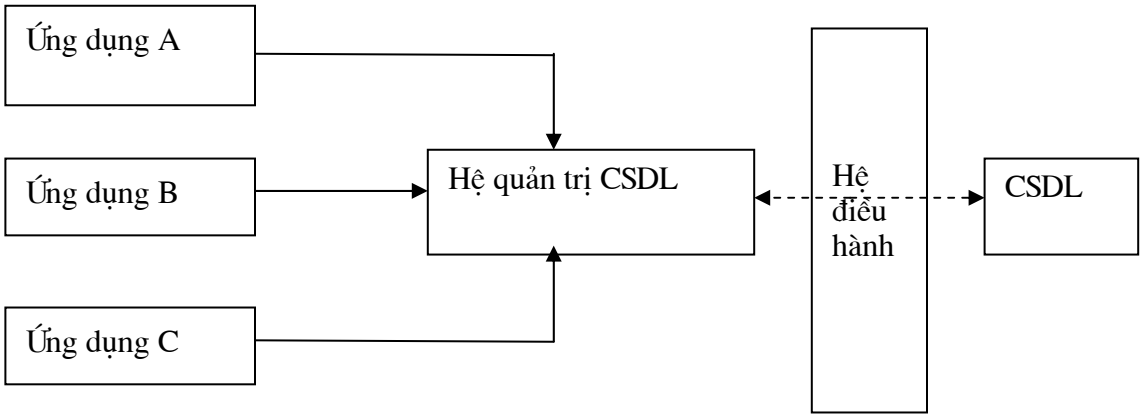
- Thu thập, tích hợp được một cơ sở dữ liệu đáp ứng rộng rãi nhu cầu của đông đảo người sử dụng;
- Đảm bảo chất lượng và tính đầy đủ của dữ liệu;
- Bảo tồn được tính riêng biệt của dữ liệu thông qua các biện pháp bảo mật trong hệ;
- Cho phép điều khiển cơ sở dữ liệu trên nguyên tắc tập trung;
- Bảo đảm tính độc lập của dữ liệu.

Trên hình 1 minh hoạ sơ đồ tổ chức cơ sở dữ liệu và hệ quản trị cơ sở dữ liệu trong máy tính.

II.3. Các mô hình quản trị cơ sở dữ liệu

Cho đến nay tồn tại nhiều mô hình quản trị cơ sở dữ liệu khác nhau, nhưng phổ biến nhất phải kể đến các mô hình sau:

1. Hệ quản trị cơ sở dữ liệu phân cấp (Hierarchial DBMS)
2. Hệ quản trị cơ sở dữ liệu mạng (Network DBMS)
3. Hệ quản trị cơ sở dữ liệu quan hệ (Relational DBMS)



Hình 1. Cơ sở dữ liệu và hệ quản trị cơ sở dữ liệu

Nhìn chung, việc xây dựng các hệ quản trị dữ liệu đều dựa trên việc lựa chọn một cấu trúc dữ liệu tối ưu, nhằm giải quyết hai yếu tố rất quan trọng là: không gian lưu trữ dữ liệu và hiệu quả của các phép xử lý. Các ví dụ dưới đây sẽ so sánh cách tổ chức các dữ liệu địa lý trong ba mô hình quản trị dữ liệu hiện đang phổ biến nhất hiện nay.

II. 3.1. Cấu trúc dữ liệu Phân cấp

Cấu trúc dữ liệu phân cấp lưu trữ dữ liệu theo một trật tự về thứ bậc được thiết lập giữa các mục của dữ liệu. Mỗi điểm nút có thể được chia ra thành một hay nhiều điểm nút con. Số các nút con tăng lên tỷ lệ thuận với số cấp, giống như sự phân nhánh trên một cái cây.

Trên hình 2.1. minh họa một thí dụ về cách tổ chức dữ liệu địa lý theo các mô hình Phân cấp và Mạng cho bản đồ M, biểu diễn hai miền I và II dưới dạng hai đa giác với các đỉnh được đánh số (1, 2, 3, 4 cho đa giác I và 4, 3, 5, 6 cho đa giác II) và các cạnh ký hiệu bằng các chữ (a, b, c, d cho đa giác I và c, e, f, g cho đa giác II).

Dữ liệu phân cấp được tổ chức theo quan hệ cha/con hoặc 1 - nhiều (Ví dụ như quản lý nhà ở dân dụng theo cấp I, cấp II, cấp III, cấp IV). Cấu trúc này tạo thuận lợi cho việc truy nhập dữ liệu. Hệ thống phân cấp chấp nhận mỗi phần của cấp đưa ra sử dụng một khóa mà nó thể hiện đầy đủ cấu trúc dữ liệu. Cho phép có một sự tương quan giữa các thuộc tính kết hợp và mục dữ liệu có thể có.

Hệ thống này cũng tiện lợi cho việc bổ sung, sửa đổi và mở rộng, tiện lợi cho việc truy nhập dữ liệu theo thuộc tính khóa, nhưng khó khăn cho những thuộc tính không phải là khóa.

Bất lợi của cấu trúc dữ liệu phân cấp là tệp chỉ số lớn cần phải được duy trì và các giá trị của thuộc tính cần phải được lặp lại nhiều lần gây ra dư thừa dữ liệu làm tăng chi phí lưu trữ và truy nhập.

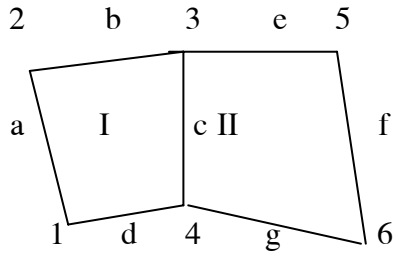
II. 3.2. Cấu trúc dữ liệu Mạng

Cấu trúc dữ liệu mạng tương tự như cấu trúc dữ liệu phân cấp, chỉ có khác là trong cấu trúc này mỗi điểm nút con có thể có nhiều hơn một điểm nút cha. Đồng thời, mỗi điểm nút lại có thể được chia ra thành một hay nhiều điểm nút con.

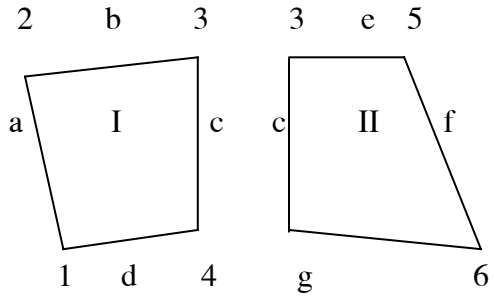
Trong cấu trúc dữ liệu địa lý, việc thể hiện các mục mà tương ứng trên bản đồ hay sơ đồ là gần nhau thì lại là các phần khác xa nhau của cơ sở dữ liệu. Hệ thống mạng rất cần thiết để thể hiện dạng này.

Cấu trúc mạng phù hợp khi quan hệ và mối liên kết đã được xác định trước, tránh được dư thừa dữ liệu. Bất tiện cho việc mở rộng bởi tổng số các điểm. Việc sửa đổi và duy trì cơ sở dữ liệu khi thay đổi cấu trúc các điểm đòi hỏi tổng chi phí lớn...

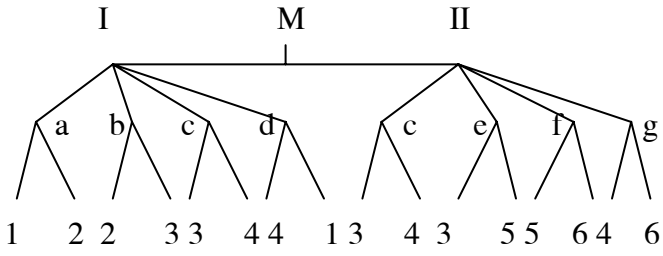
a) Bản đồ M



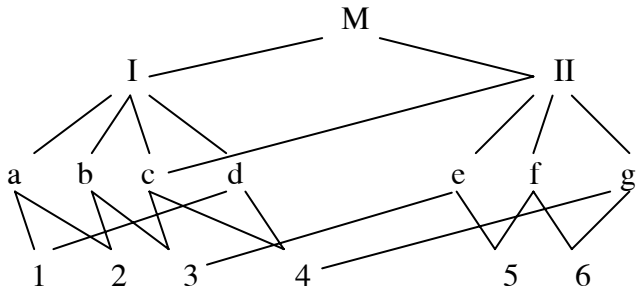
b) Vùng I và II



c) Cấu trúc dữ liệu Phân cấp



d) Cấu trúc dữ liệu Mạng



Hình 2.1. Các cấu trúc dữ liệu địa lý Mạng và Phân cấp

II.3.3. Cấu trúc dữ liệu Quan hệ

Cấu trúc dữ liệu quan hệ tổ chức dữ liệu theo dạng các bảng hai chiều, trong đó mỗi bảng là một tệp riêng biệt. Mỗi hàng của bảng là một bản ghi, và mỗi bản ghi có một tập hợp các thuộc tính. Mỗi cột của bảng biểu thị một thuộc tính. Các bảng khác nhau có thể được liên hệ với nhau thông qua một chỉ số chung thường được gọi là ***khoá.*** Các thông tin được khai thác thông qua phương thức tra vấn. Trong trường hợp bản đồ M, cách tổ chức dữ liệu theo cấu trúc quan hệ được minh họa trên hình 2.2.

Cấu trúc dữ liệu quan hệ rất mềm dẻo, nó có thể thỏa mãn được tất cả các yêu cầu mà phải được công thức hóa bởi sử dụng các luật của logic bool và các thao tác toán học. Chúng cho phép các loại dữ liệu khác nhau được tìm kiếm, so sánh. Việc bổ sung và di chuyển các mục dữ liệu dễ dàng. Có điều bất tiện là nhiều thao tác đòi hỏi tìm kiếm tuần tự. Đối với cơ sở dữ liệu lớn mất nhiều thời gian tìm kiếm. Tuy nhiên, với những máy tính có cấu hình mạnh hiện nay, đây không còn là vấn đề lớn đối với việc quản lý một cơ sở dữ liệu GIS.

- Bản đồ

M	I	II

- Vùng

I	a	b	c	d
II	c	e	f	g

- Đường

I	a	1	2
I	b	2	3
I	c	3	4
I	d	4	1
II	e	3	5
II	f	5	6
II	g	6	4
II	c	4	3

Hình 2.2. Cấu trúc dữ liệu quan hệ

III. ỨNG DỤNG CỦA HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU

Việc tổ chức một hệ thống thông tin hay xây dựng một cơ sở dữ liệu cho nghành khoa học, nghành kinh tế nào đó trong những năm 90 đã trở nên thông dụng. Điều này chứng tỏ khả năng ứng dụng rộng rãi của ngành khoa học này trong các nghành khác. Nhiều cơ sở dữ liệu ngành hay cơ sở dữ liệu quốc gia được thiết kế. Tuy không giới thiệu được hết các ứng dụng của cơ sở dữ liệu, người ta cũng có thể kể ra vài ứng dụng tiêu biểu như sau:

- Tổ chức thông tin trong các bài toán khoa học kỹ thuật,
- Kho dữ liệu trong hệ thống thông tin quản lý,
- Tổ chức dữ liệu có cấu trúc phức tạp như các dữ liệu địa lý,
- Cơ sở dữ liệu trong các hệ thống hỗ trợ công nghiệp, hỗ trợ giảng dạy,
- Tổ chức thông tin đa phương tiện, xử lý tri thức.

Thứ nhất, cơ sở dữ liệu ứng dụng trong các bài toán khoa học kĩ thuật. Các bài toán này có thuật toán khó; thường thì không đòi hỏi công cụ tốt nhất về tổ chức dữ liệu.Tuy nhiên, trong các bài toán phức tạp hơn; với nhiều dữ liệu trung gian thì cách tổ chức dữ liệu hợp lý là điều không thể không nghĩ đến.

Thứ hai, ứng dụng của cơ sở dữ liệu trong quản lý. Công tác quản lý không cần thuật toán phức tạp, nhưng đòi hỏi xử lý nhiều dữ liệu. Khối lượng lớn thông tin cần được tổ chức có khoa học để tiện cho quá trình xử lý. Hình dung như con người ta với khối lượng thông tin vừa phải còn bao quát được, chứ quá nhiều thông tin không có tổ chức, làm sao mà xem xét hết được.

Thứ ba, ứng dụng trong hệ thông tin địa lý. Các nghành khoa học không phải là công nghệ thông tin, thí dụ như vật lý, hóa học, sinh học, ngôn ngữ...cũng có các nhu cầu lưu trữ, xử lý dữ liệu. Các cơ sở dữ liệu riêng biệt này mang những đặc tính riêng của từng nghành. Các dữ liệu về địa lý, bao gồm các bảng số, các ảnh, các phương pháp truy nhập đến các kho dữ liệu...cần được tổ chức và xử lý hợp lý. Các dữ liệu địa lý, địa chất, thuỷ văn, môi trường...thường đòi hỏi các phương tiện nhớ có dung lượng lớn và được xử lý trên các bộ xử lý đặc biệt để đảm bảo tốc độ cao.

Thứ tư, cơ sở dữ liệu ứng dụng trong hệ thống hỗ trợ. Việc tổ chức lưu trữ và xử lý dữ liệu cũng có nhu cầu trong các ứng dụng có sử dụng hệ chuyên gia, người máy, xử lý các quá trình công nghiệp. Hơn nữa, trong đề án máy tính các thế hệ sau này, máy cơ sở dữ liệu có vị trí đáng kể. Riêng nhu cầu này, cơ sở dữ liệu cần có khả năng cơ giới hóa việc tìm kiếm thông tin nhờ cơ chế suy luận tự động. Vấn đề thời gian thực trong cơ sở dữ liệu được giải quyết để phù hợp với các hệ thống công nghiệp. Thời gian có thể được thực hiện trong cơ sở dữ liệu thông qua hai cách:

- Thời gian tương đối trong hệ quản trị cơ sở dữ liệu, liên quan đến thay đổi trạng thái của cơ sở dữ liệu
- Thời gian tuyệt đối của môi trường được mô tả trong cơ sở dữ liệu, liên quan đến trạng thái của môi trường.

Kiến thức về cơ sở dữ liệu còn dùng để tổ chức cơ sở tri thức, thiết lập hệ thống câu hỏi, chọn mô hình trong hệ thống hỗ trợ giảng dạy, hay trong công nghệ dạy học

Cuối cùng, cơ sở dữ liệu ứng dụng trong hệ thống đa phương tiện, xử lý tri thức. Việc xây dựng cơ sở dữ liệu đa phương tiện không thể không đề cập giao diện người dùng trong cơ sở dữ liệu, đề cập các nghiên cứu về quan hệ và sự kiện, đề cập việc tổ chức các câu hỏi cho người sử dụng. Người ta nhận thấy không có ngôn ngữ nào là đặc biệt quan trọng và ưu điểm trội hơn hẳn, ngay cả ngôn ngữ đồ thị. Một giao diện hiển thị thường được người ta ưa chuộng, với khả năng

(i) Đưa ra câu trả lời cho các câu hỏi dạng hiển thị như đồ thị, lược đồ, có tác dụng nhấn mạnh trực giác,

(ii) "Lật trang" của một cơ sở dữ liệu, tức khả năng lựa chọn thông tin nhanh một cách tự nhiên,

(iii) Tìm kiếm trong cơ sở dữ liệu theo phương thức con người đã quen thuộc, chẳng hạn theo cách tìm sách trong các tủ sách thư viện.

Trong số các giao diện người dùng, giao diện đa hình thái (multimodal), giao diện dùng ngôn ngữ tự nhiên được quan tâm và nay cũng có nhiều kết quả đáng khích lệ.

Tuy không được xây dựng như hệ thống tri thức hay hệ chuyên gia, cơ sở dữ liệu có thể mô tả và xử lý các tri thức . Một thế hệ mới của các cơ sở dữ liệu được xây dựng với các hệ thống cơ sở dữ liệu đa dạng, trong đó có cơ sở dữ liệu suy diễn. Các tri thức xử lý được thể hiện dưới các dạng:

1. Tri thức tổng quát như các luật và sự kiện
2. Các điều kiện thay đổi, hoặc kích hoạt dữ liệu
3. Suy diễn các thông tin có liên hệ với các sự kiện và luật

Ngoài ra, người ta còn đề cập khía cạnh về xử lý các tri thức không đầy đủ.

IV. LỊCH SỬ CỦA CÁC HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU

Điểm các mốc từ năm 1960 đến nay, người ta có thể thấy được lịch sử của hệ quản trị cơ sở dữ liệu:

- Những năm 60 gắn với các sản phẩm đầu tiên của hệ quản lý tệp, xuất hiện bộ nhớ ngoài như là bộ nhớ lý tưởng. Bộ nhớ này cho phép dùng chung, dễ dàng sử dụng, cho phép đánh địa chỉ trực tiếp và có dung lượng lớn. Các hệ thống dùng ngôn ngữ lập trình để xử lý dữ liệu.. Những chương trình viết ra bằng ngôn ngữ lập trình đó tạo ra hệ quản trị tệp, hay là bước đầu của hệ quản trị cơ sở dữ liệu.
- Giữa những năm 60, thế hệ đầu của hệ quản trị cơ sở dữ liệu đánh dấu bằng việc phân rã, mô tả những dữ liệu của chương trình ứng dụng và ngôn ngữ truy nhập bên trong. Bằng các lệnh hỏi phi thủ tục, người ta có thể truy nhập dữ liệu, tìm đến các bản ghi thay vì phải đi theo cấu trúc lưu trữ vật lý của các dữ liệu. Đại diện của các hệ thống này là CODASYL và IMS. Chúng dựa trên mô hình truy nhập, tức các mô hình sử dụng nhiều chức năng xử lý dữ liệu của hệ thống điều hành của máy tính và có tính đến việc tối ưu phương pháp phân phối bộ nhớ phụ, tăng tốc khai thác dữ liệu.
- Từ những năm 70, có thế hệ thứ hai của hệ quản trị cơ sở dữ liệu với mô hình quan hệ. Mô hình quan hệ giúp đơn giản hóa việc truy nhập dữ liệu của người sử dụng bên ngoài. Nó có ngôn ngữ truy nhập dữ liệu dựa trên logic, xác định được dữ liệu mà không cần mô tả cách tiếp cận. Chính hệ thống quản trị đặt kế hoạch truy nhập dữ liệu.

  Đầu năm 1980 mới xuất hiện những hệ thống quản trị cụ thể của loại này. Mô hình quan hệ có phần "bên trong" phong phú lên, nhưng đơn giản hóa mô hình ngoài để tiện cho người dùng. Những dữ liệu được thể hiện dưới dạng quan hệ với các miền

giá trị hoặc đơn giản qua các bảng. Việc tìm kiếm trong các hệ thống quan hệ là thuận lợi nhờ ngôn ngữ phi thủ tục, cho phép truy nhập dữ liệu mà không cần mô tả cách truy nhập dữ liệu. Thế hệ hai của hệ quản trị cơ sở dữ liệu cùng với việc mở rộng các hệ thống truy nhập của thế hệ thứ nhất đã góp phần tối ưu hoá việc khai thác dữ liệu.

Các hệ quản trị cơ sở dữ liệu bắt đầu được thương mại hoá từ năm 1982. Các hệ thống tiêu biểu gồm ORACLE, INGRES, SYBASE, INFORMIX, DB2 và RDB. Nhìn chung chúng có kiến trúc phân tán, tức là hoạt động theo nguyên lý các máy trạm khách hàng chuyển yêu cầu về máy chủ. Cơ sở dữ liệu được quản lý trên máy chủ.

- Thế hệ ba của hệ quản trị cơ sở dữ liệu được phát triển từ những năm 80 trong phòng thí nghiệm. Chúng dùng các mô hình dữ liệu phong phú và kiến trúc phân tán hơn so với các hệ thống trước. Kiến trúc này cho phép người dùng liên hệ với nhau tốt hơn. Thế hệ ba có thể kể ra gồm:
  - Mô hình hướng đối tượng,
  - Mô hình với các luật suy diễn như là mô hình hóa logic các dữ liệu,
  - Cơ sở dữ liệu phân tán

Chương 2

CƠ SỞ DỮ LIỆU TRONG NGHIÊN CỨU MÔI TRƯỜNG

I. HIỆN TRẠNG QUẢN LÝ DỮ LIỆU

Nhìn chung, vấn đề thu thập, lưu trữ và xây dựng cơ sở dữ liệu thường được triển khai thực hiện trong khuôn khổ các chương trình nghiên cứu khoa học và công nghệ. Mặc dù các chương trình nghiên cứu này luôn có sự tham gia của rất nhiều cơ quan nghiên cứu thuộc nhiều bộ, ngành khác nhau và của đông đảo các nhà khoa học, vấn đề quản lý các thông tin và dữ liệu theo một quy chế tập trung thường gặp rất nhiều khó khăn, đặc biệt là ở những quốc gia chưa có được những trung tâm dữ liệu với đầy đủ chức năng và cơ chế tập trung mạnh về quản lý, xử lý và trao đổi dữ liệu. Những khó khăn nêu trên thường bắt nguồn từ những nguyên nhân có thể mô tả tóm lược dưới đây.

Trước hết, cần phải nhấn mạnh đến tính *phân tán* của các dữ liệu hiện có. Các dữ liệu đo đạc, quan trắc và được tổng hợp từ những chuyến khảo sát, các chương trình, đề tài nghiên cứu, v.v… được lưu trữ rải rác và tồn tại trong khoảng thời gian dài tại các cơ sở nghiên cứu. Do hạn chế thông tin và không có những quy chế chính thức về trao đổi dữ liệu và bản quyền tác giả, các dữ liệu này do đó có thể sẽ vĩnh viễn tồn tại trong các kho lưu trữ, hoặc trở thành dữ liệu riêng của một số ít người, hay sẽ trở nên lỗi thời và mất dần giá trị sử dụng với thời gian.

Cũng vì những nguyên nhân kể trên mà hàng loạt những vấn đề nảy sinh liên quan tới *sự trùng lặp dữ liệu* và *bản quyền dữ liệu*. Do không có sự phối hợp giữa các cơ quan nên các dữ liệu đo đạc phục vụ các đề tài khác nhau nhiều khi bị trùng lặp, gây lãng phí cho nhà nước, đặc biệt là trong những trường hợp khảo sát đo đạc bằng các thiết bị đắt tiền và kéo dài nhiều ngày. Mặt khác, việc không có một quy chế chính thức về dữ liệu ở tầm cỡ quốc gia cũng sẽ dẫn đến tình trạng sao chép tuỳ tiện các dữ liệu, hay ngược lại, sẽ có quá nhiều thủ tục phiền hà, gây khó khăn cho những người sử dụng trong việc truy cập vào các cơ sở dữ liệu hiện có với những mục đích khác nhau.

Tình trạng lạc hậu, phi tin học cũng là một đặc trưng cơ bản trong công tác thu thập và quản lý dữ liệu ở nhiều nơi. Trong một thời gian dài việc kiểm kê các dữ liệu chỉ dừng lại ở các bản báo cáo, các bảng liệt kê hay bản đồ minh hoạ vẽ trên giấy.

Cuối cùng, khó khăn trong việc sử dụng và trao đổi dữ liệu có thể do các cơ sở dữ liệu được xây dựng mà không tham khảo những *khuôn dạng thống nhất và chuẩn hoá* để quản lý các thông tin dữ liệu trong khuôn khổ quốc gia, khu vực và thế giới.

II. DỮ LIỆU SỬ DỤNG TRONG NGHIÊN CỨU MÔI TRƯỜNG

Thông tin và dữ liệu cần thiết cho việc xây dựng một cơ sở dữ liệu thường hết sức đa dạng, bao gồm nhiều khuôn dạng, thể loại và hình thức lưu trữ rất khác nhau. Tuy nhiên, toàn bộ tập dữ liệu ban đầu có thể phân ra thành ba loại dữ liệu chính sau đây:

1) Thông tin về dữ liệu (Metadata), bao gồm tất cả các văn liệu, chuyên khảo hay tài liệu dạng mô tả liên quan đến khu vực nghiên cứu và đối tượng nghiên cứu. Các dữ liệu dạng này còn được gọi là dữ liệu về dữ liệu. Một Thư mục thông tin về dữ liệu sẽ giúp cho người sử dụng cơ sở dữ liệu xác định được ai có dữ liệu gì, ở đâu. Ngoài ra, thư mục này cũng cung cấp các thông tin liên quan đến chất lượng dữ liệu, phương pháp thu thập và khuôn dạng dữ liệu.

2) Dữ liệu thực (Actual Data), bao gồm các dữ liệu đo đạc và quan trắc được tại khu vực nghiên cứu;

3) Dữ liệu không gian (Spatial Data), bao gồm tư liệu ảnh, bản đồ, sơ đồ, đồ thị và các sản phẩm dữ liệu thứ sinh dưới dạng đồ hoạ của khu vực nghiên cứu. Dạng dữ liệu này có thể được gọi là *dữ liệu GIS* (GIS Data).

III. ƯU ĐIỂM CỦA CƠ SỞ DỮ LIỆU

Cơ sở dữ liệu là một hợp phần quan trọng của mỗi một dự án có khuôn khổ bao trùm những khoảng thời gian và không gian rộng lớn. Cơ sở dữ liệu không chỉ quan trọng từ góc độ lưu trữ một khối lượng lớn dữ liệu, mà còn từ góc độ đảm bảo các chuẩn mực về tính ổn định dữ liệu, cho phép dễ dàng bảo vệ và sử dụng dữ liệu. Các dữ liệu dạng ghi chép có thể tiện lợi sử dụng trong khoảng thời gian ngắn, nhưng trong thực tế, chúng không cho phép làm việc hiệu quả với các tập dữ liệu lớn hay phức tạp.

Thiết kế cơ sở dữ liệu là bước đầu tiên và cũng là một trong những bước quan trọng nhất của quy trình xây dựng một cơ sở dữ liệu. Một cơ sở dữ liệu được thiết kế tốt sẽ tạo điều kiện cho các thao tác nhập liệu dễ dàng và cho phép truy xuất dữ liệu nhanh, hiệu quả. Thiết kế cơ sở dữ liệu là một quá trình lặp đi lặp lại cho đến khi cơ sở dữ liệu thoả mãn các yêu cầu của các dữ liệu thu thập được cũng như nhu cầu của người sử dụng.

Các tập dữ liệu lớn (chứa dữ liệu thu thập được trong một phạm vi rộng lớn về không gian và thời gian) đòi hỏi một hệ thống quản trị cơ sở dữ liệu trên máy tính. Dưới đây liệt kê những ưu điểm vượt trội của một cơ sở dữ liệu được xây dựng và quản lý trên máy tính nếu đem so sánh với các tập dữ liệu được thu thập bằng các phương pháp thủ công, phi tin học (mà ta tạm gọi là các số liệu dạng ghi chép):

- ***Tính ổn định dữ liệu:*** Các cơ sở dữ liệu thường có cấu trúc xác định, sẽ giúp cho tính ổn định của các dữ liệu lưu trữ trong đó. Quá trình thiết kế cơ sở dữ liệu và phân tích sơ bộ các dữ liệu đưa vào cơ sở dữ liệu sẽ tạo ra cấu trúc cho cơ sở dữ liệu. Các cơ sở dữ liệu có cùng cấu trúc có thể được nối kết rất dễ dàng, cho phép gộp dữ liệu từ nhiều nguồn khác nhau và được thu thập trong những khoảng thời gian khác nhau về cùng một cơ sở dữ liệu lớn.
- ***Tính hiệu quả:*** Các cơ sở dữ liệu cho phép làm việc với một khối lượng lớn các dữ liệu. Các hệ cơ sở dữ liệu quan hệ có chức năng lưu trữ rất hiệu quả do loại trừ được các dữ liệu trùng lặp.
- ***Chất lượng dữ liệu:*** Nhiều đặc tính của cơ sở dữ liệu cho phép kiểm soát được chất lượng dữ liệu. Chẳng hạn, giao diện nhập liệu trên màn hình giúp cho những người nhập dữ liệu chưa có nhiều kinh nghiệm, các chương trình kiểm tra cho

phép phát hiện và loại trừ lỗi và sai số, và cấu trúc nền của cơ sở dữ liệu đảm bảo tính ổn định dữ liệu.

- ***Phân tích dữ liệu:*** Các cơ sở dữ liệu tạo ra những cổng nối tới các phần mềm đóng gói khác như các chương trình thống kê hay các phần mềm trợ giúp cho công tác văn phòng. Phần lớn các phần mềm đóng gói này cho phép làm việc trực tiếp với cơ sở dữ liệu hoặc với các tệp dữ liệu kết xuất từ cơ sở dữ liệu.
- ***Tích hợp dữ liệu:*** Cấu trúc của cơ sở dữ liệu xác lập các tiêu chuẩn cho phép nối kết nhiều tập dữ liệu khác nhau. Nhờ thế, các tập dữ liệu đơn lẻ có thể được tích hợp thành các cơ sở dữ liệu ở phạm vi khu vực hay quốc tế, dựng nên bức tranh toàn cảnh của các tập dữ liệu.

Trước đây, hình thức lưu trữ các dữ liệu dạng ghi chép đã tồn tại và được coi là rất phổ biến trong một thời gian dài. Tính linh hoạt và dễ sử dụng của các dữ liệu dạng ghi chép thường khiến cho người ta có thiên hướng dùng phương thức này để lưu trữ các dữ liệu. Mặc dù có vẻ tiện lợi khi sử dụng các dữ liệu ghi chép, chẳng hạn, bạn không phải thiết lập các bảng hay các mối quan hệ, nhưng các dữ liệu dạng ghi chép rất không thích hợp với các tập dữ liệu lớn và có thể làm ảnh hưởng đáng kể tới tính ổn định và tính tích hợp dữ liệu. Dưới đây là một vài ví dụ chứng minh những nhược điểm của các dữ liệu dạng ghi chép:

- ***Tính ổn định dữ liệu:*** Chính tính linh hoạt khiến cho các dữ liệu dạng ghi chép dễ sử dụng lại gây ra khó khăn trong việc duy trì và củng cố tính ổn định của chúng. Chẳng hạn, một bảng số liệu dạng ghi chép có thể cho phép ghi nhiều giá trị khác loại nhau trong cùng một cột (như ghi lẫn lộn các giá trị số với ngày tháng, các giá trị số với các kí tự dạng văn bản, v.v…). Trong khi đó, một cơ sở dữ liệu với một cấu trúc đã được xác lập sẽ không cho phép sự pha trộn đó, và vì thế sẽ phát hiện rất nhanh chóng các giá trị sai quy tắc và cho phép tự động kiểm tra các dữ liệu nhập vào cơ sở dữ liệu.
- ***Tích hợp dữ liệu:*** Các khó khăn trong việc bảo tồn tính ổn định dữ liệu trong trường hợp sử dụng các dữ liệu dạng ghi chép cũng gây khó khăn trong việc tích hợp các tập dữ liệu được lưu trữ ở dạng này. Các cơ sở dữ liệu tuân thủ một cấu trúc đã định trước, là nền tảng cho việc tích hợp các tập dữ liệu khác nhau về các tập dữ liệu ở phạm vi khu vực hay quốc tế.
- ***Tốc độ:*** Các cơ sở dữ liệu cho phép làm việc hiệu quả với một khối lượng lớn dữ liệu, do chúng có các chức năng thiết lập chỉ số và các thuật toán tìm kiếm chuyên biệt cho phép nhanh chóng tìm kiếm và hiển thị dữ liệu. Một tập dữ liệu dạng ghi chép không thể có các chức năng này, do vậy sẽ khiến cho người sử dụng gặp vất vả khi phải tìm kiếm dữ liệu trong một tập dữ liệu lớn. Phần lớn các cơ sở dữ liệu hiện đại có thể chứa được rất nhiều dữ liệu trong các đĩa của máy tính, trong khi điều này là hạn chế đối với các dữ liệu dạng ghi chép.
- ***Kết xuất dữ liệu:*** Sức mạnh thực sự của một cơ sở dữ liệu là khả năng truy cập dữ liệu trên cơ sở các tra vấn nhiều khi khá phức tạp. Các cơ sở dữ liệu thường chứa các ngôn ngữ tra vấn ngầm định và hỗ trợ các cấu trúc, chẳng hạn như một cơ sở dữ liệu quan hệ có thể tạo ra các tra vấn rất phức tạp, nhờ đó tạo ra khả năng truy

cập tối đa tới dữ liệu. Các dữ liệu dạng ghi chép thường không có chức năng tra vấn này.

- ***Khả năng lập trình:*** Các cơ sở dữ liệu thường có các ngôn ngữ lập trình ngầm định, bao gồm cả các ngôn ngữ tra vấn phức tạp. Chúng cũng cho phép tạo ra các màn hình nhập liệu hay báo biểu và thường kèm theo các đơn thể chương trình tính toán thống kê ngầm định. Các chứ năng ngầm định của các dữ liệu dạng ghi chép thường yếu hơn nhiều.

Chương 3

THÔNG TIN DỮ LIỆU

I. KHÁI NIỆM METADATA

Metadata là một thuật ngữ thường được sử dụng thay cho cụm từ *thông tin dữ liệu*. Đây là một khái niệm hiện đại và khá mới mẻ trong lĩnh vực nghiên cứu cơ sở dữ liệu ở nước ta. Một cách ngắn gọn nhất, *Metadata* được định nghĩa như là dữ liệu về dữ liệu, tức là sự mô tả các đặc trưng của dữ liệu được thu thập cho một lĩnh vực chuyên môn nào đó. Từ đây ta có khái niệm về cơ sở thông tin dữ liệu *(Metadatabase)*. Thông thường, các cơ sở thông tin dữ liệu trả lời cho câu hỏi "ai có dữ liệu gì, ở đâu?". Một trong những ví dụ đơn giản nhất của một cơ sở thông tin dữ liệu có thể kể đến là thư mục danh bạ điện thoại mà ta còn hay gọi là những trang vàng. Không phải ngẫu nhiên mà các thư mục thông tin dữ liệu lớn trên thế giới hiện nay thường có tên gọi như "Những trang xanh lá cây", "Những trang xanh nước biển", hay thậm chí "Những trang trắng"...

II. ƯU ĐIỂM CỦA METADATA

Metadata đang được sử dụng rộng rãi trong khu vực và trên thế giới, đặc biệt là trong lĩnh vực quản lý và trao đổi dữ liệu hải dương học và môi trường, do có những điểm mạnh sau đây:

- Metadata là công cụ vô giá để quản lý dữ liệu thông qua việc cung cấp cho người sử dụng những thông tin đầy đủ nhất liên quan đến những dữ liệu mà họ quan tâm. Thông tin trong Cơ sở dữ liệu Metadata và phần mềm quản lý được cung cấp trực tiếp đến tay người dùng mà không tốn tiền mua như đối với một số loại dữ liệu hay phần mềm khác.
- Thông tin về dữ liệu được chuyển đến người sử dụng thông qua một hệ tham chiếu, do đó sẽ không gặp phải những rắc rối về bản quyến hay trùng lặp dữ liệu.
- Việc áp dụng hệ thống Metadata sẽ tránh được những đòi hỏi về một cơ chế tập trung đối với việc quản lý các dữ liệu thực, do đó giảm nhẹ đáng kể những chi phí cho việc tổ chức hay xây dựng những Trung tâm dữ liệu lớn với cấu trúc đồ sộ mà vẫn đáp ứng được các nhu cầu sử dụng dữ liệu của nhiều đối tượng khác nhau.

III. THƯ MỤC METADATA

Thông tin về dữ liệu được lưu trữ và quản lý trong các Thư mục Metadata. Đây là thư mục chứa toàn bộ các thông tin mô tả các tập dữ liệu và việc thu thập chúng. Thư mục cũng cung cấp các thông tin chi tiết về tất cả các tập dữ liệu hiện có và ai là người cần liên hệ để có được những dữ liệu cần thiết.

Trong số các dữ liệu đã được thu thập cho một khu vực nghiên cứu, có nhiều dữ liệu không được công bố do nhiều lý do. Tuy nhiên, điều này không có nghĩa là không thể

khai thác các dữ liệu đó bằng cách này hay cách khác. Các thư mục metadata, với các công cụ tìm kiếm nhanh và hiệu quả luôn luôn có thể giúp người sử dụng dữ liệu tìm ra và khai thác các dữ liệu loại này. Thậm chí cả các dữ liệu không gian cũng có thể được tìm kiếm nhờ các công cụ tra vấn không gian, bởi các thư mục metadata thường bao hàm cả các thông tin về vị trí địa lý của các khu vực nghiên cứu.

Khi làm việc với một thư mục Metadata, người sử dụng có thể đánh giá được thông tin nào là cần thiết đối với mình và khả năng truy cập tới nguồn dữ liệu mà mình cần. Một thư mục Metadata cũng có thể được sử dụng như một phương tiện quảng bá các sản phẩm hay dịch vụ liên quan đến dữ liệu.

Quy trình xây dựng Thư mục Meatadata thường bao gồm các bước chính như sau:

1) *Thu thập thông tin dữ liệu dưới dạng các phiếu điều tra*. Các phiếu điều tra bao gồm các đề mục để trống được phổ biến tới những cơ sở hoặc cá nhân làm công tác nghiên cứu, các chuyên gia, các nhà quản lý dữ liệu liên quan tới đối tượng hay/và khu vực nghiên cứu. Tuỳ theo mức độ đầy đủ, metadata được điền vào các phiếu điều tra. Các phiếu điều tra sau khi đã điền đầy đủ sẽ được tập hợp lại để chuẩn bị nhập vào máy.
2) *Nhập và quản lý dữ liệu*. Metadata từ các phiếu điều tra được nhập vào máy, sử dụng các công cụ quản lý thông tin dữ liệu. Thông tin dữ liệu trong thư mục sẽ được cập nhật thường xuyên và cất giữ định kỳ trong khuôn dạng an toàn.

IV. KHUÔN DẠNG CHUẨN TRAO ĐỔI METADATA

Kinh nghiệm cho thấy rằng, việc giảm thiểu hay tránh được quá trình chuyển đổi dữ liệu từ một khuôn dạng này sang khuôn dạng khác có thể tiết kiệm được từ hàng vài trăm đến hàng vài nghìn giờ làm việc tại các trung tâm dữ liệu, đó là chưa kể đến các khoản chi phí khổng lồ khác.Vì thế, việc lựa chọn một khuôn dạng chuẩn để trao đổi thông tin dữ liệu đóng vai trò hết sức quan trọng. Thông thường, các thư mục metadata và công cụ quản lý chúng được thiết kế và xây dựng dựa trên cơ sở của một trong số các quy chuẩn trao đổi dữ liệu đã và đang được thế giới công nhận và sử dụng rộng rãi.

Trong số các quy chuẩn trao đổi metadata hiện đang thịnh hành trên thế giới hiện nay, đáng chú ý nhất là các quy chuẩn sau đây:

1) Quy chuẩn metadata của Mỹ, do Uỷ ban dữ liệu địa lý liên bang Hoa kỳ *(FGDC)* xây dựng. Đây là một quy chuẩn rất đồ sộ, bao gồm tới 220 mục, nhằm mô tả các dữ liệu không gian đã số hoá và sử dụng đa mục đích.
2) Quy chuẩn metadata của Ốxtrâylia-Niu Di lân, thường gọi là *ANZLIC*, do Hội đồng thông tin về đất đai của Ốxtrâylia và Niu Di lân xây dựng. Quy chuẩn này gọn nhẹ hơn nhiều so với quy chuẩn của Mỹ, chỉ gồm 67 mục, với nội dung bám sát các thông tin cô đọng và thiết thực nhất về tập dữ liệu.
3) Các quy chuẩn metadata do Ốxtrâylia xây dựng gần đây, tiêu biểu là quy chuẩn có tên gọi *Những trang Xanh nước biển (the Blue Pages),* và gần đây nhất là quy chuẩn *MEDI,* viết tắt từ tên gọi *kiểm kê dữ liệu môi trường biển (MarineEnvironmental Data Inventory),* một dự án của tổ chức quốc tế về trao đổi thông tin dữ liệu hải dương học *(IODE)*. Các quy chuẩn này đều lấy *ANZLIC* làm nền tảng, có bổ sung thêm một số mục từ các quy chuẩn trao đổi dữ liệu hải dương học khác như *GF3*.

Quy chuẩn *MEDI* đã được *IODE* công nhận là quy chuẩn metadata cho toàn khu vực Tây Thái Bình dương.

Trong bảng 1 minh hoạ quy chuẩn trao đổi thông tin dữ liệu MEDI. Các mục của quy chuẩn được sử dụng để xây dựng các trường nhập liệu trong phần mềm quản lý thư mục thông tin dữ liệu về môi trường biển áp dụng cho Việt nam.

Bảng 1. Nội dung các trường sử dụng trong phần mềm MEDI Vietnam

Phân loại	*Tên trường*	*Nội dung*
Tập dữ liệu	*Tên tập dữ liệu* *Cơ quan có dữ liệu* *Nước có dữ liệu*	Tên đầy đủ của tập dữ liệu Tên cơ quan có dữ liệu Nước (hoặc bang) của cơ quan có dữ liệu
Mô tả	*Tóm tắt* *Từ khoá tìm kiếm* *Tên vùng địa lý* *Đa giác địa lý* *Toạ độ ranh giới cực nam* *Toạ độ ranh giới cực bắc* *Toạ độ ranh giới cực tây* *Toạ độ ranh giới cực đông*	Tóm tắt nội dung tập dữ liệu. Các từ khoá phản ánh những nội dung chính của tập dữ liệu. Tên vùng địa lý, nơi dữ liệu được thu thập. Một cách mô tả khác về vùng địa lý nếu không có tên vùng địa lý phù hợp. Vĩ độ nhỏ nhất của cạnh hoặc đỉnh của đa giác chứa tập dữ liệu Vĩ độ lớn nhất của cạnh hoặc đỉnh của đa giác chứa tập dữ liệu Kinh độ nhỏ nhất của cạnh hoặc đỉnh của đa giác chứa tập dữ liệu Kinh độ lớn nhất của cạnh hoặc đỉnh của đa giác chứa tập dữ liệu
Quá trình tiến triển	*Ngày bắt đầu* *Ngày kết thúc*	Ngày đầu tiên thu thập dữ liệu. Ngày kết thúc thu thập dữ liệu.
Trạng thái dữ liệu	*Tiến trình* *Tần suất bảo trì và cập nhật*	Tiến triển của quá trình xây dựng tập dữ liệu. Tần suất bảo trì và cập nhật của tập dữ liệu.
Truy cập dữ liệu	*Định dạng dữ liệu đang lưu trữ* *Loại định dạng dữ liệu hiện có* *Hạn chế dữ liệu*	Một hay nhiều định dạng mà tập dữ liệu được lưu trữ bởi cơ quan có dữ liệu. Một hay nhiều định dạng được sử dụng trong tập dữ liệu. Những hạn chế áp dụng cho việc sử dụng tập dữ liệu.

Chất lượng dữ liệu	*Truyền thống*	Mô tả các bước xử lý được áp dụng trong quá trình xây dựng tập dữ liệu.
	Độ chính xác vị trí	Đánh giá độ chính xác về vị trí của tập dữ liệu
	Độ chính xác thuộc tính	Đánh giá độ chính xác về thuộc tính của tập dữ liệu
	Bền vững lô gích	Đánh giá độ bền vững lô gích của tập dữ liệu
	Tính đầy đủ	Đánh giá về tính đầy đủ của tập dữ liệu
Thông tin liên hệ	*OIN* *Cơ quan cần liên hệ* *Chức vụ của người cần liên hệ* *Người cần liên hệ* *Địa chỉ gửi thư* *Địa phương* *Bang* *Nước* *Mã bưu điện* *Điện thoại* *Fax* *E-mail* *WWW*	Mã số của cơ quan có dữ liệu Tên cơ quan. Chức vụ trong cơ quan Tên đầy đủ của người cần liên hệ Địa chỉ gửi thư của cơ quan có dữ liệu Lân cận hoặc vị trí Bang hoặc khu vực hành chính tương đương Tên nước Mã bưu điện Số điện thoại cần liên hệ Số Fax cần liên hệ Địa chỉ thư điện tử cần liên hệ Địa chỉ trang Web
Thông tin về metadata	*Ngày nhập Metadata* *Người nhập Metadata* *Địa chỉ thư điện tử* *Cơ quan nhập Metadata*	Ngày mà thông tin về dữ liệu được nhập vào hoặc cập nhật lần cuối. Tên người nhập hoặc cập nhật lần cuối thông tin về dữ liệu. Địa chỉ thư điện tử của người nhập thông tin về dữ liệu Tên cơ quan của người nhập thông tin về dữ liệu.
Thông tin về Chương trình	*Tên chương trình* *Điều phối viên chương trình* *Cơ quan điều phối chương trình* *Trạm thu thập dữ liệu*	Tên của chương trình dự án đã thu thập dữ liệu Tên của điều phối viên chương trình Tên của tổ chức điều phối chương trình Tên của trạm thu thập dữ liệu chính (nếu có)

Nội dung dữ liệu	*Thiết bị*	Trang thiết bị sử dụng để lấy mẫu và phân tích các dữ liệu thu thập được.
	Mô tả tham số	Mô tả các đại lượng ghi được hay đo được.
	Phương pháp lấy mẫu	Phương pháp được sử dụng để lấy mẫu
	Cường độ lấy mẫu	Số mẫu, tuyến đo, điểm đo, chu kỳ dữ liệu, ... trong tập dữ liệu.
	Mô tả các môi trường sống sinh vật	Các vùng môi trường sống sinh vật liên quan đến tập dữ liệu.
	Các nhóm độc hại	Các nhóm phân loại chính được trình bày trong tập dữ liệu.
Thông tin về xuất bản phẩm	*Tài liệu tham khảo*	Danh sách các xuất bản phẩm, báo cáo liên quan
	Nối kết trực tuyến	Địa chỉ trên Internet để tham khảo trực tuyến các thông tin chi tiết hơn.
Giám sát	*DSIN*	Mã số của tập dữ liệu
	Cơ quan chủ trì	Tên của cơ quan chủ trì
	Cơ quan tham gia chính	Tên các cơ quan tham gia chính
	Cơ quan cộng tác	Tên các cơ quan cộng tác
	Tổ chức tài trợ	Tên tổ chức tài trợ
	Mục tiêu	Mục tiêu của chương trình được giám sát
	Khách hàng	Khách hàng của chương trình được giám sát

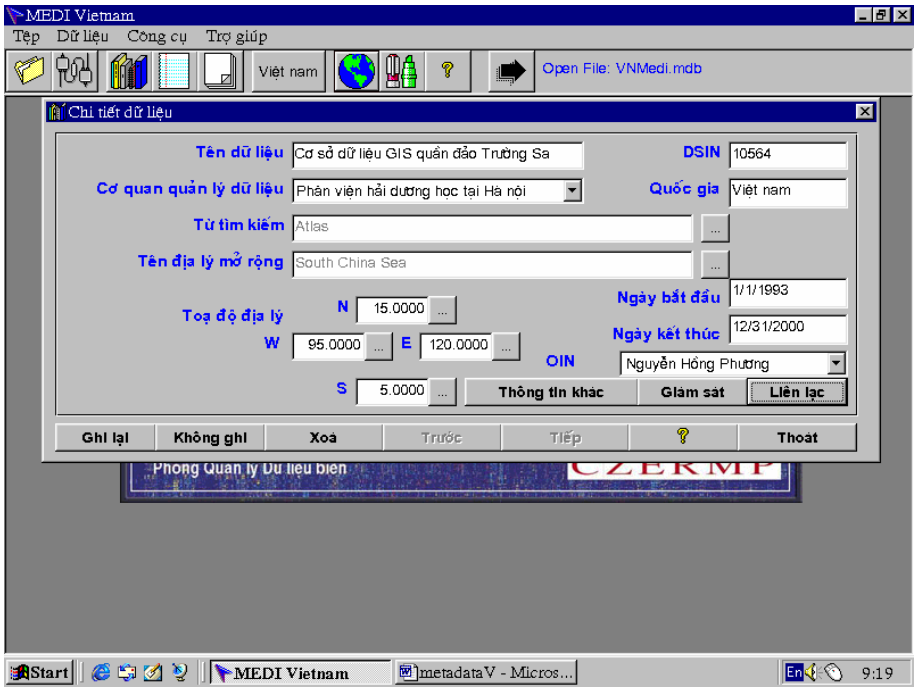
IV.5. Công cụ quản lý Metadata

Các thư mục Metadata thường được quản lý bằng một công cụ phần mềm, được thiết kế chuyên biệt cho một lĩnh vực nghiên cứu cụ thể. Ngoài việc áp dụng các chuẩn trao đổi thông tin dữ liệu đang được phổ biến rộng rãi trên trường quốc tế, công cụ này phải đảm bảo được một số chức năng quan trọng sau đây:

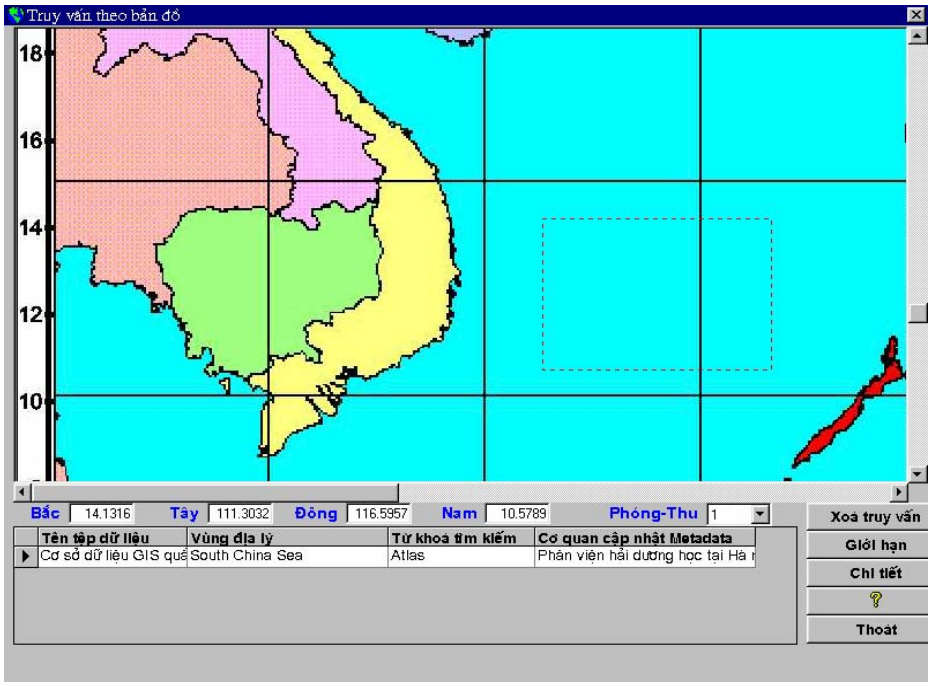
- Nhập, cập nhật dữ liệu theo khuôn dạng chuẩn ;
- Tìm kiếm, tra vấn dữ liệu nhanh, tiện lợi;
- Trao đổi, xuất-nhập khẩu dữ liệu trong khuôn khổ một số khuôn dạng chuẩn;
- Tự động tạo lập và in ấn báo biểu.

Các công cụ quản lý Metadata được xây dựng cho nhiều phạm vi sử dụng khác nhau, từ máy tính cá nhân, mạng máy tính cho đến các công cụ cho phép thao tác trên các Website trên Internet. Một trong số các phần mềm quản lý Metadata đang được sử dụng rộng rãi trên thế giới hiện nay là phần mềm MEDI, do tổ chức quốc tế về trao đổi thông tin dữ liệu hải dương học (IODE) xây dựng. Phần mềm này đã được Việt nam hoá và đưa

vào sử dụng ở Việt nam từ năm 1999 dưới tên gọi MEDI Việt nam. Trên các hình 3 và 4 minh họa một số giao diện đồ họa của phần mềm MEDI Việt nam.



Hình 3. Màn hình nhập liệu của MEDI Việt nam.



Hình 4. Màn hình truy vấn dữ liệu theo không gian của MEDI Việt nam.

Chương 4

QUẢN LÝ CÁC DỮ LIỆU THỰC

I. MỞ ĐẦU

Một trong những loại dữ liệu rất quan trọng được sử dụng trong nghiên cứu môi trường là các dữ liệu thực. Như đã nói tới ở trên, dữ liệu thực là các dữ liệu được thu thập, ghi nhận, đo đạc, quan trắc bằng máy trong các quá trình khảo sát ngoài thực địa, trên biển hay trong phòng thí nghiệm. Chúng tồn tại dưới dạng các giá trị số hoặc thông tin mô tả đặc tính của đối tượng nghiên cứu.

Công cụ tối ưu để quản lý các dữ liệu thực là các hệ quản trị cơ sở dữ liệu quan hệ. Chính vì vậy, trong chương này, mô hình cơ sở dữ liệu được xét đến một cách chi tiết. Ngoài ra, các bài tập thực hành về thiết kế một cơ sở dữ liệu thực trên Access, một trong những công cụ mạnh có sử dụng mô hình cơ sở dữ liệu quan hệ cũng được đưa vào nội dung chương.

II. CƠ SỞ DỮ LIỆU QUAN HỆ

II.1. Các khái niệm cơ bản

Để thiết kế và xây dựng các cơ sở dữ liệu dạng quan hệ, chúng ta cần làm quen với một số khái niệm cơ bản như *thực thể*, *quan hệ* và *thuộc tính*.

II.1.1. Thực thể: là sự thể hiện duy nhất của chỉ một đối tượng của thế giới thực. Thực thể được tạo bằng cách dùng các giá trị của các thuộc tính của nó theo dạng mà máy tính đọc được. (Ví dụ: ***Độ pH, Trạm đo, Chuyến khảo sát*** có thể là các thực thể trong một cơ sở dữ liệu về quan trắc môi trường).

II.1.2. Quan hệ: các quan hệ thể hiện sự liên hệ giữa hai hay nhiều thực thể. (Ví dụ: ***Quan trắc được tại*** là quan hệ liên kết hai thực thể ***Độ pH*** và ***Trạm đo***; hay ***Đo được trong chuyến khảo sát*** là quan hệ liên kết hai thực thể ***Độ pH*** và ***Chuyến khảo sát***).

II.1.3. Thuộc tính: các thuộc tính thể hiện các tính chất cơ bản của các thực thể hay các quan hệ. Mỗi thuộc tính mang một giá trị hỗ trợ cho việc định danh thực thể mà nó thuộc một phần trong đó và cho việc phân biệt thực thể đó với các phần tử khác của cùng lớp thực thể. (Ví dụ: ***Cao, Trung bình, Thấp*** là các thuộc tính của thực thể ***Độ pH***).

II.2. Mô hình cơ sở dữ liệu quan hệ

Mô hình cơ sở dữ liệu quan hệ được E.F. Codd giới thiệu lần đầu tiên năm 1970, cùng với việc đề ra những tiêu chuẩn thiết kế cấu trúc logic và một ngôn ngữ giành riêng cho các thao tác đối với loại cơ sở dữ liệu quan hệ. Cho đến nay, mô hình này đã được áp dụng khá rộng rãi, nhờ những ưu điểm chính có thể kể ra sau đây:

1. Quan hệ giữa các dữ liệu trong mô hình được hình dung trực quan dưới dạng các bảng hai chiều, trong đó mỗi loại thuộc tính được tương ứng với một cột, và mỗi tập giá trị được tương ứng với một hàng.
2. Thao tác trên các quan hệ khá đơn giản và có tính tổng hợp cao.
3. Thuận tiện trong việc ứng dụng các phép toán như đại số quan hệ, logic học, v.v..cho phép tăng đáng kể tốc độ tìm kiếm và xử lý dữ liệu.

II.3. Các tính chất của quan hệ

Mỗi bảng được coi là một quan hệ nếu có đầy đủ các tính chất sau đây:

1. Mỗi cột ứng với một thuộc tính có một tên gọi duy nhất;
2. Thứ tự các cột từ trái qua phải có thể thay đổi
3. Mỗi thuộc tính chỉ có một trị số đơn, mà không thể là một nhóm hay một mảng các trị số;
4. Các trị số nằm trong cùng một cột có cùng một tính chất;
5. Thứ tự từ trên xuống dưới các hàng cũng không bắt buộc ;
6. Giá trị của mỗi hàng là duy nhất.

Như vậy, các cấu trúc của một quan hệ có thể được hình dung một cách trực quan như là một hệ toạ độ, trong đó mỗi giá trị dữ liệu được xác định như là giao điểm của một giá trị duy nhất của hàng với một giá trị duy nhất của cột.

II.4. Các kiểu Bảng và Khoá trong cơ sở dữ liệu quan hệ

Trong một cơ sở dữ liệu dạng quan hệ, các bảng được phân loại như sau:

- *Bảng cơ sở [base table]:* là bảng chứa một hay nhiều cột mô tả tính chất của một đối tượng và chứa khóa chính được gán duy nhất cho đối tượng đó với tư cách là một thực thể dữ liệu. Mỗi bảng cơ sở phải có một khóa chính. Các bảng cơ sở thường được gọi là *bảng chính* bởi vì nó yêu cầu một khóa chính.
- *Bảng quan hệ [relation table]:* là bảng dùng để cung cấp các mối nối kết giữa các bảng khác song không phải là bảng cơ sở.

Quan hệ giữa các bảng trong cơ sở dữ liệu quan hệ đặc trưng bởi các khoá quan hệ. Các khoá là các thuộc tính hoặc tập hợp các thuộc tính đảm bảo tính duy nhất của các hàng của một bảng. Các khoá cũng được phân loại như sau:

- *Khóa chính [primary key]*. Khóa chính bao gồm một tập hợp các giá trị xác định tính duy nhất của một hàng của bảng cơ sở (bảng chính). Khoá chính không chứa các giá trị có thể bị ảnh hưởng bởi các giá trị khác.
- *Khóa dự tuyển [candidate keys]*. Tất cả các thuộc tính hay tập hợp thuộc tính thoả mãn điều kiện về tính duy nhất của mỗi hàng của bảng được gọi là các khoá dự tuyển. Nói cách khác, đây là các khoá có khả năng được chọn làm khoá chính. Chẳng hạn hai trường chứa các giá trị Tên và số chứng minh nhân dân đều là các trường khóa dự tuyển cho phép định danh một công dân, tuy nhiên số chứng minh

nhân dân là chọn lựa thích hợp hơn vì hai người có thể trùng tên nhưng không thể có cùng một số chứng minh nhân dân hợp lệ.

- *Khóa hỗn hợp [composite keys]*. Nếu cần dữ liệu từ nhiều cột trong bảng để thoả mãn yêu cầu về tính duy nhất của một khóa chính, khóa đó được mệnh danh là khóa hỗn hợp hoặc khóa ghép *[concatenated key]*. Nói cách khác, khi một thuộc tính đơn lẻ không thoả mãn tính duy nhất của hàng, một nhóm các thuộc tính sẽ được sử dụng để thoả mãn yêu cầu này.
- *Khóa lạ [foreign key]*. Khoá lạ là sự trùng lặp được kiểm soát của một thuộc tính trong một hay nhiều quan hệ. Các khoá lạ xác định các mối quan hệ giữa các bảng bằng cách chỉ ra đường dẫn lôgic hay mối liên hệ giữa các bảng này. Có thể so sánh quan hệ này như là quan hệ cha-con: một khoá lạ ở quan hệ con chính là một khoá chính trong quan hệ cha.

  Khóa lạ có thể bao gồm một trường hay nhóm trường (một khóa lạ hỗn hợp). Nếu chiều dài của một khóa lạ nhỏ hơn khóa chính tương ứng, nó sẽ được gọi là khóa lạ cắt cụt *[truncated foreign key]* hay khóa lạ từng phần *[partial foreign key]*.

II.5. Các kiểu quan hệ

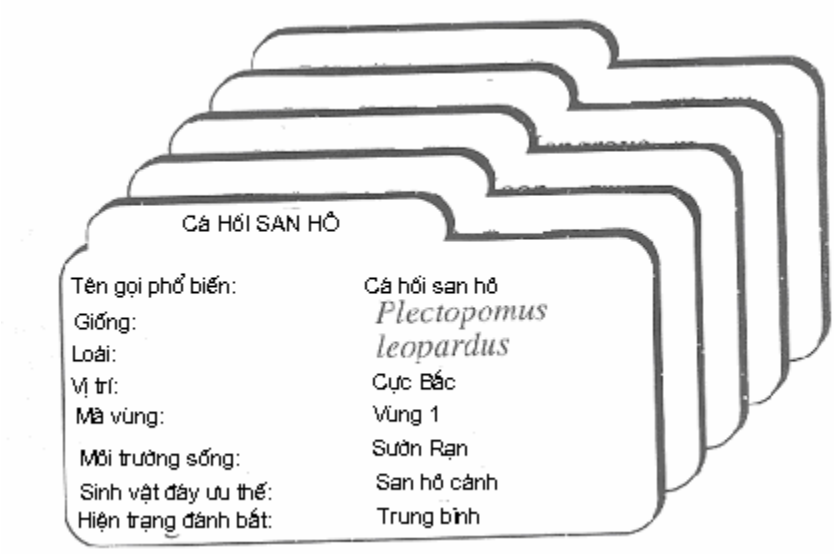
- *Mối quan hệ Một-Một:* Mối quan hệ đơn giản nhất giữa các bảng đó là mối quan hệ một-một. Trong kiểu quan hệ này, các bảng có sự tương ứng theo từng hàng một; từng hàng trong bảng không được có nhiều hàng tương ứng trong bảng kia. Các mối quan hệ một-một thường được dùng để chia các bảng cơ sở rất lớn thành các bảng nhỏ hơn.
- *Mối quan hệ Một-Nhiều:* Các quan hệ một-nhiều nối kết một hàng trong một bảng với hai hay nhiều hàng trong một bảng thông tin khác thông qua một mối quan hệ giữa khóa chính của bảng cơ sở và khóa lạ tương ứng trong bảng liên quan. Mặc dù khóa lạ trong bảng chứa các mối quan hệ phía nhiều có thể là một thành phần của một khóa chính hỗn hợp trong bảng riêng của nó, song nó vẫn là một khóa lạ cho các mục tiêu của mối quan hệ đó. Các quan hệ một-nhiều là những mối quan hệ phổ biến nhất.
- *Mối quan hệ Nhiều-Một:* Mối quan hệ nhiều-một là trường hợp đảo ngược của kiểu quan hệ một-nhiều.
- *Mối quan hệ Nhiều-Nhiều:* Các mối quan hệ nhiều-nhiều không thể diễn tả dưới dạng các mối quan hệ đơn giản giữa hai thực thể tham gia. Để xây dựng các mối quan hệ nhiều-nhiều, ta tạo một bảng có các mối quan hệ nhiều-một với hai bảng cơ sở.

II.6. Ví dụ về ưu điểm của cơ sở dữ liệu quan hệ

Có nhiều loại thiết kế cho cơ sở dữ liệu, trong đó phổ biến nhất là mô hình *tệp phẳng* và mô hình *quan hệ*. Cơ sở dữ liệu dạng tệp phẳng được xây dựng trên cơ sở cấu trúc của một tập bìa được đánh số, trong đó mỗi bìa chứa toàn bộ thông tin về một đối tượng hay sự kiện nào đó. Trong mô hình cơ sở dữ liệu quan hệ, thông tin được phản ánh

trên tất cả các bìa, và các bìa lại có mối liên hệ với nhau thông qua sự nối kết giữa các trường.

Ta hãy xét một ví dụ để so sánh hai mô hình trên đây. Giả sử bạn có một tập dữ liệu dưới dạng một tập bìa đánh số, với nội dung mô tả chi tiết về các loài cá tại một số vùng đánh bắt trong khu vực nghiên cứu. Bạn muốn nghiên cứu về các loài và mô tả chi tiết về môi trường sống của chúng. Bạn có thể đưa vào nội dung mỗi tấm bìa các thông tin sau: Tên gọi phổ biến, chi tiết về loài, vị trí xuất hiện, số vùng đánh bắt và thông tin hiện tại về các vùng đánh bắt như: loại môi trường sống, loại sinh vật đáy chiếm ưu thế và hiện trạng đánh bắt tại khu vực. Tập bìa đánh số có thể có dạng như minh hoạ trên hình 5. Thông tin trên các bìa có thể được đưa vào một bảng, trong đó mỗi mục trên bìa (tên gọi phổ biến, giống, loài, vị trí xuất hiện, v.v…) sẽ trở thành một *trường*, còn thông tin điền vào mỗi bìa sẽ trở thành một *thanh ghi* của bảng. Kết quả là một bảng được tạo ra với cấu trúc của một tệp phẳng (Bảng 4.1).



Hình 5. Tập dữ liệu dưới dạng tập bìa đánh số mô tả chi tiết về các loài cá và môi trường sống.

Các cột có tiêu đề "Tên gọi phổ biến", "Giống", "Loài", v.v… là các *trường* của cơ sở dữ liệu; các hàng bắt đầu từ "Cá hồi san hô", Cá tuyết cửa sông", v.v… là các *thanh ghi* của cơ sở dữ liệu. Nhược điểm của mô hình này là có nhiều dữ liệu được lặp đi lặp lại, gây khó khăn cho việc thay đổi hay cập nhật dữ liệu. Bạn thử hình dung một trường hợp sau đây: sau khi một trận bão xảy ra tại khu vực nghiên cứu, tại vị trí Vùng 1 người ta đã phát hiện ra là cuội sỏi đã chiếm ưu thế so với các rạn san hô. Trong mô hình tệp phẳng, và trong tập bìa đánh số, thông tin trên mỗi bìa có Mã vùng là Vùng 1 sẽ phải được cập nhật lại, và do đó bạn phải sửa lại các thông tin trên ba thanh ghi.

Có một cách khác để giải quyết vấn đề này, đó là tách dữ liệu ra thành hai bảng, một bảng chứa các thông tin chi tiết về loài, còn bảng kia chứa các thông tin chi tiết về

vùng đánh bắt. Hai bảng này phải được nối kết với nhau bằng cách nào đó sao cho thông tin từ bảng này có liên hệ với thông tin trên bảng kia. Loại cấu trúc này, với nhiều bảng nhỏ được liên kết với nhau, gọi là cơ sở dữ liệu quan hệ. Các bảng 4.2 và 4.3 minh họa việc các dữ liệu từ mô hình tệp phẳng (Bảng 4.1) được tổ chức lại theo cấu trúc của mô hình cơ sở dữ liệu quan hệ.

Bảng 4.1. Bảng có cấu trúc tệp phẳng thành lập từ các dữ liệu minh hoạ trên hình 1.

Tên gọi phổ biến	**Giống**	**Loài**	**Vị trí**	**Mã vùng**	**Môi trường sống**	**Loại sinh vật đáy ưu thế**	**Hiện trạng đánh bắt**
Cá hồi san hô	*Plectopomus*	*leopardus*	Cực Bắc	Vùng 1	Sườn rạn	San hô cành	Trung bình
Cá tuyết cửa sông	*Epinephelus*	*tauvina*	Bờ Trung	Vùng 6	Rạn Bommie	San hô bảng	Thấp
Cá mặt trăng	*Thalassoma*	*lunare*	Cực Bắc	Vùng 1	Sườn rạn	San hô cành	Trung bình
Cá Jack tràm	*Lutyanus*	*Argentimac ulatus*	Cực Nam	Vùng 4	Đáy bùn phẳng	Tảo biển	Cao
Cá đuôi vàng	*pomacentrus*	*flavicauda*	Cực Bắc	Vùng 1	Sườn rạn	San hô cành	Trung bình

Bảng 4.2. Chi tiết về các loài.

Tên gọi phổ biến	**Giống**	**Loài**	**Vị trí**	**Mã vùng**
Cá hồi san hô	*Plectopomus*	*Leopardus*	Cực Bắc	Vùng 1
Cá tuyết cửa sông	*Epinephelus*	*Tauvina*	Bờ Trung	Vùng 6
Cá mặt trăng	*Thalassoma*	*Lunare*	Cực Bắc	Vùng 1
Cá Jack tràm	*Lutyanus*	*Argentimac ulatus*	Cực Nam	Vùng 4
Cá đuôi vàng	*Pomacentrus*	*Flavicauda*	Cực Bắc	Vùng 1

Bảng 4.3. Chi tiết về môi trường sống.

Mã vùng	**Môi trường sống**	**Loại sinh vật đáy ưu thế**	**Hiện trạng đánh bắt**
Vùng 1	Sườn rạn	San hô cành	Trung bình
Vùng 6	Rạn Bommie	San hô bảng	Thấp
Vùng 1	Sườn rạn	San hô cành	Trung bình
Vùng 4	Đáy bùn phẳng	Tảo biển	Cao
Vùng 1	Sườn rạn	San hô cành	Trung bình

Hai bảng 4.2 và 4.3 có một trường chung là trường Mã vùng. Trường này có chức năng nối kết hai bảng với nhau và thường được gọi là trường khoá chính. Trường Mã vùng

xác định chính xác một vùng đánh bắt và kết nối bảng mô tả loài với bảng mô tả vùng. Chẳng hạn, để tìm ra loại sinh vật đáy ưu thế tại vùng đánh bắt cá ngừ san hô, đầu tiên bạn cần tìm thanh ghi có chứa cá ngừ san hô trong bảng mô tả loài. Mã vùng của cá ngừ san hô (Vùng 1) sau đó sẽ được sử dụng để tìm trong bảng mô tả vùng loại sinh vật đáy chiếm ưu thế (San hô cành).

Cần nhận xét rằng chỉ có một thanh ghi có mã vùng "Vùng 1" được tìm thấy trong bảng mô tả vùng, mặc dù có nhiều loài cá khác nhau có mặt tại vùng đánh bắt này được ghi nhận trong bảng mô tả loài. Loại quan hệ này gọi là quan hệ một-nhiều. Nếu có hai thanh ghi với mã vùng là "Vùng 1" được tìm thấy trong bảng mô tả vùng thì chắc chắn là có sai sót, vì khi đó bạn sẽ không thể xác định được thanh ghi nào ứng với bảng mô tả vùng. Mã vùng phải được trỏ từ bảng mô tả loài tới một thanh ghi duy nhất trong bảng mô tả vùng. Điều này có nghĩa là các giá trị trong trường Mã vùng của bảng mô tả vùng phải là các giá trị duy nhất.

Trong ví dụ trên, bạn có thể cảm thấy cấu trúc cơ sở dữ liệu quan hệ chứa nhiều dữ liệu hơn và có vẻ phức tạp hơn cấu trúc tệp phẳng. Sự khác biệt giữa hai cấu trúc này là ở chỗ, trường khóa chính "Mã vùng" có mặt trong cả hai bảng của cơ sở dữ liệu Quan hệ, nhưng chỉ xuất hiện một lần trong cấu trúc tệp phẳng. Tuy nhiên, cấu trúc quan hệ rõ ràng là có hiệu quả cao hơn, bởi bảng mô tả vùng chỉ chứa có ba thanh ghi. Nếu bạn thử cộng số loài vào, bạn sẽ thấy đối với cùng một tập dữ liệu, cấu trúc tệp phẳng chứa 40 lần nhập liệu (8 trường*5 thanh ghi), trong khi cấu trúc quan hệ chỉ chứa 37 lần nhập liệu [Bảng mô tả loài:(5 trường*5 thanh ghi)+Bảng mô tả vùng (4 trường*3 thanh ghi)]. Kích thước bảng càng lớn, sự khác biệt sẽ càng trở nên đáng kể hơn.

Dưới đây liệt kê những ưu điểm của cấu trúc quan hệ so với cấu trúc tệp phẳng:

- *Ưu thế về sức chứa dữ liệu.*

So sánh: Trong ví dụ trên, đối với các bảng rất nhỏ, cấu trúc tệp phẳng đòi hỏi 40 vị trí lưu dữ liệu, trong khi đối với cùng tập dữ liệu, cấu trúc quan hệ chỉ cần 37 vị trí.

- *Giảm số lượng dữ liệu cần nhập vào cơ sở dữ liệu*

So sánh: Trong cấu trúc tệp phẳng (bảng 4.1), để thêm một loài mới vào bảng, cần phải thêm vào 8 trường, trong khi đối với cùng tập dữ liệu, cấu trúc quan hệ chỉ yêu cầu thêm vào 5 trường.

- *Dễ cập nhật bảng.*

So sánh: Để thay đổi hiện trạng đánh bắt cho vùng "Vùng 1" từ trung bình đến cao, cấu trúc tệp phẳng (bảng 4.1) đòi hỏi cập nhật tất cả các thanh ghi có mã vùng là "Vùng 1", tức là phải cập nhật ba thanh ghi. Trong cấu trúc quan hệ, chỉ có một thanh ghi trong bảng mô tả vùng phải cập nhật, bất kể số loài cá được ghi nhận tại "Vùng 1" là bao nhiêu.

- *Dễ thay đổi cấu trúc cơ sở dữ liệu.*

So sánh: Để thêm một cột mới mô tả độ sâu đánh bắt vào bảng có cấu trúc tệp phẳng, bạn phải thay đổi tất cả các thanh ghi, trong khi đối với cấu trúc quan hệ, chỉ có các thanh ghi trong bảng mô tả vùng phải cập nhật. Điều này cho phép bạn sửa đổi, mở rộng và quản lý cơ sở dữ liệu một cách dễ dàng hơn nhiều.

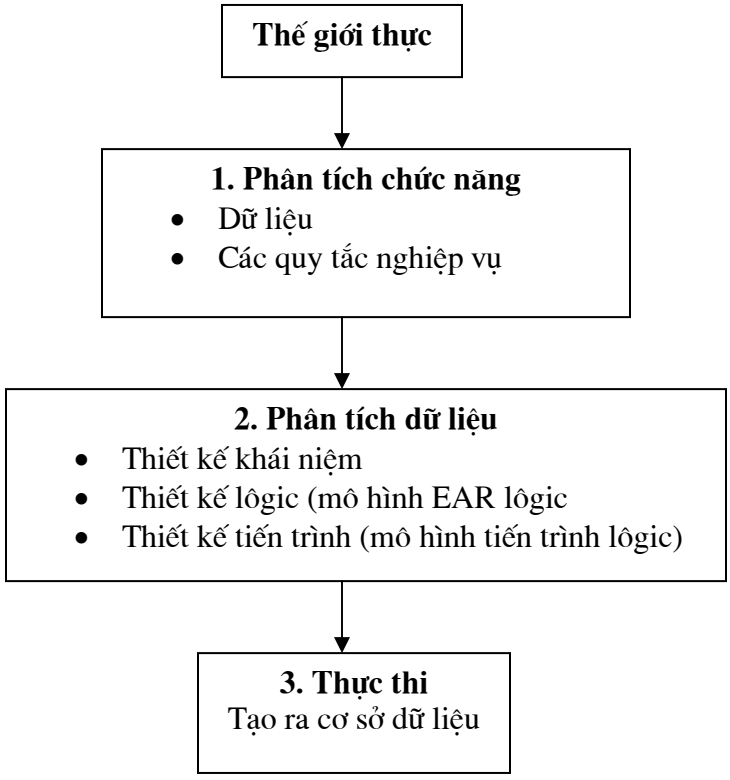
Nhược điểm duy nhất của hệ thống cơ sở dữ liệu quan hệ là nó đòi hỏi các chức năng của công cụ máy tính trong việc liên kết các bảng và kết xuất dữ liệu.

III. THIẾT KẾ CƠ SỞ DỮ LIỆU

III.1. Quy trình

Thiết kế cơ sở dữ liệu là một trong những bước quan trọng nhất của quy trình xây dựng một cơ sở dữ liệu. Một cơ sở dữ liệu được thiết kế tốt sẽ tạo điều kiện cho các thao tác nhập liệu dễ dàng và cho phép truy xuất dữ liệu nhanh, hiệu quả. Thiết kế cơ sở dữ liệu là một quá trình lặp đi lặp lại cho đến khi cơ sở dữ liệu thoả mãn các yêu cầu của các dữ liệu thu thập được cũng như nhu cầu của người sử dụng.

Quy trình thiết kế cơ sở dữ liệu được minh hoạ trên hình 6, bao gồm ba bước chính: Phân tích chức năng, Phân tích dữ liệu và Thực thi.



Hình 6. Các bước thiết kế cơ sở dữ liệu.

1. Phân tích chức năng: Đây là quá trình mô hình hóa các chức năng xử lý của thế giới thực đối với dữ liệu. Phân tích chức năng là xác định các thể loại dữ liệu và các qui tắc nghiệp vụ cần cho việc xử lý dữ liệu.

Trong giai đoạn phân tích chức năng, các vấn đề được xem xét, mô hình hóa và phân ra thành các thực thể, các quá trình xử lý và dữ liệu. Mô hình này được trình bày dưới dạng sơ đồ luồng dữ liệu, với các ký hiệu được minh hoạ trên hình 7:

- *Thực thể:* Biểu tượng này mô tả một thực thể bất kỳ, được nối với quá trình xử lý. Thực thể cung cấp đầu vào hay là nơi tiếp nhận thông tin của các quá trình xử lý. Tên của thực thể được đưa vào trong ký hiệu
- *Một quá trình xử lý:* Biểu tượng này biểu diễn một quá trình xử lý của dữ liệu. Mô tả của tiến trình được đưa vào trong ký hiệu
- *Lưu dữ liệu:* Biểu tượng này mô tả vị trí lưu dữ liệu
- *Luồng dữ liệu:* Biểu tượng này mô tả dữ liệu được truyền tới từ một thực thể, một quá trình hay nơi lưu dữ liệu. Mô tả của dữ liệu được đặt cạnh biểu tượng.

a) Thực thể b) Quá trình xử lý c) Lưu dữ liệu d) Luồng dữ liệu

Hình 7. Các ký hiệu sử dụng trong sơ đồ luồng dữ liệu ở giai đoạn phân tích chức năng

Sơ đồ luồng dữ liệu cho phép xác định những dữ liệu có liên quan, vị trí có thể tìm thấy dữ liệu, các qui tắc nghiệp vụ liên quan trong các tiến trình. Dữ liệu và các qui tắc nghiệp vụ được xác định trong giai đoạn này sẽ được dùng làm đầu vào cho giai đoạn phân tích dữ liệu.

2. Phân tích dữ liệu: Trong giai đoạn này, mô hình dữ liệu được tạo ra trên cơ sở các dữ liệu và các qui tắc nghiệp vụ đã xác định được trong giai đoạn trước. Giai đoạn phân tích dữ liệu được phân thành 3 giai đoạn nhỏ hơn, trong đó kết quả của mỗi giai đoạn này sẽ được sử dụng làm đầu vào cho giai đoạn kế tiếp. Các quyết định trong mỗi bước giai đoạn thực hiện có thể làm thay đổi thiết kế ở giai đoạn trước, vì thế cần phải cân nhắc thận trọng trước khi ra các quyết định tại mỗi giai đoạn thực hiện.

a) Giai đoạn thiết kế khái niệm

Đầu vào của giai đoạn thiết kế khái niệm chính là sơ đồ luồng dữ liệu - kết quả của giai đoạn phân tích chức năng. Sơ đồ luồng dữ liệu cung cấp cho chúng ta vị trí của dữ liệu, tập hợp các qui tắc nghiệp vụ để giúp chúng ta xây dựng cấu trúc dữ liệu.

Các bước cần làm trong giai đoạn thiết kế khái niệm

1. Xác định dữ liệu. Điều này có ý nghĩa đặc biệt nếu cách tiếp cận Bottom-up được dùng.

2. Bước tiếp theo là phân lớp dữ liệu. Hai cách tiếp cận được sử dụng là cách tiếp cận từ trên xuống *(Top-down)* và cách tiếp cận từ dưới lên *(Bottom-up)*. Các dữ liệu đã xác định được phân lớp thành các thực thể, thuộc tính hay giá trị. Các qui tắc nghiệp vụ được xác định trong giai đoạn phân tích chức năng cũng có thể được sử dụng để xác định các quan hệ giữa các thực thể.

Cách tiếp cận từ trên xuống thường được dùng để xây dựng những cơ sở dữ liệu mới hay các cơ sở dữ liệu chủ thể. Cách này được thực hiện theo trình tự như sau:

- Xác định thực thể
- Xác định quan hệ giữa các thực thể
- Xác định thuộc tính của các thực thể

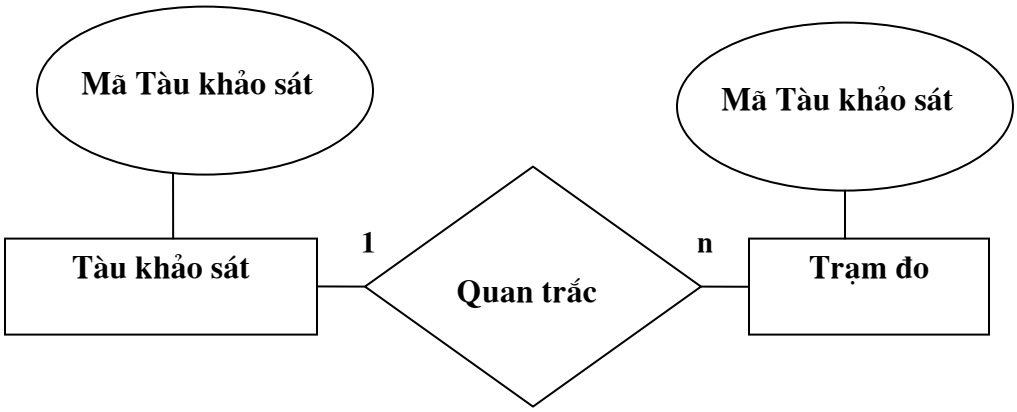
Cách tiếp cận từ dưới lên thường dùng khi dữ liệu đã tồn tại được tổ chức lại, hay xây dựng các cơ sở dữ liệu ứng dụng. Cách này được thực hiện theo trình tự như sau:

- Thu thập dữ liệu
- Xác định sự phụ thuộc trong dữ liệu (có nghĩa là xác định các thực thể và các thuộc tính)
- Xác định các quan hệ giữa các thực thể.

Các quan hệ cơ bản giữa các thực thể là: một-một, một-nhiều và nhiều-nhiều.

3. Tất cả các thông tin ở trên được sử dụng để tạo ra *mô hình khái niệm quan hệ-thực thể-thuộc tính*, thường được biểu diễn dưới dạng các sơ đồ mô hình dữ liệu. Một trong những sơ đồ được sử dụng rộng rãi là loại sơ đồ Thực thể-Quan hệ do Peter Chen đề xuất năm 1976. Trong các sơ đồ loại này, các thực thể dữ liệu được biểu thị bằng các khung chữ nhật, các thuộc tính dữ liệu được biểu thị bằng các hình trái xoan, còn các quan hệ được biểu thị bằng các khung hình thoi. Hình 8 minh họa sơ đồ Thực thể-Quan hệ dạng một-nhiều cho các thực thể *Tàu khảo sát* và *Trạm đo*.

4. Mô hình khái niệm quan hệ-thực thể-thuộc tính được kiểm tra bởi các chuyên gia và những người sử dụng dữ liệu.



Hình 8. Sơ đồ Thực thể-Quan hệ của mối quan hệ một- nhiều

b) Thiết kế logic

Mục đích của giai đoạn thiết kế lôgic tập trung vào việc nâng cao tính hiệu quả của mô hình, thực hiện các phép chuẩn hóa và tạo ra mô hình dữ liệu logic.

Các bước thực hiện trong giai đoạn thiết kế lôgic bao gồm:

- *Tinh lọc mô hình khái niệm quan hệ-thực thể-thuộc tính*
- Xác định khóa chính cho mỗi thực thể: Xác lập một thuộc tính hay nhóm các thuộc tính xác định duy nhất một sự xuất hiện của thực thể.
- Xác định khóa lạ: Xác lập một thuộc tính nào đó có chứa giá trị tương ứng với giá trị của thuộc tính trong một bảng khác.
- Giải quyết quan hệ một - một cho hiệu quả hơn
- Giải quyết quan hệ nhiều - nhiều
- Tinh lọc các thuộc tính.

Trong giai đoạn này, các qui tắc chuẩn hóa được áp dụng để tinh lọc mô hình khái niệm quan hệ-thực thể-thuộc tính. Chuẩn hoá *(Normalization)* là một thủ tục hình thức hoá, qua đó các thuộc tính dữ liệu được gom nhóm thành các bảng và các bảng được gom nhóm thành cơ sở dữ liệu. Tuỳ theo độ phức tạp của cơ sở dữ liệu, quá trình chuẩn hoá được thực hiện nhiều hay ít nhằm đưa dữ liệu về các dạng chuẩn sau:

Dạng chuẩn 1: Mô hình dữ liệu thoả mãn dạng chuẩn 1 nếu tất cả các thuộc tính lặp bị loại khỏi các thực thể.

Dạng chuẩn 2: Mô hình dữ liệu thoả mãn dạng chuẩn 2 nếu nó thoả mãn dạng chuẩn 1 và tất cả các thuộc tính chỉ phụ thuộc vào một phần khóa chính bị loại ra khỏi các thực thể. Ở dạng chuẩn này dữ liệu trong mọi cột phi khoá của một bảng phải lệ thuộc hoàn toàn vào khoá chính và từng phần (cột) của khoá chính nếu đó là một khoá chính hỗn hợp.

Dạng chuẩn 3: Mô hình dữ liệu thoả mãn dạng chuẩn 3 nếu nó thoả mãn dạng chuẩn 2 và tất cả các thuộc tính phụ thuộc vào thuộc tính khác không phải khóa chính bị loại ra khỏi các thực thể. Nói cách khác dạng chuẩn này yêu cầu tất cả các cột phi khoá của một bảng phải lệ thuộc vào khoá chính của bảng và độc lập với nhau.

c) Thiết kế tiến trình

Mục đích của giai đoạn thiết kế tiến trình là xác định khối lượng dữ liệu và tần suất các giao dịch, phân tích các giao dịch và tạo ra mô hình dữ liệu tiến trình.

Các công việc chủ yếu cần làm trong giai đoạn thiết kế tiến trình là ghi lại những thông tin vật lý về cách dữ liệu được lưu và truy nhập. Kết quả là tạo ra mô hình tiến trình logic chứa thông tin về khối lượng dữ liệu, dạng xử lý của một số các giao dịch mẫu, khối lượng dữ liệu được truy nhập và tần suất các giao dịch sẽ thực hiện. Các thông tin này sẽ hỗ trợ cho giai đoạn thực thi ra các cơ sở dữ liệu một cách có hiệu quả.

3. Thực thi: Tạo ra cơ sở dữ liệu thực phản ánh dữ liệu đã được mô hình hóa và có cấu trúc mềm dẻo, dễ thay đổi, chỉnh sửa và tiện lợi cho người sử dụng.

Các công việc tiêu biểu được thực hiện trong giai đoạn này bao gồm:

- Phân tách hay nhập các bảng lại để tăng hiệu suất (phi chuẩn).
- Bổ sung các loại dữ liệu phụ để làm tăng hiệu suất (phi chuẩn).
- Xác định kích cỡ của cơ sở dữ liệu.
- Xác định vị trí vật lý của cơ sở dữ liệu.
- Xác định cơ sở dữ liệu và các bảng.
- Tải các dữ liệu người dùng.

III.2. Kinh nghiệm

Trên thực tế, không có một quy tắc cụ thể nào tồn tại cho việc thiết kế cơ sở dữ liệu. Tuy nhiên, những kinh nghiệm đúc kết từ thực tiễn liệt kê dưới đây có thể giúp ích cho những người bước đầu thiết kế cơ sở dữ liệu:

- Đầu tiên, hãy cố gắng làm quen với các dữ liệu đang được thu thập. Chẳng hạn, bạn cần tìm hiểu thêm về các quy trình lấy mẫu, nội dung các thí nghiệm (bao gồm cả định nghĩa các mẫu, các bản sao và các mẫu trùng lặp), các dữ liệu sẽ được thu thập, điều kiện lấy mẫu và các yêu cầu đối với việc phân tích mẫu. Sẽ rất bổ ích nếu bạn có mặt trong một chuyến thực địa để tham quan quá trình lấy mẫu.
- Bước tiếp theo là việc thiết kế các bảng ghi dữ liệu để sử dụng ngoài thực địa và thiết kế cấu trúc cơ sở dữ liệu. Các công việc này có thể tiến hành song song, vì cách tổ chức ghi chép dữ liệu ngoài thực địa có liên quan chặt chẽ với cấu trúc của cơ sở dữ liệu. Các bảng ghi chép dữ liệu ngoài thực địa phải đảm bảo sao ghi nhận được tất cả các biến số cần thiết, đảm bảo độ tin cậy và tính đầy đủ của các dữ liệu được thu thập, và đảm bảo các dữ liệu này có thể có ích với nhiều người sử dụng trong tương lai.
- Xác định thể loại, độ dài và các quy tắc cho các giá trị dữ liệu trong từng trường. Thông thường, một trường có thể thuộc loại Số, Ký tự hay Ngày tháng. Độ dài trường xác định giá trị cực đại cho phép đối với một giá trị dữ liệu. Việc xác lập các quy tắc nhập liệu cho các trường cũng sẽ là cơ sở để xây dựng các chương trình kiểm tra chất lượng dữ liệu trong cơ sở dữ liệu. Đồng thời, điều này cũng hạn chế các sai sót trong quá trình nhập liệu. Hãy mô tả mỗi bảng dữ liệu theo mục đích sử dụng của chúng, và mô tả tên trường, loại, độ dài và quy tắc nhập liệu cho từng trường. Sự mô tả này còn được gọi là Từ điển dữ liệu đơn giản.
- Cần khẳng định chắc chắn rằng tất cả các biến số cần thiết cho các thao tác với dữ liệu và cho các phép phân tích tiếp theo đã có đủ trong cơ sở dữ liệu. Nếu bạn muốn kết xuất hay sắp xếp dữ liệu theo ngày tháng, bạn cần lưu ý đến điều này ngay trong giai đoạn thiết kế cơ sở dữ liệu.
- Cần khẳng định chắc chắn rằng các bản sao của dữ liệu sẽ được phát hiện chính xác trong cơ sở dữ liệu. Các bản sao dữ liệu phải được xác định ngay trong cơ sở dữ liệu dưới hai dạng: nhóm các bản sao dữ liệu và các bản sao dữ liệu đơn lẻ.
- Lúc này, các bảng dữ liệu có thể được xây dựng thành một bộ phần mềm để đưa vào sử dụng. Có nhiều công cụ thích hợp cho công việc này, chẳng hạn như các phần mềm đóng gói cho phép quản lý các cơ sở dữ liệu trên máy tính.

- Cần xây dựng giao diện nhập liệu sao cho ngay cả những người sử dụng không có nhiều kinh nghiệm cũng có thể thao tác dễ dàng. Nếu có thể, các đơn thể chương trình kiểm tra chất lượng dữ liệu cũng cần được đưa vào các chương trình nhập liệu theo các quy tắc đã xác lập ở trên. Luôn luôn nhớ kiểm tra các dữ liệu sau khi chúng đã được nhập vào cơ sở dữ liệu. Các quy tắc nhập liệu chỉ có thể kiểm tra xem dữ liệu nhập vào có hợp lệ hay không, chúng không thể kiểm tra tính đúng đắn của các dữ liệu nhập vào.
- Khi côngtác thu thập dữ liệu đã được bắt đầu, có thể phải hiệu chỉnh lại cơ sở dữ liệu. Đây là công việc thiết thực, nhằm kiểm tra xem cơ sở dữ liệu có phản ánh được đầy đủ các dữ liệu đang được thu thập theo các yêu cầu đề ra hay không. Những ý kiến phản hồi từ phía người sử dụng là rất bổ ích.
- Phương pháp thử nghiệm hoạt động của cơ sở dữ liệu tốt nhất là thử kết xuất dữ liệu càng sớm càng tốt. Các trục trặc có thể được phát hiện và chỉnh sửa sớm mà chỉ cần thử nghiệm với một tập dữ liệu nhỏ trong cơ sở dữ liệu. Đây cũng là công việc có ý nghĩa quan trọng trong việc trả lời những ý kiến đóng góp của những người sử dụng, đồng thời giúp phát hiện các điểm yếu trong thiết kế cơ sở dữ liệu hay trong bản thân dữ liệu.

IV. QUẢN LÝ CÁC CƠ SỞ DỮ LIỆU

Trong thực tế, không phải dữ liệu nào cũng có thể nhập được ngay vào cơ sở dữ liệu, các dữ liệu cần phải được quản lý. Quản lý dữ liệu là một quá trình bao gồm nhiều giai đoạn, từ việc kiểm soát quá trình nhập dữ liệu, mô tả cơ sở dữ liệu đến việc sao lưu dữ liệu. Việc tiêu phí nhiều tiền của cho công tác lấy số liệu sẽ trở nên vô nghĩa nếu các dữ liệu lưu trong cơ sở dữ liệu là không chính xác, hoặc không ai biết được dữ liệu nằm ở đâu và biểu thị cái gì.

Các điểm chính cần chú trọng trong công tác quản lý dữ liệu bao gồm :

- Quy định *trách nhiệm* về quản lý dữ liệu.
- Xác lập quy trình *thu thập dữ liệu*.
- Xác lập quy trình *thao tác với các dữ liệu*.
- Có ý thức về *chất lượng dữ liệu*.
- Có ý thức về việc *mô tả dữ liệu*.
- Lưu giữ các *dữ liệu thực*.

I. Trách nhiệm

Xác định trách nhiệm tại các công đoạn khác nhau của toàn bộ quá trình quản lý dữ liệu là rất quan trọng. Việc quy định trách nhiệm tại các bước khác nhau trong quy trình xử lý sẽ đảm bảo chất lượng của dữ liệu. Người thu thập dữ liệu sẽ có thể được gán luôn trách nhiệm nhập và kiểm tra dữ liệu, trong khi trách nhiệm sao lưu và lưu trữ dữ liệu có thể phân cho các cán bộ làm việc tại trung tâm máy tính. Cần xây dựng các văn bản mô tả đầy đủ cách sử dụng cơ sở dữ liệu và quy định rõ trách nhiệm của các cá nhân đối với từng công đoạn quản lý dữ liệu.

Thu thập dữ liệu

Thu thập dữ liệu là bước quan trọng nhất trong bất kỳ một chương trình giám sát môi trường nào. Dữ liệu được thu thập dưới dạng các bảng ghi chép là sự phản ánh hiện trạng về một biến số được đo đạc hay quan trắc tại một thời điểm nào đó. Tất cả các bước tiếp theo chẳng qua chỉ là các quy trình chuyển đổi, thao tác và phân tích "thực trạng" này. Sau đây là một số điều cần lưu ý khi thu thập dữ liệu :

- Áp dụng các phương pháp được công nhận rộng rãi, đã được mô tả trong các văn liệu.
- Đào tạo, nâng cao nghiệp vụ cho những cán bộ làm côngtác thu thập dữ liệu.
- Sử dụng các bảng ghi số liệu để dễ dàng kiểm tra các dữ liệu thu thập.
- Xây dựng quy trình lưu giữ các bảng ghi chép dữ liệu và các mẫu vật thu thập được.

Thao tác với các dữ liệu

Cách duy nhất để khẳng định rằng các dữ liệu thu thập được trên thực địa đã được đưa vào cơ sở dữ liệu một cách chính xác là áp dụng một quy trình nghiêm ngặt cho tất cả các bước thao tác với dữ liệu.

Chẳng hạn, cần xây dựng quy trình cho các công đoạn sau đây :

- Gán mã cho các mẫu.
- Nhập dữ liệu.
- Kiểm tra tính hợp lệ của dữ liệu.
- Quy trình bổ sung số liệu quan trắc và các bảng.
- Sao lưu và lưu trữ dữ liệu.
- Lưu trữ các ghi chép thực địa và các mẫu vật thu được.

Mỗi bước thao tác với dữ liệu cần được mô tả trong một văn bản hướng dẫn sử dụng. Văn bản này cần mô tả chi tiết tất cả các quy trình thao tác với cơ sở dữ liệu, các phương pháp sử dụng để kiểm tra dữ liệu, danh sách mã sử dụng trong cơ sở dữ liệu, cách sao lưu và lưu trữ dữ liệu, và phân công trách nhiệm trong mỗi công đoạn thao tác với dữ liệu.

Chất lượng dữ liệu

Chất lượng dữ liệu không chỉ bao hàm trong việc áp dụng các quy trình thao tác với dữ liệu. Nó bao hàm cả độ chính xác dữ liệu, sự lặp lại các phép đo ngoài thực địa và chất lượng dữ liệu lưu trong cơ sở dữ liệu. Có thể đảm bảo được chất lượng dữ liệu nhờ sử dụng các quy trình thao tác dữ liệu viết thành văn bản, bằng những dự đoán trước về nguồn gốc sai số và bằng cách kiểm tra thường xuyên.

Để đảm bảo chất lượng dữ liệu, cần lưu ý tới những điểm sau:

- Sự hiểu biết về dữ liệu và thiết kế thực nghiệm là rất quan trọng.

- Việc sử dụng các bảng ghi dữ liệu ngoài thực địa sẽ làm tăng độ tin cậy của các dữ liệu, khẳng định việc các dữ liệu được thu thập và được ghi nhận bằng một phương thức nhất quán.
- Các dữ liệu thu thập được phải được nhập vào cơ sở dữ liệu càng sớm càng tốt. Điều này sẽ cho phép loại trừ những sai sót trong khi những người lấy dữ liệu vẫn còn nhớ, đồng thời cũng cho phép thu thập lại các dữ liệu bị mất hoặc nghiên cứu kỹ lại vùng lấy mẫu.
- Các chương trình nhập liệu trong cơ sở dữ liệu sẽ đảm bảo để các giá trị dữ liệu nằm trong phạm vi cho phép. Các phép kiểm tra ngầm định trong cơ sở dữ liệu có thể bao gồm: kiểm tra khoảng giá trị hợp lí, kiểm tra tính đúng đắn của ngày tháng, mã số và kiểm tra xem các tài liệu tham khảo của các dữ liệu nhập vào có tồn tại đầy đủ hay không.
- Kiểm tra dữ liệu là một phần quan trọng của công đoạn nhập liệu và đảm bảo chất lượng dữ liệu.
- Một khi dữ liệu đã được nhập vào cơ sở dữ liệu, có thể cho chạy các chương trình kiểm tra dữ liệu để phát hiện những lỗi lô gich, các giá trị bị bỏ sót hay các giá trị vượt quá giới hạn cho phép.
- Sao lưu và lưu trữ dữ liệu là những công việc nhằm bảo toàn dữ liệu. Dữ liệu cần được sao lưu định kỳ trên đĩa hay băng và phải được bảo quản ở những nơi an toàn cách xa nguồn dữ liệu gốc. Bạn hãy nhớ ba quy tắc sau: sao lưu, sao lưu và sao lưu. Sẽ là không thể tha thứ được nếu bạn không sao lưu dữ liệu. Ngoài việc sao lưu dữ liệu, sổ sách ghi chép và các mẫu vật thu thập được ngoài thực địa cũng phải được lưu trữ ở nơi an toàn.
- Cuối cùng, cần phải có một người nào đó chịu trách nhiệm về chất lượng dữ liệu. Bằng cách giao trách nhiệm cho một người cụ thể, có thể thực hiện nhiều công việc liên quan đến chất lượng dữ liệu, bao gồm cả việc viết các quy phạm. Bằng cách giao quyền hạn cho người đó, những người khác sẽ phải hoàn thành bổn phận của mình trong việc đảm bảo chất lượng dữ liệu.

Từ điển dữ liệu

Việc mô tả cấu trúc cơ sở dữ liệu và cách sử dụng nó quan trọng không kém gì việc xây dựng cơ sở dữ liệu. Một tài liệu mô tả đầy đủ sẽ giúp người sử dụng hiểu được cấu trúc cơ sở dữ liệu, thiết kế và các mối quan hệ giữa các tập dữ liệu. Sự mô tả này sẽ khiến cho cơ sở dữ liệu trở thành phổ dụng và làm tăng giá trị của dữ liệu.

Trong việc mô tả dữ liệu cần lưu ý tới một số vấn đề sau:

- Mô tả đầy đủ về dự án, bao gồm cả các vấn đề chung, mục tiêu và các chi tiết khác.
- Mô tả đầy đủ về các phương pháp lấy số liệu sử dụng, hoặc trích dẫn các tài liệu có liên quan khác.
- Mô tả các bảng và mối quan hệ giữa chúng trong cơ sở dữ liệu.

- Đối với mỗi trường trong mỗi bảng, cần có mô tả chi tiết về tên trường, loại trường, độ dài trường, quy tắc nhập liệu của trường và mô tả vắn tắt về dữ liệu chứa trong trường.
- Mô tả vắn tắt tất cả các phần mềm sử dụng, toàn bộ mã nguồn của các chương trình nhập và kiểm tra dữ liệu, và mô tả các hệ máy tính sử dụng trong khuôn khổ dự án.

Sẽ là lí tưởng nếu phần mô tả dữ liệu được đưa vào văn liệu chung mô tả quy trình sử dụng và quản lý cơ sở dữ liệu. Văn bản này cũng cần được phổ biến cho các cơ quan khác.