

Thủ phác thảo kho dữ liệu Thống kê cho nghiên cứu và đào tạo

*Hoàng Minh Thiện**

Viện Khoa học Thống kê (KHTK) đã có hơn 35 năm hoạt động, đồng hành cùng các hoạt động nghiệp vụ của ngành Thống kê, do đó Viện KHTK đã hình thành nên nguồn dữ liệu đa dạng cho hoạt động tác nghiệp, nghiên cứu và học tập. Tuy nhiên, nguồn dữ liệu đa dạng này đang phân tán tại nhiều nơi, tin học hoá chưa được áp dụng có hệ thống, dẫn đến tình trạng người tác nghiệp, nghiên cứu, học tập gặp nhiều khó khăn khi muốn tiếp cận nguồn dữ liệu này; hay tệ hơn nữa, nhiều nghiên cứu đã lặp lại những gì đã thực hiện trước đây hoặc các hoạt động tác nghiệp trong ngành không tận dụng được những gì đã nghiên cứu. Điều này đương nhiên dẫn đến suy nghĩ về sự hình thành một kho dữ liệu tin học hoá phục vụ cho tác nghiệp, nghiên cứu và học tập (từ đây gọi tắt là kho dữ liệu). Nhằm tạo thuận lợi cho những người cần đến nguồn tin này có thể khai thác ngay trên bàn làm việc của mình.

Bài viết này giới thiệu những ý tưởng ban đầu về kho dữ liệu như thế làm cơ sở cho việc triển khai xây dựng trong thực tiễn tại Viện KHTK. Như thể chắc là sơ lược, song chúng tôi vẫn mạnh dạn đưa ra nhằm thu thập thêm ý kiến các chuyên gia thống kê và tin học để quá trình thực tiễn sẽ hoàn thiện hơn!

Như lời dẫn ở trên, theo suy luận logic bình thường, ai cũng sẽ đặt ra mấy câu hỏi tự nhiên sau:

- Kho dữ liệu làm ra để phục vụ ai? - Kho dữ liệu gồm có những dữ liệu gì?

- Kho dữ liệu này được quản trị như thế nào?

- Công cụ nào ưu việt cho hoạt động của kho dữ liệu?

Để dễ dàng hình dung mô hình kho chúng ta tạm chấp nhận cách trả lời đơn giản sau:

- Kho dữ liệu này làm ra để phục vụ những người tác nghiệp, nghiên cứu, học tập liên quan tới thống kê - gọi là người dùng tin (User);

- Kho này gồm các dữ liệu thống kê cần cho người tác nghiệp, nghiên cứu, học tập tạm gọi là các dữ liệu cơ sở (DLCS);

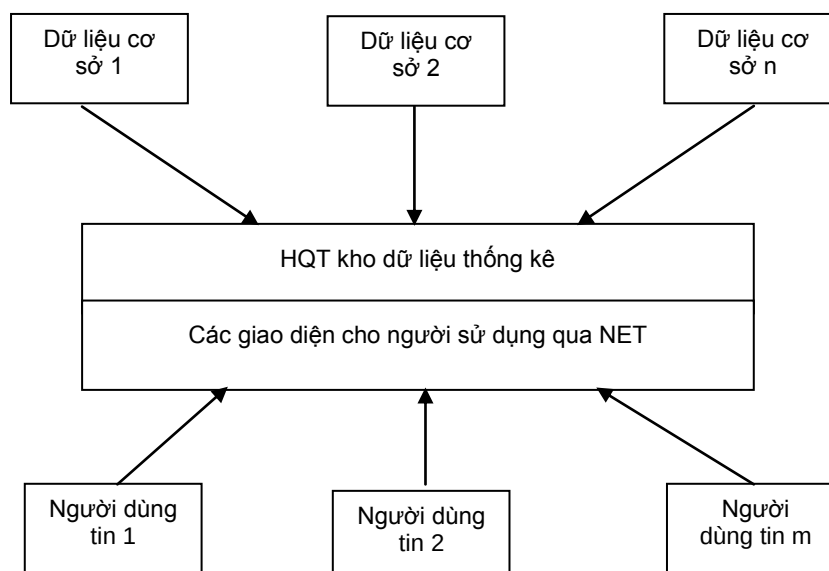
- Kho này được quản trị bằng tin học để an toàn, thuận tiện nhất cho người dùng tin và người quản trị kho trong hoạt động - gọi là hệ quản trị (HQT);

- Kho này cần dùng kỹ thuật mạng (NET) để cung cấp tận bàn làm việc cho người dùng tin và cũng là nguồn cung cấp dữ liệu lâu dài cho kho.

Những phân tích chi tiết hơn về người dùng tin, nguồn dữ liệu cơ sở, hoạt động quản trị và công cụ kỹ thuật NET sẽ được trình bày sau. Với các định nghĩa đơn giản này chúng ta có thể hình dung ngay được mô hình cơ bản của kho dữ liệu này, như sau:

* Viện Khoa học Thống kê

1. Phác thảo mô hình kho dữ liệu



Đọc một cách đơn giản mô hình thì kho dữ liệu này hoạt động như sau: kho có một số DLCS đã được thu thập và có một hệ quản trị kho; người dùng tin thông qua NET với một số thủ tục sẽ khai thác, sử dụng được các dữ liệu có trong kho này. Trong thực tiễn thì hoạt động một kho dữ liệu dạng này sẽ phức tạp hơn, nhưng cách nhìn giản đơn cũng cho thấy khái quát về hệ thống dự định xây dựng.

Cho đến đây mọi góc nhìn đều đơn giản, tuy vậy để có thể xây dựng kho dữ liệu này thì chúng ta cần xác định chi tiết hơn và có thể khả thi!? Như vậy chúng ta lại cần phải trả lời 4 câu hỏi trên thật cụ thể cho khi triển khai: Thu thập những DLCS nào? Mỗi DLCS thu thập những gì có thể và có khả năng quản trị nó? Nhu cầu thực tiễn của người dùng tin cụ thể là gì? Đòi hỏi với họ cụ thể ra sao khi tiếp cận (giao diện) kho này? cập nhật, an toàn và thuận tiện thế nào của hệ quản trị với người sử dụng và người khai thác kho?... Chúng tôi không thể qua bài viết này cụ thể được tất cả nhưng cũng vẫn muốn đưa ra các định nghĩa chung để có thể triển khai thực tiễn mà không bỏ lỡ. Dưới đây sẽ định nghĩa về các thành phần cơ bản của kho dữ liệu này; cách tiếp cận ở đây xuất phát

từ người dùng tin xem như là: ***“Cung cấp cho người dùng tin cái họ cần chứ không phải cái ta có”***.

2. Người dùng tin là ai hay là kho dữ liệu phục vụ cái gì?

Như đã định nghĩa và giới hạn về người dùng tin ở trên, cùng với chức năng nhiệm vụ của Viện KHTK, chúng tôi xác định người dùng tin của kho này gồm: những người đang tác nghiệp, nghiên cứu, học tập có liên quan đến thống kê và khoa học thống kê. Vậy những người tác nghiệp, nghiên cứu, học tập có liên quan đến thống kê cần những dữ liệu gì? Theo kinh nghiệm chúng tôi và một số khảo sát của Viện gần đây, thì nhu cầu này tập trung ở mấy chủ đề chính sau:

- Các cuộc điều tra thống kê (cả mẫu lẫn tổng điều tra), các số liệu thống kê qua hệ thống báo cáo, các số liệu thống kê tổng hợp, các phân tích thống kê, các super data (danh mục, định nghĩa dữ liệu...). Ở đây nghiêng về sử dụng số liệu;

- Các đề tài nghiên cứu về thống kê trong nước đã và đang thực hiện; các phương pháp, giải pháp thống kê được nghiên cứu ứng dụng...;

- Các bài viết về thống kê đã đăng trên tờ Thông tin Khoa học Thống kê của Viện; của quốc tế;
- Các tư liệu thống kê khác (nghe về tư liệu văn bản) như: sách (thư viện) và các tư liệu văn bản khác (như văn bản pháp lý, báo cáo hành chính các cuộc điều tra, tài liệu nghiệp vụ như: từ chuẩn thống kê, hướng dẫn nghiệp vụ, tổng luận, chuyên đề, tài liệu dịch...);
- Tư liệu thống kê nước ngoài (chắc là về lâu dài và thông qua liên kết);
- Tư liệu về các nhà nghiên cứu thống kê trong nước, các cơ sở đang hoạt động và nghiên cứu thống kê;

Qua nói về nhu cầu của người dùng tin cũng là nói về yêu cầu cần thu thập DLCS; tuy vậy định nghĩa chặt chẽ một dữ liệu cơ sở để có thể quản trị và đáp ứng nhu cầu người dùng tin thì chúng ta cần xác định chi tiết hơn với từng loại tư liệu. Điều này cần thiết để hệ quản trị có thể hoàn thiện các chức năng như: tìm kiếm, bổ sung, loại bỏ, an toàn hay như cung cấp các dịch vụ tiện ích phù hợp với người dùng tin và khả năng cung cấp... Nói thế có nghĩa chúng ta cần định nghĩa cụ thể hơn về dữ liệu cơ sở của kho dữ liệu này.

3. Định nghĩa dữ liệu cơ sở của kho

(1) Dữ liệu về điều tra thống kê

Khác với 1 CSDL bình thường, dữ liệu cơ sở (DLCS) về 1 cuộc điều tra thống kê có thể lại có cả CSDL trong đó. Một dữ liệu cơ sở như vậy được chúng tôi dự định thu thập như sau:

- Báo cáo hành chính của cuộc điều tra nêu có;
- Các quyết định hành chính, pháp lý của cuộc điều tra;
- Phương án điều tra, thiết kế mẫu, phiếu điều tra, thời điểm điều tra...;
- Các định nghĩa, khái niệm được dùng trong điều tra;
- Từ khoá chính cho cuộc điều tra;

- Các báo cáo tổng hợp và phân tích chính thức về cuộc điều tra;
- Các tư liệu có liên quan khác;
- Các CSDL của cuộc điều tra, dữ liệu vi mô, dữ liệu tổng hợp...;
- Phần mềm quản trị CSDL, phần mềm tổng hợp, phân tích điều tra...;
- Các sản phẩm mềm liên quan khác.

Từ những tư liệu trên chúng ta sẽ xác định bản ghi quản lý cho một cuộc điều tra để hệ quản trị có thể kiểm soát.

(2) Số liệu báo cáo thống kê tổng hợp và phân tích

Đây là mảng số liệu đa dạng, quan trọng và rời rạc, tuy nhiên trong nghiên cứu và học tập nó không có vị trí chủ chốt và nó lại có thể khai thác từ nguồn khác trong Tổng cục Thống kê; vì thế chúng tôi đang bản khoản hình thức lưu trữ nó như một dữ liệu số hay chỉ như một dữ liệu văn bản dùng cho tham khảo khi nghiên cứu, học tập.

(3) Đề tài nghiên cứu

Sau đây là các dữ liệu cơ bản về một đề tài cần dự định thu thập để cung cấp cho người dùng tin:

- Tên đề tài
- Chủ nhiệm
- Địa chỉ liên hệ
- Thời gian nghiên cứu
- Phân loại đề tài (cấp)
- Tổng kinh phí
- Chủ đề (Tags)
- Nội dung chính
- Báo cáo tóm tắt
- Báo cáo đầy đủ
- Hội đồng nghiệm thu
- Các tư liệu liên quan (các văn bản pháp lý, tài liệu sử dụng, nguồn trích...)
- Các sản phẩm kèm theo (file số liệu, chương trình máy tính, sơ đồ kỹ thuật...)

Hiện nay Viện KHTK đang thử nghiệm một hệ quản trị các đề tài nghiên cứu, khi thành công các tư liệu trên có thể chiết xuất từ đó; hoặc giả đơn giản hơn là một đường link của hệ quản trị kho đến hệ quản trị đó; điều này đang là thực tế.

(4) Thư viện điện tử và tư liệu thông kê văn bản khác

Kho sẽ liên kết với thư viện điện tử của Viện KHTK đang được xây dựng, người dùng tin sẽ có thể sử dụng thư viện như một nhánh của kho dữ liệu; hiện nay kho của thư viện có hơn 2000 đầu tài liệu, đã được quản trị bằng hệ ISIS. Một thư viện điện tử hoàn thiện có thể được tính đến với sách, tài liệu tập trung vào thông kê, thiết thực với người dùng tin dự định là tác nghiệp, nghiên cứu và học tập thông kê.

(5) Các bài viết về khoa học thông kê

Đây chính là nội dung tờ Thông tin Khoa học Thông kê hiện đã được đăng thường kỳ trên trang WEB của Viện; việc liên kết hay đưa về kho dữ liệu coi như không có vấn đề gì. Nguồn tin đã có các bài viết từ hàng chục năm nay và dự định sẽ cung cấp đến mức toàn văn bài viết.

(6) Tư liệu về thông kê nước ngoài

Nguồn tư liệu này sẽ nằm rải trong thư viện qua tài liệu dịch, sách hay tư liệu nước ngoài trong thư viện; Việc truy tìm thông qua từ khoá hay chủ đề mà thôi. Tất nhiên sẽ có thêm các đường Link liên kết với các nguồn tư liệu nước ngoài khác mà người dùng tin phải tuân thủ theo các điều kiện của các nơi cung cấp đó.

(7) Tư liệu về các nhà nghiên cứu, các cơ sở nghiên cứu thông kê khác

Trong nghiên cứu việc tham khảo công trình, kiến thức, kinh nghiệm của các nhà nghiên cứu khác, hay tư liệu, hướng nghiên cứu của các tổ chức khác là rất cần thiết để không lặp lại những gì đã được nghiên cứu, hay như tìm ra hay theo một hướng nghiên cứu

mới. Vì thế chúng tôi dự định sẽ thiết lập hai CSDL về các nhà nghiên cứu thông kê và các đơn vị hoạt động nghiên cứu thông kê.

Như vậy đến đây chúng ta nhận chân những dữ liệu cơ bản về thông kê phục vụ cho nghiên cứu, học tập và tác nghiệp. Mô tả như vậy về nguồn dữ liệu thật là còn rất giản đơn, song trong một bài thông tin giới thiệu ý tưởng về kho dữ liệu cũng nên giành cho các thành phần khác của nó mô tả cơ bản.

4. Hệ quản trị kho dữ liệu

Hệ quản trị phải đạt các yêu cầu như sau:

- Đảm bảo khả năng cập nhật kho: bổ sung, loại bỏ, sửa các thành phần DLCS;
- Khả năng tìm kiếm các DLCS theo chủ đề (Tag) và phục hoạt các CSDL cho user khai thác;
- Đảm bảo khả năng đa CSDL cùng hoạt động (đáp ứng nhiều user cùng khai thác - đặc tính của khai thác mạng);
- An toàn cho kho dữ: lưu trữ, phục hồi kho, chống thâm nhập phá hoại kho...;
- Có các giao diện đơn giản và thân thiện với user mạng hiện nay.

Trên đây là các yêu cầu kỹ thuật tối thiểu khi xây dựng hệ quản trị phải có để các cán bộ tin học có thể triển khai và người dùng tin hiểu các đáp ứng và đòi hỏi của hệ thống.

5. Các dịch vụ cho người dùng tin

Trên đây đã giới thiệu các thông tin cung cấp cho người dùng tin, ở đây chúng tôi xin giới thiệu các hình thức đáp ứng nhu cầu thông tin của kho dữ liệu, bao gồm các hình thức dịch vụ sau:

(Xem tiếp trang 41)

Handbook of Statistics
////////////////////

(tiếp theo trang 31)

- Cung cấp tất cả các dữ liệu thống kê như định nghĩa ở mục 2. về DLCS;
- Cung cấp khả năng khai thác số liệu của DLCS (nếu có) cho các thử nghiệm nghiên cứu và học tập;
- Cung cấp khả năng sao chép các dữ liệu trong kho theo thủ tục;
- Giới thiệu các sản phẩm mềm đã sử dụng phân tích, tổng hợp, dự báo... trong các CSDL;
- Các dịch vụ khác...

Trên đây là phác thảo ban đầu sơ lược về kho các CSDL điều tra thống kê dự định xây dựng, nay mạnh dạn giới thiệu mong có đóng góp ý kiến của độc giả để có thể tiến tới xây dựng một mô hình hoàn thiện và khả thi.

Tài liệu tham khảo:

1. Thông tin Khoa học thống kê Số 4-2007, bài “Củng cố kho thông tin tư liệu KHTK kết hợp tin học hóa”, tác giả: Hoàng Minh Thiện.
2. Đề tài: “Nghiên cứu xây dựng kho dữ liệu đầu vào của Tổng cục Thống kê”, chủ nhiệm: ThS. Nguyễn Văn Đoàn.