VÀI Ý KIẾN VỀ XÂY DỰNG CÁC CƠ SỞ DỮ LIỆU CỦA NGÀNH

Hoàng Minh Thiện

Gần đây, Tổng cục thống kê đã có những bước đi quan trọng theo hướng đẩy mạnh ứng dụng các kỹ thuật tin học vào các hoạt động nghiệp vụ của ngành, một trong những hướng đi chủ yếu đó là xây dựng các cơ sở dữ liệu (CSDL) cho ngành, các lĩnh vực. Các kế hoạch đã được vạch ra và các khoản chi phí đã được chuẩn bị sẵn sàng cho việc xây dựng một nền tảng quan trọng của ngành trong thời đại tin học đó là xây dựng nguồn dữ liệu thống tin thống kê qua các CSDL và từng bước đưa các CSDL này phục vụ cho các yêu cầu về thông tin thống kê đang ngày một tăng lên. Tuy nhiên, việc nắm chắc đòi hỏi về các tính năng của CSDL, cũng như những kỹ thuật tin học cần thiết, hay sự phối hợp chung toàn ngành trong một tiến trình vẫn chưa rõ ràng và thách thức đối với các vụ nghiệp vụ chuyên ngành. Chúng tôi không hy vọng làm sáng tỏ mọi kỹ thuật này hay đưa ra giải pháp chung cho tất cả các vụ nghiệp vụ, song chúng tôi hy vọng sẽ cùng bàn với các cán bộ nghiệp vụ về những nghi vấn đó và qua đó tìm ra các giải pháp thích ứng cho mỗi đơn vị đã và đang bắt tay vào công việc không mới nhưng cũng còn xa lạ này.

Trước tiên, chúng ta hay cùng xem xét việc xây dựng CSDL là nhằm tạo ra sản phẩm gì? Theo kinh nghiệm đúc kết của các nơi đã phát triển thành công các CSDL và theo lý thuyết, thì CSDL là một tập hợp các dữ liệu căn bản về lĩnh vực hay đối tượng nào đó; nằm dưới sự quản lý của một hệ quản trị, có khả năng truy nhập vào dữ liệu cơ sở để tìm kiếm, lấy ra dữ liệu, thực hiện các tính toán, trình bày thống kê; có thể bổ sung, loại bỏ hay sửa đổi theo một thời hạn định kỳ và được bảo vệ an toàn; nhằm giúp cho người sử dụng có thể khai thác dữ liệu một cách có hệ thống, thuận tiện và lâu dài. Như vậy, chúng ta có thể thấy ngay các vấn đề cần xác lập trước khi bắt tay vào xây dựng các CSDL, hơn nữa là hệ thống các CSDL của ngành, đó là:

- Các dữ liệu cơ sở về lĩnh vực, ngành hay đối tượng dự định xây dựng CSDL là gì? - nó có thể là các số liệu thống kê, các danh mục, văn bản (ở ngành thống kê chủ yếu là các loại dữ liệu này), hay hình ảnh, đôi khi cả âm thanh,... mà các tính toán, trình bày thống kê,... về lĩnh vực hay đối tượng đó đều xuất phát từ các dữ liệu này; các dữ liệu này là cần phải có đối với đối tượng xây dựng CSDL và không thể tính toán hay biến đổi từ các dữ liệu khác. Quan trọng hơn nữa là giữa chúng có mối quan hệ với nhau. Đây chính là nội dung thông tin cốt lõi của hệ thống các CSDL. Hiện nay các dữ liệu cơ sở của ngành còn nhiều khiếm khuyết như không có dữ liệu ở khoảng thời gian nào đó, hay do chuyển

đổi mà không thống nhất, so sánh được, các danh mục không ổn định, văn bản pháp quy thiếu hệ thống, và gần như không có các lưu trữ hình ảnh, âm thanh,... đó là những bất cập thực tế; lại nữa, do công việc lưu trữ không được tiến hành có hệ thống dẫn đến thất lạc và đặc biệt lưu trữ trên thiết bị tin học càng không theo chuẩn mực, lại do lạc hậu về kỹ thuật làm cho việc sử dụng dữ liệu lịch sử đưa vào hệ thống mới rất khó khăn. Vì thế, việc xem xét lại dữ liệu cơ sở, nên bắt đầu từ thời gian nào, dữ liệu gì có thể sử dụng được trong hệ thống mới, dữ liệu gì còn thiếu,... đều là những việc cần thiết phải làm trước khi có thể bắt tay vào xây dựng CSDL.

- Hệ quản trị dữ liệu, là sản phẩm mềm quản lý dữ liệu được lựa chọn hay các bộ chương trình được viết ra làm nhiệm vụ quản trị dữ liệu - nhằm giúp người quản lý và sử dụng dễ dàng thực hiện các hoạt động trên CSDL như: tìm kiếm, sao chép, phân tổ, lập biểu, hay bổ sung, loại bỏ, sửa đổi dữ liệu,... Trước tình hình có đa dạng các sản phẩm mềm ứng dụng như hiện nay việc chọn lựa này là rất quan trọng, vì liền theo đó là các vấn đề như: có phù hợp với tính chất dữ liệu của ngành không, với trình độ đội ngũ cán bộ nghiệp vụ hiện nay không, các dịch vụ của nhà cung cấp có đảm bảo ổn định và tiên tiến không, bản quyền và chi phí duy trì hệ quản trị, hay tính chất bảo mật của một phần dữ liệu có được khẳng định không,... Trường hợp viết một bộ chương trình quản trị riêng bằng cách thuê nhà thầu hay tự viết lấy cũng lại nẩy sinh những vấn đề khác không kém phần phức tạp, đặc biệt do tính đa dạng dữ liệu, các vấn đề luôn luôn nẩy sinh trong ngành, cũng như các tính toán, trình bày dựa trên dữ liệu đó, nên không thể một sớm một chiều có ngay được một hệ quản trị như ý; hoặc khi sử dụng các nhà thiết kế nước ngoài thì khó mà đáp ứng được đòi hỏi chi phí cao. Trên nền tảng hệ thống thông tin hiện có, dù quyết định phát triển các CSDL thế nào, nhất thiết chúng ta cũng phải sử dụng tư vấn kỹ thuật cao cấp của ngoài ngành vì thực lực cán bộ kỹ thuật tin học của ngành hiện rất mỏng, có ít kinh nghiệm về kỹ thuật này, thậm chí có cả cán bộ nghiệp dư do không được đào tạo căn bản về kỹ thuật tin học nên thường nhìn nhận lệch lạc vấn đề.

- Vấn đề an toàn CSDL ở đây là hệ thống các quy định về khai thác, cung cấp hay duy trì CSDL, cũng như các giải pháp kỹ thuật bảo vệ an toàn hệ thống trong khai thác,... nhằm làm cho CSDL có thể tồn tại dài lâu trong đáp ứng yêu cầu của người sử dụng. Như chúng ta đều biết là chúng ta chưa có các quy định mang tính pháp lý này và sớm muộn chúng ta cũng cần phải xây dựng để quá trình đưa các CSDL vào hoạt động có cơ sở pháp lý và xác lập trách nhiệm của những người quản lý, khai thác. Sự an toàn của hệ thống các CSDL cũng phụ thuộc vào các kế hoạch quản lý hoạt động, chế độ lưu trữ dữ liệu,... mà chúng ta cần xác lập và phân công trách nhiệm.

- Thời hạn cập nhật (up-date) của CSDL là nói đến khả năng thay đổi dữ liệu cơ sở (bổ sung, loại bỏ, sửa đổi) trong CSDL theo thời gian, mà thông thường là: ngày, tháng, quý, năm hay vài năm tuỳ theo tính chất của loại dữ liệu mà CSDL cung cấp. Thời hạn cập nhật dữ liệu càng ngắn thì chi phí duy trì CSDL sẽ càng cao. Hiện nay, nhiều số liệu thống kê chuyển sang thu thập bằng phương pháp điều tra, các số liệu thu thập thông qua báo cáo cũng chưa ổn định, ngay đến một danh mục ban hành chính thức cũng không bao giờ có thời hạn mà nó sẽ được cập nhật

tiếp theo; cho nên xác lập đúng và thực tiễn thời hạn cập nhật cũng là rất khó khăn và đôi khi cần sự ủng hộ của cả các đơn vị ngoài ngành như tài chính (cung cấp kinh phí đều kỳ cho các cuộc điều tra) hay các bộ ngành khác trong thực hiện báo cáo thống kê,...

Như vậy, bắt tay vào xây dựng các CSDL trước tiên chúng ta cần phân tích và xác lập rõ ràng những tính năng trên của mỗi CSDL cần có; bước này thường gọi là phân tích hệ thống; ở đây chưa nói tới cấu trúc của hệ thống các CSDL cũng như những đảm bảo khác về nhân lực, tiềm lực máy móc kỹ thuật hay kinh phí,... Đối với hầu hết các lĩnh vực, đối tượng hiện nay trong thống kê thì bước này hầu như chưa được làm cẩn thận; lại thêm vào đó sự chồng chéo, trùng lặp không cần thiết giữa các vụ hay thiếu hụt dữ liệu chung của cả hệ thống cũng chưa được tính đến. Vì thế nếu bắt tay ngay vào các việc biên tập dữ liệu thì người phát triển ứng dụng sẽ nhanh chóng rơi vào tình trạng không kiểm soát được công việc, vì mỗi nơi sẽ triển khai theo nhu cầu riêng của mình mà không tính đến toàn thể; sự lúng túng này đã nhiều lần xảy ra trong quá khứ.

Để giải quyết tình trạng này chúng ta nên thực hiện theo hai hướng sau:

- Giải quyết các vấn đề chung của hệ thống các CSDL, cấu trúc chung như thế nào, phối hợp các vụ chuyên ngành, phân tích và thiết kế hệ thống để giải quyết các vấn đề quan hệ tới nhiều lĩnh vực, trong đó có dữ liệu cơ sở, sự phân cấp và hệ quản trị như là môi trường chung cho trao đổi và làm việc của hệ thống, phối hợp thời gian cập nhật các CSDL thống nhất, và cuối cùng là đặt ra các quy định cho an toàn chung của hệ thống các CSDL. Như thế tức là cần một ban chỉ đạo chung có thể tác động được đến từng lĩnh vực cụ thể dự kiến xây dựng CSDL.

- Các lĩnh vực nghiệp vụ riêng rẽ cần thực hiện xác định chi tiết dữ liệu cơ sở của mình thống nhất với toàn bộ hệ thống, xác định thời gian cập nhật định kỳ cho mỗi CSDL cụ thể để hoà nhập vào hệ thống chung hay trách nhiệm trước các cấp dữ liệu khác.

Hoạt động của hai hướng này ban đầu sẽ rà soát lại toàn bộ hệ thống và các trường hợp trùng lặp và đi tới sự thống nhất cần thiết về dữ liệu cũng như sự phân cấp cần thiết. Tuy nhiên dữ liệu cơ sở sẽ thiếu hụt ở những thời gian nào đó, hệ quản trị chung sẽ không phù hợp với bộ phận cán bộ nào đó (vì chưa được đào tạo) thì đó cũng là điều bình thường. Tất nhiên sẽ có kế hoạch biên tập dữ liệu thống nhất (bắt đầu từ thời gian nào đó rồi mở rộng về quá khứ và tương lai của thời điểm đó) và sẽ có các lớp đào tạo để các cán bộ nghiệp vụ nắm được, quản lý và khai thác tốt CSDL mới được xây dựng,...

Theo dòng phân tích như thế mà tiếp tục chúng ta sẽ thấy hoạt động xây dựng CSDL này chắc chắn sẽ tác động đến hầu hết các bộ phận trong ngành thống kê; một hoạt động có tác động rộng lớn như thế cần được kiểm soát và có kế hoạch cụ thể, thậm chí là nên xây dựng thành đề án khả thi chi tiết, với các kế hoạch công việc, kế hoạch tổ chức nhân sự, kế hoạch tài chính, kế hoạch thời gian,... ngược lại vội vã lao vào các công việc cụ thể ngay chúng ta sẽ tự làm khó cho phát triển ứng dụng và gây ra mất lòng tin vào tính ưu việt của các kỹ thuật mới.

Theo thời gian phát triển kỹ thuật tin học, tới hiện nay mới bắt đầu xây dựng các CSDL ở một cơ quan thông tin kinh tế xã hội thì đã là muộn, song điều đó không

đồng nghĩa với việc làm ngay mà bỏ qua các bước kỹ thuật cần thiết tối thiểu; chính vì thế bài viết này chỉ hy vọng đưa ra một thông điệp rằng hãy cẩn trọng khi bắt đầu một công việc mà chúng ta chưa hiểu thấu đáo bản chất công việc đó cũng như những tác động có thể có mà nó sẽ gây ra. ở đây, việc xây dựng hệ thống các CSDL đối với ngành thống kê cũng như vậy. Song nếu chúng ta phân tích kỹ lưỡng công việc, tận dụng tốt các thành quả khoa học về phát triển CSDL của trong nước và thế giới, thì chúng tôi tin chắc hướng phát triển này sẽ mang lại nhiều kết quả lạc quan và lâu dài cho hoạt động đổi mới ngành hiện nay■