NGHIÊN CỨU XÂY DỰNG CƠ SỞ DỮ LIỆU VÀ MÔ HÌNH QUẢN LÝ KHAI THÁC SỐ LIỆU TỪ KẾT QUẢ ĐIỀU TRA TIẾN SĨ NĂM 2000

Lê Đỗ Mạch

Hiện nay nước ta đang tiến hành công nghiệp hoá và hiện đại hoá đất nước, đồng thời cũng đang hướng đến một nền kinh tế trí thức. Khoa học và công nghệ đóng một vai trò quan trọng hàng đầu trong cuộc cách mạng này. Đội ngũ các nhà khoa học cần phải đủ mạnh về số lượng và chất lượng để phục vụ cho sự phát triển kinh tế và xã hội cũng như nghiên cứu khoa học công nghệ. Trong đó, tiến sĩ và tiến sĩ khoa học là những người có trình độ khoa học cao, được đào tạo cơ bản, hệ thống, có năng lực nghiên cứu và ứng dụng là rất quan trọng. Nhiều người đang giữ những vị trí quan trọng trong các cơ quan, nhà nước và các tổ chức khoa học công nghệ. Ý thức được vấn đề này năm 2000 Viện Khoa học Thống kê đã tiến hành Cuộc điều tra tiến sĩ. Mục đích của cuộc điều tra là nắm được thực trạng đội ngũ tiến sĩ, phục vụ cho chính sách phát triển và khai thác đội ngũ này tốt hơn nữa, có chiến lược đào tạo lâu dài bù đắp cho những thiếu hụt do một số người tuổi cao sức yếu đã về nghỉ.

Nguồn dữ liệu thu được lưu trên các file Foxpro với hơn 30 tiêu chí khác nhau, từ nhân thân của tiến sĩ đến quá trình đào tạo, trình độ ngoại ngữ và những đóng góp của họ về khoa học trong 5 năm gần nhất kể từ ngày khai báo dữ liệu. Đây là một nguồn dữ liệu rất quí và đáng tin cậy không chỉ về mặt tổng thể mà đối với cả từng cá thể, lần đầu tiên là tương đối đầy đủ nhất ở nước ta. Chính vì thế cần phải có kế hoạch lưu giữ, bảo quản lâu dài nguồn số liệu này, và định kỳ cặp nhật để phục vụ người dùng dưới các góc độ khai thác khác nhau. Đề tài cấp cơ sở "Nghiên cứu xây dựng cơ sở dữ liệu và mô hình quản lý khai thác số liệu từ kết quả điều tra tiến sĩ năm 2000" hình thành trên cơ sở những ý tưởng như vậy.

Cơ sở dữ liệu tiến sĩ được định hướng phát triển trên mô hình quan hệ và cài đặt nó trong môi trường Access 2000. Access là một ứng dụng trong bộ Office rất thông dụng, nhiều người biết và đang sử dụng trong công tác hàng ngày. Vì thế việc khai thác cơ sở dữ liệu là dễ dàng, thuận lợi.

Cũng xin nói thêm là cơ sở dữ liệu dù có tốt đến đâu nhưng nguồn dữ liệu trong nó không đáng tin cậy thì cũng chẳng mang lại lợi ích gì. Chính vì thế trước khi chuyển chúng thành cơ sở dữ liệu chúng tôi rất quan tâm đến vấn đề kiểm tra, làm sạch và hoàn thiện nguồn dữ liệu. Mặc dù cuộc điều tra rất công phu và bài bản, việc kiểm tra và hiệu chỉnh số liệu trong lúc xử lý là cẩn thận, nhưng số liệu dùng cho xử lý theo một nghĩa nào đó vẫn chưa đủ hoàn thiện để dùng ngay vào xây dựng cơ sở dữ liệu. Vì rằng kết quả điều tra thống kê chỉ quan tâm nhiều đến con số tổng thể và các sai sót về cá thể có thể bù trừ nhau, nhưng trong cơ sở dữ liệu, dữ liệu về một cá thể là rất quan trọng, phải đúng đắn. Chúng tôi phải tiếp tục kiểm tra hoàn thiện nguồn dữ liệu này trên từng cá thể. Việc kiểm tra được tiến hành trên một số khía cạnh như sự trùng lắp thông lin, nội dung thông tin, khoảng xác định số liệu và phạm vi mã hoá, sự ràng buộc giữa các dữ liệu. Sau khi làm sạch và hoàn thiện, dữ liệu của từng cá thể trở nên hợp lý hơn, đáng tin cậy hơn.

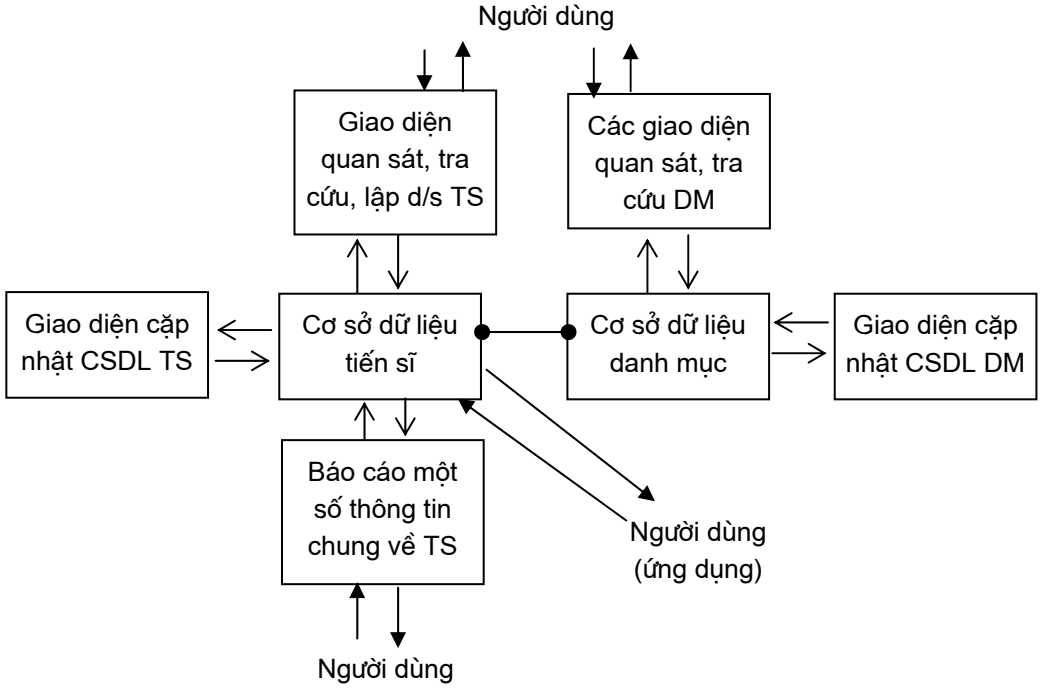
Dựa vào mục đích của CSDL tiến sĩ và nguồn dữ liệu thu được từ cuộc điều tra tiến sĩ năm 2000, chúng tôi liến hành xây dựng CSDL. Mô hình hệ thống cơ sở dữ liệu, quản lý và khai thác dữ liệu tiến sỹ có ba chức năng chính:

- Tìm kiếm, tra cứu dữ liệu. Lập danh sách tiến sĩ

- Cặp nhật dữ liệu.

- Lập một số báo cáo thống kê về đội ngũ tiến sĩ.

Dưới đây là sơ đồ tổng quát của hệ thống



- CSDL Tiến sĩ bao gồm một số bảng dữ liệu của tiến sĩ

- CSDL danh mục bao gồm các bảng danh mục và mã hoá

- Giao diện quan sát và tra cứu lập danh sách tiến sĩ giúp người dùng xem, tra cứu thông tin tiến sĩ, lập danh sách tiến sĩ theo một yêu cầu nào đó

- Giao diện quan sát và tra cứu danh mục và mã hoá giúp người dùng xem và tra cứu các thông tin về các bảng danh mục và mã hoá

- Báo cáo một số thông tin chung về tiến sĩ là một số bảng thống kê đơn giản về những khía cạnh chung nhất của đội ngũ tiến sĩ để người dùng tra cứu

- Giao diện cặp nhật CSDL tiến sĩ giúp cho việc sửa chữa, bổ sung hoặc xoá các dữ liệu về tiến sĩ

- Giao diện cặp nhật CSDL danh mục giúp cho việc sửa chữa, bổ sung hoặc xoá các dữ liệu về danh mục

CSDL tiến sĩ tuân theo các nguyên tắc thiết kế của một CSDL quan hệ. Nó thoả mãn được các yêu cầu:

- Dễ sử dụng và đáp ứng những thay đổi theo yêu cầu thiết kế tương lai

- Cách trình bày và các mối quan hệ dễ hiểu

- Dễ nhìn và sử dụng dung lượng lưu trữ hợp lý

Trong thiết kế CSDL tiến sĩ, chúng tôi rất quan tâm đến tính loàn vẹn tham chiếu của các mối quan hệ giữa các bảng dữ liệu nhằm đảm bảo tính nhất quán của cơ sở dữ liệu. Còn mối quan hệ ràng buộc giữa các dữ liệu trong CSDL cũng luôn luôn được chú ý để bảo đảm tính đúng đắn của dữ liệu.

Việc tra cứu, tìm kiếm thông tin được thực hiện trên các giao diện. Có thể tìm kiếm thông tin trên một trường hoặc nhiều trường. Có thể sử dụng các kí tự thay thế trong thông tin tìm kiếm như "*" để tìm một số kí tự ở đầu hoặc cuối trường, "?" để tìm một kí tự đơn lẻ, "#" để tìm một chữ số đơn lẻ, [ ] để tìm một số kí tự đơn lẻ chỉ định trước trong dấu ngoặc. Có thể sử dụng các phép toán so sánh quan hệ như <, <=, >, >=, <> và các phép toán logic OR, AND, BETWEEN,... để kết nối các thông tin tìm kiếm thành một biểu thức.

Kết quả tìm kiếm được hiện trong giao diện giúp tra cứu các thông tin cần thiết. Có thể in các thông tin đã tìm được, hoặc chuyển chúng sang Word hoặc Excel để xử lý tiếp.

Việc cặp nhật cơ sở dữ liệu tiến sĩ cũng được thực hiện qua các giao diện giống như quan sát và tra cứu dữ liệu, lập danh sách tiến sĩ. Việc cặp nhật được tiến hành trên ba mặt:

- Sửa đổi dữ liệu hoặc thay thế những dữ liệu cũ bằng dữ liệu mới thu được

- Xoá dữ liệu, xoá tiến sĩ hoặc danh mục ra khỏi CSDL.

- Thêm dữ liệu, nhập một tiến sĩ mới, hoặc một danh mục mới vào cơ sở dữ liệu.

Nhờ có thiết kế tính toàn vẹn của các mối quan hệ khi tham chiếu nên hành động cặp nhật luôn luôn bảo đảm được tính nhất quán của CSDL. Trong CSDL cũng thiết kế sẵn một số báo cáo thống kê đơn giản mô tả những mặt chung nhất về đội ngũ tiến sĩ hiện có trong cơ sở dữ liệu. Những thông tin này giúp người dùng nắm được thực trạng đội ngũ tiến sĩ hiện thời. Có thể xem, in những báo cáo này hoặc chuyển chúng sang Word hoặc Excel để tiếp tục xử lý.

Trên đây là những nét khái quát về cơ sở dữ liệu tiến sĩ. Để kết thúc bài viết này, chúng tôi có một vài đề nghị. Thứ nhất, cơ sở dữ liệu tiến sĩ là một nguồn

dữ liệu quí, cần có kế hoạch định kỳ cặp nhật để phục vụ người dùng. Hướng cặp nhật là mở rộng nội dung thông tin, cặp nhật những thông tin thay đổi về tiến sĩ, tình trạng tiến sĩ như đang làm việc hay đã nghỉ, còn sống hay đã chết, bổ xung tiến sĩ mới. Đồng thời qua cặp nhật cũng hoàn thiện thêm một số thông tin đã thu được trong cuộc điều tra tiến sĩ năm 2000.

Thứ hai, ngành Thống kê là một ngành có nhiều số liệu điều tra rất quí, chỉ khai thác một lần chưa hết. Cần phải có kế hoạch lưu trữ để tiếp tục khai thác, phục vụ cho những phân tích sâu hơn về số liệu và tìm kiếm thông tin. Nhất là một số cuộc điều tra mà số liệu của từng cá thể là rất có ý nghĩa trong việc cung cấp và khai thác thông tin thì nên xây dựng thành cơ sở dữ liệu để sử dụng lâu dài và định kỳ cặp nhật■