

NGUỒN DỮ LIỆU, GIẢI PHÁP HOÀN THIỆN ĐỂ XÂY DỰNG CÁC CSDL PHỤC VỤ NGHIÊN CỨU VÀ ĐÀO TẠO THỐNG KÊ

Đình Bá Hiên, Hoàng Minh Thiện

Hiện nay, Viện KHTK thực hiện 5 chức năng chính: Nghiên cứu khoa học; quản lý khoa học toàn ngành và quản lý chất lượng hoạt động thống kê; phổ biến thông tin KHTK; đào tạo, bồi dưỡng kiến thức thống kê và quản lý nhà nước trong lĩnh vực thống kê; xây dựng và tổ chức thực hiện chiến lược phát triển thống kê Việt Nam giai đoạn 2011-2020 và tầm nhìn đến năm 2030. Nhằm thực hiện tốt các chức năng và các nhiệm vụ được giao, Viện KHTK đã hình thành Đề án “Xây dựng các cơ sở dữ liệu phục vụ nghiên cứu và đào tạo thống kê”. Bài viết này giới thiệu kết quả nghiên cứu ban đầu của Đề án nói trên.

1. Các dữ liệu cần thiết để xây dựng các CSDL phục vụ nghiên cứu và đào tạo

Trên cơ sở khảo sát nhu cầu và tham khảo các CSDL trong và ngoài nước, Viện KHTK tiến hành xây dựng 06 CSDL: 1) CSDL một số cuộc điều tra thống kê; 2) CSDL Hệ thống chỉ tiêu thống kê quốc gia; 3) CSDL đề tài nghiên cứu khoa học; 4) CSDL sách và tài liệu thống kê; 5) CSDL các nhà nghiên cứu thống kê; 6) CSDL các tổ chức nghiên cứu và đào tạo thống kê. Để có thể xây dựng được các CSDL nói trên cần phải có các loại dữ liệu cụ thể tương ứng với từng CSDL. Các loại dữ liệu có thể ở nhiều định dạng khác nhau, có cấu trúc, không có cấu trúc, như: văn bản, bảng số liệu tổng hợp, dữ liệu vi mô, ...

(1) CSDL một số cuộc điều tra thống kê, dữ liệu liên quan bao gồm các văn bản pháp lý quyết định việc thực hiện cuộc điều tra; phương án điều tra; phiếu điều tra; tài liệu hướng dẫn điều tra, sổ tay hướng dẫn điều tra; dữ liệu vi mô (dữ liệu gốc); ấn phẩm phản

ánh kết quả điều tra (kết quả chủ yếu, kết quả toàn bộ, báo cáo tổng hợp, báo cáo phân tích...); các bảng danh mục sử dụng trong cuộc điều tra; tài liệu khác có liên quan tới cuộc điều tra.

(2) CSDL hệ thống chỉ tiêu thống kê quốc gia, dữ liệu cần thiết là dãy số liệu của từng chỉ tiêu đã được công bố chính thức theo các phân tổ và kỳ công bố đã được quy định trong Hệ thống chỉ tiêu thống kê quốc gia (Ban hành theo Quyết định số 43/2010/QĐ-TTg ngày 02 tháng 6 năm 2010 của Thủ tướng Chính phủ).

(3) CSDL kết quả nghiên cứu khoa học, dữ liệu cần thiết của CSDL này bao gồm các thông tin chung về đề tài (tên đề tài; mã số đề tài; chủ nhiệm đề tài; năm thực hiện đề tài; cấp đề tài (cấp Bộ, cấp cơ sở); lĩnh vực nghiên cứu; chủ đề nghiên cứu; nội dung nghiên cứu; đánh giá kết quả đề tài; đơn vị chủ trì thực hiện; đơn vị quản lý đề tài); các văn bản hành chính liên quan tới thực hiện và triển khai đề tài nghiên cứu (quyết định tên gọi và chủ nhiệm đề tài; thuyết minh đề tài; các văn bản khác liên quan tới việc triển khai, điều hành của đơn vị quản lý đề tài, đơn vị thực hiện đề tài); kết quả nghiên cứu của đề tài (báo cáo tổng hợp; báo cáo tóm tắt; báo cáo tổng thuật tài liệu; báo cáo tổng hợp chuyên đề); số liệu đã thu thập phục vụ nghiên cứu (các bảng biểu kết quả tổng hợp của đề tài, dữ liệu thu thập, điều tra phục vụ nghiên cứu của đề tài; chương trình, phần mềm máy tính do đề tài xây dựng, thiết kế); các tài liệu khác có liên quan tới đề tài.

(4) CSDL sách và tài liệu, tại thư viện KHTK, các thông tin và dữ liệu về sách và tài liệu được quản lý

bằng hệ quản trị CDS/ISIS for Windows. Các thông tin này đã có sẵn tại thư viện KHTK.

(5) CSDL đội ngũ chuyên gia thông kê, tổ chức nghiên cứu và đào tạo thống kê.

- Đối với đội ngũ chuyên gia, cần thu thập các thông tin và dữ liệu: Họ và tên; cơ quan công tác; địa chỉ liên hệ; học hàm và bằng cấp; quá trình công tác; tên các công trình nghiên cứu; tên các ấn phẩm đã xuất bản; tên các bài viết đã công bố;

- Đối với các tổ chức nghiên cứu và đào tạo thống kê, cần thu thập các thông tin và dữ liệu: Tên tổ chức; cơ quan chủ quản (nếu có); chức năng, nhiệm vụ; lĩnh vực hoạt động; người đứng đầu tổ chức; địa chỉ giao dịch; cơ sở vật chất; nguồn lực tài chính; nhân lực của tổ chức.

2. Quy trình chuẩn hóa, chuyển đổi các dữ liệu đã thu thập

- Chuẩn hóa dữ liệu: Như trên đã đề cập, các dữ liệu phục vụ cho việc xây dựng các cơ sở dữ liệu nói trên có thể ở nhiều định dạng khác nhau, do đó, cần phải chuẩn hóa các dữ liệu này theo quy trình 9 bước dưới đây.

Bước 1: Thu thập dữ liệu (từ các nguồn dữ liệu đã được xác định trước đó);

Bước 2: Lưu trữ nguyên bản các dữ liệu thu thập được;

Bước 3: Phân loại dữ liệu và xác định các định dạng của các dữ liệu;

Bước 4: Điền các thông tin có liên quan của dữ liệu vào Phiếu quản lý dữ liệu của CSDL;

Bước 5: Chuyển đổi các dữ liệu sang các định dạng theo yêu cầu của các CSDL;

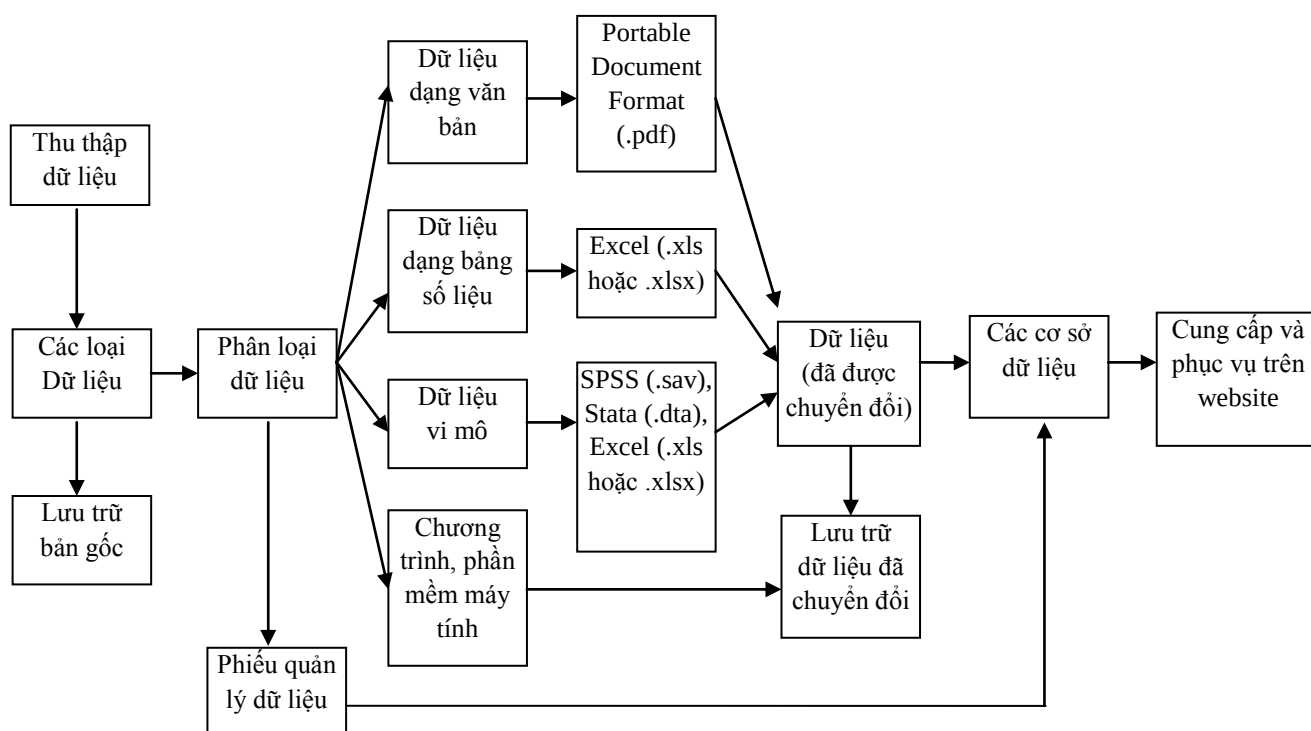
Bước 6: Tập hợp các dữ liệu đã được chuyển đổi;

Bước 7: Lưu trữ các dữ liệu đã được chuyển đổi (mục đích là để dự phòng trong trường hợp có sự cố đối với các CSDL);

Bước 8: Cập nhật các dữ liệu đã được chuyển đổi định dạng cho phù hợp vào các CSDL;

Bước 9: Cài đặt các CSDL lên website và thực hiện cung cấp dữ liệu cho người sử dụng (xem Sơ đồ).

Sơ đồ quy trình chuẩn hoá dữ liệu



– Chuyển đổi dữ liệu, do dữ liệu thu thập được có rất nhiều loại với các dạng khác nhau nên cần thiết phải chuyển đổi các dữ liệu cùng loại theo định dạng thống nhất.

Loại tài liệu, dữ liệu văn bản chuyển đổi sang định dạng .pdf (Portable Document Format). Định dạng này được các website sử dụng phổ biến trong việc cung cấp tài liệu cho người sử dụng.

Loại tài liệu, dữ liệu là các bảng biểu số liệu thuần túy thì chuyển đổi sang định dạng của file Excel (.xls hoặc .xlsx).

Loại dữ liệu vi mô thì chuyển đổi sang định dạng file của các phần mềm thống kê thông dụng là SPSS (.sav), Stata (.dta) và Excel (.xls hoặc .xlsx). Việc chuyển sang các định dạng này nhằm tạo sự thuận tiện cho người dùng có thể sử dụng các phần mềm phân tích và xử lý số liệu thống kê thông dụng.

Loại dữ liệu là chương trình, phần mềm máy tính, không thực hiện chuyển đổi định dạng. Trong thông tin quản lý dữ liệu của CSDL vẫn sẽ có trường quản lý với thông báo là có phần mềm, chương trình máy tính.

Loại CSDL sách và tài liệu của thư viện hiện đang được quản trị bằng hệ quản trị CDS/ISIS for Windows thì không cần thiết phải chuẩn hóa các dữ liệu đầu

vào của CSDL này vì bản thân các dữ liệu khi được cập nhật trên CDS/ISIS for Windows đã được chuẩn hóa rồi. Loại dữ liệu, tài liệu thu thập được ở dạng văn bản như bảng phân loại, phiếu điều tra, tài liệu hướng dẫn điều tra, các báo cáo và sản phẩm của cuộc điều tra thực hiện chuyển đổi sang định dạng .pdf.

Để đáp ứng được các yêu cầu về xây dựng các CSDL phục vụ nghiên cứu và đào tạo thống kê, đồng thời với việc chuẩn hóa và chuyển đổi dữ liệu, cần tiếp tục thu thập dữ liệu liên quan tới các cuộc điều tra thống kê và các chỉ tiêu thuộc Hệ thống chỉ tiêu thống kê quốc gia. Tổ chức lưu trữ tốt hơn nữa các tài liệu liên quan và các sản phẩm KQNC của các đề tài, nhiệm vụ nghiên cứu. Chuyển hoá CSDL sách và tài liệu của thư viện chuyển dẫn dữ liệu từ bản in giấy sang file điện tử toàn văn và cập nhật dữ liệu điện tử toàn văn mới cho thư viện. Tổ chức thực hiện khảo sát thu thập thông tin và dữ liệu cần thiết về các chuyên gia thống kê, các tổ chức nghiên cứu và đào tạo thống kê phục vụ nghiên cứu và đào tạo thống kê.

Nguồn: Chuyên đề *Nguồn dữ liệu, giải pháp hoàn thiện để xây dựng các CSDL phục vụ nghiên cứu và đào tạo thống kê* thuộc đề tài “Nghiên cứu các cơ sở dữ liệu phục vụ nghiên cứu và đào tạo thống kê”.

